ABSTRACT

Title of Dissertation: INVESTIGATING THE USE OF MAZE-CBM FOR HIGH SCHOOL STUDENTS

Marisa Ann Mitchell, Doctor of Philosophy, 2016

Dissertation directed by: Dr. Jade Wexler, Department of Counseling, Higher Education, and Special Education
University of Maryland, College Park

Recent legislation and initiatives set forth high academic expectations for all high school graduates in the area of reading (National Governors Association Center for Best Practices, 2010; Every Student Succeeds Act, 2015). To determine which students need additional support to meet these reading standards, teachers can conduct universal screening using formative assessments. Maze Curriculum-Based Measurement (Maze-CBM) is a commonly used screening and progress monitoring assessment that the National Center on Intensive Intervention (2013) and the Center on Instruction (Torgesen & Miller, 2009) recommend. Despite the recommendation to use Maze-CBM, little research has been conducted on the reliability and validity of Maze-CBM for measuring reading ability for students at the secondary level (Mitchell & Wexler, 2016).

In the papers included in this dissertation, I present an initial investigation into the use of Maze-CBM for secondary students. In the first paper, I investigated prior studies of Maze-CBM for students in Grades 6 through 12. Next, in the second paper, I investigated the alternate-form reliability and validity for screening students in Grades 9 and 10 using signal detection theory methods. In the third paper, I examined the effect of genre on Maze-CBM scores with a sample of students in Grades 9 and 10 using multilevel modeling.

When writing these three papers, I discovered several important findings related to Maze-CBM. First, there are few studies that have investigated the technical adequacy of Maze-CBM for screening and progress monitoring students in Grades 6 through 12. Additionally, only two studies (McMaster, Wayman, & Cao, 2006; Pierce, McMaster, & Deno, 2010) examined the technical adequacy of Maze-CBM for high school students. A second finding is that the reliability of Maze-CBM is often below acceptable levels for making screening decisions or progress monitoring decisions (.80 and above and .90 and above, respectively; Salvia, Ysseldyke, & Bolt, 2007) for secondary students. A third finding is that Maze-CBM scores show promise of being a valid screening tool for reading ability of secondary students. Finally, I found that the genre of the text used in the Maze-CBM assessment does impact scores on Maze-CBM for students in Grades 9 and 10.

INVESTIGATING THE USE OF MAZE-CBM FOR HIGH SCHOOL STUDENTS

by

Marisa Ann Mitchell

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2016

Advisory Committee:
Assistant Professor Jade Wexler, Chair
Associate Professor Rebecca Silverman
Assistant Professor Kelli Cummings
Associate Professor Laura Stapleton
Principal Researcher Rebecca Zumeta Edmonds

Acknowledgements

First, I would like to thank my advisor, Jade Wexler. You have not only taught me an immense amount about conducting research and teaching, but you have also taught me a lot about myself. You have encouraged me along the way, given me guidance, and repeatedly illustrated to me how perseverance is key! I would also like to thank Rebecca Silverman for advising me in the first year of this program and allowing me to be such an active member of your research team from the beginning. Kelli Cummings, Laura Stapleton, and Rebecca Zumeta Edmonds—my other amazing committee members—I appreciate all of your support and guidance. You have pushed me to think critically, which caused me to be a better researcher, writer, and teacher. I could not have done it without all of your support and guidance.

Thank you to the various professors in the University of Maryland Department of Special Education for initially accepting me into your "family" and teaching me what it means to be a special education teacher. You taught me the basics and also deepened my understanding of the field through research. A special thank you to Debbie Speece for my initial exposure to educational assessment and CBM, as well as giving me my initial glimpse into the world of research. You all have helped shape me to be the teacher and researcher I am today.

This dissertation would not have been possible without the support of fellow University of Maryland students, both past and present. First, I would like to thank Erin Clancy, who has been a daily support for me throughout the program. Thank you for not only being a great fellow student but also for assisting with data collection, being a shoulder to cry on, making me laugh, and being an all-around

amazing friend. Thank you to Peet Smith for taking time away from your own studies to help me with data collection for this dissertation and for being another great peer throughout this program. Kathleen McPartland and Paige Hawkins, thank you for all of your meticulous work helping me score and enter the data necessary for this dissertation. Thank you to Sarah Fisher for all of your help with editing. I also could not have made it through this program without the support of prior students, especially Katryna Andrusik and Beth Montanaro. You listened to my struggles, had work sessions with me, provided a school where I could conduct my study, and gave a lot of great advice along the way.

Additionally, I need to thank both Kaitlyn Colleli and Ali Rendulic. You were both there from the very beginning and served as my support system during coursework, student teaching, our first year in the classroom, and the years that followed. You constantly remind me of what it’s like to be an incredible special education teacher and you inspire me to make a difference in the field. I only hope that someday we get a chance to work together again.

I would also like to thank my family, who supported me throughout this program. First, thank you to my mom for leading the path of “CBM” research in our family, always letting me vent or cry, helping me with my writing, and always having thoughtful conversations with me about the field of special education. To my sister, Daniela, thank you for supporting me and showing me through your own career path how to follow your passions in life. Thank you to my incredible husband, Steve, for leading the way in the pursuit of a Ph.D., dealing with me and my emotions throughout the journey, and helping me find time to relax with a good beer in hand.

You have been an incredible emotional support for me and have seen me at my best and worst and stood by me regardless.

Last, but certainly not least, I would like to thank my former students for showing me the joy, and also the struggle, of being a special education teacher. You have touched me and inspired me in ways you might never know. I would not be where I am today without each and every one of you.

Table of Contents

Introduction

Currently, there is a significant nationwide emphasis on standards-based learning and large-scale accountability assessment (e.g., Common Core State Standards [CCSS], National Governors Association Center for Best Practices, 2010; Every Student Succeeds Act [ESSA], 2015). The standards provide stringent guidelines for schools to help all students, including those with and at risk for disabilities, meet these high academic standards. The assumption is that when students reach these high expectations, they will be prepared for post-secondary success.

One of the most critical standards, according to these reforms, is that students become proficient readers. To reach a level of proficiency in reading, students must be able to illustrate that they can read, understand, evaluate, and synthesize a variety of text, including expository and narrative text. The CCSS set forth English Language Arts standards to provide guidance for teachers to help students in developing the skills necessary to achieve reading proficiency. It is important to note that the CCSS English Language Arts standards are also designed to be implemented across History, Social Studies, Science, and Technical subjects. Due to the implementation of English Language Arts standards across these content areas, students will be exposed to, and expected to comprehend, complex texts presented across their content-area classes.

Teachers use end-of-year, statewide assessments to determine which students are meeting these standards. Teachers use these statewide tests, along with the new CCSS-aligned assessments developed by Smarter Balanced Assessment Consortium (n.d.) and the Partnership for Assessment of Readiness for College and Careers (2014), as assessment tools to provide information about how well a student, or multiple groups of students, has met the necessary standards for his or her grade level. These assessments represent summative assessments and are considered a measurement *of* learning because they take place at the end of the year and assess whether students have learned the standards teachers taught them throughout the year. Additionally, they are considered high-stakes tests because they are linked to requirements of ESSA and, recently, to teacher evaluations and pay (Pullin, 2013).

Although summative information can be valuable in evaluating the effectiveness of an education program, it does not allow teachers the opportunity to adjust their instruction for those students who are at risk of not achieving the college and career readiness standards. A considerable problem with summative assessments is that teachers often do not receive students' scores until after the student has moved to the next grade level; thus, the results cannot help teachers adjust the instruction for students based on the scores. In order to drive instruction to meet the current needs of students, it is critical for teachers to gauge students' progress on these standards on an ongoing basis throughout the school year.

Formative assessment may provide a feasible compliment to summative assessment that would allow teachers to adjust their instruction *for* learning. Formative assessment is a broad term for assessments that are intended to support learning (Van der Kleij, Vermeulen, Schildkamp, and Eggen, 2014). Formative assessment is used to measure how well, how much, and what students are learning in response to instruction. Formative assessment can be informal (e.g., student work samples) or formal in nature (i.e., standardized and validated assessments). They are commonly given throughout the school year to yield timely information for teachers about which students are not making progress, what skills students are still struggling with, and how far behind students are in meeting academic standards. Due to the ongoing use of formative assessment, teachers can use these data to make adjustments to their instruction to meet the needs of students who are not making progress on standards. Teachers can also use formative assessment to assign students, who are not meeting the academic standards, to supplemental interventions that will address their areas of need. Then, they can use formative assessment to continue monitoring student progress in supplemental interventions before the end of the year. In fact, it has been shown at the elementary level that teachers who use standardized and validated formative assessment to examine students' growth during instruction yield higher gains for their students than those who do not use formative evaluation (e.g., Black & Wiliam, 1998; Fuchs, Fuchs, Hamlett, & Ferguson, 1992; Stecker, Fuchs, & Fuchs, 2005). It is therefore plausible that the use of formative assessment will be efficacious for secondary level teachers to adjust their instruction in order to meet student needs.

Problem-Solving Model

Teachers and school personnel often use standardized and validated formative assessments that meet rigorous technical standards to make data-based decisions about their students. These assessments help teachers to make ongoing decisions regarding the instruction they are providing to students. The process of making data-based decisions involves teachers and school personnel systematically collecting and analyzing data in order to guide instructional decisions for their students (Torgesen & Miller, 2009). Educational data-based decisions are often made within the context of a Problem-Solving Model (Deno, 1989). This model is recursive, wherein teachers and school personnel take a series of steps to meet the educational needs of students in their classroom or school. The steps in the problem-solving model are (a) problem identification, (b) problem definition, (c) exploring solutions, (d) implementing solutions, and (e) problem solution (Deno, 1989) and are illustrated in Figure 1.

In the context of reading instruction, the first step of the model—problem identification—involves teachers determining which students are below expected levels of reading compared to their same-grade peers. This process is commonly referred to as screening. During the screening process, all students, or at the secondary level the students who are suspected of having a reading difficulty, are given a reading assessment. Students who score below the expected level are considered at risk for reading difficulty. The second step of the model involves teachers defining the reading problem. In this step, teachers use the assessment used for screening, in addition to other data sources (e.g., diagnostic tests, work samples, informal observation), to determine which reading skills the student is deficient in. In the third step, teachers use the data to determine what instruction would remediate the student’s deficits. During this process, teachers might consider what type of reading skills the instruction should include, who will deliver the instruction, and when the instruction will occur. Next, in the fourth step of the model, teachers implement the instruction and monitor students’ progress using a formative reading assessment. For the fifth and final step, teachers use the data from the formative assessments to determine if the instruction has been effective or if the student fails to reach required levels of reading. If the student is not making adequate progress towards meet the reading standards, then they return to be first step in the problem-solving model. At that point, teachers will attempt to further clarify the problem with reading that the student struggles with using the additional data they have collected throughout the intervention. The process then is repeated until the student has received instruction that enables them to reach expected reading standards.

Curriculum-Based Measurement

Curriculum-based measurement (CBM) is one of the most commonly used standardized formative assessment tools for making data-based decisions in the area of reading. Experts recommend CBM for use in screening and progress monitoring students as part of a Response to Intervention (RTI) or Multi-Tiered System of Support (MTSS) model of reading (National Center on Response to Intervention, 2010; Torgesen & Miller, 2009). RTI and MTSS are schoolwide models of tiered instruction that involve teachers making data-based decisions in order to provide the appropriate level of support to students so that they will reach academic standards. There are two types of CBM: (a) general outcome measures (GOMs) and (b) mastery measures (MM). GOMs measure general ability by sampling performance across several subskills (Hosp, Hosp, & Howell, 2007). MM are used to measure discrete and easily identified sets of items that are related to a common skill (Hosp, Hosp, & Howell, 2007). Advantages of using CBM for screening and progress monitoring are that the use of CBM is (a) quick and efficient, (b) cost effective, (c) involves alternate forms that can be administered over time, allowing the results of the assessments to guide data-based decision making, (d) aligned to the curriculum, (e) validated, and (f) technically adequate (Hosp et al., 2007). Therefore, formative assessments such as CBM allow for the teacher to adjust instruction during the school year in a way that the end-of-year statewide assessments were not designed to do.

Maze-CBM

Maze-CBM is the most commonly utilized CBM at the secondary level because it has shown potential for being more sensitive to growth over time in reading than other types of CBM (i.e., oral reading or word reading) for students at this age (McMaster, 2010; Torgesen & Miller, 2009; Wayman, Wallace, Wiley, Tichá, & Espin, 2007). Additionally, Maze-CBM has been viewed as a measure of reading comprehension, a vital skill needed at the secondary level. During a Maze-CBM assessment, students are required to silently read a passage for a fixed number of minutes (e.g., 3 minutes) in which words are deleted at fixed ratios and replaced by a choice (i.e., a maze) of three words, one of which is the correct word. As students read the passage, they are instructed to circle the word that best completes the sentence. The score on Maze-CBM is typically calculated by totaling the correct number of words the students was able to replace in the allotted time. Maze-CBM is also a beneficial assessment for students at this age because it can be group-administered, allowing for minimal impact to instruction, and appears to measure reading comprehension in addition to accurate decoding (Fuchs & Fuchs, 1992; Wiley & Deno, 2005).

One large disadvantage of using Maze-CBMs for progress monitoring and screening at the secondary level is the lack of research on its use, and thus the validity of its use. This is particularly true for high-school students (Stecker, Fuchs, & Fuchs, 2005; Wayman et al., 2007). Although researchers and practitioners may be able to draw from the current research on the technical adequacy of Maze-CBM for elementary students, we need to consider that these conclusions might not generalize to older students. In fact, Schatschneider et al (2004) have shown that verbal knowledge and reasoning skills become increasingly more important as students move from elementary grades to high school. Further research needs to be conducted to provide evidence of reliability and validity for the use of Maze-CBM in making screening and progress monitoring decisions for high school students.

Variability of Scores

A key assumption of Maze-CBMs is that alternate forms are parallel in nature and can be used interchangeably. In the early development of CBM, researchers would choose passages at random from the students' curriculum without consideration of the equivalent nature of the passages (Shinn, 1989). Although it is now common for test developers of CBM to more carefully choose passages based on various readability formulas or to equate passages after development (Baker et al., 2015; Christ & Ardoin, 2009; Cummings, Park, & Schaper, 2012; Francis et al., 2008; Petscher & Kim, 2011), the research on passage variance has primarily occurred with elementary school students on CBM measures of reading fluency rather than Maze-CBM. Furthermore, at the elementary level, there is evidence that student scores across passages differ considerably despite being explicitly controlled (Hintze & Christ, 2004). There are many potential sources of variability in scores, including administrator error, student error, and error due to the testing environment; however, the focus of this dissertation will be sources of error that are a result of the development of alternate forms of Maze-CBM.

The use of alternate forms containing different passages may lead to differing levels of text complexity as the dimensions of text complexity may differ across passages. The CCSS outlines a model of text complexity that contains three main dimensions (Figure 2): (a) qualitative dimensions of text, (b) quantitative dimensions of text, and (c) reader and task considerations (Appendix A, National Governors Association Center for Best Practices, 2010). Qualitative dimensions of text refer to "the aspects of text complexity best measured or only measureable by an attentive human reader, such as levels of meaning or purpose; structure; language conventionality and clarity; and knowledge demands" (Appendix A, National Governors Association Center for Best Practices, 2010). Quantitative dimensions of text refer to "the aspects of text complexity, such as word length or frequency, sentence length, and text cohesion, that are…today typically measured by computer software" (Appendix A, National Governors Association Center for Best Practices, 2010). There are several formulas for measuring the quantitative dimensions of text, such as the Flesch-Kincaid readability formula (Kincaid, Fishburne, Rodgers, & Chissom, 1975), the Dale-Chall readability formula (Chall & Dale, 1995), the Coh-Mextrix text easability assessor (McNamara, Louwerse, Cai, & Grasser, 2005), and the Lexile framework (MetaMetrics, 2015). These formulas examine features of text such as word length, sentence length, word frequency, and the cohesiveness of a text (i.e., how tightly the text holds together) to determine the quantitative complexity of text. The final component of the text complexity model in the CCSS is the reader and task considerations, which refers to "variables specific to a particular reader and tasks" (Appendix A, National Governors Association Center for Best Practices, 2010). Examples of variables related to a particular reader include motivation, knowledge, and experiences that the reader has and brings with him to the reading experience (RAND Reading Study Group, 2002). According to the CCSS model of text complexity, the interaction between these three dimensions determines how difficult a particular text is for a reader.

Thus, according to the CCSS model of text complexity, there are several factors that interact when a student reads a passage on a Maze-CBM assessment and that will impact the variability of scores across alternate forms. For instance, a student may have a great deal of background knowledge of a particular science topic and, thus, will score higher on Maze-CBM passages that include information about that topic as opposed to passages on a science topic they have little background knowledge of, even though the two passages have similar readability scores. It may not be reasonable to control all aspects of text complexity in alternate forms of Maze-CBM at the high school level because students will have differing levels of background knowledge and skills that they bring to each passage. This indicates that, in order to achieve truly parallel alternative forms of Maze-CBM passages, test developers would have to account for both text characteristics and also consider student variables such as prior knowledge and reading ability that potentially interact with the text characteristics. One option used previously with oral reading CBM is to retroactively equate the passages (Baker et al., 2015; Christ & Ardoin, 2009; Cummings et al., 2012; Francis et al., 2008; Petscher & Kim, 2011). While this is a suitable method for oral reading CBM it may be particularly challenging to conduct with Maze-CBM which has an additional item difficultly parameter that is associated with each item within the passage itself. Although this is an ambitious goal, it is imperative that researchers begin to investigate potential sources of this variance in scores and how this variance impacts the reliability and validity of scores.

Overview of Three Articles

Drawing from the current literature on Maze-CBM and text complexity, my dissertation includes three articles designed to answer questions about the reliability and validity of Maze-CBM for high school students. The overarching question of these three articles is: What is the reliability and validity of scores on Maze-CBM for high school students and are these psychometric characteristics influenced by passage differences or student ability?

The first article in this dissertation presents a synthesis of the current literature on the technical adequacy of Maze-CBM for secondary students. In that article, I synthesized information across existing studies of Maze-CBM for students who are in Grades 6 through 12. Specifically, I address the following research questions:

1. What are the features of Maze-CBM administration (e.g., duration of testing session and scoring procedures) used in studies of Maze-CBM for students in Grades 6 through 12?
2. What are the features of the texts used in studies of Maze-CBM for students in Grades 6 through 12, including text length and level?
3. What is the technical adequacy as reported in studies of Maze-CBM for students in Grades 6 through 12, including reliability and validity of static

scores for screening, estimates of slope, and reliability and validity of the slope for determining response to instruction?

The second article in the dissertation is an empirical study on the alternate-form reliability of Maze-CBM scores and the validity of their use for screening high school students. In that article, I answer the following research questions:

1. What is the alternate-form reliability of scores on Maze-CBM of high school students?
2. What is the validity of scores on Maze-CBM in the prediction of high school students' reading risk status?

The final article in the dissertation is an empirical study investigating the effect of the genre of text on high school students' Maze-CBM scores. In that article I also initially explore student characteristics, such as oral reading ability, that influence Maze-CBM scores. The aim of that article was to answer the following research questions:

1. What proportion of variance in Maze-CBM scores is due to between and within student differences?
2. What proportion of variance in Maze-CBM scores is attributable to genre effects?
3. Does OR ability interact with genre in interpreting Maze-CBM scores?

Definition of Key Terms

Alternate-form reliability: The consistency of scores across two similar forms of a test administered to the same group of examinees within a very short period of time (Crocker & Algina, 2008).

Concurrent validity: The relationship between test scores and criterion measurements made at the time the test was given (Crocker & Algina, 2008).

Curriculum-based measurement (CBM): A type of formative assessment used within a research-based set of procedures for teachers to simply and frequently monitor student progress toward instructional goals (Deno, 1985).

Expository text: A type of text that is intended to inform the reader about a topic. Examples include textbooks, newspapers, and magazine articles (Sáenz & Fuchs, 2002).

Genre: A broad category of text type that students may encounter. In this dissertation, the two main genres of text are narrative and expository (Sáenz & Fuchs, 2002).

Narrative text: A type of text that commonly involves stories that are written to entertain and contain text elements such as characters, a sequence of events, morals, and themes (Sáenz & Fuchs, 2002).

Predictive validity: The degree to which test scores predict criterion measurements that will be made at some point in the future (Crocker & Algina, 2008).

Reading risk status: The determination of students' reading ability based on a particular cut point on a criterion measure, which is meant to indicate if the student is meeting the expected level of reading or is not and thus is in need of supplemental reading instruction.

Reliability: The desired consistency, or reproducibility, of test scores (Crocker & Algina, 2008).

Test–retest reliability: The consistency of scores on a test administered to a group of examinees on two occasions (Crocker & Algina, 2008).

Validity: The extent to which a test developer and test user have collected evidence to support the types of inferences that can be made from test scores (Cronbach, 1971).

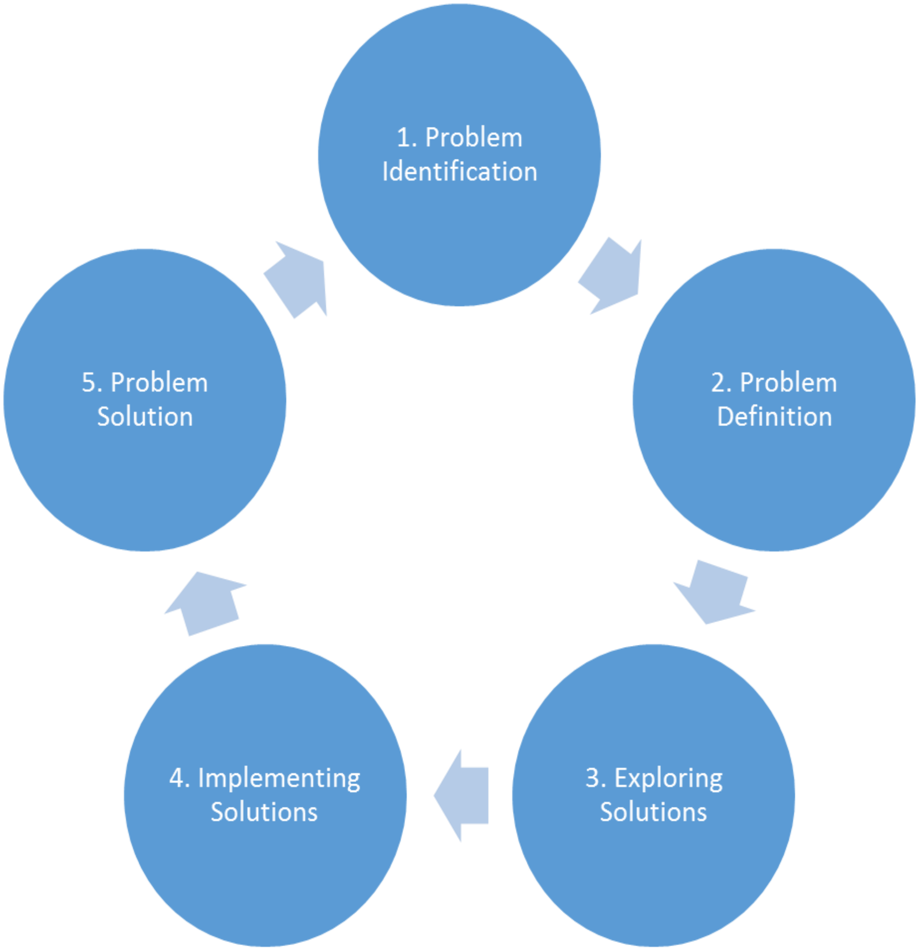


Figure 1. An illustration of the cyclical steps in a problem-solving model. Adapted from "Curriculum-Based Measurement and Special Education Services: A Fundamental and Direct Relationship" by S.L. Deno, 1989, in M.R. Shinn, Curriculum-Based Measurement: Assessing Special Children, p.12. Copyright 1989 by The Guilford Press.

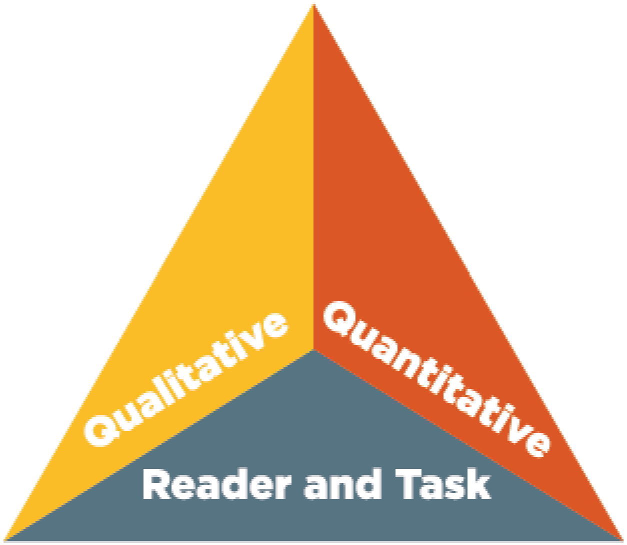


Figure 2. The approach that the CCSS takes to text complexity. It involves the interaction between qualitative dimensions of the text, quantitative dimensions of the text, and reader and task considerations. Adapted from “Appendix A: Research Supporting Key Elements of the Standards,” by National Governors Association Center for Best Practices and Council of Chief State School Officers, 2010. Retrieved from http://www.corestandards.org

Article 1: A Literature Synthesis of the Technical Adequacy of Maze-CBM for Secondary Students

A Literature Synthesis of the Technical Adequacy of Maze-CBM for Secondary Students

Marisa Mitchell

University of Maryland, College Park

Abstract

In this paper I present a synthesis of the extant research on Maze Curriculum-Based Measurement (Maze-CBM) for students in Grades 6 through 12. Fourteen studies, published between March 1993 and August 2014, met the criteria for inclusion in this study. Features of Maze-CBM administration, features of text used, and the technical adequacy of Maze-CBM are synthesized. Results suggest that although we have preliminary support of the technical adequacy for use of Maze-CBM for screening decisions, we have much less support for its use in making progress monitoring decisions. Additionally, many of the synthesized studies did not provide sufficient information on the administration of and features of text used in the Maze-CBM passages. Implications for practitioners as well as guidance for future research regarding the use of Maze-CBM with secondary students are discussed.

Keywords: curriculum-based measurement, Maze, reading, secondary students

A Literature Synthesis of the Technical Adequacy of Maze-CBM for Secondary Students

Recent educational legislation and initiatives, such as the Every Student Succeeds Act (ESSA; 2016) and the Common Core State Standards (CCSS; National Governors Association Center for Best Practices, 2010), set forth high academic expectations for all students, including those with and or at risk for disabilities, to meet prior to high school graduation. The reforms emphasize the need for students to become proficient readers—meaning that they should be able to read, understand, evaluate, and synthesize expository and narrative text. Teachers can help students meet these expectations by implementing evidence-based instruction in reading across all content areas and in supplemental reading classes. Theoretically, by meeting these standards, students will be prepared for success in post-secondary endeavors such as college and the work force.

Despite the national emphasis on high academic expectations for all students, data suggest that many students are still not learning to read at the level expected before graduation. The 2013 Nation's Report Card reports that as many as 64% of students in the United States are below proficient in reading by 8th grade (National Center for Education Statistics [NCES], 2014). These statistics demonstrate the urgent need to provide intensive, supplemental reading intervention to many students. Formative assessments can be used to (a) determine which students need reading interventions (i.e., screening), (b) measure the effectiveness of these interventions (i.e., progress monitoring), (c) make data-based decisions for instructional purposes (i.e., diagnostic), and (d) determine if students are meeting grade-level expectations.

Formative assessments are potentially useful for these reasons yet little is known about their use with high school students. The purpose of this synthesis is to examine the current literature related to the use of Maze Curriculum-Based Measurement (Maze-CBM), one type of formative assessment, at the secondary level. I provide some information about formative assessments, and specifically Maze-CBM, followed by a review of the previous literature on Maze-CBM.

Formative Assessments

Formative assessments are used to measure how well, how much, and what students are learning in response to instruction. Students commonly take these brief assessments throughout the school year to provide teachers with information about which students are not making adequate progress toward end-of-the-year goals, what skills students are still struggling with, and how far students are from meeting academic standards. Teachers can use these data to monitor student progress and make adjustments to their instruction for students who continue to fall behind. Research has confirmed that, at the elementary level, teachers who use formative assessments to examine students' growth during instruction yield higher gains for their students than those who do not use formative evaluation (e.g., Fuchs, Fuchs, Hamlett, & Ferguson, 1992; Stecker, Fuchs, & Fuchs, 2005), but this has not been investigated empirically at the secondary level. At the secondary level teachers and school personnel often use extant data such as grades, attendance, disciplinary referrals, and summative test scores to identify students who are not making adequate progress toward end-of-the-year goals. After identifying these students, schools will still need an additional tool, which is sensitive to growth, to measure the progress of students receiving supplemental instruction.

Curriculum-Based Measurement. Curriculum-Based Measurement (CBM) is one of the most widely used standardized and validated formative assessment tools for measuring students' responses to the curriculum or an intervention, and it has proliferated over the past decade (Stecker et al., 2005; Tindal, 2013; Wayman, Wallace, Wiley, Tichá, & Espin, 2007). Developed in the 1970s by Stanley Deno and colleagues as a method of measuring special education students' progress in their instructional programs (Deno, 1985), CBM was primarily used in the context of special education. However, with the advent of ESSA and Reading First initiatives (U.S. Department of Education, Office of Elementary and Secondary Education, 2002), CBM is now used to evaluate progress of all students. It is the tool most often used for evaluating students' instructional level and rate of progress in an instructional program for both screening and progress monitoring (Deno, 2003; Shinn, 1989; Tindal, 1992).

CBM measures have several characteristics that make them more useful for guiding instruction than traditional summative assessments. These measures are designed to (a) link to the curriculum; (b) serve as quick and efficient measures that are indicators of larger, more comprehensive skills; (c) include alternate, equivalent forms to be used over time to measure student progress; (d) be sensitive to student progress over time; and (e) meet high standards for reliability and validity (Shinn, 1989). Furthermore, researchers have developed CBMs in a variety of academic areas, such as early literacy, reading fluency (i.e., CBM-R), reading comprehension (e.g., Maze), writing, spelling, and mathematics (e.g., Deno, Mirkin, & Chiang, 1982; Deno, Mirkin, & Marston, 1980; Fuchs et al., 1992; Fuchs, Fuchs, Hamlett, & Stecker, 1990; Fuchs et al., 1994).

Maze-CBM. One commonly used type of CBM at the secondary level is Maze. Maze-CBM, designed to measure comprehension, is typically administered to students in Grade 4 and above (National Center on Intensive Intervention [NCII], n.d.), which is when there is a shift in our expectations of students from learning to read to reading for comprehension. The Maze-CBM requires students to silently read a passage in which words are deleted at fixed ratios and then replaced with a choice of three words (one correct replacement and two distractors) for a set number of minutes (e.g., 3 minutes). At the deletion point, students select the word that they determine is the correct replacement. An advantage associated with Maze-CBM, over other CBM types, is that it is typically administered to students as a group, which minimizes the use of instructional time for assessment purposes. Scores on Maze-CBM can be calculated in many ways, but the most common is counting the correct number of replacements in the given testing time (e.g., Espin, Deno, Maruyama, & Cohen, 1989; Fuchs et al., 1992; Jenkins & Jewell, 1993) or counting the correct number of replacements with a correction for guessing (e.g., Espin, Wallace, Lembke, Campbell, & Long, 2010; Pierce, McMaster, & Deno, 2010).

Despite development of published Maze-CBMs (e.g., Dynamic Indicators of Basic Early Literacy Skills, AIMSweb, EdCheckup, and System to Enhance Educational Performance), which typically undergo rigorous evaluation processes to determine the reliability and validity of the assessment before publishing, and current research on Maze-CBM as a tool for screening and progress monitoring in reading, there are several areas in which current Maze-CBM measures differ from one another. These differences may lead to inconsistencies of technical adequacy and thus interpretation and use of scores. For example, choice of text was not originally a large concern in early studies of CBM so a variety of sources, difficulty levels, and genres were utilized. This resulted in a large range of text difficulties and genres used across studies, published measures, and teacher-created passages. Additionally, although Fuchs and Fuchs (1992) set forth procedures for the creation of Maze-CBM passages, there is not yet consensus on procedures for development of Maze-CBM, including the interval in which words are deleted, manner in which distractor items are created, length of passage, duration of administration, and scoring procedures. As research develops regarding the use of Maze-CBM, particularly for students in Grades 6 through 12, these are important factors to consider since they may impact the technical adequacy.

Technical adequacy of Maze-CBM. In order to use Maze-CBM as a measurement tool for screening and progress monitoring for students in Grade 6 and above, we need to ensure technical adequacy of the measure for students at this age. Two important measures of technical adequacy are reliability and validity. Reliability is the desired consistency, or reproducibility, of test scores (Crocker & Algina, 2008). Current recommendations in the field of educational assessments suggest that reliability should be above .60 for data used to make decisions about a group of individuals, above .80 for screening decisions of individual students, and above .90 for high-stakes decisions involving educational placement (e.g., placement in supplemental reading intervention or special education services) of an individual student (Salvia, Ysseldyke, & Bolt, 2007). Because alternate, equivalent forms are an essential feature of Maze-CBM, it is particularly important to examine the consistency of test scores across each form—which is also known as alternate-form reliability (Salvia et al., 2007).

Validity is the extent to which a test developer and test user have collected evidence to support the types of inferences that can be made from test scores (Cronbach, 1971). Although validity of an assessment is not determined by one study alone but through an ongoing and recursive process (Messick, 1989), Wayman et al. (2007) suggest guidelines for interpreting the strength of validity coefficients for reading measures. They also suggest that validity coefficients above .70 suggest strong relations, coefficients between .50 and .70 suggest moderate relations, and coefficients below .50 suggest weak relations.

Finally, if Maze-CBM is meant to be used as a repeated assessment in order to monitor the progress of students in reading, it is important to establish slope or growth estimates for students in Grades 6 through 12. These estimates will allow practitioners to determine if students are making adequate growth in reading and thus no longer need intervention, or if students are not making adequate growth and need continued or more intensive intervention. The reliability of the slope estimate is also important to consider. It is critical that we consistently measure slope to properly determine adequate progress in an intervention. Additionally, slopes may be used to predict summative measures of college and career readiness.

Previous Reviews of CBM in the Area of Reading. There are no previous literature reviews or syntheses that focus exclusively on Maze-CBM; however, two reviews have been conducted on CBM in the general area of reading (i.e., Madelaine & Wheldall, 2004; Wayman et al., 2007). Authors of these reviews have examined the technical adequacy, utility, and state of research in the area of CBM since its development in the mid to late 1970s at the University of Minnesota's Institute for Research on Learning Disabilities. Although the authors examined all types of reading CBMs and not specifically Maze-CBM, many of the issues that relate to other reading CBM measures also impact Maze-CBM—such as principles of reliability and validity, the parallel nature of the alternative forms, and the selection of text to use on each measure.

Madelaine and Wheldall (2004) conducted a review of empirical research of reading CBM. This review examined the technical adequacy, uses, and conceptual issues in the CBM research. The authors reported that scores on CBM-R have shown high validity for screening and progress monitoring, and reliability, although they noted that the majority of studies were conducted with students at the elementary level both in special education and general education. They also found that reading CBM is used commonly to screen students, to monitor students' progress, to help determine referral for special education, and as a teacher accountability measure. Madelaine and Wheldall (2004) discussed the lack of face validity that reading CBM has with teachers. A conceptual issue they examined is the nature in which passages are selected for use in CBM measures. The authors identified that while an initial advantage of reading CBM is that passages can be taken directly from curriculum utilized in the classroom, this may cause several disadvantages—such as variations between readability of passages and familiarity of passages. Despite these important findings, the authors of this review did not evaluate each study in the review through a systematic process. Additionally, even though Maze-CBM was included, it was not the primary focus of the review.

In 2007, another review of the reading CBM literature was conducted by Wayman et al. This review examined the technical adequacy, effects of text materials, and issues around measuring growth of three commonly used reading CBM measures for students in kindergarten through Grade 12—word identification fluency (WIF), CBM-R, and Maze-CBM. The authors of this review noted that although CBM-R may be a good measure of reading growth for elementary-level students, Maze-CBM may be more sensitive to student growth in middle and high school. They also mentioned that there is a scarcity of studies examining the technical adequacy of CBM at the secondary level. In regards to text selection, Wayman et al. (2007) also noted that the text used in the creation of reading CBM measures varies across studies. Some researchers studied CBM created by using text from students' instruction, while others have used text that the students have never seen before. Wayman et al. (2007) concluded that teachers do not necessarily have to use passages from the reading material used during instruction in order to develop reliable CBM measures that are valid in measuring student reading ability. Furthermore, it is important to establish the passage equivalence for CBM progress monitoring, which may be a source of variability of data points around the slope of individual students. Although this study reported findings about the features of passages utilized in reading CBM, the researchers did not systematically code for these features. They also noted that they did not focus on Maze-CBM, but rather on other types of reading CBM, when they reported findings about features of text used.

Rationale and Research Questions

Since the reviews by Madelaine and Wheldall (2004) and Wayman et al. (2007), researchers have conducted additional empirical research on the use of CBM, particularly focusing on the use of Maze-CBM with students above Grade 5; however, the adequacy of Maze-CBM for students above Grade 5 has not been thoroughly reviewed in a synthesis to date. Additionally, few studies have examined the features of Maze-CBM test administration and features of the text utilized in the Maze-CBM assessment that may impact technical adequacy of Maze-CBM across studies. With limited research in this area and a pressing need to better understand how to measure students' progress toward meeting college and career readiness, we need to further investigate Maze-CBM as a viable tool at the secondary level. This study synthesized information across existing studies of Maze-CBM for students who are in Grades 6 through 12 to examine the features of the assessment, features of text used, and technical adequacy. I addressed the following research questions:

1. What are the features of Maze-CBM administration (e.g., duration of testing session and scoring procedures) used in studies of Maze-CBM for students in Grades 6 through 12?
2. What are the features of the text used in studies of Maze-CBM for students in Grades 6 through 12, including text length and level?

3. What is the technical adequacy as reported in studies of Maze-CBM for students in Grades 6 through 12, including reliability and validity of static scores for screening, estimates of slope, and reliability and validity of the slope for determining response to instruction?

Method

Search Procedure and Criteria

I identified studies through a multistep process that included an electronic search, a hand search, and an ancestral search of the literature. These methods have been used by several other authors in prior syntheses in the area of assessment (Reed, 2011; Reed, Cummings, Schaper, & Biancarosa, 2014). I first conducted an electronic search of the literature using ERIC, Education Research Complete, and PsychINFO databases. I used the following keywords: *curriculum based measur**, *curriculum-based measure**, *general outcome measure**, *progress monitoring*, *mastery measure**, *mainstream consultation agreement**, *interim assessment**, *mastery monitor**, *formative assessment**, *universal screening*, *benchmarking*, *norming*, and *data based decision.* I also included these additional age and grade keywords: *secondary, adolescent*, *middle school*, *high school, grade 6*, *grade 7*, *grade 8*, *grade 9*, *grade 10*, *grade 11*, *grade 12*, and *older students.* There were no date restrictions on the search in order to obtain all of the extant research in this area. Additionally, because CBM originated in the 1970s, I wanted to capture all studies of Maze-CBM since its inception.

I followed up the electronic search with a hand search to identify any articles that may not have appeared in the electronic search. I searched the following journals from January 2012 through August 2014: *School Psychology Review, Journal of School Psychology*, *Exceptional Children*, *Journal of Special Education*, *Assessment for Effective Intervention*, *Learning Disabilities Research & Practice*, *Journal of Learning Disabilities*, and *Learning Disability Quarterly*. Additionally, I conducted an archival search of the literature, including previous syntheses on reading CBM (Madelaine & Wheldall, 2004; Stecker et al., 2005; Tindal, 2013; Wayman et al., 2007), to find any studies that may not have been captured in the electronic or hand searches.

I reviewed more than 2,000 abstracts for inclusion in the synthesis based on the following criteria:

1. Studies were published in a peer-reviewed journal.
2. The articles included students in Grades 6 through 12, or included younger students but disaggregated results for students in Grades 6 through 12.
3. Studies included the use of Maze-CBM for the assessment of reading achievement. Studies may have also included other CBM measures in reading, such as CBM-R, but data from these measures were excluded from this synthesis.
4. Studies reported technical characteristics of the assessment as well as psychometric properties of reliability, validity, or slope. Studies that were solely an investigation of the effects of teachers' use of the assessment were excluded.
5. Studies were published in English and assessed students' reading skills in English.

Studies where Maze-CBM measures were used for content-specific learning, describing case studies, or analyzing qualitative studies were excluded from this review. Ultimately, through the search process, 14 studies met the criteria for inclusion in this synthesis.

Coding Procedures

I developed a comprehensive coding sheet to examine the technical characteristics of the Maze-CBM included in each study. The coding sheet contained the following information: (a) participants, (b) measures that were included (Maze-CBM plus any criterion measures), (c) features of Maze-CBM administration, (d) reliability results, (e) validity results, and (f) results regarding the slope of Maze-CBM scores.

Participants and Measures. When coding for participants and measures, I coded for the overall sample size of each study, number of students in Grades 6 through 12, number of students who were identified as having special education services, number of students who were identified as struggling readers, and number of English Language Learners (ELLs). I also reported the criterion measures used, the number of administered Maze-CBMs, and what time of year the researchers administered both the Maze-CBM and criterion measures. I included the duration of administration, manner in which deletions were created, procedures for creating distractors, and scoring guidelines as codes for features of Maze-CBM administration. I included length of text in words, source of text, if the text was on or off grade level, difficulty of text, cohesion of text, and the genre of text as codes for text features. I reported text features according to what the authors outlined; therefore, difficulty, cohesion, and genre of text are reported based on metrics described in the original studies. Furthermore, I determined the grade level of text, on or off, in one of two ways. If the author directly stated in their study that the text was on or off grade level, it was coded as such. Otherwise, I coded the text on or off grade level depending on the level of text difficulty that the authors reported. Then, I compared this level to the grade level of the participants in the study.

Reliability and Validity. I coded for several types of reliability that the authors reported in the included studies. Reliability for Maze-CBM was coded in three categories: alternate-form, delayed alternate-form, and test-retest. Although there are several types of validity that can be investigated, I coded each study for predictive and concurrent criterion validity, which are commonly investigated types of validity in this area of literature. I adopted the interpretation of reliability and validity coefficients that was suggested by Wayman et al. (2007) to examine the validity of Maze-CBM for secondary students (i.e., coefficients below .50 indicate weak relations, coefficients between .50 and .70 indicate moderate relations, and coefficients above .70 indicate strong relations).

Finally, because a common use of Maze-CBM is to measure progress of students in reading over time, I coded results of the slope or growth of Maze-CBM. Specifically, I coded for the slope of Maze-CBM, reliability of the slope, and validity of the slope.

Results

A total of 14 studies, reported in peer-reviewed journals between the March 1993 and August 2014, met criteria for inclusion in this synthesis. The majority of the studies were conducted after 2006 (*n*=11), with the remaining three studies conducted between 1993 and 1996. In these results, I report the participants and measures included in the studies, followed by features of the Maze-CBM administration and text features. Next, I report the findings on reliability and validity of the static Maze-CBM scores. Finally, I conclude by reporting the results related to the calculation of slope and its reliability and validity.

Sample and Measures

The corpus of 14 studies included 5,839 students from Grades 6 through 12 (ranging from 25 to 1,343 participants per study). Authors of several studies (Codding, Petscher, & Truckenmiller, 2014; Espin & Foegen, 1996; Espin et al., 2010; Fore, Boon, Burke, & Martin, 2009; Fore, Boon, & Martin, 2007; Fuchs, Fuchs, Hamlett, Walz, & Germann, 1993; McMaster, Wayman, & Cao, 2006; Tichá, Espin, & Wayman, 2009) included descriptions of the number of participating students with disabilities, which totaled a minimum of 202 across all students in this synthesis. The remaining authors either (a) did not report the number of participants with disabilities (Jenkins & Jewell, 1993; Silberglitt, Burns, Madyun, & Lail, 2006; Tolar, Barth, Fletcher, Francis, & Vaughn, 2014; Tolar et al., 2012) or (b) did not disaggregate this information (Pierce et al., 2010; Yeo, Fearrington, & Christ, 2012). Of the 14 total studies, authors of three studies reported the inclusion of students who were ELLs (*n*=46 at a minimum; Codding et al., 2014; McMaster et al., 2006; Tichá et al., 2009).

The majority of the studies focused on participants in Grades 6 through 8, with nine studies including students in Grade 6 (*n*=934; Espin & Foegen, 1996; Fore et al., 2009; Fore et al., 2007; Fuchs et al., 1993; Jenkins & Jewell, 1993; Pierce et al., 2010; Tolar et al., 2014; Tolar et al., 2012; Yeo et al., 2012), nine studies including students in Grade 7 (*n*=1,212; Codding et al., 2014; Espin & Foegen, 1996; Fore et al., 2009; Fore et al., 2007; Pierce et al., 2010; Silberglitt et al., 2006; Tolar et al., 2014; Tolar et al., 2012; Yeo et al., 2012), and 11 studies including students in Grade 8 (*n*=2,066; Espin & Foegen, 1996; Espin et al., 2010; Fore et al., 2009; Fore et al., 2007; McMaster et al., 2006; Pierce et al., 2010; Silberglitt et al., 2006; Tichá et al., 2009; Tolar et al., 2014; Tolar et al., 2012; Yeo et al., 2012). Only two studies included students in Grades 9, 10, and 11 (*n*=17; McMaster et al., 2006; Pierce et al., 2010), and only one study included students in Grade 12 (*n*=7; McMaster et al., 2006). Nine of the 14 studies contained samples that included students in multiple grade levels (Espin & Foegen, 1996; Fore et al., 2009; Fore et al., 2007; McMaster et al., 2006; Pierce et al., 2010; Silberglitt et al., 2006; Tolar et al., 2014; Tolar et al., 2012; Yeo et al., 2012).

Criterion measures included a variety of assessments, such as researcher-developed measures and state achievement tests across the studies. Espin and Foegen (1996) used three researcher-created measures involving multiple-choice comprehension questions as their criterion measures to compare with scores on Maze-CBM. Authors of four studies (Fore et al., 2009; Jenkins & Jewell, 1993; Tolar et al., 2014; Tolar et al., 2012) used standardized norm-referenced assessments of reading as their criterion measure. Authors of four other studies (Codding et al., 2014; Espin et al., 2010; Fore et al., 2007; Yeo et al., 2012) used results from their state achievement tests as their criterion measure. Pierce et al. (2010) used both a standardized norm-referenced assessment of reading and another CBM (CBM-R) as their criterion measures. Authors of three other studies (McMaster et al., 2006; Silberglitt et al., 2006; Tichá et al., 2009) used both a standardized norm-referenced assessment of reading and the results from the state achievement test. Fuchs et al. (1993) did not utilize criterion measures in their study.

The number of Maze-CBM forms the researchers administered to each student in each study ranged from one (Fore et al., 2009; Fore et al., 2007; Silberglitt et al., 2006) to 15 (Tolar et al., 2014; Tolar et al., 2012). Fuchs et al. (1993) did not report the number of Maze-CBM forms they administered to each student. Across all studies, the average number of Maze-CBM forms researchers administered to each student was approximately five. The time of year the researchers administered Maze-CBM ranged from fall through spring, with eight studies where researchers gave Maze-CBM to their participants in the fall (Codding et al., 2014; Espin et al., 2010; Fuchs et al., 1993; Jenkins & Jewell, 1993; Pierce et al., 2010; Tolar et al., 2014; Tolar et al., 2012; Yeo et al., 2012), six studies where researchers administered the Maze-CBM in the winter (Codding et al., 2014; Fuchs et al., 1993; McMaster et al., 2006; Tolar et al., 2014; Tolar et al., 2012; Yeo et al., 2012), and eight studies where researchers administered the Maze-CBM in the spring (Codding et al., 2014; Fuchs et al., 1993; Jenkins & Jewell, 1993; McMaster et al., 2006; Pierce et al., 2010; Tolar et al., 2014; Tolar et al., 2012; Yeo et al., 2012). In four studies, authors did not report the time of year the researchers administered Maze-CBM (Espin & Foegen, 1996; Fore et al., 2009; Fore et al., 2007; Tichá et al., 2009). Silberglitt et al. (2006)

reported that they administered Maze-CBM within 2 months of the date of the corresponding criterion measures, which were administered in the winter and spring.

Authors also administered criterion measures at different times of the year. They administered a criterion measure in the fall in four studies (Jenkins & Jewell, 1993; McMaster et al., 2006; Pierce et al., 2010; Tolar et al., 2012); in the winter in three studies (Espin et al., 2010; McMaster et al., 2006; Silberglitt et al., 2006); and then, most commonly, in the spring in seven studies (Codding et al., 2014; Jenkins & Jewell, 1993; McMaster et al., 2006; Pierce et al., 2010; Silberglitt et al., 2006; Tolar et al., 2012; Yeo et al., 2012). Additionally, authors did not report when criterion measures were administered in five studies (Espin & Foegen, 1996; Fore et al., 2009; Fore et al., 2007; Tichá et al., 2009; Tolar et al., 2014). I report detailed information about the participants and measures used across the included studies in Table 1.

Maze-CBM Features

Features of Maze-CBM Administration. The time of administration of the Maze-CBM task ranged from 1 minute (Jenkins & Jewell, 1993) to 4 minutes (Espin et al., 2010; Tichá et al., 2009)—with 3 minutes as the most common length of administration. Espin et al. (2010) and Tichá et al. (2009) examined administration of the Maze-CBM task at 2-, 3-, and 4-minute intervals.

In addition, there was some variety in the frequency of words that were deleted and replaced by the multiple-choice items. Fore et al. (2007) reported deleting every few words after the first sentence but did not outline a more precise method for this deletion. Fore et al. (2009) reported deleting every fifth word after the first sentence. Deletion of every seventh word after the first sentence was the most common frequency, with authors of nine studies following this procedure (Espin & Foegen, 1996; Espin et al., 2010; Fuchs et al., 1993; Jenkins & Jewell, 1993; McMaster et al., 2006; Silberglitt et al., 2006; Tichá et al., 2009; Tolar et al., 2012; Yeo et al., 2012). Finally, authors of three studies did not report the frequency in which words were deleted and replaced by the multiple-choice items (Codding et al., 2014; Pierce et al., 2010; Tolar et al., 2014).

Although the number of word choices at each deletion in the Maze-CBM appears to be common with three choices across studies, the description of the method used to create the two distractor items (the incorrect answer choices) was dissimilar across studies. Authors did not report the manner in which they created distractors in half of the studies (Codding et al., 2014; Fore et al., 2009; Fore et al., 2007; McMaster et al., 2006; Pierce et al., 2010; Tolar et al., 2014; Yeo et al., 2012). Authors of two studies provided minimal information about the manner in which they created distractors—for example, that they were incongruous with the context of the story (Jenkins & Jewell, 1993) or that they were comprised of one near and one far distractor (Silberglitt et al., 2006)—but did not supply further information. Espin and Foegen (1996) created distractors that were not semantically correct, did not rhyme, sound similar, or look visually similar to the correct word choice, and were no more than two letters longer or shorter than the correct answer. Fuchs et al. (1993) constructed distractors in which one was semantically correct but neither sounded or looked similar to the correct choice. In both studies by Espin and Foegen (1996) and Fuchs et al. (1993), authors also kept the length of the distractors similar to the correct choice, meaning each distractor was the same length or within one letter of the correct choice. Similarly, Espin et al. (2010) selected distractors that were within a one letter difference from the correct choice; they also ensured that distractors started with different letters and were different parts of speech than the correct choice. Tichá et al. (2009) reported that the distractors were approximately the same in length as the distractor but did not provide more specific information as to guidelines for how close in length. Tichá et al. (2009) also selected distractors that were not semantically correct and did not sound or look like the correct choice, similar to procedures by Espin and Foegen (1996) and Fuchs et al. (1993). The Tolar et al. (2012) study utilized a slightly different approach to selecting distractors. In their study, authors created one distractor which was the same part of speech as the correct word and another distractor that was not. Additionally, they created both distractors by randomly selecting from words that appeared in the rest of the passage.

The most common way to score the Maze-CBM across these studies was to calculate the number of correct choices the student circled. This method was implemented by authors in 12 studies (Codding et al., 2014; Espin & Foegen, 1996; Espin et al., 2010; Fore et al., 2009; Fore et al., 2007; Fuchs et al., 1993; Jenkins & Jewell, 1993; McMaster et al., 2006; Pierce et al., 2010; Silberglitt et al., 2006; Tichá et al., 2009; Yeo et al., 2012). Additionally, authors of two studies examined multiple methods of calculating scores to determine the impact of different scoring procedures on the Maze-CBM. Espin et al. (2010) also calculated scores on the Maze-CBM by calculating the number correct minus the number incorrect. Pierce et al. (2010) examined four discrete scoring options: (a) the number correct minus the number incorrect, (b) the number correct minus half of the number incorrect, (c) scores based on a two-error rule (i.e., discontinuing scoring after the student made two consecutive errors), and (d) scores based on a three-error rule (i.e., discontinuing scoring after the student made three consecutive errors). Pierce et al. (2010) examined these methods of scoring to investigate if they would control for the effects of guessing and thus more accurately measure the students' reading ability.

Authors of several studies also included the use of a discontinuation rule in which scoring stopped after a set number of consecutive incorrect answers. Discontinuation of scoring after three consecutive incorrect answers occurred in three studies (Espin et al., 2010; Jenkins & Jewell, 1993; Tichá et al., 2009). Pierce et al. (2010) implemented the discontinuation rule after both two and three consecutive incorrect answers to compare the two methods. The authors of the remaining ten studies did not report if they used a discontinuation rule when scoring their Maze-CBMs (Codding et al., 2014; Espin & Foegen, 1996; Fore et al., 2009; Fore et al., 2007; Fuchs et al., 1993; McMaster et al., 2006; Silberglitt et al., 2006; Tolar et al., 2014; Tolar et al., 2012; Yeo et al., 2012). In Table 2, I report a summary of the features of Maze-CBM administration across all studies.

Text Features. Authors did not report the length of text used in the Maze-CBM passages in five of the 14 studies (Codding et al., 2014; Fore et al., 2009; Fore et al., 2007; McMaster et al., 2006; Silberglitt et al., 2006). The authors who did report the length of text of Maze-CBM passages, however, used a Maze-CBM passage that was 400 words or less (Espin & Foegen, 1996; Fuchs et al., 1993; Jenkins & Jewell, 1993; Tolar et al., 2014; Tolar et al., 2012). Pierce et al. (2010) used Maze-CBM passages that were 350 words or longer and Espin et al. (2010) used ones that were 800 words or longer. Tichá et al. (2009) used passages that were, on average, 750 words in length.

AIMSweb Maze-CBM probes were the most commonly used source of text for the Maze-CBM passages in six studies (Espin & Foegen, 1996; Fore et al., 2007; Silberglitt et al., 2006; Tolar et al., 2014; Tolar et al., 2012; Yeo et al., 2012). Fore et al. (2009) and Fuchs et al. (1993) did not report the source of text for the passages they used. In two studies (Espin et al., 2010; Tichá et al., 2009), the authors developed Maze-CBM passages using newspaper articles that were both expository and narrative. Silberglitt et al. (2006) used passages that were developed using text from the Silver Burdett & Ginn reading series (Pearson et al., 1989) in addition to Maze-CBM passages from AIMSweb. The remaining researchers used a variety of other text sources to develop the Maze-CBM passages utilized in their studies (Espin & Foegen, 1996; Jenkins & Jewell, 1993; McMaster et al., 2006; Pierce et al., 2010).

When authors reported the grade level of the text used in Maze-CBM passages, below grade-level text was used most often (Espin & Foegen, 1996; Espin et al., 2010; McMaster et al., 2006; Tichá et al., 2009). Authors of three studies used on grade-level text in their Maze-CBM passages (Fuchs et al., 1993; Tolar et al., 2014; Tolar et al., 2012). Codding et al. (2014) and Pierce et al. (2010) used a combination of on grade-level and off grade-level passages for students. In these same two studies, there was a range of grade level of participants. Additionally, the authors provided on grade-level text to at least one grade of students, while the rest of the participants were provided off grade-level text. Silberglitt et al. (2006) did not report the grade level of text used in their study but did report that the difficulty of passages was standardized and equated based on Lexile scores and student performance data. In the remaining four studies, authors did not report grade level of text used in the Maze-CBM passages (Fore et al., 2009; Fore et al., 2007; Jenkins & Jewell, 1993; Yeo et al., 2012).

Authors reported information about the difficulty of the passages used in the Maze-CBM in only three studies (Espin et al., 2010; Jenkins & Jewell, 1993; Tichá et al., 2009). Jenkins and Jewell (1993) examined the difficulty of their Maze-CBM passages using the Spache readability formula. They reported a mean Spache readability of 2.3. Espin et al. (2010) and Tichá et al. (2009) reported both the Flesch-Kincaid readability level (ranging from 5th to 8th grade) and the Degrees of Reading Power level (ranging from 51 to 61). The authors of the remaining studies did not report any measures of difficulty of the Maze-CBM passages they used. Text cohesion was not reported by authors in any of the 14 studies included in the review.

Authors did not commonly report the genre of text in the description of the Maze-CBM passages in these studies. Espin and Foegen (1996) reported using expository text to create their Maze-CBM passages and Espin et al. (2010) used both narrative and expository texts. The authors of the other 12 studies (Codding et al., 2014; Fore et al., 2009; Fore et al., 2007; Fuchs et al., 1993; Jenkins & Jewell, 1993; McMaster et al., 2006; Pierce et al., 2010; Silberglitt et al., 2006; Tichá et al., 2009; Tolar et al., 2014; Tolar et al., 2012; Yeo et al., 2012) did not report the text used to create their passages. I provide a full description of the text features used in the Maze-CBM passages across all studies in Table 3.

Reliability

The authors of only one study reported alternate-form reliability (Espin et al., 2010). Espin et al. (2010) investigated the differences between scoring measures and duration of test administration (i.e., 2, 3, and 4 minutes). Alternate-form reliability coefficients ranged from .79 to .96 with a median of .86 (Espin et al., 2010).

In four studies (Codding et al., 2014; McMaster et al., 2006; Tichá et al., 2009; Yeo et al., 2012), authors reported delayed alternate-form reliability results that ranged from .52 to .89 with a median of .74.

Tolar et al. (2012) were the only authors who reported test-retest reliability coefficients. They investigated two types of Maze-CBM conditions—familiar and novel—across Grades 6, 7, and 8. In the familiar condition, students repeatedly read the same passage with each Maze-CBM administration. The students in the novel condition read new passages each time. The minimum test-retest reliability across all grades and conditions was .64 with a maximum of .91.

In the majority of studies (Espin & Foegen, 1996; Fore et al., 2009; Fore et al., 2007; Fuchs et al., 1993; Jenkins & Jewell, 1993; Pierce et al., 2010; Silberglitt et al., 2006; Tolar et al., 2014), authors did not report calculations of reliability of the Maze-CBM they used. In most of these situations, this was not expected because the focus of the study was on validity or slope calculations.

Validity

In six studies (Codding et al., 2014; Espin & Foegen, 1996; Espin et al., 2010; Fore et al., 2007; Tolar et al., 2012; Yeo et al., 2012), authors estimated the predictive validity of Maze-CBM with the correlation coefficients ranging from .44 to .81 with a median of .62.

In a total of nine studies (Codding et al., 2014; Fore et al., 2009; Jenkins & Jewell, 1993; McMaster et al., 2006; Pierce et al., 2010; Silberglitt et al., 2006; Tichá et al., 2009; Tolar et al., 2012; Yeo et al., 2012), authors reported the concurrent validity of Maze-CBM with coefficients ranging from .22 to .90 with a median of .63. In Table 4, I provide detailed information about the reliability as well as the validity reported by each study.

Slope Properties

Authors investigated growth or slope estimates of Maze-CBM in seven studies (Codding et al., 2014; Fuchs et al., 1993; McMaster et al., 2006; Tichá et al., 2009; Tolar et al., 2014; Tolar et al., 2012; Yeo et al., 2012). The authors used three different methods to estimate the slope on Maze-CBM within these studies. In three studies (Fuchs et al., 1993; McMaster et al., 2006; Tolar et al., 2012), authors used ordinary least squares (OLS) regression to calculate slope. Tichá et al. (2009) used multilevel modeling instead. Additionally, authors of three studies utilized latent growth models to determine slope of Maze-CBM (Codding et al., 2014; Tolar et al., 2014; Yeo et al., 2012).

In the studies that used OLS, the authors reported slopes ranging from 0.17 to 0.46 with a median of 0.27 correct Maze-CBM replacements per week. Tolar et al. (2014) and Yeo et al. (2012), who both used a latent growth method, reported slopes on Maze-CBM that had a larger range (–0.49 to 1.82) and a higher median (0.58) of correct replacements per week. Codding et al. (2014) also used a latent growth model—they reported a slope of 4.98 standard deviation increase of correct Maze-CBM replacements per assessment period (fall, winter, spring). Tichá et al. (2009)

calculated slope of Maze-CBM using multilevel modeling and reported a slope of 1.29 correct Maze-CBM replacements per week.

In addition to calculating the slope of Maze-CBM scores, authors of three studies (Tolar et al., 2014; Tolar et al., 2012; Yeo et al., 2012) investigated the reliability of the slope calculation by determining the standard error of the estimated slope within the latent growth model or the ratio between model-estimated true slope variance and observed variance in OLS-estimated individual slope. Tolar et al. (2012) and Yeo et al. (2012) reported standard errors that ranged from 0.04 to 0.23 with a median value of 0.17. Tolar et al. (2014) reported reliability of slope as the ratio between model-estimated true slope variance and observed variance in OLS-estimated individual slope, which ranged from .30 to .46 with a median of .33. In these three studies, researchers also examined the predictive validity of slope estimates on criterion measures. Tolar et al. (2014) and Tolar et al. (2012) reported correlations of slope to criterion measures that ranged between .11 and .55 with a median of .35 across both conditions (familiar and novel passages) and across Grades 6 through 8. Yeo et al. (2012) reported this relationship in unstandardized direct effect coefficients, which cannot be compared to the coefficients reported in Tolar et al. (2014) and Tolar et al. (2012). In Table 5, I report detailed information on the slope outcomes reported in each study.

Discussion

The purpose of this synthesis was to examine the features of administration, features of the text utilized, and the technical adequacy reported in studies of Maze-CBM for students in Grades 6 through 12. Given that many secondary students continue to perform below basic levels of reading proficiency, these students require continued reading intervention. Essential components of providing reading intervention are processes of identifying which students need the intervention and then monitoring their progress to determine if the intervention was effective. Due to the importance of using formative assessments (such as Maze-CBM) for these purposes, schools are looking for guidance on which measures to use and how to use them effectively. The National Center on Intensive Intervention (n.d.) recommends Maze-CBM for use as a formative assessment in screening and progress monitoring of students in Grade 4 and above; however, they have partial evidence to support its use in Grades 6 through 8 and no evidence of its use above Grade 8.

Despite the need for a more robust understanding of the features of Maze-CBM, which is used for students in Grades 6 through 12, we continue to lack research on it. The research that does exist is limited and lags behind the more robust evidence base we have for CBM-R in reading, which dates back over 30 years. For example, the majority of work investigating the use of Maze-CBM for students in Grades 6 through 12 has been conducted only in the past 8 years, as reflected by the majority of studies identified in this synthesis. Also, many of these studies focus on the technical adequacy of the static score, which Fuchs (2004) describes as the first stage of CBM research. Furthermore, while authors consistently reported some features of the Maze-CBM (e.g., duration of assessment and scoring methods) in the studies included in this synthesis, more evidence about several other critical features of Maze-CBM is needed.

First, authors of this corpus of studies did not consistently report critical information about participants, either in total or disaggregated by demographics of interest. For example, many authors did not disaggregate information by grade level. Although both students with disabilities and students who were ELLs were represented in the samples across the studies in this synthesis, the authors did not report specific sample numbers of these student populations. Due to this absence of reporting or disaggregation of students in special populations (e.g., students with disabilities and students who are ELLs), we lack confidence in the generalizability of the results to these students. Because Maze-CBM is a potentially useful tool to identify students in need of reading intervention and those not making progress in interventions, it is essential that we extend our knowledge about the utility of Maze-CBM to apply to all populations. This is especially important for students with disabilities and/or ELLs, as these subgroups of students commonly demonstrate reading difficulties or disabilities (NCES, 2014).

Researchers of these studies primarily focused on students in Grades 6 through 8, with only two studies that involved students above Grade 8. The scarcity of research above Grade 8 indicates that, although we can draw some reasonable conclusions about Maze-CBM for students in Grades 6 through 8, we do not yet have sufficient evidence on the technical adequacy of Maze-CBM for students in high school. This is a crucial area for future research as it is common for high school students to require continued reading intervention, and progress monitoring tools will be vital to help evaluate the effectiveness—or lack thereof—of these interventions.

Next, there were many unreported aspects of the Maze-CBM administration in the synthesized studies. For example, the manner in which distractor items were created was often vaguely discussed or not reported at all. In most of the studies, the authors did not include information about the difficulty level or genre of the text, and in no study did authors include information about the cohesion of text. The lack of information regarding the text used in Maze-CBM makes it difficult to understand what features of the measurement could potentially lead to higher reliability and validity outcomes. It is possible that variability of scores between alternate test forms is due to features of the text itself, such as the difficulty level, genre, or cohesion. If we can control these factors across alternate forms of Maze-CBM, then scores across forms may be more stable for individual students and will allow us to measure growth free of measurement error.

Reliability coefficients ranged greatly across the studies. The median alternate-form reliability was only calculated for one study and did not yield a median reliability coefficient above the level acceptable for making individual decisions regarding placement (i.e., .90; Salvia et al., 2007). The range of coefficients fell between .79 and .96, indicating that the coefficients were sometimes acceptable for making individual decisions (>.90) and other times were only acceptable for program evaluation (>.60; Salvia et al., 2007). Delayed alternate-form reliability had even lower values, showing a range of coefficients between .52 and .89 with a median of .74. Due to the fact that, at times, Maze-CBM fell below the .60 level recommended by Salvia et al. (2007), it may not be acceptable even for group decisions. The highest value of delayed alternate-form reliability was also below the acceptable reliability criterion for making individual decisions about placement. Only one study reported test-retest reliability, yet it did not yield coefficients consistently above .90. The results of reliability for Maze-CBM for students in Grades 6 through 12 indicate that we can sometimes, but not always, be sure that scores are consistent over time or across alternate forms to the level that is acceptable for making individual decisions about placement in special education or supplemental reading interventions (Salvia et al., 2007). If the primary use of Maze-CBM is to determine which students need supplemental interventions and to determine student growth in response to those interventions, we need to utilize reading measures that have levels of reliability that are at least .80 or above. Because the reliability authors reported in these studies was not above this level, we have reason to be concerned about the use of Maze-CBM for determining reading growth of students in Grades 6 through 12. There is some preliminary evidence that Maze-CBM may reach acceptable reliability levels to make group and screening decisions for students in this age group, but at this point we cannot be confident that the scores on Maze-CBM are consistent across time or alternate test forms in order to make high stakes decisions about placement of individual students.

Findings reported for validity of Maze-CBM also varied greatly across the included studies. The reported median for both predictive and concurrent validity was approximately .60, which indicates that Maze-CBM may have strong validity. Reported validity coefficients also had ranges that were quite large, with some concurrent validity coefficients falling below .30. These large ranges might be a result of the use of many different criterion measures, indicating that Maze-CBM scores may correlate strongly with some criterion measures while correlating weakly with others. The criterion measures used in these studies spanned all aspects of reading and made it difficult to determine which construct of reading Maze-CBM can be expected to correlate with. Additionally, the use of researcher-developed measures as the criterion measure in some studies and normative standardized measures in other studies makes it difficult to draw conclusions about what component of reading Maze-CBM is actually measuring. In order to use Maze-CBM as a quick and efficient measure that can indicate the global reading ability of students in Grades 6 through 12, we need to be confident in the criterion validity and in which measures of global reading ability Maze-CBM relates more strongly to. At this point, we do not have evidence of a strong relationship between Maze-CBM and global measures of reading ability for students in Grades 6 through 12. Additionally, with the current emphasis on statewide summative assessments, it would be helpful to determine the strength of correlation between Maze-CBM and the statewide summative assessments for students in Grades 6 through 12. This will help educators successfully monitor student progress toward end-of-the-year standards using Maze-CBM.

Average reported weekly slopes ranged dramatically across studies, which made it difficult to understand the estimated growth rate for students in Grades 6 through 12. Authors reported slopes as low as –.49 and as high as 1.82 correct Maze-CBM replacements per week. These scores may have ranged so dramatically due to the different calculation methods utilized across studies; they also may be due to features of the alternate forms of the Maze-CBM, differing student skill levels, or whether the students were participating in an intervention at the time they were assessed. Without a solid understanding of the estimated growth rates of students on Maze-CBM, we are not able to successfully determine which students are responding adequately to reading intervention versus those who are not, which is one of the primary purposes of the use of Maze-CBM and CBM in general. The few studies that also calculated the reliability of slope estimates within latent growth models reported a wide range of reliability estimates, which indicates a lack of consensus around the reliability of the calculation of slope within this methodology. Furthermore, the slopes calculated in these models did not lead to high levels of correlation with criterion measures. Just as we need reliability of the scores across alternate forms of Maze-CBM, we also want to ensure we are measuring individual student growth consistently over time. If we are unsure of the consistency of our slope calculations, then we cannot be sure our growth estimates are accurate. Inaccuracy of slope estimates could lead to students being incorrectly identified as responding or not responding to intervention. Incorrect identifications could then potentially result in the delivery of services when inappropriate or in situations where we neglect to provide services to students who do in fact need them. This could cause detrimental effects on students' progress toward meeting reading standards.

Limitations

There were several limitations of the current synthesis. In most studies, authors did not report information about the passages utilized in creation of Maze-CBMs. It could be that the authors of these studies did determine the level of passage difficulty or cohesion, for example, but they did not include this information in their manuscripts. Alternatively, it could be that the authors did not investigate these features of the passage when developing their Maze-CBM. Although this lack of reporting makes it difficult to synthesize results, it is important to investigate initial findings so that we might guide practitioners in their use of Maze-CBM. Additionally, results of reliability and validity estimates were not always disaggregated by grade level, making it impossible to synthesize results by grade level. Also itt is difficult to synthesize results of validity reported across studies because the criterion measures varied, typically because the authors of these studies used a criterion measure specific to the state or region. Finally, this synthesis did include reports and technical manuals that are written by test publishers. These psychometric evaluations of published assessments are typically not included in peer reviewed journals and did not meet the criteria for inclusion in this synthesis

Implications and Directions for Future Research

There are several implications for both practitioners and researchers as a result of the findings of this synthesis. The results indicate that we are at the early stages of research on Maze-CBM, and the technical adequacy for students in Grades 6 through 12 has not been well documented in the literature. The results of this review also parallel the academic progress monitoring tool chart produced by NCII, which contains evidence to support use of only three Maze-CBM assessments at or above Grade 6, with no evidence to support use of Maze-CBM above Grade 8. Although preliminary evidence suggests that Maze-CBM may be a reliable and valid indicator of reading performance for students in Grades 6 through 8, we know much less about its reliability and validity above Grade 8, which makes it more difficult to understand the technical adequacy for high school age students.

Considering that practitioners are encouraged to use Maze-CBM for both screening and progress monitoring decisions, we need to consider the standards of reliability and validity for each decision. Practitioners should exercise caution when making individual decisions regarding placement in supplemental reading interventions or special education services based on Maze-CBM scores alone. Furthermore, since there is not agreement in the literature regarding an expected growth rate for students in Grades 6 through 12, it will be difficult to determine whether an individual student is making adequate improvement in reading compared to an expected rate of growth based on Maze-CBM scores. For this reason, teachers should use a variety of measures, including Maze-CBM, when making placement decisions for an individual student. Maze-CBM, however, may be a useful tool in screening students who are at risk of not meeting grade-level reading expectations and need additional support, by using it in combination with other sources of data, such as state reading assessments.

There are several areas that researchers can also improve upon when conducting empirical studies of Maze-CBM. It would be helpful for future studies to report findings by grade level so that these results could be synthesized. Researchers should report features of text (e.g., difficulty level) and features of Maze-CBM administration (e.g., manner in which they create distractors) in future studies of Maze-CBM to allow for future examination of these Maze-CBM qualities across studies. Researchers should also disaggregate results of reliability and validity by grade level so that we can conduct an examination of patterns.

Future research needs to further investigate the reliability and validity of Maze-CBM for students in Grades 6 through 12 and to examine what features of Maze-CBM passages or Maze-CBM administration can lead to higher reliability and validity. The wide range of results for both reliability and validity indicate that Maze-CBM can be highly reliable or valid at times but less reliable and valid at other times. We should examine what features of Maze-CBM—such as difficulty level, cohesion, genre of passage, and procedures for creation of distractors—can lead to higher technical adequacy of Maze-CBM. Research should also examine the results of technical reports written by test publishers in addition to peer reviewed studies as these might provide a larger context of findings. Additionally, we need to further study validity of Maze-CBM, especially at the high school level. It would be important to also examine the degree to which different features of Maze-CBM passages relate to measures of the college and career readiness expectations set forth by ESSA and CCSS as many states transition to utilization of the Partnership for Assessment of Readiness for College and Careers (PARCC; 2014) and Smarter Balanced Assessment Consortium (n.d.) assessments. Finally, further research should examine the growth rates of Maze-CBM for students in Grades 6 through 12 and the reliability of these rates because very little is known at this time about expected growth on these measures.

Although little is known about Maze-CBM and its use for screening and progress monitoring students for reading in Grades 6 through 12, teachers are still faced with making these instructional decisions every day. Therefore, we need to proceed with investigating this assessment since it is one of few assessments that can be used for both purposes. With further research on Maze-CBM for students in Grades 6 through 12, we can continue to build evidence for use of Maze-CBM for screening and progress monitoring decisions for all students. A solid understanding of the reliability and validity of Maze-CBM static scores and growth scores will allow research to shift to the investigation of the instructional utility of Maze-CBM for both practitioners and students.

Table 1

Sample and Measures

	Sample						Measures			
Study	Total Sample	Total Students in Grades 6–12	Special Education	Struggling Readers	ELLs	Grade of Students	Criterion Measures	# of Maze-CBM Probes Administered	Time of Year Maze-CBM Administered	Time of Year Criterion Measures Administered
Codding, Petscher, & Truckenmiller, 2014	279	279	53	N/A	20	7th grade	Massachusetts Comprehensive Assessment System- English Language Arts	3	Fall, winter, & spring	Spring
Espin & Foegen, 1996	184	184	13	41	N/A	6th–8th grades	10 multiple-choice comprehension questions 10 multiple-choice item daily test from timed reading posttest of 25 multiple-choice questions	2	NR	NR
Espin, Wallace, Lembke, Campbell, & Long, 2010	236	236	21	0	N/A	8th grade	Minnesota Basic Standards Test in reading	2	Fall	Winter
Fore, Boon, Burke, & Martin, 2009	55	55	55	N/A	N/A	6th–8th grades	Woodcock Johnson III Passage Comprehension subtest Woodcock	1	NR	NR

							Johnson III Reading Fluency subtest			
Fore, Boon, & Martin, 2007	50	50	50	N/A	N/A	6th–8th grades	Criterion Referenced Competency Test (CRCT), Georgia's accountability test	1	NR	NR
Fuchs, Fuchs, Hamlett, Walz, & Germann, 1993	257	45	5	N/A	N/A	1st–6th grades	N/A	NR	Fall to spring	N/A
Jenkins & Jewell, 1993	335	125	NR	N/A	N/A	2nd–6th grades	Gates-MacGinitie Reading Tests Total Reading Comprehension	3	Fall & spring	Fall
							Metropolitan Achievement Tests (MAT) Total Reading Comprehension			Spring
McMaster, Wayman, & Cao, 2006	25	25	0	25	25	8th–12th grades	Test of Emerging Academic English (academic language proficiency)	5	Winter to spring (every 3 weeks)	Fall
							Woodcock Johnson III Letter Word ID Subtest			Winter & Spring

							Woodcock Johnson III Word Attack Subtest			Winter & Spring
							Comprehensive Reading Assessment Battery Oral Reading			Winter & Spring
							Comprehensive Reading Assessment Battery Comprehension			Winter & Spring
							Comprehensive Reading Assessment Battery Maze-CBM			Winter & Spring
							Minnesota Basic Skills Test Total			Winter
							Minnesota Basic Skills Test Literal			Winter
							Minnesota Basic Skills Test Inferential			Winter
Pierce, McMaster, & Deno, 2010	199	54	Not dis-aggregated	54	Not dis-aggregated	1st–11th grades	Kaufman Test of Educational Achievement-Second Edition Letter Word Identification	4	Fall & spring	Spring
							Kaufman Test of Educational Achievement-			Spring

							Second Edition Reading Comprehension			
							Oral Reading Fluency CBM			Fall & Spring
Silberglitt, Burns, Madyun, & Lail, 2006	5,472	1,310	NR	NR	NR	3rd, 5th, 7th, & 8th grades	Minnesota Comprehensive Assessments-Reading (state achievement test)	1	Within 2 months of date of correspondin g criterion measure	Spring
							Basic Standards Test-Reading (criterion references standardized achievement test)			Winter
Tichá, Espin, & Wayman, 2009	35	35	5	9	1	8th grade	Woodcock Johnson III Broad Reading Cluster	10	NR	NR
							Minnesota Basic Skills Test			NR
Tolar et al., 2012	1,343	1,343	NR	NR	NR	6th–8th grades	Test of Word Reading Efficiency (TOWRE) Sight Word Efficiency & Phonemic Decoding	15	Fall, winter, & spring	Beginning of year, End of year
							Woodcock Johnson III Passage Comprehension			Beginning of year, End of year

							Group Reading Assessment and Diagnostic Evaluation Passage Comprehension (GRADE)			Beginning of year, End of year
Tolar, Barth, Fletcher, Francis, & Vaughn, 2014	1,343	1,343	N/A	755	N/A	6th–8th grades	Woodcock Johnson III Passage Comprehension Test of Word Reading Efficiency (TOWRE) Sight Word Efficiency & Phonemic Decoding	15	Fall, winter, & spring	NR
Yeo, Fearrington, & Christ, 2012	1,528	755	Not dis-aggregate d	N/A	N/A	3rd–8th grades	Tennessee Comprehensive Assessment Program (TCAP)	3	Fall, winter, & spring	Spring

Note. Numbers in the sample columns for Total Sample, Total Students in Grades 6–12, Special Education, Struggling Readers, and ELLs indicate the number of students reported in each study. ELLs=English Language Learners; N/A=not applicable to the study; NR=not reported.

Table 2

Features of Maze-CBM Administration

Study	Time of Administration (e.g., 1 minute)	How Are Words Deleted?	How Many Choices at Each Deletion?	How Are Distractors Created?	Scoring Procedures	Discontinuation Rule Used? How?
Codding, Petscher, & Truckenmiller, 2014	3 minutes	NR	NR	NR	Number correct	NR
Espin & Foegen, 1996	2 minutes	Every 7th word deleted after 1st sentence	3	They were not semantically correct, did not rhyme or sound similar to the correct answer, were not visually similar to the correct choice, and were no more than two letters longer or shorter than the correct answer	Number correct	NR
Espin, Wallace, Lembke, Campbell, & Long, 2010	2, 3, and 4 minutes	Every 7th word deleted	3	Were within one letter in length of correct word, started with different letters than correct word, and had different parts of speech from correct word	Number correct and number correct minus number incorrect	Yes, discontinue scoring after 3 consecutive incorrect
Fore, Boon, Burke, & Martin, 2009	3 minutes	Every 5th word deleted after 1st sentence	3	NR	Number correct	NR

Fore, Boon, & Martin, 2007	3 minutes	Every few words deleted after 1st sentence	3	NR	Number correct	NR
Fuchs, Fuchs, Hamlett, Walz, & Germann, 1993	2.5 minutes	1st sentence was intact, every 7th word deleted after	3	One was semantically correct, not auditorally or graphically similar, and was same length or within one letter of the correct replacement	Number correct	NR
Jenkins & Jewell, 1993	1 minute	Every 7th word deleted	3	Clearly incongruous with the context of the story	Number correct	After three consecutive incorrect choices
McMaster, Wayman, & Cao, 2006	2 minutes	Every 7th word deleted after 1st sentence	3	NR	Number correct	NR
Pierce, McMaster, & Deno, 2010	2 minutes	NR	NR	NR	Number correct, Number correct minus number incorrect, Number correct minus 1/2 incorrect, Two-error rule, Three-error rule	Discontinued after two consecutive errors and after three consecutive errors
Silberglitt, Burns, Madyun, & Lail, 2006	3 minutes	Every 7th word deleted after 1st sentence	3	One near and one far distractor	Number correct	NR

Tichá, Espin, & Wayman, 2009	2, 3, and 4 minutes	1st sentence intact, every 7th word deleted after	3	Not semantically correct, auditorally and graphically different from the correct choice but approximately the same length	Number correct	After three consecutive incorrect choices
Tolar, Barth, Fletcher, Francis, & Vaughn, 2014	3 minutes	NR	3	NR	Number correct minus number incorrect	NR
Tolar et al., 2012	3 minutes	Every 7th word deleted after 1st sentence	3	One distractor is a word that is the same part of speech, the other is not the same type and is randomly selected from the passage	Number correct minus number incorrect	NR
Yeo, Fearrington, & Christ, 2012	3 minutes	Every 7th word deleted after 1st sentence	3	NR	Number correct	NR

Note. NR=not reported.

Table 3

Text Features

Study	Length of Text in Words	Source of Text	On or Off Grade Level	Readability Formula Used	Readability Level	Text Cohesion	Type of Text (Genre)
Codding, Petscher, & Truckenmiller, 2014	NR	AIMSweb	Off for all except 7th grade	NR	NR	NR	NR
Espin & Foegen, 1996	400	Timed Readings Book 2	Below grade level	NR	NR	NR	Expository
Espin, Wallace, Lembke, Campbell, & Long, 2010	≥800	Human interest stories in the local newspaper	Below grade level	Flesch-Kincaid formula Degrees of Reading Power	5th to 7th grade 51 to 61	NR	Newspaper articles (both narrative and expository)
Fore, Boon, Burke, & Martin, 2009	NR	NR	NR	NR	NR	NR	NR
Fore, Boon, & Martin, 2007	NR	AIMSweb	NR	NR	NR	NR	NR
Fuchs, Fuchs, Hamlett, Walz, & Germann, 1993	400	NR	On grade level	NR	NR	NR	NR
Jenkins & Jewell, 1993	226–313	Basic Academic Skills Samples	NR	Mean Spache	2.3	NR	NR
McMaster, Wayman, & Cao, 2006	NR	Standard Reading Passages	Below grade level	NR	NR	NR	NR

Pierce, McMaster, & Deno, 2010	≥350	Project PROACT MAZEN/ACBM Reading Passages	Below or above grade level for all but 2nd grade	NR	NR	NR	NR
Silberglitt, Burns, Madyun, & Lail, 2006	NR	From the Silver Burdett & Ginn reading series and AIMSweb	N/A	NR	NR	NR	NR
Tichá, Espin, & Wayman, 2009	Average of 750	Articles from a newspaper	Below grade level	Flesch-Kincaid formula Degrees of Reading Power	5th to 8th grade 51 to 61	NR	NR
Tolar, Barth, Fletcher, Francis, & Vaughn, 2014	150–400	AIMSweb	On grade level	NR	NR	NR	NR
Tolar et al., 2012	150–400	AIMSweb	On grade level	NR	NR	NR	NR
Yeo, Fearrington, & Christ, 2012	150–400	AIMSweb	NR	NR	NR	NR	NR

Note. N/A=not applicable to the study; NR=not reported.

Table 4

Reliability and Validity

		Reliability			Validity	
Study	Criterion Measures	Delayed Alternate-Form Reliability	Alternate-Form Reliability	Test-Retest Reliability	Predictive Validity	Concurrent Validity
Codding, Petscher, & Truckenmiller, 2014	Massachusetts Comprehensive Assessment System-English Language Arts (MCAS-EL), Fall to MCAS-EL	.80	N/A	N/A	.65	N/A
	Winter to MCAS-EL				.65	N/A
	Spring to MCAS-EL				N/A	.70
Espin & Foegen, 1996	10 multiple-choice comprehension questions	N/A	N/A	N/A	.56	N/A
	10 multiple-choice daily test from timed reading	N/A	N/A	N/A	.59	N/A
	Posttest of 25 multiple-choice questions	N/A	N/A	N/A	.62	N/A
Espin, Wallace, Lembke, Campbell, & Long, 2010	Minnesota Basic Standards Test in reading Correct choices 2 minutes	N/A	.80	N/A	.75	NR
	Correct choices 3 minutes	N/A	.86	N/A	.77	NR
	Correct choices 4 minutes	N/A	.88	N/A	.80	NR
	Correct minus incorrect 2 minutes	N/A	.79	N/A	.77	NR
	Correct minus incorrect 3 minutes	N/A	.86	N/A	.78	NR
	Correct minus incorrect 4 minutes	N/A	.96	N/A	.81	NR

Fore, Boon, Burke, & Martin, 2009	Woodcock Johnson III Passage Comprehension subtest	N/A	N/A	N/A	N/A	.89
	Woodcock Johnson III Reading Fluency subtest	N/A	N/A	N/A	N/A	.22
Fore, Boon, & Martin, 2007	Criterion-referenced competency test (CRCT)	N/A	N/A	N/A	.44	N/A
Fuchs, Fuchs, Hamlett, Walz, & Germann, 1993	N/A	N/A	N/A	N/A	N/A	N/A
Jenkins & Jewell, 1993	Gates-MacGinitie Reading Tests Total Reading	N/A	N/A	N/A	N/A	.71
	Comprehension					.68
	Metropolitan Achievement Tests (MAT) Total Reading					.67
	Comprehension					.65
McMaster, Wayman, & Cao, 2006	Test of Emerging Academic English (academic language proficiency)	Median=.80 (.75– .89)	NR	NR	NR	.59
	Woodcock Johnson III Letter Word ID Subtest	Time 1 to Time 5=.74			NR	.74
	Woodcock Johnson III Word Attack Subtest				NR	.68
	Comprehensive Reading Assessment Battery Oral Reading				NR	.85
	Comprehensive Reading Assessment Battery Comprehension				NR	.60
	Comprehensive Reading Assessment Battery Maze-CBM				NR	.90

	Minnesota Basic Skills Test Total				NR	.68
	Minnesota Basic Skills Test Literal				NR	.64
	Minnesota Basic Skills Test Inferential				NR	.48
Pierce, McMaster, & Deno, 2010	Kaufman Test of Educational Achievement-Second Edition Letter Word Identification 2 errors	N/A	NR	N/A		.70
	3 errors					.72
	All correct Maze-CBM choices					.71
	Correct Maze-CBM choices minus incorrect Maze-CBM choices					.75
	Correct Maze-CBM choices minus 1/2 incorrect Maze-CBM choices					.74
	Kaufman Test of Educational Achievement-Second Edition Reading Comprehension 2 errors					.63
	3 errors					.64
	All correct Maze-CBM choices					.61
	Correct Maze-CBM choices minus incorrect Maze-CBM choices					.69
	Correct Maze-CBM choices minus 1/2 incorrect Maze-CBM choices					.66
	Oral Reading Fluency CBM 2 errors					.72
	3 errors					.74
	All correct Maze-CBM choices					.76
	Correct Maze-CBM choices minus incorrect Maze-CBM choices					.73
	Correct Maze-CBM choices minus 1/2 incorrect Maze-CBM choices					.75
Silberglitt, Burns, Madyun, & Lail, 2006	Minnesota Comprehensive Assessments-Reading (state achievement test) 7th grade	N/A	N/A	N/A	N/A	.54

	Basic Standards Test-Reading (criterion references standardized achievement test) OR Minnesota Comprehensive Assessments-Reading (state achievement test) 8th grade				N/A	.48
Tichá, Espin, & Wayman, 2009	Woodcock Johnson III Broad Reading Cluster		N/A	N/A	N/A	
	2 minutes	.84				.86
	3 minutes	.87				.88
	4 minutes	.88				.88
	Minnesota Basic Skills Test					
	2 minutes					.80
	3 minutes					.82
	4 minutes					.85
Tolar, Barth, Fletcher, Francis, & Vaughn, 2014	Woodcock Johnson III Passage Comprehension Familiar Typical	N/A	N/A	N/A	N/A	N/A
	Familiar Struggling No Intervention	N/A	N/A	N/A	N/A	N/A
	Familiar Struggling Intervention	N/A	N/A	N/A	N/A	N/A
	Novel Typical	N/A	N/A	N/A	N/A	N/A
	Novel Struggling No Intervention	N/A	N/A	N/A	N/A	N/A
	Novel Struggling Intervention	N/A	N/A	N/A	N/A	N/A
	Test of Word Reading Efficiency (TOWRE) Sight Word Efficiency & Phonemic Decoding Familiar Typical	N/A	N/A	N/A	N/A	N/A
	Familiar Struggling No Intervention	N/A	N/A	N/A	N/A	N/A
	Familiar Struggling Intervention	N/A	N/A	N/A	N/A	N/A
	Novel Typical	N/A	N/A	N/A	N/A	N/A

	Novel Struggling No Intervention	N/A	N/A	N/A	N/A	N/A
	Novel Struggling Intervention	N/A	N/A	N/A	N/A	N/A
Tolar et al., 2012	Test of Word Reading Efficiency (TOWRE) Grade 6 Familiar	N/A	N/A	Median=.86 (.80–.87)	Wave 1=.60, Wave 2=.64, Wave 3=.63, Wave 4=.64	Wave 1=.57, Wave 5=.61
	Grade 7 Familiar	N/A	N/A	Median=.87 (.82–.91)	Wave 1=.53, Wave 2=.57, Wave 3=.57, Wave 4=.49	Wave 1=.58, Wave 5=.55
	Grade 8 Familiar	N/A	N/A	Median=.88 (.68–.78)	Wave 1=.55, Wave 2=.53, Wave 3=.54, Wave 4=.56	Wave 1=.45, Wave 5=.54
	Grade 6 Novel	N/A	N/A	Median=.75 (.69–.79)	Wave 1=.57, Wave 2=.61, Wave 3=.62, Wave 4=.58	Wave 1=.56, Wave 5=.63
	Grade 7 Novel	N/A	N/A	Median=.75 (.64–.79)	Wave 1=.50, Wave 2=.59, Wave 3=.63, Wave 4=.60	Wave 1=.48, Wave 5=.53
	Grade 8 Novel	N/A	N/A	Median=.74 (.68–.78)	Wave 1=.54, Wave 2=.58, Wave 3=.62, Wave 4=.58	Wave 1=.51, Wave 5=.62
Tolar et al., 2012	Woodcock Johnson III Passage Comprehension Grade 6 Familiar	N/A	N/A	N/A	Wave 1=.65, Wave 2=.68, Wave 3=.67, Wave 4=.67	Wave 1=.69, Wave 5=.68

Grade 7 Familiar	N/A	N/A	N/A	Wave 1=.54, Wave 2=.58, Wave 3=.60, Wave 4=.60	Wave 1=.53, Wave 5=.62
Grade 8 Familiar	N/A	N/A	N/A	Wave 1=.63, Wave 2=.65, Wave 3=.66, Wave 4=.67	Wave 1=.61, Wave 5=.68
Grade 6 Novel	N/A	N/A	N/A	Wave 1=.61, Wave 2=.61, Wave 3=.61, Wave 4=.65	Wave 1=.56, Wave 5=.63
Grade 7 Novel	N/A	N/A	N/A	Wave 1=.54, Wave 2=.63, Wave 3=.62, Wave 4=.65	Wave 1=.54, Wave 5=.61
Grade 8 Novel	N/A	N/A	N/A	Wave 1=.64, Wave 2=.70, Wave 3=.68, Wave 4=.73	Wave 1=.60, Wave 5=.69
Group Reading Assessment and Diagnostic Evaluation Passage Comprehension (GRADE) Grade 6 Familiar	N/A	N/A	N/A	Wave 1=.63, Wave 2=.69, Wave 3=.63, Wave 4=.64	Wave 1=.56, Wave 5=.62
Grade 7 Familiar	N/A	N/A	N/A	Wave 1=.59, Wave 2=.65, Wave 3=.66, Wave 4=.65	Wave 1=.56, Wave 5=.66
Grade 8 Familiar	N/A	N/A	N/A	Wave 1=.62, Wave 2=.62, Wave 3=.65, Wave 4=.68	Wave 1=.61, Wave 5=.70

	Grade 6 Novel	N/A	N/A	N/A	Wave 1=.60, Wave 2=.64, Wave 3=.62, Wave 4=.65	Wave 1=.52, Wave 5=.63
	Grade 7 Novel	N/A	N/A	N/A	Wave 1=.60, Wave 2=.56, Wave 3=.59, Wave 4=.65	Wave 1=.50, Wave 5=.60
	Grade 8 Novel	N/A	N/A	N/A	Wave 1=.51, Wave 2=.61, Wave 3=.60, Wave 4=.63	Wave 1=.59, Wave 5=.63
Yeo, Fearrington, & Christ, 2012	Tennessee Comprehensive Assessment Program (TCAP) Grade 6 Fall to TCAP	Grade 6 Fall to Winter=.63	N/A	N/A	.47	N/A
	Grade 6 Winter to TCAP	Grade 6 Fall to Spring=.69			.50	N/A
	Grade 6 Spring to TCAP	Grade 6 Winter to Spring=.66			N/A	.49
	Grade 7 Fall to TCAP	Grade 7 Fall to Winter=.52			.51	N/A
	Grade 7 Winter to TCAP	Grade 7 Fall to Spring=.74			.45	N/A
	Grade 7 Spring to TCAP	Grade 7 Winter to Spring=.64			N/A	.60
	Grade 8 Fall to TCAP	Grade 8 Fall to Winter=.74			.56	N/A

Grade 8 Winter to TCAP	Grade 8 Fall to Spring=.63	.59	N/A
Grade 8 Spring to TCAP	Grade 8 Winter to Spring=.72	N/A	.61

Note. N/A=not applicable to the study; NR=not reported.

Table 5

Growth/Slope Results

Study	Method of Slope Calculation	Average Slope per Week	Reliability of Slope	Correlations of Growth and Criterion Measures
Codding, Petscher, & Truckenmiller, 2014	Latent growth model	4.98 standard deviations per assessment period	NR	N/A
Espin & Foegen, 1996	N/A	N/A	N/A	N/A
Espin, Wallace, Lembke, Campbell, & Long, 2010	N/A	N/A	N/A	N/A
Fore, Boon, Burke, & Martin, 2009	N/A	N/A	N/A	N/A
Fore, Boon, & Martin, 2007	N/A	N/A	N/A	N/A
Fuchs, Fuchs, Hamlett, Walz, & Germann, 1993	OLS regression	0.27	NR	N/A
Jenkins & Jewell, 1993	N/A	N/A	N/A	N/A
McMaster, Wayman, & Cao, 2006	OLS regression	0.41	NR	N/A
Pierce, McMaster, & Deno, 2010	N/A	N/A	N/A	N/A
Silberglitt, Burns, Madyun, & Lail, 2006	N/A	N/A	N/A	N/A
Tichá, Espin, & Wayman, 2009	Multilevel modeling	1.29	N/A	N/A

Tolar, Barth, Fletcher, Francis, & Vaughn, 2014	Model determined slope in unconditional model	Familiar Typical=1.82 Familiar Struggling No Intervention=1.04 Familiar Struggling Intervention=1.32 Novel Typical=0.83 Novel Struggling No Intervention=0.58 Novel Struggling Intervention=0.46	Familiar Typical=0.33 Familiar Struggling No Intervention= 0.46 Familiar Struggling Intervention=0.35 Novel Typical=0.32 Novel Struggling No Intervention=NR Novel Struggling Intervention=0.3	WJ-III Passage Comprehension: Familiar Typical=0.52 Familiar Struggling No Intervention=0.43 Familiar Struggling Intervention=0.46 Novel Typical=0.4 Novel Struggling No Intervention=NR Novel Struggling Intervention=0.52 Test of Word Reading Efficiency (TOWRE) Sight Word Efficiency & Phonemic Decoding: Familiar Typical=0.36 Familiar Struggling No Intervention=0.45 Familiar Struggling Intervention=0.26 Novel Typical=0.42 Novel Struggling No Intervention=NR Novel Struggling Intervention=0.24
Tolar et al., 2012	OLS regression	Grade 6 Familiar=0.46 Grade 7 Familiar=0.46 Grade 8 Familiar=0.43	Grade 6 Familiar=0.17 Grade 7 Familiar=0.16 Grade 8 Familiar=0.18	Test of Word Reading Efficiency (TOWRE): Grade 6 Familiar=0.38 Grade 7 Familiar=0.25 Grade 8 Familiar=0.34

		Grade 6 Novel=0.21	Grade 6 Novel=0.19	Grade 6 Novel=0.30
		Grade 7 Novel=0.22	Grade 7 Novel=0.21	Grade 7 Novel=0.11
		Grade 8 Novel=0.17	Grade 8 Novel=0.23	Grade 8 Novel=0.28
				WJ-III Passage Comprehension:
				Grade 6 Familiar=0.42
				Grade 7 Familiar=0.43
				Grade 8 Familiar=0.47
				Grade 6 Novel=0.30
				Grade 7 Novel=0.17
				Grade 8 Novel=0.28
				Group Reading Assessment and Diagnostic Evaluation Passage Comprehension (GRADE):
				Grade 6 Familiar=0.32
				Grade 7 Familiar=0.42
				Grade 8 Familiar=0.55
				Grade 6 Novel=0.29
				Grade 7 Novel=0.14
				Grade 8 Novel=0.33
Yeo, Fearrington, & Christ, 2012	Model determined in linear univariate growth model	Grade 6=0.38	Grade 6=0.05	*
		Grade 7=0.49	Grade 7=0.04	
		Grade 8=–0.49	Grade 8=0.07	

Note. N/A=not applicable to the study; NR=not reported; OLS=ordinary least squares. *=Yeo et al. (2012) reported predictive validity of the Maze-CBM slope by calculating the unstandardized direct effect coefficient of their latent growth model.

Article 2: The Validity and Reliability of Maze-CBM for Reading Screening of High School Students

The Validity and Reliability of Maze-CBM for Reading Screening of High School Students

Marisa Mitchell

University of Maryland, College Park

Abstract

A growing number of high schools are adopting Response to Intervention and Multi-Tiered Systems of Support models to support students in need of reading intervention. It is critical that schools have access to measures of reading ability that can be used to quickly determine which students are at risk in reading and thus need additional reading intervention. Teachers commonly use Maze Curriculum-Based Measurement (Maze-CBM) for screening students in reading, but little research has been conducted on its utility at the high school level. This study investigated the use of Maze-CBM for screening students in Grade 9 and 10. Specifically, I examined the concurrent validity of Maze-CBM for determining which students are at risk in reading by using signal detection methods. In this paper, I also examine the alternate-form reliability of Maze-CBM. Implications and directions for future research are discussed.

Keywords: curriculum-based measurement, Maze, reading, secondary students

The Validity and Reliability of Maze-CBM for Reading Screening of High School Students

It is becoming a growing trend in the United States for secondary schools to implement schoolwide models of reading instruction, commonly known as Response to Intervention or Multi-Tiered Systems of Support (Vaughn & Fletcher, 2012). These models have three major components: (a) determining which students are at risk of reading failure (i.e., screening), (b) providing tiered levels of interventions to students who are at risk, and (c) monitoring the effectiveness of the interventions provided to students (i.e., progress monitoring). Curriculum-based measurement (CBM) is a common assessment tool utilized in both screening and progress monitoring processes, but little is known about the reliability, validity, and utility of CBM for making these decisions with high school students (Stecker, Fuchs, & Fuchs, 2005; Wayman, Wallace, Wiley, Tichá, & Espin, 2007). The purpose of this study is to examine the utility of CBM as a screening tool for high school students.

Maze-CBM

Although there are many CBMs available for use to screen and monitor the progress of students' reading ability and growth, the most commonly recommended CBM in the area of reading for students in high school is Maze-CBM (National Center on Intensive Intervention, n.d.; Torgesen & Miller, 2009). Maze-CBM is a standardized and validated formative assessment designed to measure students' general reading ability. Maze administration requires students to silently read a passage for several minutes (e.g., most commonly 3 minutes). As students read the passage, they encounter deleted words at fixed ratios; those words are then replaced with a choice of three words (one correct response and two distractors). At the deletion point, students are asked to select the word that they believe is the correct replacement in the sentence. Scores on Maze-CBM are calculated in a variety of ways, although the most common scoring protocol is to count the correct number of replacements the student made in the given testing time. Maze-CBM is intended to reflect the curricular expectations for students, which is why a variety of text types are used in the passages—including both expository and narrative. Although intended to reflect overall reading comprehension ability, Maze-CBM has been criticized for lacking face validity because it does not involve answering complex questions about the deeper meaning of text.

Using CBM to Identify Risk

At the elementary level, schools commonly use CBM to initially identify students who might be at risk and in need of supplemental intervention. At the high school level, however, the question remains whether formal screening for reading risk using Maze-CBM, or another formative assessment, is necessary. Those opposed to additional screening using CBM at the high school level would argue that by the time students are in high school, we have sufficient historical data (e.g., course grades, state assessment results, and behavior referrals) to identify students who are in need of supplemental intervention without dedicating resources for additional screening using CBM (Balfanz, Herzog, & Mac Iver, 2007; Kennelly & Monrad, 2007).

Others, such as Torgesen and Miller (2009), advocate that a formal screening process for reading is an essential component of a comprehensive assessment plan for high school students. In fact, Torgesen and Miller (2009) indicate that at the high school level, we have a greater need to determine not just who is at risk but also which students need a more intensive intervention versus a less intensive intervention. Because resources for providing supplemental intervention are often limited at the high school level, due to students' requirements to take particular courses for graduation, it is important to correctly identify those students who are truly at risk and also to what degree they need intervention. When schools can correctly identify students truly at risk, then they can direct their resources in the most efficient manner. Furthermore, screening may be necessary at the high school level when schools do not have adequate extant data on some students—for example, students who are new to a school or a school district. In these cases, schools need to conduct some type of assessment upon arrival to determine whether students need intervention. Lastly, some schools may use screening data in conjunction with existing data such as prior year's state assessment scores, class grades, attendance records, to confirm decisions made using the schools' existing data about which students need intervention. In this case schools might begin making decisions based on existing data about the students and then screen only students which are suspected of needing additional support based on this data. This confirmation is important to avoid providing additional instruction to students who do not really need it and potentially missing the opportunity to provide intervention for students who may have regressed over the summer. In sum, there are a myriad of reasons for high schools to conduct screening at the beginning of each school year.

Importance of Technical Adequacy

If schools plan to use Maze-CBM measures, it is critical that those measures are technically adequate (i.e., reliable across forms and valid for measuring reading ability). Because Maze-CBM is often administered multiple times throughout the year, multiple forms are required. We assume that these alternate forms are equivalent and therefore students' scores will remain stable regardless of the form. For this reason, before we examine the use of these alternate forms across the school year, it is important to establish the technical adequacy of the static score (i.e., the performance of a student across all forms taken at one particular point in time) to ensure that the alternate forms of the Maze-CBM function in a similar manner free of measurement error. This is particularly vital for measures like the Maze-CBM, which can be used for both screening and progress monitoring because alternate forms are commonly used. In practice, teachers administer the assessment across different times throughout the school year and need to ensure scores measure true student growth over the year without error introduced by form differences.

Previous Studies of Maze-CBM for High School Students

Despite the recommendation to use Maze-CBM to screen students at the middle and high school levels, very few studies exist that investigate the technical adequacy of Maze-CBM for students of those ages. In a recent synthesis, Mitchell & Wexler (2016) found only 14 studies that examined the reliability and validity of Maze-CBM for students in Grades 6 through 12. Of those 14 studies, only two included samples of students in high school grade levels (i.e., Grades 9–12). McMaster, Wayman, and Cao (2006) conducted a study investigating the reliability and validity of Maze-CBM for English learners in Grades 8 through 12. Their study found that delayed alternate-form reliability across 3-week administrations ranged from .75 to .89 with a median of .80. Some forms of Maze-CBM fell into the range that would be considered acceptable for making group decisions (i.e., >.60), but they did not yet reach the level acceptable for making individual decisions about placement (>.90; Salvia, Ysseldyke, & Bolt, 2007). Additionally, the concurrent correlations between Maze-CBM scores and various criterion measures of reading ranged from weak (.48) to strong (.90), indicating that Maze-CBM may be highly related to some measures of reading and not others (<.50 indicating weak relation, between .50 and .70 indicating moderate relation, and >.70 indicating strong relation; Wayman et al., 2007. Finally, using ordinary least squares regression, McMaster et al. (2006) found that the average slope per week was 0.41 correct Maze replacements.

Pierce, McMaster, and Deno (2010) also conducted a study focusing on the technical adequacy of Maze-CBM with a focus on examining the impact of scoring methods on the criterion-related validity of scores. They conducted their study with 199 students in Grades 1 through 11. Results were disaggregated for students in Grades 6 through 11 but not further disaggregated for high school age students. Additionally, all student participants were already at risk in reading or identified with a disability in reading. Due to this sample of students already identified as at risk or with a reading disability it is likely that their estimates were deflated. Pierce et al. (2010) found that for students in Grades 6 through 11, the correlation coefficients for Maze-CBM with their measure of reading comprehension (Kaufman Test of Educational Assessment–Second Edition Reading Comprehension subtest)—across all scoring methods examined—ranged from .61–.69, which indicates a moderate level of validity (.50-.70; Wayman et al., 2007).

The fact that only two studies have examined this measure for students in this age range illustrates the dearth of research that exists regarding the technical adequacy of Maze-CBM for making screening decisions about the reading abilities of students in high school. Additionally, authors of these studies utilized participants who were previously identified as at risk in reading, but they did not examine the ability of Maze-CBM to distinguish between students who are at risk and those who are not, which is the core of screening decisions. Furthermore, authors of both studies reported that they administered the Maze-CBM passages across a period of time. This makes it difficult to determine how the static score functions and whether differences in scores on each Maze-CBM are due to individual differences across times of the year or due to the impact of the passage itself. Also, neither study reported the type of text used in the creation of the Maze-CBM passages. In order to support the recommendations by the National Center on Intensive Intervention (n.d.) and Torgesen and Miller (2009), we need to further investigate the technical adequacy of Maze-CBM and its utility for predicting students' reading risk status.

Purpose

The purpose of the present study was to contribute to the limited literature base on the use of Maze-CBM for high school students. In this study, I examined the reliability of Maze-CBM passages and their validity in making screening decisions. Specifically, I aimed to answer the following research questions:

1. What is the alternate-form reliability of scores on Maze-CBM for high school students?
2. What is the validity of scores on Maze-CBM in the prediction of high school students' reading risk status?

Method

Setting and Participants

The participants in this study were students in one urban public charter school in a Mid-Atlantic state. I recruited all 9th and 10th grade students within the school to participate in the study. This school has nearly 1,000 students from pre-k through 12th grade and serves approximately 47% Hispanic students, 37% African American students, 9% White students, 4% multi-racial students, and less than 3% Asian students. Approximately 73% of the students are eligible for free or reduced-priced lunch in the school. Additionally, 14% of students receive special education services and 19% of students are English Language Learners. The school serves students in need of reading support through inclusive, co-taught English Language Arts classes led by both a general education and a special education teacher.

I reported the demographics of the 98 9th and 10th grade students who participated in this study in Table 1. The 98 students ranged in age from 13 to 17 years old and were 60% female. The sample was comprised of 53% and 47% 9th and 10th graders, respectively. Approximately 19% of the students received special education services and 12% of the students were English Language Learners. The school determined which students were English Language Learners and reported this information to me. The predominant race of students in this study was Latino (56%), with another large portion of students who were Black or African American (35%).

Measures

In order to examine the validity and reliability of Maze-CBM for high school age students, I administered nine Maze-CBM passages and obtained three criterion measures of general reading comprehension.

Maze-CBM Measures. Participating students took nine Maze-CBM passages from Content Area Reading Indicators (CARI; Abbott, Stollar, Good, & McMahon, 2014). The nine passages were taken from three triads (i.e., established sets of three Maze-CBM passages) and contained three prose passages, three science passages, and three social studies passages (see Appendix A for an example of a triad). For the purpose of this study, I used only one triad (i.e., one prose passage, one science passage, and one social studies passage) in the analysis for validity of screening and used all triads (i.e., all nine passages) for the analysis of reliability. To conduct the analysis for validity of screening, I used the fall triad that CARI recommends for the beginning of the school year because I administered these assessments to participating students at that time.

During the Maze-CBM assessment, I presented students with a reading passage in which approximately every seventh word was replaced by a multiple-choice box that included the correct word and two distractors. In these measures, all distractors were the same part of speech as the correct answer. Additionally, any verb distractors were the same tense as the correct answer and any noun distractors were singular/plural to match the correct answer. The students read the passage silently for three minutes and selected the words that best fit the sentences. I gave students two practice items prior to the nine Maze-CBM passages to ensure they understood the task.

Students were randomly assigned one of nine possible orders of the Maze-CBM passages (see Appendix B for the outline of the passage orders). I specifically ordered the Maze-CBM passages in a counterbalanced fashion. The counterbalancing of the Maze-CBM passages was done in order to control for order effects (Kline, 2009). First, I created blocks of three passages according to the genre of the passage and then assigned the blocks in a counterbalanced fashion so that some students received the prose passages first, some received the science passages first, and others received the social studies passages first. Then, I counterbalanced the three passages within each block. This counterbalanced fashion resulted in each passage appearing in each position (i.e., first, second, etc.). Between eight and 13 students took each of the nine possible passage orders.

The CARI Maze-CBM passages I used in this study are, on average, 573 words in length (range=546–602). The passages have Lexile levels between 1120L and 1300L with the mean Lexile level being 1193L. The mean Lexile level of 1193L falls into the range the Common Core State Standards recommends for students in Grade 9 and 10 (National Governors Association Center for Best Practices, 2010). I administered and scored Maze-CBM passages according to the Early Release CARI Assessment Manual (Abbott et al., 2014). According to these procedures, I did not use a discontinuation rule. I calculated adjusted Maze-CBM scores for each passage by calculating the number of correct responses minus half the number of incorrect responses and then rounding to the nearest whole number as specified in the assessment manual (Abbott et al., 2014). In Appendix C, I report full details of each of the texts used in the nine Maze-CBM measures that I administered to the participating students.

Criterion Measures. I also obtained three criterion measures of reading comprehension for each of the participating students: the Test of Silent Reading Efficiency and Comprehension (TOSREC), the Scholastic Reading Inventory (SRI), and the Preliminary SAT/National Merit Scholarship Qualifying Test (PSAT). I administered the TOSREC to the participating students and obtained scores on the SRI and PSAT from the school, which had already administered these assessments to the students.

Test of Silent Reading Efficiency and Comprehension (TOSREC). The TOSREC is a brief assessment of silent reading comprehension (Wagner, Torgesen, Rashotte, & Pearson, 2010). During this group-administered assessment, students were given three minutes to silently read sentences and determine whether the sentence was true or not. Students then circled "yes" if they thought the sentence was true or "no" if they thought it was false. I calculated raw scores on the TOSREC by subtracting the number of incorrect items from the number of correct items the student answered in three minutes. Then I assigned students a percentile rank score based on their raw scores. Publishers report the alternate-form reliability for TOSREC to be .88 for students in Grade 9 and .86 for students in Grades 10–12. They also report scorer reliability to be .99 for students in Grades 9–12 (Wagner et al., 2010).

Scholastic Reading Inventory (SRI). The SRI is a computer adaptive test, aligned with the Common Core State Standards, that is intended to measure a student's ability to comprehend both narrative and expository text (Scholastic, 2007). For this assessment, students read short passages and answered cloze format multiple-choice questions about the passages for approximately 20 to 30 minutes. Publishers calculated the reliability for the SRI using a reader consistency coefficient that was reported to be .85 for students in Grade 9 and .90 for students in Grade 10 (Scholastic, 2007).

The Preliminary SAT/National Merit Scholarship Qualifying Test (PSAT) Reading Subtest. The PSAT Reading test is a subsection of the norm-referenced PSAT test that measures students' college readiness in reading (College Board, 2015). During this assessment, students were presented with passages from history, social studies, and science fields in addition to informational graphics. They are then asked to answer multiple-choice questions about what they read. Scores on the PSAT Reading subtest range from eight to 38 points. I also obtained national percentile ranks for each participating student. PSAT Reading scores have been shown to be highly correlated to AP exam scores in English Language and English Literature (Zhang, Patel, & Ewing, 2014). Dressel-KR20 coefficients for the PSAT critical reading subtest range from .54 to .73 (Kim, Hendrickson, Patel, Melican, & Sweeney, 2014). For an overview of the measures see Appendix D.

Procedures

Assessment training and reliability. I administered both the TOSREC assessment and Maze-CBM assessments to all students. Prior to administration, I reviewed standard administration directions for both assessments as well as scoring rules. No fidelity information was collected prior to or during the administration of either test since I was the only test administrator. To aid in scoring, I hired and trained two undergraduate students. I conducted a 1-hour training session with these students prior to them scoring both tests. These research assistants were required to reach 95% accuracy of scoring with two practice assessments prior to scoring the data.

Data collection. I administered the TOSREC and Maze-CBM tests during October and November, within a 3-week time frame. I administered these assessments to all participating students during one class period. School personnel administered the PSAT test to students during their regularly scheduled time in mid-October. During the school day, at a time scheduled by the school in the month of October, school personnel also administered the SRI. I subsequently obtained the scores on PSAT and SRI from the school.

Scoring. The school scored the SRI electronically and The College Board scored the PSAT. The two undergraduate research assistants and I conducted scoring of all Maze-CBM passages and the TOSREC. We scored Maze-CBM passages according to the directions provided by Abbott et al. (2014), and we scored the TOSREC according to the administration and scoring manual (Wagner et al., 2010). Initially, I scored all Maze-CBM passages and the TOSREC; after that, the undergraduate research assistants double checked the calculated scores. Research assistants agreed with initial scoring at 95% accuracy or above. I then obtained students' percentile rank scores on the TOSREC using their raw scores according to the scoring manual.

In addition to the data I collected, I obtained Lexile levels and percentile rank scores on the SRI test and subtest and national percentile rank scores on the reading portion of the PSAT from school personnel.

Data Analysis

Descriptive statistics. Data analysis began by examining the descriptive statistics. I calculated the means and standard deviations of the adjusted Maze-CBM scores for each of the Maze-CBM measures, the adjusted scores for the combined fall triad (i.e., the sum of all three of the fall triad measures), and each criterion measure. Next, I determined the alternate-form reliability using correlational analysis. Finally, I examined the validity of Maze-CBM for screening reading by conducting correlational analysis and using signal detection methods. The details of the reliability and validity analysis are described in detail below.

Reliability. In order to examine reliability of Maze-CBM for high school students, I calculated alternate-form reliability across all nine passages which were administered in one sitting. The alternate-form reliability was computed using the Pearson product moment formula, corrected with the Fisher r to Z transformation, with each of the nine forms being correlated to each of the other eight forms (see Table 1).

Validity. In order to investigate the validity of scores on Maze-CBM in the prediction of high school students' reading ability, I conducted a correlational analysis and used signal detection methods. First, I correlated the adjusted fall triad scores with each of the criterion measures. Next, I created two dummy variables in which I dichotomized each of the criterion variables using percentile ranks. The first dummy variable dichotomized students into severe risk (below the 20th percentile rank) or not (at or above the 20th percentile rank) groups for each criterion measure. The second dummy variable dichotomized students into some risk (below the 40th percentile rank) or not (at or above the 40th percentile rank) groups for each criterion measure. Then I generated receiver operating characteristic (ROC) curves for each criterion measure, both at severe risk and some risk, and calculated the area under the curve, *A*, with its 95% confidence interval to evaluate overall accuracy with respect to each of the criterion variables of reading. This yielded six total ROC curves for this analysis.

An excellent screener will have a value for *A* at or above .950, a good screener will have an *A* value between .850 and .949, and reasonable screener will have an *A* value between .750 and .849 (Swets, 1988). Poor screeners will have *A* values below .75, as it is believed that reading screeners with this *A* value are not more valuable than teacher judgments about students' reading ability (Smolkowski, Cummings, & Strycker, 2016).

Next, I used signal detection methods to generate optimal cut scores on Maze-CBM in order to classify students who have severe risk (below the 20th percentile rank) or some risk (between the 20th and 40th percentile ranks) based on each criterion variable of reading. For the purpose of this study, I generated optimal cut scores based on the sensitivity, or the ability to detect students who truly belong to the severe and moderate risk groups. I chose to use sensitivity values to generate cut scores as opposed to the negative predictive value (NPV) because NPV is dependent on the base rate and sensitivity is not. To generate these cut scores, I set sensitivity levels to at or above .80 to ensure that the cut score correctly identified 80% of students who truly belonged to the severe or some risk groups. I made this decision a priori so that no more than 20% of students who were in need of intervention were incorrectly classified as not being at risk. This method is advantageous for schools since it is more ethical to provide reading interventions to some students who might not truly need it than to fail to provide this instruction to students who are actually at risk (Smolkowski et al., 2016). This will result in cut scores for students who are classified as having severe risk vs. not severe risk and some risk vs. not at risk for each criterion variable of reading.

In addition to the *A* values with their 95% confidence intervals and optimal cut scores for each level of risk, I calculated the sensitivity, specificity, NPV, positive predictive value (PPV), ρ, and τ of the Maze-CBM scores' prediction of each criterion measure. The specificity is the proportion of students correctly identified as not at risk on the screener who are also not at risk on the outcome. The NPV is the probability that a student who tested as not at risk on the screener is truly not at risk. The PPV then refers to the probability that a student who tested in the at risk group on the screener is truly at risk. Values for ρ indicate the base rate or the proportion of students from the entire population who are in the reading risk group. The τ value represents the proportion of students screened as positive, or below the threshold of risk.

Results

Because I used multiple assessments in this study, some of which school personnel collected, there was some missing data that was unavoidable. To examine the need for imputing missing data I used procedures similar to those of Cummings, Park, & Bauer Schaper (2012). First I examined the amount of missing data and I found that I had relatively little missing data overall (<2%). Based on the recommendation by Aday and Cornelius (2006) that it is not necessary to impute values unless 10% or more of the values are missing, I decided not to impute missing data. All of the following results, therefore, are based on analysis of complete data.

In this results section, I first present the results of descriptive analysis. Next, I present results related to the alternate-form reliability. I complete the section by reporting on the ROC curve analysis and signal detection methods.

Descriptive Statistics

Average Maze-CBM adjusted scores ranged from 11.70 to 21.16 correct Maze replacements in 3 minutes. The average fall triad adjusted Maze-CBM score was 50.99 with a standard deviation of 26.20. TOSREC raw scores averaged 23.19 with a standard deviation of 7.64. Lexile levels from the SRI averaged 956.63 with a standard deviation of 249.67. The average for PSAT Reading subtest scores was 19.72 with a standard deviation of 3.58. To summarize these details, I provide descriptive information for all nine Maze-CBM measures, the fall triad, TOSREC, SRI, and PSAT in Table 2.

Reliability

I report the alternate-form correlations for all nine Maze-CBM measures in Table 3. All correlations were significant (*p*<.01). The median alternate-form correlations across all nine Maze-CBM passages was .76 with a range of .68 to .87. The median alternate-form correlations across the three prose Maze-CBM passages was .85 with a range of .83 to .85. The median alternate-form correlations across the three science Maze-CBM passages was .81 with a range of .81 to .87. The alternate-form correlations of all three social studies Maze-CBM passages were .76. Of the 36 alternate-form correlations, only eight correlations (22%) were equal to or greater than .80.

Validity

I report the concurrent validity correlations of the fall triad scores with each of the criterion measures in Table 4. All correlations were significant (*p*<.01). The fall triad had a concurrent correlation with the TOSREC of .64, a correlation with the SRI of .70, and a correlation with the PSAT of .65.

Next, to further investigate the concurrent validity of Maze-CBM in screening students for reading risk, I plotted a ROC curve analysis. The *A* values are reported in Table 5. In Figure 1, I display the six ROC curves. *A* values ranged from .72 to .86 for the severe risk classification and from .82 to .89 for the some risk classification.

I also report the results of the signal detection analysis in Table 5. I selected optimal cut scores in this table based on the score where 80% sensitivity or greater was achieved. For the prediction of TOSREC severe risk students, the *A* value was .77 with a confidence interval of .67 to .86. The optimal cut point is 58.80, indicating that students with a score of 58.80 and below on the Maze-CBM fall triad are predicted to score below the 20th percentile rank (severe risk) on the TOSREC. The sensitivity value of that cut score is .80, indicating that 80% of students who fell below the 20th percentile rank on the TOSREC were also classified as severe risk based on the fall triad Maze-CBM scores. The specificity value of .50 for this cut point indicates that for students who scored at or above the 20th percentile rank on the TOSREC, 50% scored at or above the cut score of 58.80. The NPV of .71 indicates that 71% of students who scored above the cute point of 58.80 on the Maze-CBM were actually not below the 20th percentile rank on the TOSREC. In contrast, the PPV of .63 indicates that 63% of students who scored below the cut point of 58.80 were actually below the 20th percentile rank on the TOSREC. The base rate of .51 from the severe risk classification of the TOSREC indicates that 51% of students in the sample would be classified as having severe reading risk on the TOSREC. Finally, the positive screen fraction of .65 shows that 65% of students in the sample were screened as belonging to the severe risk group based on their adjusted Maze-CBM scores.

Discussion

In this study, I investigated the reliability and validity of scores on Maze-CBM in predicting high school students' reading risk status. I examined reliability by calculating alternate-form reliability across nine measures of Maze-CBM. Then, I examined validity of scores for making screening decisions by correlating Maze-CBM with three criterion measures of reading given at approximately the same time of year. I concluded by examining the validity of Maze-CBM scores.

I examined the median and range of the reliability estimates by comparing the values to thresholds set forth by Salvia et al. (2007). I interpreted the reliability coefficient as acceptable for making screening decisions if it was above .80 (Salvia et al., 2007). With only 22% of the alternate-form reliability correlations equal to or greater than .80, alternate forms of Maze-CBM are not at the adequate levels of stability across forms required of screening tools. Additionally, the median alternate-form correlation across all measures was .76, which is not above the adequate level for screening decisions. It is, however, interesting to note that across the alternate-form reliabilities among just prose and just science Maze-CBM passages, the median and the range of alternate-form reliabilities were acceptable for making screening decisions (.85 and .81, respectively). In contrast, the median and the range of alternate-form reliabilities among just social studies Maze-CBM passages were all below the .80 reliability level required of screening tools. This may indicate that there is sufficient stability of Maze-CBM passages for screening when they are all the same genre of text, particularly prose or science. Some variation in test scores though, and thus instability across forms, may be introduced by using multiple genres of text across Maze-CBM forms.

The results I report are similar to the delayed alternate-form correlations found by McMaster et al. (2006) with English learners. These findings may be similar due to the high percentage of English learners in the present study. However, in comparison to the alternate-form reliabilities reported by Pierce et al. (2010), the findings of this analysis were lower. This difference may be due to the fact that Pierce et al. (2010) reported alternate-form reliabilities across participants in Grades 3 through 11 so the comparison is not equivalent and his correlations may be artificially inflated due to the range of grade levels represented. Additionally, Pierce et al.

(2010) conducted their study with students who were already identified as being at risk or having disabilities in reading, which may have caused deflated reliability correlations.

In order to examine the concurrent validity of Maze-CBM for screening high school students for reading risk, I first correlated scores on the fall triad to scores on each of the criterion measures. The correlations resulting from my analysis ranged from .64 to .70. These results were consistent with those of both McMaster et al. (2006) and Pierce et al. (2010) for students in approximately the same grade levels on measures of reading comprehension (correlation coefficients were reported for Grades 8 through 12 and Grades 6 through 11, respectively, and were not disaggregated further). The results indicate that Maze-CBM scores are moderately related to each of these criterion measures of reading. These moderate correlations suggest that Maze-CBM scores are tapping into constructs of reading comprehension that are also measured by the TOSREC, SRI, and PSAT assessments.

No prior study of Maze-CBM has looked at the validity of making screening decisions of high school students in the area of reading by examining ROC curves. According to the above results, Maze-CBM has promise for adequately measuring how students would perform on other, longer, standardized criterion measures of reading comprehension. At this point we have minimal evidence in the use of Maze-CBM for classifying students into severe risk and some risk groups. The large confidence intervals indicate that practitioners should be wary of using Maze-CBM alone to determine reading risk as the *A* values dip below the minimum level of .75, which is the level where screeners are considered no more valuable than teacher judgements (Martin & Shapiro, 2011). Furthermore, due to the overlap of confidence intervals of the *A* values, it appears that Maze-CBM predicts classification on all three criterion measures equally at both the severe and some risk levels.

Limitations and Future Research

There are several limitations of this study. Although this study of Maze-CBM contains a sample of nearly three times the number of students previously studied in Grades 9 through 12, the small sample size is still a limitation. The sample size potentially explains the large confidence intervals around the estimated *A* values. The large confidence intervals at times fell below minimal acceptable levels of *A* (.75), which makes it hard to determine how well Maze-CBM functions as a screener for reading risk of high school students. Furthermore, the students who participated in this study were from one charter school from only two grade levels. Although this sample was diverse, it would be important for future research to be conducted with a larger sample of students who are representative of schools across the country (STARD Statement, 2008). Additionally, there was a high percentage of students who were at risk in reading or were ELL which may have caused a deflation of correlations reported due to a restriction of range. It is also important to investigate whether the results of this study would be similar to findings with a sample of students in Grades 11 and 12 and if Maze CBM scores would be able to predict graduation rates of high school students.

In addition, the results of this study indicate that even with Maze-CBM measures that have been carefully constructed by a team of educational measurement experts, the level of reliability across forms is not adequate for use in screening decisions. Alternate-form reliability correlations in this study indicate that forms are not exactly parallel; therefore, the form(s) a teacher administers to students could potentially impact their scores, and ultimately their classification of risk. Future research should be done to examine how features of the Maze-CBM (e.g., genre, item difficulty, background knowledge, or vocabulary) are impacting scores on Maze-CBM. Further information about sources of variance could help test developers create alternate forms that are more reliable. Researchers should consider conducting generalizability studies in order to examine the reliability of alternate forms of Maze-CBM. Additionally, further research should investigate how equating methods (Santi, Barr, Khalaf & Francis, 2015) may be used to equate test forms as a method of dealing with the variance associated with alternate forms of Maze-CBM.

Finally, in this study I focused on the concurrent validity of Maze-CBM for screening high school students in reading at the beginning of a school year. In practice, it might be more valuable for teachers to know how Maze-CBM administered at the beginning of the school year can predict scores on reading assessments at the end of the school year. Also, although I used three criterion measures that are reliable, valid for measuring reading ability, and commonly used by practitioners, it would be advantageous to investigate the predictive validity to end-of-the-year, high-stakes assessments such as state tests. Furthermore, it would be valuable to examine the predictive validity to the new Partnership for Assessment of Readiness for College and Careers (PARCC; 2014) and Smarter Balanced Assessment Consortium (n.d.) assessments as more schools are adopting these tests and using them to determine whether students are learning necessary reading skills each year.

Implications for Practice

This study has several implications for practitioners, researchers, and test developers. First and foremost, it is important to note that, currently, the literature on the use of Maze-CBM for screening high school students is extremely limited (Mitchell &Wexler, 2016). At this point, teachers and school personnel should be cautious about making decisions regarding student placement in supplemental reading classes based on results of Maze-CBM alone. Practitioners should use multiple sources of data when making these decisions. Of equal importance is that teachers should use multiple Maze-CBM measures since alternate-form reliability is low. For example, teachers could use the adjusted Maze-CBM score across three measures as was conducted in this study. The use of a median or combined score may be more accurate for measuring a student's true reading ability than one passage alone.

Because there are few Maze-CBM measures commercially available for schools to use, it may be tempting for teachers to create their own Maze-CBM assessment passages. Deno (1985) suggested that teachers could create their own measures using text from the class curriculum. Results of this study would indicate that teachers might be adding even further variation across alternate forms, thus lowering the reliability across passages even more. This is of particular concern when making moderately high-stakes decisions like providing a change of placement or supplemental intervention for students. The results of this study do, however, indicate that Maze-CBM (particularly well-constructed passages such as the CARI measures) is likely to be more effective for screening students for reading risk than teacher judgement alone. Additional research needs to be conducted on Maze-CBM's utility as a screening assessment. Further, educational measurement publishers should consider these issues when they publish such assessments for high school teachers' use. Finally, once we have a solid literature base on the use of Maze-CBM for screening high school students, we need to examine how teachers use and see value in Maze-CBM. This is of particular importance since Maze-CBM is often criticized because it lacks face validity.

In sum, findings of this study, along with the work of others, provide some preliminary evidence that Maze-CBM is a useful tool for the reading screening of high school students, but teachers should take caution when using Maze-CBM for instructional decisions. It may be necessary for teachers to use other measures of reading ability (e.g., state assessments or district assessments) in addition to Maze-CBM when making these important decisions regarding student services.

Table 1

Student Demographics

		n	%
Gender			
	Male	39	39.8
	Female	59	60.2
Grade			
	9	52	53.1
	10	46	46.9
Age			
	13	8	8.2
	14	42	42.9
	15	42	42.9
	16	4	4.1
	17	1	1.0
Special education		19	19.4
English Language Learners		12	12.2
Race/ethnicity			
	American Indian	1	1.0
	Asian	5	5.1
	Black or African American	34	34.7
	Latino	55	56.1
	Native Hawaiian or Pacific Islander	1	1.0
	White	2	2.0

Table 2

Descriptive Statistics

Measure	*n*	*M*	*SD*
Maze-CBM adjusted scores			
Passage 1	98	20.24	10.36
Passage 2	98	15.59	9.05
Passage 3	98	15.66	8.88
Passage 4	98	19.56	11.01
Passage 5	98	13.60	9.64
Passage 6	98	13.94	8.81
Passage 7	98	21.16	10.23
Passage 8	98	18.59	10.73
Passage 9	98	11.70	7.55
Fall triad	98	50.99	26.30
TOSREC percentile rank	98	25.03	21.89
TOSREC raw scores	98	23.19	7.64
SRI percentile rank	96	40.53	29.64
SRI Lexile	96	956.63	249.67
PSAT Reading percentile rank	83	26.59	22.10
PSAT Reading scores	83	19.72	3.58

Note. Fall triad is the sum of adjusted scores for Passages 1, 2, and 3. Maze-CBM = Maze Curriculum Based Measurement; TOSREC = Test of Silent Reading Efficiency and Comprehension; SRI = Scholastic Reading Inventory; PSAT = Preliminary SAT/National Merit Scholarship Qualifying Test.

Table 3

Alternate-Form Correlations for Maze-CBM Measures

Maze-CBM Form	1	2	3	4	5	6	7	8	9
1									
2	.79								
3	.80	.74							
4	.85	.79	.76						
5	.80	.87	.72	.79					
6	.79	.78	.76	.75	.76				
7	.85	.74	.77	.83	.72	.74			
8	.73	.81	.68	.68	.81	.74	.69		
9	.78	.74	.76	.78	.69	.76	.73	.70	

Table 4

Concurrent Validity Correlations

	Fall Triad	TOSREC	SRI	PSAT
Fall Triad				
TOSREC	.64			
SRI	.70	.51		
PSAT	.65	.49	.46	

Note. TOSREC = Test of Silent Reading Efficiency and Comprehension; SRI = Scholastic Reading Inventory; PSAT = Preliminary SAT/National Merit Scholarship Qualifying Test.

Table 5

Optimal Cut Scores for Maze-CBM

Statistic	TOSREC (*N*=98)	SRI (*N*=96)	PSAT (*N*=83)
Severe risk			
A (CI)	.77 (.67–.86)	.86 (.78–.94)	.72 (.61–.83)
Optimal cut point	58.50	49.00	56.50
Sensitivity	.80	.86	.81
Specificity	.50	.67	.53
NPV	.71	.92	.78
PPV	.63	.53	.57
ρ	.51	.30	.43
τ	.65	.49	.61
Some risk			
A (CI)	.86 (.73–.99)	.82 (.74–.91)	.89 (.83–.96)
Optimal cut point	63.50	54.50	61.50
Sensitivity	.83	.80	.81
Specificity	.73	.61	.75
NPV	.44	.74	.56
PPV	.95	.69	.91
ρ	.85	.52	.76
τ	.74	.60	.67

Note. TOSREC=Test of Silent Reading Efficiency and Comprehension; SRI=Scholastic Reading Inventory; PSAT=Preliminary SAT/National Merit Scholarship Qualifying Test. Optimal cut points are based on criterion measure values of 20th percentile rank for severe risk and 40th percentile rank for some risk. *A*=the area under the ROC curve; CI=confidence interval. NPV=negative predictive value; PPV=positive predictive value; ρ=base rate; and τ=proportion screened positive (scored below the optimal cut point).

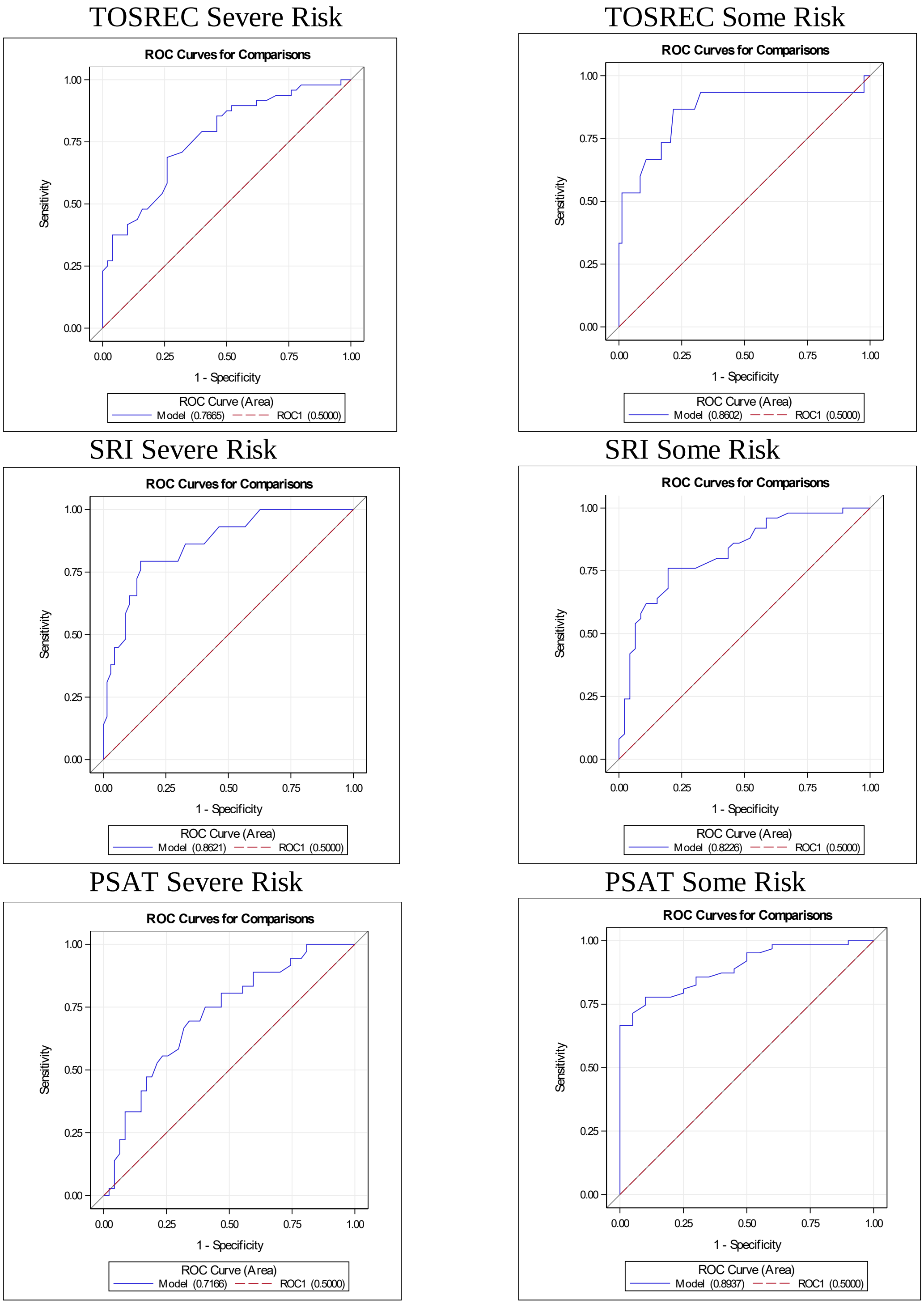


Figure 1. Receiver operating characteristic (ROC) curves for the severe risk and some risk categories of each reading assessment. TOSREC=Test of Silent Reading Efficiency and Comprehension; SRI=Scholastic Reading Inventory; PSAT=Preliminary SAT/National Merit Scholarship Qualifying Test.

Article 3: The Effect of Text Genre on Maze-CBM Scores

The Effect of Text Genre on Maze-CBM Scores

Marisa Mitchell

University of Maryland, College Park

Abstract

With many states across the nation recently adopting the Common Core State Standards (National Governors Association Center for Best Practices, 2010), teachers and schools are responsible for helping students become proficient readers across all types of text, including narrative and expository text. One way teachers can measure student progress toward meeting these standards is by administering the Maze Curriculum-Based Measurement (Maze-CBM) task to students. Despite the growing research on Maze-CBM at the elementary level, little is known about the use of the measure for high school students. In this study, I examine the influence of text genre on Maze-CBM scores for high school students. The 97 students who participated in this study were 9th and 10th grade students in one public charter school. I use a multilevel model to examine the passage effects of genre as well as student oral reading ability on Maze-CBM scores. Results, implications, and directions for future research are discussed.

Keywords: curriculum-based measurement, Maze, reading, secondary students

The Effect of Text Genre on Maze-CBM Scores

The Common Core State Standards (CCSS; National Governors Association Center for Best Practices, 2010) set new college and career ready standards for reading instruction. Although not a new expectation, the CCSS explicitly emphasize that students read a variety of text types, including both narrative and expository text. The CCSS also outline reading standards for history/social studies, science, and technical subjects for students in high school, which highlight the necessity for students to be able to read complex texts across multiple content areas (National Governors Association Center for Best Practices, 2010). An important implication of adopting these standards is that schools not only implement a rigorous reading curriculum across all content areas in high school, but they must also implement assessments that can monitor which students are making progress toward meeting these standards. One common measure used to progress monitor general reading achievement is curriculum-based measurement (CBM).

Maze-CBM

CBM, a type of standardized and validated formative assessment, is the most common tool for evaluating students' level and rate of progress in a curriculum (Deno, 2003; Shinn, 1989; Tindal, 1992). CBM is particularly useful for progress monitoring as it is a quick and efficient measure that can be used repeatedly over time due to its multiple alternate forms. Although there are several types of CBMs to monitor progress in reading, Maze Curriculum-Based Measurement (Maze-CBM) is recommended most commonly for progress monitoring in high schools because it shows the most promise for measuring growth over time for students at that age

(McMaster, 2010; Torgesen & Miller, 2009; Wayman, Wallace, Wiley, Tichá, & Espin, 2007).

Maze-CBM is recommended for progress monitoring high school students in reading for a variety of reasons. One reason is that Maze-CBM is viewed as a measure that taps into reading comprehension skills, which are the primary focus at the high school level. In addition to Maze-CBM being quick and efficient, it can be administered to a group, which may save instructional time for high school teachers who often teach many groups of students throughout the day.

Despite these advantages, however, the majority of the studies of Maze-CBM have been conducted at the elementary level (Stecker, Fuchs, & Fuchs, 2005; Wayman et al., 2007), with less information known about the technical adequacy of Maze-CBM for students above Grade 5. In fact, a recent synthesis of Maze-CBM revealed that only two studies (McMaster, Wayman, & Cao, 2006; Pierce, McMaster, & Deno, 2010) have examined the technical adequacy of Maze-CBM for students in high school (i.e., Grades 9 through 12; Mitchell & Wexler, 2016).

Another developing concern about Maze-CBM measures is the equivalency of the alternate forms (Wayman et al., 2007). The equivalency of alternate forms is often dependent on the complexity of the passages, which can be influenced by both text characteristics and student abilities. Equivalency of alternate forms of Maze-CBM passages is important to ensure that researchers are measuring student growth free of error caused by the measurement itself.

Passage Effects

To date, no study has examined the variation of Maze-CBM scores due to passage differences for students in high school; however, several studies at the elementary and middle school levels have examined passage effects in Oral Reading (OR) CBM passages (Baker et al., 2015; Christ & Ardoin, 2009; Cummings, Park, & Schaper, 2012; Francis et al., 2008; Petscher & Kim, 2011). Researchers can use this knowledge as a guide, or starting point, for future investigation of these issues at the high school level.

All of the aforementioned studies found passage effects for OR CBM and suggested controlling for these effects by equating passages. Additionally, Baker et al. (2015) attempted to study the passage effects of OR CBM in students in Grades 7 and 8. In that study, the authors found that there were form effects on OR CBM despite the attempt to control for passage differences in the creation of the alternate forms. These results parallel those of Francis et al. (2008), which found significant form effects on OR CBM measures of 2nd grade students.

Together, these studies suggest that passage effects are present in OR CBM and also Maze-CBM for students in elementary and middle school. Equating may be an option for test developers when controlling for these passage effects but this would not be a feasible option for teachers or schools who are creating their own Maze-CBM measures by selecting passages of text from their own curriculum. At this point researchers have not yet examined whether these passage effects can be generalized to students in high school. Additionally, little has been done to examine which features of the text contribute to these passage effects.

Text genre as a passage effect. There are many facets of a text that could yield these passage effects for Maze-CBM. Some features of text that could yield these passage effects are its difficulty, cohesion, topic, vocabulary, and structure. Given these potential sources of passage effects for Maze-CBM, genre of the passage text might be a preliminary place to begin investigation since it is a larger text feature that may include some of these more specific elements of text. Text genre can be defined in variety of nuanced ways. In this paper, I define genre as the content area the text comes from (i.e., prose, social studies, or science). Researchers have found differences in the text genre to impact the comprehension of the text (Denton et al., 2015; McNamara, Ozuru, & Floyd, 2011; Yoo, 2015). This relationship is particularly important at the high school level when we expect students to read and understand a variety of texts across content-area courses where they will encounter different genres. Because of this expectation, all genre texts should be represented on assessments of reading. However, inclusion of various genres of text may be problematic in assessment such as Maze-CBM where alternate passages are used in the creation of alternate forms. The differing genres across alternative Maze-CBM passages may lead to potential variation in scores on Maze-CBM—due to passage differences rather than individual ability or growth over time.

To date, no study has examined the impact of different genres of text on the scores of Maze-CBM. In fact, a synthesis of Maze-CBM for secondary students (Mitchell & Wexler, 2016) could not examine reliability of scores and validity of their use to predict reading ability in relation to genre of text since many of the included studies did not report the genre of text used in creation of the Maze-CBM passages. Furthermore, no study has empirically studied the impact of genre of text on the reliability of Maze-CBM scores and the scores' validity to predict student reading ability.

Oral Reading and Its Relationship to Reading Comprehension. In addition to variation in scores due to passage genre, it is also important to consider the variation in scores that is due to student differences. It has been shown that students' OR ability is highly predictive of reading comprehension (Fuchs, Fuchs, Hosp, & Jenkins, 2001; Schatschneider et al., 2004; Yovanoff, Duesbery, Alonzo, & Tindal, 2005). When reading text, students must be able to decode words accurately in order to access the meaning of (i.e., comprehend) text. Due to this relationship, it is likely that students' OR ability will influence students' scores on Maze-CBM in addition to the genre of text. Students' OR ability, however, should remain stable regardless of the Maze-CBM passage presented to students; for this reason, it is used as a student level attribute in this study. Thus, as the genre and topic of the Maze-CBM passage changes between each alternate form, it is important to understand how the individual student's OR ability impacts Maze-CBM scores.

Purpose

Very little research has been conducted on Maze-CBM for high school students. Furthermore, little is known about the potential passage effects, particularly due to genre of text, that impact student scores on Maze-CBM. Students' OR ability and its impact on Maze-CBM scores for high school students are also unexplored. These are important dimensions to investigate when developing Maze-CBM measures and interpreting Maze-CBM scores for decisions regarding progress monitoring of reading ability. Therefore, the purpose of this study is to answer the following questions:

1. What proportion of variance in Maze-CBM scores is due to between- and within-student differences?
2. What proportion of variance in Maze-CBM scores within individual students is attributable to genre effects?
3. Does OR ability interact with genre in interpreting Maze-CBM scores?

Method

Setting and Participants

The participants in this study were recruited from 9th and 10th grade classes in one urban public charter school in a Mid-Atlantic state. Of the nearly 1,000 pre-K through 12th grade students in the school approximately 47% are Hispanic students, 37% are African American students, 9% are White students, 4% are multiracial students, and less than 3% are Asian students. Approximately 73% of the students are eligible for free or reduced-priced lunch in the school. Additionally, the school reports that 14% of students receive special education services and 19% of students are limited/non-English proficient. Students in the school who are in need of reading support receive instruction through inclusive, co-taught English Language Arts classes taught by general education and special education teachers.

In Table 1, I report the demographics of the 97 9th and 10th grade students who participated in this study. The 97 students ranged in age from 13 to 17 years old and were 60% female. Of the students in the sample 54% were in 9th grade and 46% were in 10th grade. Approximately 20% of the students received special education services and 11% of the students were English Language Learners. A majority of the students were Latino (57%) and Black or African American (35%).

Measures

Outcome variable. Participating students took nine Maze-CBM passages from the Content Area Reading Indicators (CARI; Abbott, Stollar, Good, & McMahon, 2014). The nine passages were from three triads that CARI recommended for benchmarking, and they contained three prose passages, three science passages, and three social studies passages (see Appendix A for an example of a triad). During the Maze-CBM assessment, a reading passage was presented to students in which approximately every seventh word was replaced by a multiple-choice box that included the correct word and two distractors. In these measures, all distractors were the same part of speech as the correct answer. Additionally, any verb distractors were the same tense as the correct answer and any noun distractors were singular/plural to match the correct answer. The students read the passage silently for 3 minutes and selected the words that best belonged in the sentences. I gave students two practice items prior to the nine Maze-CBM passages. Then, they were randomly assigned one of nine possible orders of the Maze-CBM passages (see Appendix B for the outline of the passage orders). I specifically ordered the Maze-CBM passages in a counterbalanced fashion in order to avoid order effects (Kline, 2009). First, I created blocks of three passages according to the genre of the passage and then assigned the blocks in a counterbalanced fashion so some students received the prose passages first, some received the science passages first, and others received the social studies passages first. Then, I counterbalanced the three passages within each block. This counterbalanced fashion resulted in each passage appearing in each position (i.e., first through ninth).

In Appendix C, I report full details of each of the text passages used in the nine Maze-CBM measures that I administered to the participating students. The CARI Maze-CBM passages I used in this study are an average of 573 words in length (range=546–602). The passages have Lexile levels between 1120L and 1300L with a mean Lexile level of 1193L. I administered and scored Maze-CBM passages according to the Early Release CARI Assessment Manual (Abbott et al., 2014). According to these procedures, I did not use a discontinuation rule. I calculated adjusted Maze-CBM scores for each passage by subtracting half of the number of correct responses from the number of correct responses and then rounding to the nearest whole number, as specified in the Assessment Manual. I used adjusted scores on each of the nine Maze-CBMs as the outcome variable.

Predictor variables. Considering the purpose of this study is to examine the impact of OR fluency and passage genre on Maze-CBM scores, these are both used as predictors.

Content Area Reading Indicators (CARI) Oral Reading (OR). I administered three OR measures from the CARI OR benchmark materials (see Appendix E for a sample triad). The three measures contained a prose passage, a science passage, and a social studies passage. During this individually-administered assessment, research team members presented each student with three 350–400 word passages and gave 90 seconds for the student to read each passage aloud while the test administrator noted errors that the student made. I obtained scores on each passage of OR by calculating the total words the student read correctly in 90 seconds. Then, I averaged the scores from each of the three passages for each student, creating an average OR score. The average OR score was used as a student-level predictor of Maze-CBM scores. For an overview of the measures see Appendix D.

Text Genre. I assigned each passage a categorical value for text genre. The possible values were prose, social studies, and science. Genres were assigned to each passage by CARI and were used as such in this analysis. I used these categorical values to recode the variable of genre into two dummy-coded variables representing science and social studies, with prose passages as the referent category.

Procedures

Assessment training and reliability. To help with administration of assessments and scoring, I hired four research assistants. Two of the research assistants were doctoral students, both with extensive experience in reading research. These two research assistants helped me administer the OR assessments. I personally administered the Maze-CBM measure to participating students. The remaining two research assistants were undergraduate students who helped with scoring and data entry.

I trained both doctoral students in the standard administration of OR as well as scoring rules. We followed all procedures outlined by the Early Release CARI Assessment Manual (Abbott et al., 2014) in the administration and scoring of OR. Then, each doctoral student practiced administration with the other. Finally, the doctoral research assistants administered the OR assessment to each other as I used a fidelity sheet to determine the reliability of administration and scoring of the assessments. Both doctoral research assistants completed the practice administration with a reliability score of 100% before they administered the OR assessment to students participating in the study. Additionally, I observed each doctoral research assistant administering one OR assessment in the field, and they both obtained 100% reliability in this field administration.

The two undergraduate research assistants participated in a 1-hour training session prior to helping with scoring and data entry of OR and Maze-CBM.

Data collection. I administered all assessments in the months of October and November, with the help of research team members. On 1 day of the week, in one class period, I administered Maze-CBM to students in a group setting. On the following days of the week, the hired doctoral research assistants and I administered the OR assessments individually to students in the hallway outside of their classroom. I administered follow-up assessments the subsequent week for any students who were absent or who we were unable to complete testing with.

Scoring. The members of the research team conducted scoring of all measures. They scored OR and Maze-CBM measures according to the directions in the CARI Assessment Manual (Abbott et al., 2014). Once OR and Maze-CBM scoring was complete, a second member of the research team double checked the scoring.

Data Analysis

In the first step of data analysis, I ran descriptive statistics to examine means and standard deviations of Maze-CBM scores for each passage as well as the average OR scores. Next, due to the nested nature of the Maze-CBM passages within students, I conducted a two-level hierarchical linear model (HLM) analysis in several steps. The first model I ran was an unconditional model with Maze-CBM scores as the outcome. This model was used to calculate the Intraclass Correlation Coefficient (ICC) to determine the proportion of variance due to student differences using the following equation:

$$\text{ICC} = \frac{\tau}{\tau + \sigma^2}$$

After this initial model, I ran several two-level models to examine genre effects and OR fluency effects on Maze-CBM scores. I conducted all analyses in SAS PROC MIXED using maximum likelihood estimation. In these models, I used genre as a passage level predictor and mean centered mean OR scores were used as a student level predictor. I used a variance components covariance structure for Models 1–3 and 6. I used an unstructured covariance structure for Models 4, 5, and 7 (See Figure 1 for covariance structures).

Model 1

$$y_{ij} = \beta_{0j} + e_{ij}$$
$$\beta_{0j} = \gamma_{00} + u_{0j}$$

Model 2

$$y_{ij} = \beta_{0j} + e_{ij}$$
$$\beta_{0j} = \gamma_{00} + \gamma_{01}(OR_j) + u_{0j}$$

Model 3

$$y_{ij} = \beta_{0j} + \beta_{1j}(SCI_{ij}) + \beta_{2j}(SS_{ij}) + e_{ij}$$
$$\beta_{0j} = \gamma_{00} + \gamma_{01}(OR_j) + \mu_{0j}$$
$$\beta_{1j} = \gamma_{10}$$
$$\beta_{2j} = \gamma_{20}$$

Model 4

$$y_{ij} = \beta_{0j} + \beta_{1j}(SCI_{ij}) + \beta_{2j}(SS_{ij}) + e_{ij}$$
$$\beta_{0j} = \gamma_{00} + \gamma_{01}(OR_j) + \mu_{0j}$$
$$\beta_{1j} = \gamma_{10} + \mu_{1j}$$
$$\beta_{2j} = \gamma_{20} + \mu_{2j}$$

Model 5

$$y_{ij} = \beta_{0j} + \beta_{1j}(SCI_{ij}) + \beta_{2j}(SS_{ij}) + e_{ij}$$
$$\beta_{0j} = \gamma_{00} + \gamma_{01}(OR_j) + \mu_{0j}$$
$$\beta_{1j} = \gamma_{10} + \mu_{1j}$$
$$\beta_{2j} = \gamma_{20}$$

Model 6

$$y_{ij} = \beta_{0j} + \beta_{1j}(SCI_{ij}) + \beta_{2j}(SS_{ij}) + e_{ij}$$
$$\beta_{0j} = \gamma_{00} + \gamma_{01}(OR_j) + \mu_{0j}$$
$$\beta_{1j} = \gamma_{10} + \mu_{1j}$$
$$\beta_{2j} = \gamma_{20} + \mu_{2j}$$

Model 7

$$y_{ij} = \beta_{0j} + \beta_{1j}(SCI_{ij}) + \beta_{2j}(SS_{ij}) + e_{ij}$$
$$\beta_{0j} = \gamma_{00} + \gamma_{01}(OR_j) + \mu_{0j}$$
$$\beta_{1j} = \gamma_{10} + \gamma_{11}(OR_j) + \mu_{1j}$$
$$\beta_{2j} = \gamma_{20} + \gamma_{21}(OR_j) + \mu_{2j}$$

Results

I provide descriptive information for all nine Maze-CBM measures and OR in Table 2. Using the information from the unconditional model, I calculated the ICC, which is the proportion of between-group variance in Maze-CBM scores. The ICC (.67) indicates that 67% of variance in Maze-CBM scores is accounted for by individual students and 33% of variance is due to within-student differences.

In Model 2, I added OR scores at Level 2 (student level). The coefficient for the intercept was 16.74 *(p*<.001) and could be interpreted as the mean Maze-CBM score for those students at the mean OR score. The slope estimate for mean centered mean OR was 0.11 *(p*<.001), indicating that OR significantly predicted average Maze-CBM scores of students. This means that we can expect students who read an additional word read correctly on OR would be predicted to increase their Maze-CBM adjusted score by 0.11. Additionally, by comparing the variance of the intercept (τ_{00}) estimates from Model 2 with Model 1 we see that OR scores explained 58% of variance in student average Maze-CBM scores. Model 2 was statistically different from Model 1 with χ^2(1)=77.3, *p*<.001, thus Model 2 is favored over Model 1.

Next, in Model 3, I added social studies and science genre dummy codes as fixed on Level 1 (Maze-CBM passage level). The estimate for the intercept of 20.38

(*p*<.001) indicated that, across students, the average prose Maze-CBM adjusted score was 20.38, for those at the mean score of OR. The estimate for social studies was –4.37 (*p*<.001), which indicated that, within students, scores on average on social studies Maze-CBM passages were 4.37 points lower than those on prose passages, for those at the mean score of OR. The estimate for science was –6.55 (*p*<.001), which indicated that scores were 6.55 points lower than those on prose passages, for those at the mean score of OR. These differences were statistically significant for both social studies and science. Furthermore, the addition of genre (social studies and science) explained 25% of variance in average Maze-CBM scores within students. Model 3 was significantly different from Model 2 with $\chi^2(2)=222.6$, *p*<.001, which is why Model 3 is favored over Model 2.

In Model 4, I added the social studies and science genre dummy codes as random effects on Level 1 (Maze-CBM passage level) and allowed the covariance to be unstructured. The estimates for the intercept and fixed effects remained the same as in Model 3. Estimates for the random effects of science and social studies slopes were 16.03 (*p*<.001) and 9.32 (*p*=.00), respectively. This indicated that the difference in average Maze-CBM scores between science and prose and social studies and prose differed across students. Model 4 was significantly different from Model 3 with $\chi^2(5)=50$, *p*<.001, and thus Model 4 is a better fitting model than Model 3.

Next, in Model 5, I removed the social studies dummy code from the random effects, maintaining the unstructured covariances. All estimates for the fixed effects remained the same as in Model 4. The random effect for science on the slope of 9.96 was significant (*p*<.001). Comparing Model 5 to Model 4 showed that they were significantly different, $\chi^2(3)=29$, $p<.001$, so the more complex Model 4 is favored over Model 5.

Model 6 was run to see whether a variance components covariance structure would be a better fit than the unstructured variance structure. In this model, I maintained the science and social studies dummy codes as random effects as in Model 4 and fixed the variance structure. Model 6 was significantly different from Model 4, $\chi^2(3)=28.2$, $p<.001$. Because these models were significantly different, I selected to return to the unstructured variance structure in Model 4.

Finally, I investigated the impact of OR on the science and social studies slopes in Model 7. In this model, I retained all elements of Model 4 and added in OR as a predictor of science and social studies slopes. The estimate of OR on the science slope of –0.02 was not statistically significant (p=.09); however, the estimate of OR on the social studies slope of –0.03 was statistically significant (p<.00). Additionally, the OR explained only 5% of variance in the science slope and only 25% of the variance in the social studies slope. Model 7 was also statistically different from Model 4, $\chi2(2)=10.5$, p=.005. Thus, Model 7 is the final model that I will interpret in the discussion. For a summary of the results for all seven models see Table 3.

Discussion

Although no studies have examined passage differences on Maze-CBM for high school students, several studies have shown that there are passage differences for younger students in the area of OR CBM (Baker et al., 2015; Christ & Ardoin, 2009; Cummings et al., 2012; Francis et al., 2008; Petscher & Kim, 2011). In the present study, I aimed to investigate the passage effects in Maze-CBM scores by examining the impact of genre on the scores. This was a first step to understanding the impact of text features on Maze-CBM scores. Further, I examined whether OR explains some of the variance between and within students. I conducted a series of two-level HLMs to examine these questions.

Similar to studies that have investigated passage differences in OR CBM (Baker et al., 2015; Christ & Ardoin, 2009; Cummings et al., 2012; Francis et al., 2008; Petscher & Kim, 2011), my results indicate that there are passage differences for Maze-CBM. The results I present in this study suggest that the variability in passages for Maze-CBM conducted with high school students may be greater than that of OR CBM with students in younger grades. In comparison to the Baker et al. (2015) study of OR with 7th and 8th grade students, which found that 84 to 90% of variance was due to student differences and 8 to 10% of variance was due to passage differences, I found less variance (67%) due to differences in students and more variance (33%) due to passage differences with Maze-CBM administered to high school students. Although not directly comparable to this study, it is important to note that the CBM passages Baker et al. (2015) used in their study were comprised of only one genre of text (narrative). The stability of this factor across passages may have led to the lower amount of variance due to the passages they found, in comparison to the present study which included several genres of text.

Even after controlling for the effects of students' OR ability, the genre of the text explained approximately a quarter of the variance in average Maze-CBM scores within students. On average, students scored lower on both science and social studies passages in comparison to prose. These genre effects were also different across students. Based on the estimates from the variance in the science slope in Model 7, some students scored 12 points lower on science passages than they did on prose, while other students scored approximately 3 points higher on science passages versus prose. Taking into account the estimated variance in slope on social studies passages in Model 7, some students scored approximately 12 points lower on social studies passages in comparison to prose, while others had scores on social studies passages that were only about a point lower than prose. This indicates that genre effects are present within Maze-CBM passages and that those effects differ across students. These differences may be due to variables associated with students such as vocabulary knowledge or background knowledge of the particular text presented.

Another aim of this study was to examine how students' OR ability impacted Maze-CBM scores and interacted with text genre as OR ability has been linked to reading comprehension (Denton, Barth, & Fletcher, 2011). Results of this study indicate that OR does explain nearly 60% of variance in Maze-CBM scores between students. Furthermore, OR had a significant interaction with social studies compared to prose passages, but it did not have an interaction effect with science compared to prose passages. This may indicate that there are additional student characteristics, not measured by OR, that might explain the differences on Maze-CBM scores that occur within students across the genres of text. These differences might be knowledge of the topic within the passage or understanding of the meaning of vocabulary words. These are student characteristics that greatly impact the comprehension of text. Science texts in particular are often rich in content specific vocabulary, require background knowledge of topics, and are generally more complex in nature (RAND Reading Study Group, 2002; Snow, 2010).

Limitations and Future Research

There are several limitations of this study. Despite using a sample of students that is nearly three times the number of students that have previously been studied with Maze-CBM in Grades 9 through 12, the sample size is still relatively small. Additionally, although the sample was diverse, all students in the study attended the same school and many were struggling readers. This sample may have led to a restriction of range and thus underestimate the effects yielded in this analysis. In order to investigate whether the results of this study are generalizable, more research should be conducted with a larger sample of students from a variety of areas across the country. Furthermore, this sample included only Grade 9 and 10 students and not students in the upper high school grades. Future research should examine whether these results are replicable with students in Grades 11 and 12 as well.

Another limitation of this study was the narrow focus of passage effects. Although there are a multitude of text characteristics that could lead to passage effects on Maze-CBM, the focus of this study was genre of text. More specifically, it was whether the text was social studies, science, or prose. Because no previous studies of passage effects have been conducted with Maze-CBM and high school age students, in this study I focused on a general category of text type (i.e., genre) as a preliminary factor that might cause passage effects. Future research should examine whether other text characteristics (e.g., frequency of content-specific vocabulary words, cohesion of text, or text structure) also contribute to passage effects. The results I present in this study may also indicate the need to statistically equate Maze-CBM passages using item response theory methods to control for the differences, as suggested by other authors who have examined passage effects (Cummings et al., 2012; Francis et al., 2008; Petscher & Kim, 2011).

Implications for Practice

This study has several implications for practitioners, researchers, and test developers. One of the primary uses of Maze-CBM is to help teachers evaluate the level and rate of progress of student reading ability in a curriculum. In order to measure the rate of progress, teachers have to administer multiple alternate forms of Maze-CBM over the course of the year. Thus, teachers need to be confident that the alternate forms of Maze-CBM are equivalent and measure students' growth free of passage differences. The results I present in this study indicate that alternate forms of Maze-CBM are, in fact, not equivalent and that genre of the text impacts student scores.

High school teachers have few options when it comes to commercially-produced Maze-CBM measures, so they may be tempted to create their own. Deno (1985) initially proposed that teachers could create their own Maze-CBM measures using text from their class curriculum. However, the creators of CARI carefully selected and edited text for use in their Maze-CBM passages, so it is likely that Maze-CBM passages created by busy teachers will be even less equivalent. Additionally, if teachers do create their own Maze-CBM passages, they may consider using passages all chosen from the same genre of text in an attempt to create higher alternate-form equivalency. The concern with the lack of equivalency of alternative forms is that it will lead teachers to make incorrect conclusions about how students are progressing in reading throughout the year. The risk of making these incorrect conclusions is that students who need further reading support may not receive it.

Additionally, because there are very few options when it comes to commercially-produced Maze-CBM measures. Test developers need to work on developing and validating Maze-CBM measures for high school students. It is of high importance for test developers to examine form equivalence or equating methods that may help teachers better examine students' results on Maze-CBM.

Researchers should also conduct studies of the use of Maze-CBM and the effects of both text and student characteristics that lead to variation in Maze-CBM scores. I found that genre and OR did significantly impact Maze-CBM scores but there are many more specific features of the text (e.g., text structure, cohesion, or topic) and the student (e.g., vocabulary knowledge and background knowledge) that should also be investigated for effects on Maze-CBM scores. Future research should also examine what teachers and school personnel use currently in light of limited accessibility to Maze-CBM assessments for high school students. Potentially it is not important to have Maze-CBM assessments specifically for high school students and Maze-CBM intended for middle school age students would be adequate. It is also possible that teachers are able to use extant data or informal assessments to determine the progress of students throughout the year.

Currently, teachers are faced with the responsibility of assessing how students are progressing in reading throughout the year. Maze-CBM might be a viable option for doing so, but extreme caution should be taken when interpreting student scores.

Teachers should be aware of the potential passage effects, even in commercially-produced Maze-CBM measures. Finally, teachers should interpret growth on these measures with caution and use a variety of measures in addition to Maze-CBM when making decisions about the instruction students should receive.

Table 1

Student Demographics

		n	%
Gender			
	Male	39	40.2
	Female	58	59.8
Grade			
	9	52	53.6
	10	45	46.4
Age			
	13	8	8.2
	14	42	43.3
	15	42	43.3
	16	4	4.1
	17	1	1.0
Special education		19	19.6
English Language Learners		11	11.3
Race/ethnicity			
	American Indian	1	1.0
	Asian	5	5.2
	Black or African American	34	35.1
	Latino	55	56.7
	Native Hawaiian or Pacific Islander	1	1.0
	White	2	2.1

Table 2

Descriptive Statistics

Measure	*n*	*M*	*SD*
Maze-CBM adjusted scores			
Prose (33.33%)			
Passage 1	97	20.33	10.38
Passage 4	97	19.61	11.06
Passage 7	97	21.23	10.27
Science (33.33%)			
Passage 2	97	15.70	9.03
Passage 5	97	13.68	9.66
Passage 8	97	18.68	10.75
Social Studies (33.33%)			
Passage 3	97	15.67	8.93
Passage 6	97	14.01	8.82
Passage 9	97	11.82	7.49
OR total words read correctly	97	190.56	55.14

Table 3

Estimates From Multilevel Models

	Model 1		Model 2		Model 3		Model 4		Model 5		Model 6		Model 7	
	Estimate	*p*	Estimate	*p*	Estimate	*p*	Estimate	*p*	Estimate	*p*	Estimate	*p*	Estimate	*p*
Intercept (β_0)														
Intercept (γ_{00})	16.75	<.001	16.74	<.001	20.38	<.001	20.38	<.001	20.38	<.001	20.38	<.001	20.38	<.001
OR (γ_{01})			0.11	<.001	0.11	<.001	0.10	<.001	0.12	<.001	0.12	<.001	0.13	<.001
Slope on SCI (β_1)														
Intercept (γ_{10})					–4.37	<.001	–4.37	<.001	–4.37	<.001	–4.37	<.001	-4.37	<.001
OR (γ_{11})													-0.02	.09
Slope on SS (β_2)														
Intercept (γ_{20})					–6.55	<.001	–6.55	<.001	–6.55	<.001	–6.55	<.001	-6.55	<.001
OR (γ_{22})													-0.03	<.001
Random effects														
	Estimate	*p*	Estimate	*p*	Estimate	*p*	Estimate	*p*	Estimate	*p*	Estimate	*p*	Estimate	*p*
Variance in intercepts (τ_{00})	68.65	<.001	28.90	<.001	29.82	<.001	43.80	<.001	29.23	<.001	29.80	<.001	41.49	<.001
Variance in SCI slope (τ_{11})							16.03	<.001	9.96	<.001	9.99	<.001	15.17	<.001

Covariance of intercepts and SCI slope (τ_{10})							–10.17	.02	–.35	.90			-8.76	.03
Variance in SS slope (τ_{22})							9.32	<.001			1.85	.20	6.96	.01
Covariance of intercepts and SS slope (τ_{20})							–16.01	<.001					-13.67	<.001
Covariance of SCI and SS slopes (τ_{21})							7.41	.02					5.98	.03
Variance within students (σ^2)	33.49		33.49		25.14		20.65		22.65		22.13		20.65	
Model fit														
-2LL (deviance)	5830.6		5753.3		5530.7		5480.7		5509.7		5508.9		5470.2	

Note. In a reanalysis of Model 7 with social studies passages as the referent group there was a significant difference between the estimate of the fixed effects between social studies and science passages. OR=oral reading; SCI=science; SS=social studies.

Variance Components
Covariance Matrix

$$\begin{bmatrix} \tau_{00} & & \\ 0 & \tau_{11} & \\ 0 & 0 & \tau_{22} \end{bmatrix}$$

Unstructured
Covariance Matrix

$$\begin{bmatrix} \tau_{00} & & \\ \tau_{10} & \tau_{11} & \\ \tau_{20} & \tau_{21} & \tau_{22} \end{bmatrix}$$

Figure 1. The two covariance structures used within the multilevel model-building process.

Conclusion

All students, including those with and at risk for disabilities, must demonstrate proficient reading ability to meet increasingly high national academic expectations. To improve students' reading ability, secondary level teachers need to provide targeted instruction so that students are able to read and comprehend a variety of narrative and expository texts across the content areas (i.e., science, social studies, and English Language Arts). In order for teachers to determine which students are below proficient levels of reading, design instruction to meet their reading needs, and determine whether the instruction is increasing the students' reading proficiency, they need to collect data using formative assessments. Maze Curriculum-Based Measurement (Maze-CBM) is a commonly recommended standardized formative assessment to aid secondary teachers in making these data-based decisions, but little is known about its technical adequacy (i.e., reliability and validity for screening and progress monitoring) for secondary students.

In this dissertation, I attempted to examine the reliability and validity of Maze-CBM and its use for high school students. The overarching questions of the three articles were: What is the reliability and validity of scores on Maze-CBM for high school students and are the psychometric characteristics of Maze-CBM influenced by passage differences or student abilities? In the first article in this dissertation, I provided a synthesis of the current literature on Maze-CBM for students in Grades 6 through 12. Specifically, I synthesized and presented information about the features of Maze-CBM administration (e.g., scoring procedures) that were reported in studies of Maze-CBM, the features of the text used in studies of Maze-CBM, and the technical adequacy reported in studies of Maze-CBM for students in Grades 6 through 12. In the second article in this dissertation, I investigated the alternate-form reliability of Maze-CBM scores and the validity of their use for screening high school students for reading risk. In the third and final article in this dissertation, I reported on a study investigating the effect of text genre on Maze-CBM scores. In that article, I also initially explored the effect of students' oral reading ability on those scores.

In this final chapter, I summarize the findings across the three articles relating to the use of Maze-CBM for students in high school. Next, I discuss the limitations of these articles. Finally, I report the implications of the findings for teachers and school staff and discuss potential areas of future research in the area of Maze-CBM for high school students.

Findings

One major finding from Article 1 of this dissertation is that very few studies have investigated the technical adequacy of Maze-CBM for screening and progress monitoring students in Grades 6 through 12 (Mitchell & Wexler, 2016). In fact, only two studies (McMaster, Wayman, & Cao, 2006; Pierce, McMaster, & Deno, 2010) have examined the technical adequacy of Maze-CBM for high school students (i.e., Grade 9 and above). The shortage of research on Maze-CBM for students in Grades 6 through 12 is also noted by other review authors in the area of reading CBM (Madelaine & Wheldall, 2004; Wayman, Wallace, Wiley, Tichá, & Espin, 2007). This is particularly troublesome because teachers are increasingly expected to use formative assessments—such as Maze-CBM—for screening and progress monitoring students in reading, yet there is a lack of empirical evidence to suggest that Maze-CBM is reliable and valid for these purposes for students in Grades 6 through 12.

A second finding from both Articles1 and 2 is that the level of reliability of Maze-CBM is often found to be below the acceptable level for making screening decisions (.80 and above; Salvia, Ysseldyke, & Bolt, 2007) or individual decisions about placement through the process of progress monitoring (.90 and above; Salvia et al., 2007). In previous studies of Maze-CBM, reliability coefficients ranged from .52 to .96 (Mitchell & Wexler, 2016), indicating that we cannot be sure that scores on Maze-CBM are consistent over time or across alternate forms at the levels that are acceptable for making important decisions in screening or progress monitoring processes. I reported further support for this finding in Article 2 of this dissertation, in which I examined the alternate-form reliability of Maze-CBM with a sample of students in Grades 9 and 10. In that article, I found that only 22% of alternate-form reliability coefficients were equal to or greater than .80. This indicates that, in terms of screening (a low-stakes decision), Maze-CBM scores are not as parallel in nature across alternate forms as we would require of a measure used for screening.

A third important finding from this dissertation is that Maze-CBM scores show promise for being a valid measure of students reading comprehension in reading for students in Grades 6 through 12. In Article 1, I found that the medians of both predictive and current criterion validity coefficients were in the moderate range but these coefficients had a large range of values. Researchers calculated these coefficients using a variety of criterion measures so they indicate that Maze-CBM is strongly related to some reading measures but less related to others. I further examined the validity of Maze-CBM for screening students in Grades 9 and 10 in Article 2. Here, I found moderate correlations between scores on Maze-CBM and three criterion measures of reading. These results were consistent with that of McMaster et al. (2006) and Pierce et al. (2010) for students in Grades 6 through 12. Furthermore, in Article 2, I added evidence of the validity of Maze-CBM for making screening decisions by using receiver operating characteristic curves to examine the overall accuracy of using Maze-CBM to classify students into reading risk categories. Through this method, I found that it is unclear if Maze-CBM has adequate ability to classify students. In this analysis I reported large confidence intervals for the area under the curve values, *A*, so we cannot be certain of the accuracy of this result. No prior study of Maze-CBM for students in high school used this technique to examine the validity for making screening decisions. Overall, the results of this analysis provide support that Maze-CBM is not adequate for classifying students into risk categories and caution must be taken when using Maze-CBM as the only reading measure.

Finally, I was able to explore the impact of text used in Maze-CBM on students' scores in this dissertation. In Article 1, I found that in most of the prior studies of Maze-CBM for students in Grades 6 through 12, the authors did not include information related to the genre of the text. This is an important feature of Maze-CBM to investigate since students in Grades 6 through 12 are expected to proficiently read a variety of texts in multiple genres across all content-area classes (National Governors Association Center for Best Practices, 2010). In Article 3, I began preliminary investigation of the effects of text genre on Maze-CBM scores for students in Grades 9 and 10. I found there were passage effects on Maze-CBM that paralleled passage effects on other CBM measures (Baker et al., 2015; Christ & Ardoin, 2009; Cummings, Park, & Schaper, 2012; Francis et al., 2008; Petscher & Kim, 2011). Additionally, I found that text genre had significant effects on Maze-CBM scores after controlling for student oral reading ability and that these effects were different across students. This supports the theories that the genre of the passage used in the Maze-CBM measure impacts a student's score on that particular measure and that genre effects differ across students.

Ultimately, the three articles in this dissertation illustrate that research on Maze-CBM for students in Grades 6 through 12 is still in the preliminary stage. The evidence from my dissertation adds to the preliminary idea that there is some support for the use of Maze-CBM for students at the secondary level; however, there are limitations and important implications that need to be considered.

Limitations

One limitation of this dissertation is that there are few prior studies of the use of Maze-CBM for students in Grades 9 through 12. Thus, I wrote Article 1 on literature related to Maze-CBM in Grades 6 through 12 and used this corpus of studies, though limited, to design the empirical studies conducted for Articles 2 and 3. Additionally, the authors of studies that I reviewed in Article 1 did not always disaggregate results by grade level so I was not able to parse out results specific to Grades 9 through 12. Furthermore, authors of many studies included in Article1 did not report features of the text used in their Maze-CBM passages, which made it difficult to synthesize results regarding text features. Because of this, I had to use evidence from studies on other types of CBM or text features in general when designing Article3, which investigated genre effects of Maze-CBM.

A second limitation that impacted the empirical articles in this dissertation (i.e., Articles 2 and 3) was the limited sample size. Although the sample size used in these articles was nearly three times the number of students in Grades 9 through 12 than previously studied in the Maze-CBM literature, the sample size is still considered limited and, therefore, effects the results of these studies. For example, the small sample size potentially explains the large confidence intervals around the estimated *A* values in Article 2. This led to confidence intervals that had large overlaps with each other and that fell below the minimal acceptable levels of *A*. This made it difficult to determine how well Maze-CBM functions as a screening tool for reading risk of high school students. Additionally, the students who participated in this study were all from one charter school and only two grade levels. Despite the ethnically diverse sample, it is questionable whether the results from these articles could be replicated with a sample of students that is more representative of schools across the country.

Finally, in Article 3, I maintained a narrow focus into the potential passage effects in Maze-CBM. I chose to examine just one dimension of text complexity that had potential for causing passage effects; however, text complexity can involve an interaction between various dimensions of text, such as qualitative dimensions, quantitative dimensions, and reading and task considerations (Appendix A, National Governors Association Center for Best Practices, 2010). In Article 3, I attempted to conduct a preliminary examination of passage effects for Maze-CBM by choosing genre as the text characteristic impacting passages effects. Furthermore, in that same article, I was focused on comparing social studies and science genres to prose and not to each other. Because I found effects of genre, a general category of text complexity, it is likely that there are more specific categories of text complexity that could also be a source of the passage effects.

Implications

The results of the three articles in this dissertation have several important implications for researchers, test developers, and researchers. The results indicate that the research on Maze-CBM for high school students is still in the early stages. Prior to this dissertation, there were only two studies (McMaster et al., 2006; Pierce et al., 2010) that investigated the technical adequacy of Maze-CBM for high school students, and those studies contained only 38 students in Grades 9 through 12. Although this dissertation contributed to the empirical literature in this area, the overall sample of students with whom we have investigated Maze-CBM still remains very small.

Due to the limited literature on the use of Maze-CBM for high school students much more research needs to be done in this area. Researchers need to examine further the reliability and validity of the use of Maze-CBM for both screening and progress monitoring high school students in reading. Researchers may want to examine the potential effects of text and student characteristics on Maze-CBM scores and begin investigation into possible ways to control for these effects in the creation of Maze-CBM passages as well as equating methods that could be used similarly to that done with OR.

The largest implication for test developers is that there is a great need to develop and validate Maze-CBMs for use with students in high school. Test developers need to consider the existing research in Maze-CBM for high school students and consider the potential challenges of using differing genres of text on alternate forms. Once test developers have validated their Maze-CBM measures they then need to insure that this information is provided to teachers and administrators so that they may properly select the tools for use.

The limited empirical literature and lack of availability of Maze-CBM measures indicates that teachers who are using Maze-CBM for screening and progress monitoring high school students should use caution when interpreting scores on Maze-CBM, as there is a dearth of evidence to support its use in these important processes. At this point, the evidence supports that Maze-CBM has promise of being technically adequate for use but more research still needs to be conducted in this area.

Teachers should use data when making decisions regarding which students are not meeting grade level standards and how the students are progressing within supplemental support. Maze-CBM can be one source of data. Due to the minimal evidence of technical adequacy of Maze-CBM, teachers need to use it along with other sources of data on students' reading ability when making screening and progress monitoring decisions. Examples of alternative sources teachers might use along with Maze-CBM data are state test scores, district assessments, work samples, grades, and informal observation data when making these important decisions.

Finally, although Deno (1985) initially proposed that teachers could create their own Maze-CBM measures using text from their class curriculum, teachers should be very cautious when exploring this option. In the studies across this dissertation, passages were carefully created by researchers who attempted to create equivalent forms of Maze-CBM passages. It is likely that if teachers selected text from their curriculum without carefully accounting for the dimensions of text complexity, then they would create Maze-CBM passages with even less equivalency than the ones reported in this dissertation. This may lead teachers to make incorrect conclusions about how students are progressing in reading, which could result in students not receiving the reading support that they need. Thus, it is recommended that teachers use published Maze-CBM passages when they are available.

Future Research

The present dissertation presents some important findings and leads to several key areas for future research. Potential future research could investigate the following research questions:

1. What is the influence of students' background knowledge and vocabulary on Maze-CBM scores?
2. What is the influence of other text dimensions (e.g., text cohesion) on the reliability and validity of Maze-CBM scores?
3. How well do Maze-CBM scores predict which students are at risk for failing new state assessments, such as the ones by the Partnership for Assessment of Readiness for College and Careers (2014) and the Smarter Balanced Assessment Consortium (n.d.)?
4. How can Maze-CBM be used in addition to extant data within a multiple gating procedure for screening in reading?

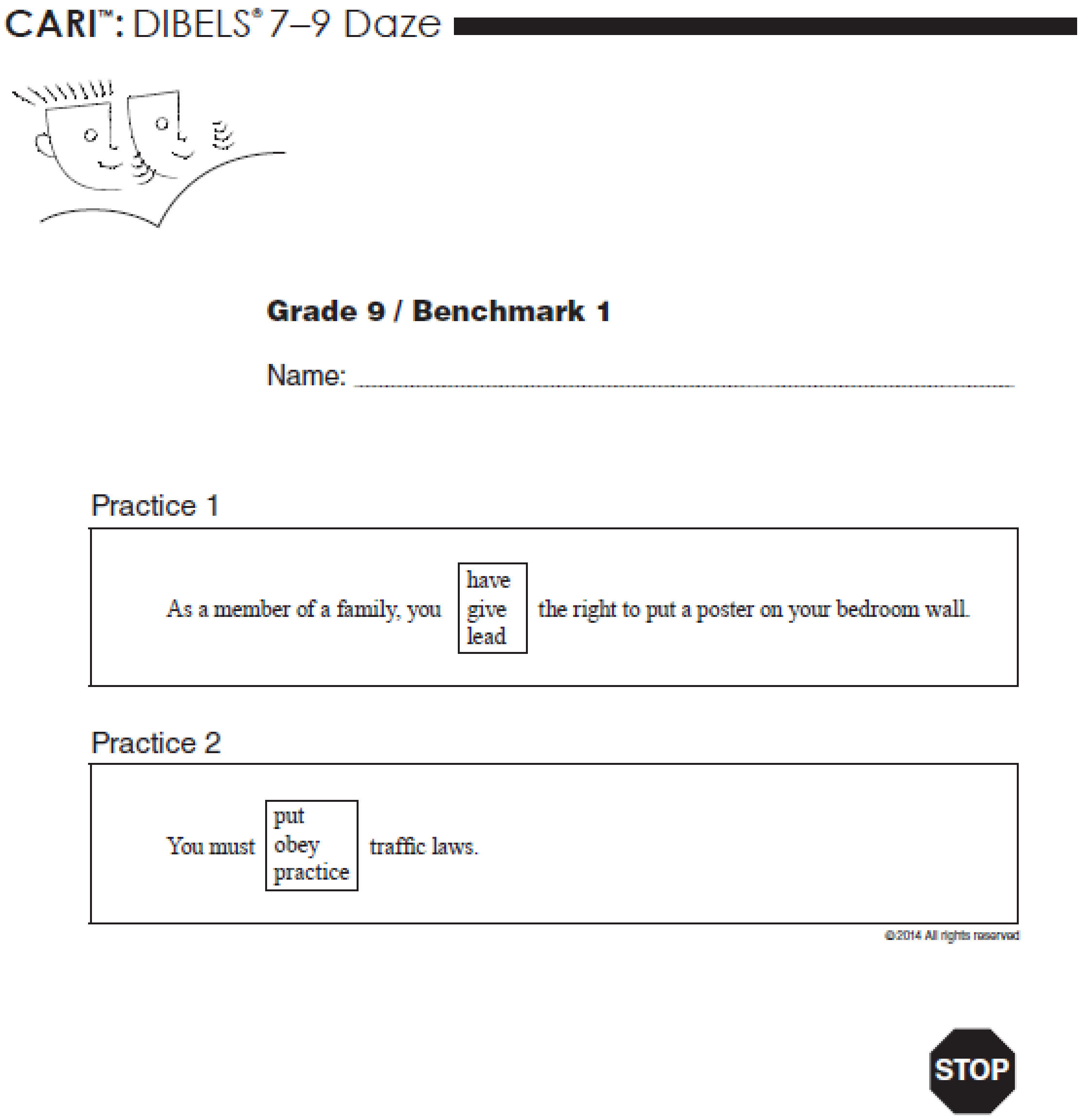
5. What are current data collection sources and practices of high school teachers for screening and progress monitoring students in reading?

With the current emphasis on standards-based learning and having all students reach proficient levels of reading, high school teachers are faced with many important instructional decisions. Maze-CBM is a promising formative assessment that can help high school teachers make decisions about which students need additional instruction. Once teachers identify these students and begin to provide them with targeted interventions, teachers can then use Maze-CBM to monitor these students' progress. The current research on Maze-CBM is limited for use with high school students, so it is important to continue research in this area. It is necessary to have a substantial research base on the reliability and validity of Maze-CBM scores in screening and progress monitoring of high school students so that teachers can make more confident decisions in response to their data. Additionally, these short assessments may save important instructional time if research supported they are just as effective for screening and progress monitoring as other, longer assessments of reading. In conclusion, the research in this dissertation provides a foundation for future research in the use of Maze-CBM for high school students.

Appendices

Appendix A

Sample Maze-CBM Triad



CARI™: DIBELS® 7–9 Daze

Grade 9 / Benchmark 1

Name: ______________________________

Practice 1

As a member of a family, you | have / give / lead | the right to put a poster on your bedroom wall.

Practice 2

You must | put / obey / practice | traffic laws.

STOP

The Accidental Opera Fan

Recently, at a drawing at my school's fall festival, I won two tickets to the upcoming opera at the local theater. Since I'm not much of an [arrangement / opera / excitement] fan, I didn't think I would [wait / go / burst]. But then I remembered that my [granddad / scale / lobby] was a devout fan of the [class / opportunity / opera], plus, I learned that I could [begin / travel / get] extra credit for music class if I [embraced / rose / attended]. I decided it might be good for me to [attempt / broaden / decide] my horizons and give opera a [story / try / bias].

It turned out that my granddad was quite [ornate / excited / beautiful] at the opportunity to go to an [opera / art / intermission] with his favorite grandson, me. For [fans / years / performances], he attempted to convert me into an opera [lover / moment / silence]. Winning the tickets seemed to Granddad an [extra / opening / unmistakable] omen that the moment had come for his [minute / grandson / theater] to embrace the classic art form [known / played / assured] as opera.

Life of a star

A star is an incredibly hot sphere of gas that radiates light and heat energy and is bound together by its own gravitational attraction. Although a star is not a quiet / living / hot organism, the period during which a barren / particular / higher star exists is referred to as its fuel / process / lifespan. The lifespan of most stars stretches billions of series / dwarfs / years, and a star will undergo several elements / transformations / copies over the duration of its life.

A balance / sequence / star is "born" deep within space in a cloud / struggle / supply of dust and gas called a force / nebula / temperature. Within the nebula, particles are drawn together by light / gravity / entity; as these particles accumulate, it creates the core / duration / series of a forming star, or protostar. As additional / staggering / classified matter is drawn in, the temperature at the sun / core / space of the protostar rises, creating a struggle / helium / majority between gravity pulling atoms in and earth / layer / gas pressure pushing light and heat out. In order to become / escape / pull a star, the protostar must reach hydrogen / transformation / equilibrium, a

Strength in Diversity

Beginning in about 1500 B.C., two cultures met and merged in the region of Asia now known as India. One of these cultures belonged to books people practices who had occupied the region for almost a thousand religions years worlds. The other was brought to the marriage region doctrine by people migrating from the north. As the households goddesses cultures blended, so did their religions, and the general established resulting hybrid, or mixture, was the religion known done divided as Hinduism.

Hinduism is unique among world central ultimate religions because of its remarkable diversity. It boasts does blends not have a central body of beings homes teachings, or doctrines. It has neither a single various nurturing book of scripture nor a founder. Hinduism is shaped restored made up of many different peacefully coexisting mixtures incense sects, or groups, each having their own people beliefs bodies and worshipping their own gods.

Appendix B

Counterbalanced Order of Maze-CBM Passages

	Passage Number								
	Prose			Science			SS		
Form A	1	4	7	2	5	8	3	6	9
Form B	4	7	1	5	8	2	6	9	3
Form C	7	1	4	8	2	5	9	3	6
	Science			SS			Prose		
Form D	2	5	8	3	6	9	1	4	7
Form E	5	8	2	6	9	3	4	7	1
Form F	8	2	5	9	3	6	7	1	4
	SS			Prose			Science		
Form G	3	6	9	1	4	7	2	5	8
Form H	6	9	3	4	7	1	5	8	2
Form I	9	3	6	7	1	4	8	2	5

Appendix C

Summary of Maze-CBM Measures

Triad #	Passage #	Title	Genre	Lexile Level	Passage Length
1	1	The Accidental Opera Fan	Prose	1170	588
1	2	Life of a star	Science	1180	592
1	3	Strength in Diversity	SS	1150	555
2	4	The Sting of Truth	Prose	1120	602
2	5	Animal Adaptations	Science	1300	566
2	6	The Strait of Hormuz	SS	1250	550
3	7	Exchanging Dreams	Prose	1150	574
3	8	Life in the Tropical Rainforest	Science	1210	585
3	9	Canada's Journey to Independence	SS	1220	549

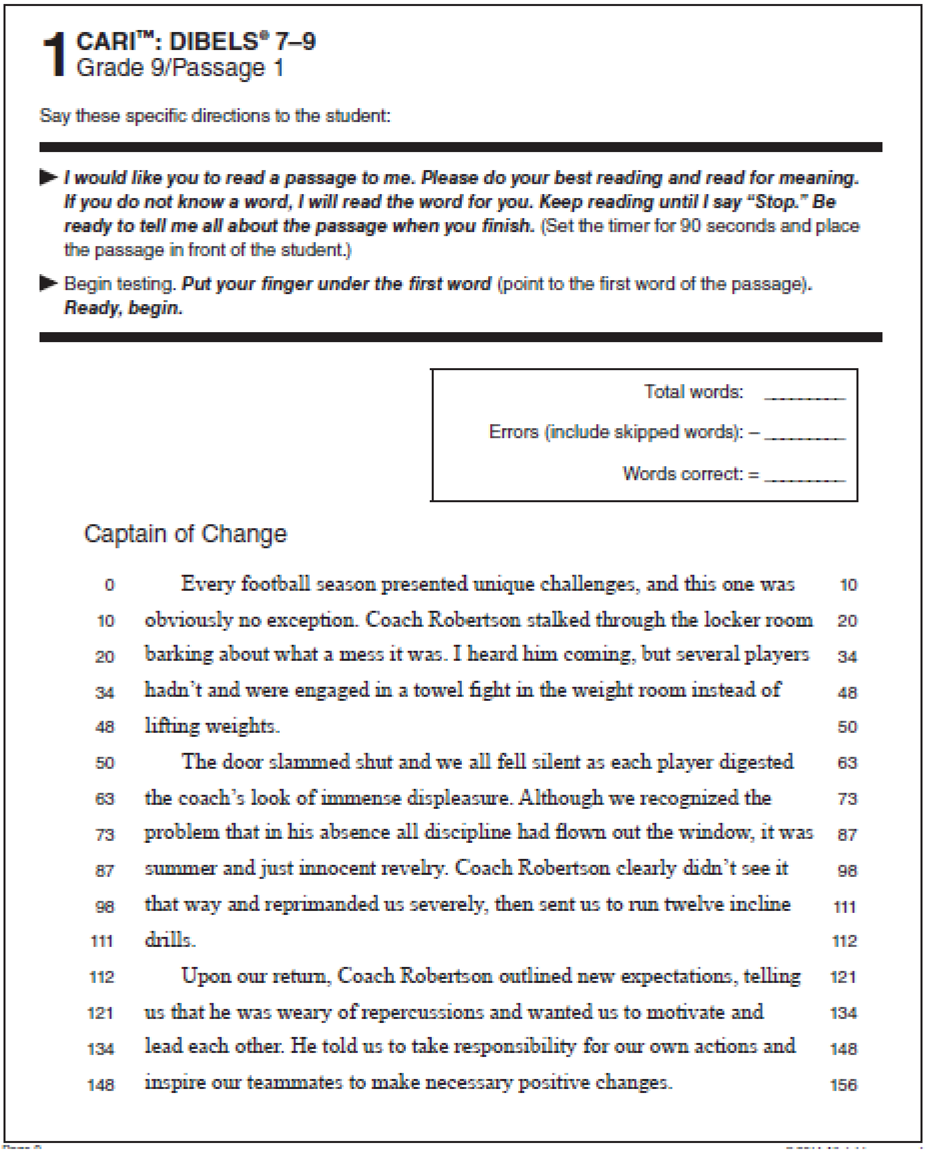
Appendix D

Overview of Measures

Measure	Approximate Time to Administer	How Administered	Use in Article 2	Use in Article 3
Maze CBM- 9 measures	30 minutes	Group administered by research team	Predictor variable	Outcome variable
TOSREC	5 minutes	Group administered by research team	Criterion variable indicating reading risk status	N/A
PSAT Reading	60 minutes	Administered by school staff	Criterion variable indicating reading risk status	N/A
SRI	30 minutes	Computer administered by school staff	Criterion variable indicating reading risk status	N/A
Oral Reading	5 minutes	Individually administered by research team	N/A	Student level predictor variable

Appendix E

Example CARI OR Triad



1 CARI™: DIBELS® 7–9
Grade 9/Passage 1

Say these specific directions to the student:

▶ ***I would like you to read a passage to me. Please do your best reading and read for meaning. If you do not know a word, I will read the word for you. Keep reading until I say "Stop." Be ready to tell me all about the passage when you finish.*** (Set the timer for 90 seconds and place the passage in front of the student.)

▶ Begin testing. ***Put your finger under the first word*** (point to the first word of the passage). ***Ready, begin.***

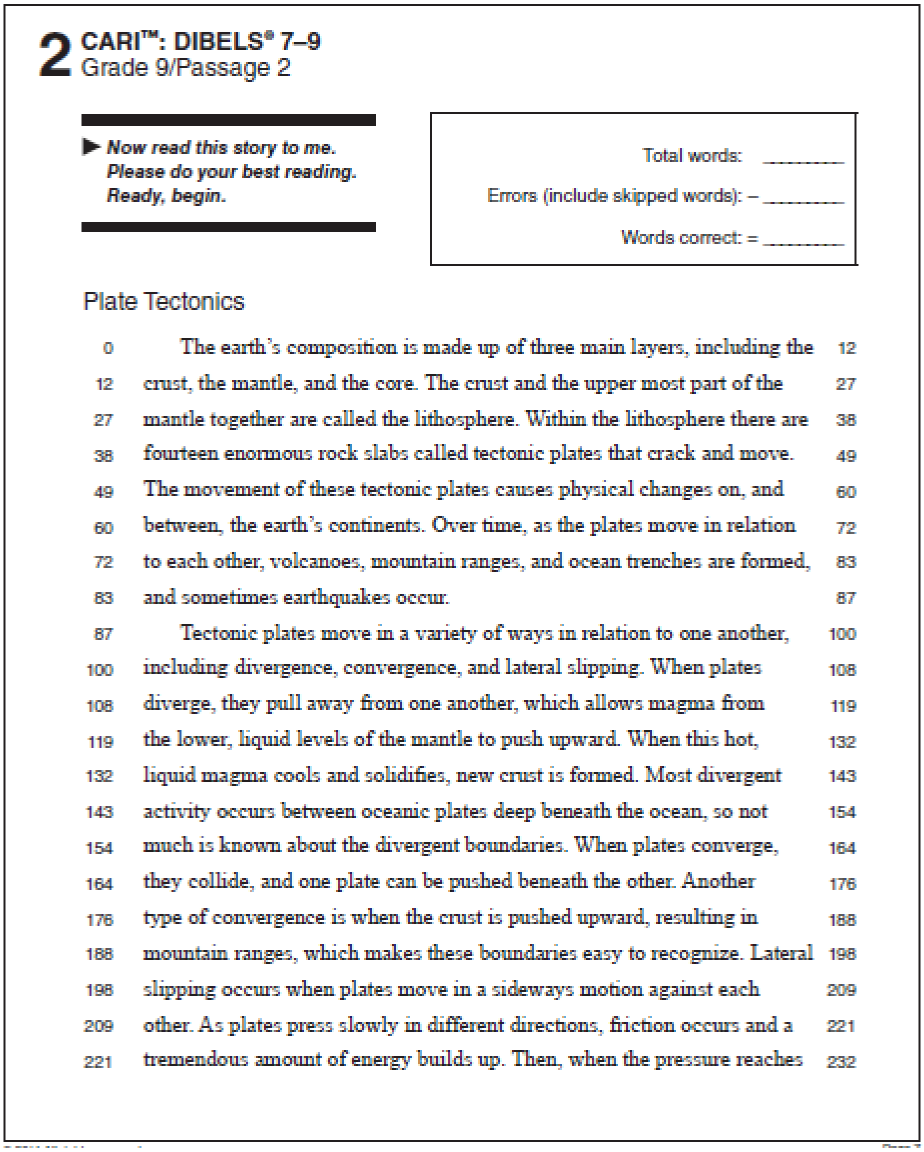
Total words: ________

Errors (include skipped words): – ________

Words correct: = ________

Captain of Change

0 Every football season presented unique challenges, and this one was 10
10 obviously no exception. Coach Robertson stalked through the locker room 20
20 barking about what a mess it was. I heard him coming, but several players 34
34 hadn't and were engaged in a towel fight in the weight room instead of 48
48 lifting weights. 50
50 The door slammed shut and we all fell silent as each player digested 63
63 the coach's look of immense displeasure. Although we recognized the 73
73 problem that in his absence all discipline had flown out the window, it was 87
87 summer and just innocent revelry. Coach Robertson clearly didn't see it 98
98 that way and reprimanded us severely, then sent us to run twelve incline 111
111 drills. 112
112 Upon our return, Coach Robertson outlined new expectations, telling 121
121 us that he was weary of repercussions and wanted us to motivate and 134
134 lead each other. He told us to take responsibility for our own actions and 148
148 inspire our teammates to make necessary positive changes. 156



2 CARI™: DIBELS® 7–9
Grade 9/Passage 2

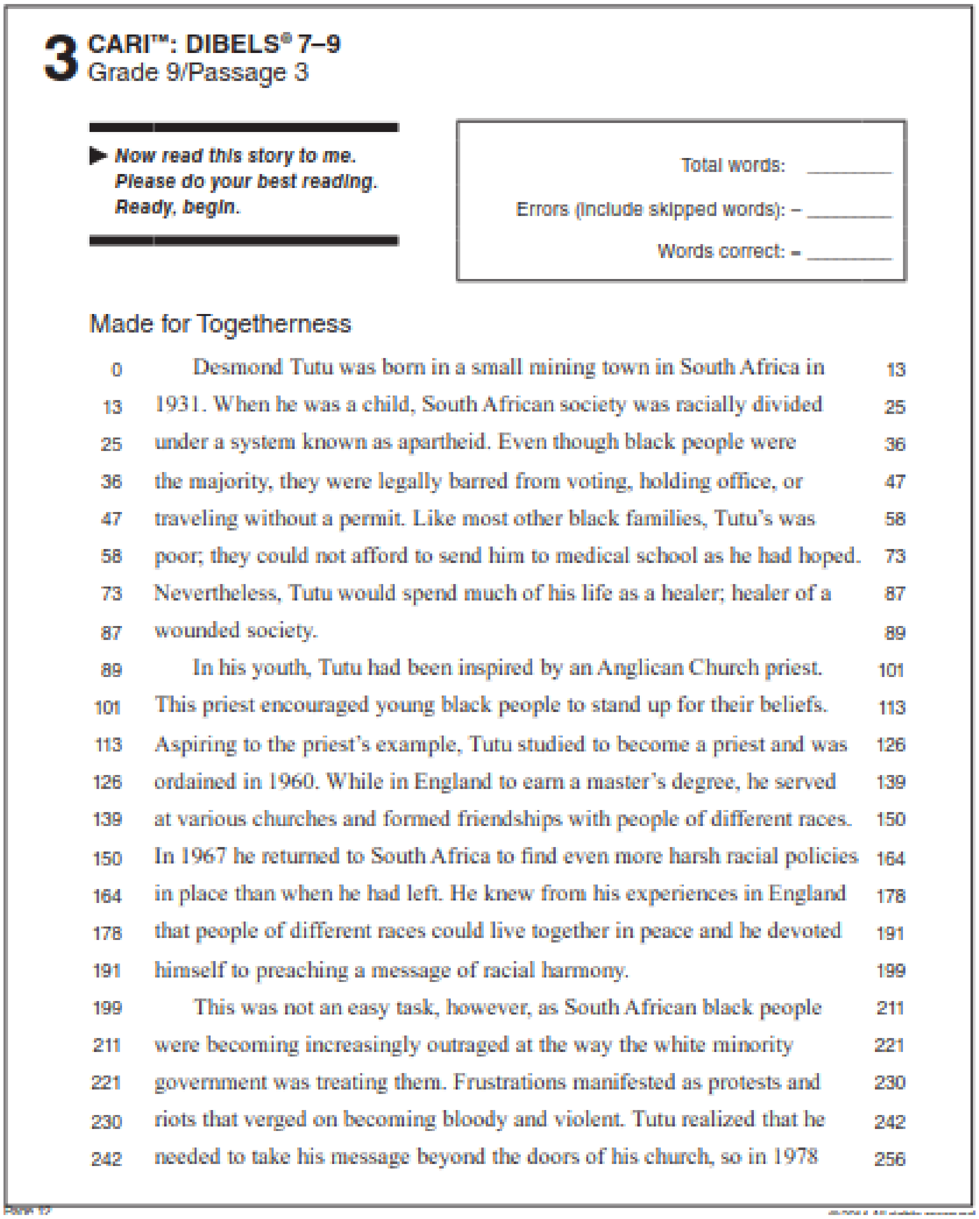
▶ *Now read this story to me.*
Please do your best reading.
Ready, begin.

Total words: _________
Errors (include skipped words): – _________
Words correct: = _________

Plate Tectonics

0 The earth's composition is made up of three main layers, including the 12
12 crust, the mantle, and the core. The crust and the upper most part of the 27
27 mantle together are called the lithosphere. Within the lithosphere there are 38
38 fourteen enormous rock slabs called tectonic plates that crack and move. 49
49 The movement of these tectonic plates causes physical changes on, and 60
60 between, the earth's continents. Over time, as the plates move in relation 72
72 to each other, volcanoes, mountain ranges, and ocean trenches are formed, 83
83 and sometimes earthquakes occur. 87

87 Tectonic plates move in a variety of ways in relation to one another, 100
100 including divergence, convergence, and lateral slipping. When plates 108
108 diverge, they pull away from one another, which allows magma from 119
119 the lower, liquid levels of the mantle to push upward. When this hot, 132
132 liquid magma cools and solidifies, new crust is formed. Most divergent 143
143 activity occurs between oceanic plates deep beneath the ocean, so not 154
154 much is known about the divergent boundaries. When plates converge, 164
164 they collide, and one plate can be pushed beneath the other. Another 176
176 type of convergence is when the crust is pushed upward, resulting in 188
188 mountain ranges, which makes these boundaries easy to recognize. Lateral 198
198 slipping occurs when plates move in a sideways motion against each 209
209 other. As plates press slowly in different directions, friction occurs and a 221
221 tremendous amount of energy builds up. Then, when the pressure reaches 232



3 CARI™: DIBELS® 7–9
Grade 9/Passage 3

▶ ***Now read this story to me.***
Please do your best reading.
Ready, begin.

Total words:	________
Errors (include skipped words): –	________
Words correct: =	________

Made for Togetherness

0 Desmond Tutu was born in a small mining town in South Africa in 13
13 1931. When he was a child, South African society was racially divided 25
25 under a system known as apartheid. Even though black people were 36
36 the majority, they were legally barred from voting, holding office, or 47
47 traveling without a permit. Like most other black families, Tutu's was 58
58 poor; they could not afford to send him to medical school as he had hoped. 73
73 Nevertheless, Tutu would spend much of his life as a healer; healer of a 87
87 wounded society. 89

89 In his youth, Tutu had been inspired by an Anglican Church priest. 101
101 This priest encouraged young black people to stand up for their beliefs. 113
113 Aspiring to the priest's example, Tutu studied to become a priest and was 126
126 ordained in 1960. While in England to earn a master's degree, he served 139
139 at various churches and formed friendships with people of different races. 150
150 In 1967 he returned to South Africa to find even more harsh racial policies 164
164 in place than when he had left. He knew from his experiences in England 178
178 that people of different races could live together in peace and he devoted 191
191 himself to preaching a message of racial harmony. 199

199 This was not an easy task, however, as South African black people 211
211 were becoming increasingly outraged at the way the white minority 221
221 government was treating them. Frustrations manifested as protests and 230
230 riots that verged on becoming bloody and violent. Tutu realized that he 242
242 needed to take his message beyond the doors of his church, so in 1978 256

Page 12

References

Abbott, M., Stollar, S., Good, R., & McMahon, R. (2014). *CARI: DIBELS 7th–9th grade early release administration and scoring guide*. Retrieved from http://dibels.org/cariearlyrelease/

Aday, L. A., & Cornelius, L. J. (2006). *Designing and conducting health surveys: A comprehensive guide* (3rd ed.). San Francisco, CA: Jossey-Bass.

Baker, D. L., Biancarosa, G., Park, B. J., Bousselot, T., Smith, J. L., Baker, S. K., ... 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.

Balfanz, R., Herzog, L., & Mac Iver, D. J. (2007). Preventing student disengagement and keeping students on the graduation path in urban middle-grades schools: Early identification and effective interventions. *Educational Psychologist, 42*(4), 223–235. doi:10.1080/00461520701621079College Board. (2015). *2015 PSAT/NMSQT fact sheet*. Retrieved from https://collegereadiness.collegeboard.org/pdf/psat-nmsqt-fact-sheet.pdf

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Educational Assessment: Principals, Policy and Practice*, *5*, 7–74.

Chall, J. S., & Dale, E. (1995). *Readability revisited: The new Dale-Chall readability formula*.

Cambridge, MA: Brookline Books.

Christ, T. J., & Ardoin, S. P. (2009). Curriculum-based measurement of oral reading: Passage equivalence and probe-set development. *Journal of School*

Psychology, *47*(1), 55–75.

Crocker, L., & Algina, J. (2008). *Introduction to classical and modern test theory*. Mason, OH: Cengage Learning.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 443–507). Washington, DC: American Council on Education.

Codding, R. S., Petscher, Y., & Truckenmiller, A. (2014). CBM reading, mathematics, and written expression at the secondary level: Examining latent composite relations among indices and unique predictions with a state achievement test. *Journal of Educational Psychology*, *107*(2), 437–450. doi:10.1037/a0037520

Cummings, K. D., Park, Y., & Bauer Schaper, H. A. (2012). Form effects on DIBELS Next oral reading fluency progress-monitoring passages. *Assessment for Effective Intervention*, *38*(2), 91–104. doi:10.1177/1534508412447010

Deno, S. L. (1985). Curriculum-Based Measurement: The Emerging Alternative. *Exceptional Children*, *52*(3), 219–232.

Deno, S. L. (1989). Curriculum-based measurement and alternative special education services: A fundamental and direct relationship. In M. R. Shinn (Ed.), *Curriculum-based measurement: Assessing special children* (pp. 1–17). New York, NY: Guilford Press.

Deno, S. L. (2003). Developments in curriculum-based measurement. *Journal of Special Education*, *37*(3), 184–192.

Deno, S. L., Mirkin, P. K., & Chiang, B. (1982). Identifying valid measures of reading. *Exceptional Children*, *49*, 36–45.

Deno, S. L., Mirkin, P. K., & Marston, D. (1980). *Relationships among simple measures of written expression and performance on standardized achievement tests* (Research Report No. 22). Minneapolis: University of Minnesota Institute for Research on Learning Disabilities.

Denton, C. A., Barth, A. E., & Fletcher, J. M. (2011). NIH Public Access. *Research in Education*, *4*(3), 208–230. doi:10.1080/19345747.2010.530127

Denton, C. A., Enos, M., York, M. J., Francis, D. J., Barnes, M. A., Kulesz, P. A., ... Carter, S. (2015). Text-processing differences in adolescent adequate and poor comprehenders reading accessible and challenging narrative and informational text. *Reading Research Quarterly*, *50*(4), 393–416.

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. A., & Foegen, A. (1996). Validity of general outcome measures for predicting. *Exceptional Children, 62*(6), 497–514.

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. doi:10.1111/j.1540-5826.2010.00304.x

Fore, C., III, Boon, R. T., Burke, M. D., & Martin, C. (2009). Validating curriculum-based measurement for students with emotional and behavioral disorders in middle school. *Assessment for Effective Intervention*, *34*(2), 67–73.

Fore, C., III, 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.

Francis, D. J., Santi, K. L., Barr, C., Fletcher, J. M., Varisco, A., & Foorman, B. R. (2008). Form effects on the estimation of students' oral reading fluency using DIBELS. *Journal of School Psychology*, *46*(3), 315–342.

Fuchs, L. S. (2004). The past, present, and future of curriculum-based measurement research. *School Psychology Review*, *33*(2), 188–192.

Fuchs, L. S., & Fuchs, D. (1992). Identifying a measure for monitoring student reading progress. *School Psychology Review*, *21*(1), 45–58.

Fuchs, L. S., Fuchs, D., Hamlett, C. L., & Ferguson, C. (1992). Effects of expert system consultation within curriculum-based measurement using a reading maze task. *Exceptional Children, 58*, 436–450.

Fuchs, L. S., Fuchs, D., Hamlett, C. L., & Stecker, P. M. (1990). The role of skills analysis in curriculum-based measurement in math. *School Psychology Review, 19*, 6–22.

Fuchs, L. S., Fuchs, D., Hamlett, C. L., Thompson, A., Roberts, P. H., Kubek, P., & Stecker, P. M. (1994). Technical features of a mathematics concepts and applications curriculum-based measurement system. *Assessment for Effective Intervention*, *19*(4), 23–49.

Fuchs, L. S., Fuchs, D., Hamlett, C. L., Walz, L., & Germann, G. (1993). Formative evaluation of academic progress: How much growth can we expect? *School Psychology Review, 22*(1), 27–48.

Fuchs, L., Fuchs, D., Hosp, M., & Jenkins, J. R. (2001). Oral reading fluency as an indicator of reading competence: A theoretical, empirical, and historical analysis. *Studies of Reading*, *5*(3), 239–256. doi:10.1207/S1532799XSSR0503_3

Hintze, J. M., & Christ, T. J. (2004). An examination of variability as a function of passage variance in CBM progress monitoring. *School Psychology Review*, *33*(2), 204–217.

Hosp, M. K., Hosp, J. L., & Howell, K. W. (2007). *The ABCs of CBM. A practical guide to curriculum-based measurement.* New York, NY: Guilford Press.

Jenkins, J. R., & Jewell, M. (1993). Examining the validity of two measures for formative teaching: Reading aloud and maze. *Exceptional Children*, *59*(5), 421–432.

Kennelly, L., & Monrad, M. (2007). *Approaches to dropout prevention: Heeding early warning signs with appropriate interventions*. Washington, DC: National High School Center.

Kim, Y., Hendrickson, A., Patel, P., Melican, G., & Sweeney, K. (2014). *Development of a new ReadiStep™ scale linked to the PSAT/NMSQT® scale* (No. 2013-4). College Board Statistical Report.

Kincaid, J. P., Fishburne, R. P., Jr., Rogers, R. L., & Chissom, B. S. (1975). *Derivation of new readability formulas (automated readability index, fog count and Flesch reading ease formula) for navy enlisted personnel* (No. RBR-8-75). Millington, TN: Naval Technical Training Command Millington TN Research Branch.

Kline, R. (2009). *Becoming a behavioral science researcher: A guide to producing research that matters*. New York, NY: Guilford Press.

Madelaine, A., & Wheldall, K. (2004). Curriculum-based measurement of reading: Recent advances. *International Journal of Disability, Development and Education*, *51*(1), 57–82.

Martin, S. D., & Shapiro, E. S. (2011). Examining the accuracy of teachers' judgments of DIBELS performance. *Psychology in the Schools*, *48*(4), 343–356.

McMaster, K. L. (2010). *High school response to intervention: Progress monitoring* [PowerPoint slides]. Retrieved from http://www.rti4success.org/sites/default/files/HSTII_Progress_Monitoring_Webinar_5-12-10.pdf

McMaster, K. L., 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*(4), 17–31.

McNamara, D. S., Louwerse, M. M., Cai, Z., & Graesser, A. (2005). Coh-Metrix version 1.4. Retrieved from http//:cohmetrix.memphis.edu

McNamara, D. S., Ozuru, Y., & Floyd, R. G. (2011). Comprehension challenges in the fourth grade: The roles of text cohesion, text genre, and readers' prior knowledge. *International Electronic Journal of Elementary Education*, *4*(1), 229–257.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York, NY: Macmillan.

MetaMetrics. (2015). *Lexile-to-grade correspondence.* Retrieved from http://www.lexile.com/about-lexile/grade-equivalent/

Mitchell, M., & Wexler, J. (2016). *A Literature Synthesis of the Technical Adequacy of Maze-CBM for Secondary Students.* Manuscript submitted for publication.

National Center for Education Statistics. (2014). *The nation's report card: Reading 2013. National assessment of educational progress at grades 4 and 8* (NCES 2012-457). Washington, DC: U.S. Department of Education.

National Center on Intensive Intervention. (n.d.). *Using academic progress monitoring for individualized instructional planning* [PowerPoint slides]. Retrieved from http://www.intensiveintervention.org/resource/using-academic-progress-monitoring-individualized-instructional-planning-dbi-training

National Center on Intensive Intervention. (2013). Data-based individualization: A framework for intensive intervention. Washington, DC: Office of Special Education, U.S. Department of Education.

National Center on Response to Intervention. (2010). *Essential components of RTI – A closer look at response to intervention.* Washington, DC: U.S. Department of Education, Office of Special Education Programs, National Center on Response to Intervention.

National Governors Association Center for Best Practices, Council of Chief State School Officers. (2010). *Common Core State Standards*. Retrieved from http://www.corestandards.org

Every Student Succeeds Act (ESSA) of 2015, Pub. L. No. 114-95 (2015).

Partnership for Assessment of Readiness for College and Careers. (2014). *The PARCC assessment*. Retrieved from http://www.parcconline.org/parcc-assessment

Pearson, P. D., Johnson, D. D., Clymer, T., Indrisano, R., Venezky, R. L., Baumann, J. F., … Toth, M. (1989). *World of reading*. Needham, MA: Silver Burdett, & Ginn.

Petscher, Y., & Kim, Y. S. (2011). The utility and accuracy of oral reading fluency score types in predicting reading comprehension. *Journal of School Psychology*, *49*(1), 107–129.

Pierce, R. L., McMaster, K. L., & Deno, S. L. (2010). The effects of using different procedures to score Maze-CBM measures. *Learning Disabilities Research & Practice, 25*(3), 151–160.

Pullin, D. (2013). Legal issues in the use of student test scores and value-added models (VAM) to determine educational quality. *Education Policy Analysis Archives, 21*(6), 1–27.

RAND Reading Study Group. (2002). *Reading for understanding: Toward an R & D program in reading comprehension*. Santa Monica, CA: RAND.

Reed, D. K. (2011). A review of the psychometric properties of retell instruments. *Educational Assessment*, *16*(3), 123–144.

Reed, D. K., Cummings, K. D., Schaper, A., & Biancarosa, G. (2014). Assessment fidelity in reading intervention research: A synthesis of the literature. *Review of Educational Research*, *84*(2), 275–321.

Sáenz, L. M., & Fuchs, L. S. (2002). Examining the reading difficulty of secondary students with learning disabilities: Expository versus narrative text. *Remedial and Special Education*, *23*(1), 31–41.

Salvia, J., Ysseldyke, J. E., & Bolt, S. (2007). *Assessment in special and inclusive education*. Boston, MA: Houghton Mifflin Company.

Santi, K., Barr, C., Khalaf, S., & Francis, D. (2016). Different approaches to equating oral reading fluency passages. In K. D. Cummings & Y. Petscher (Eds.), *The fluency construct* (pp. 223–265). New York, NY: Springer.Scholastic. (2007). *Scholastic Reading Inventory technical guide*. New York, NY: Scholastic Inc.

Schatschneider, C., Buck, J., Torgesen, J. K., Wagner, R. K., Hassler, L., Hecht, S., & Powell-Smith, K. (2004). *A multivariate study of factors that contribute to individual differences in performance on the Florida Comprehensive Reading Assessment Test* (Vol. 5, Technical Report).

Shinn, M. R. (Ed.). (1989). *Curriculum-based measurement: Assessing special children*. New York, NY: 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*(5), 527–535. doi:10.1002/pits.20175

Smarter Balanced Assessment Consortium. (n.d.). *Smarter balanced assessments*. Retrieved from http://www.smarterbalanced.org/smarter-balanced-assessments/

Smolkowski, K., Cummings, K., & Strycker, L. (2016). An introduction to the statistical evaluation of fluency measures with signal detection theory. In K. D. Cummings & Y. Petscher (Eds.), *The fluency construct* (pp. 187–221). New York, NY: Springer.

STARD Statement. (2008). *Standards for the reporting of diagnostic accuracy studies*. Retrieved from http://www.stard-statement.org

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*(8), 795–819.

Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. *Science, 240*(4857), 1285–1293.

Tichá, R., Espin, C. A., & Wayman, 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. doi:10.1111/j.1540-5826.2009.00287.x

Tindal, G. (1992). Evaluating instructional programs using curriculum-based measurement. *Preventing School Failure: Alternative Education for Children and Youth*, *36*(2), 39–42.

Tindal, G. (2013). Curriculum-based measurement: A brief history of nearly everything from the 1970s to the present. *ISRN Education, 2013,* 1–29.

Tolar, T. D., Barth, A. E., Fletcher, J. M., Francis, D. J., & Vaughn, S. (2014). Predicting reading outcomes with progress monitoring slopes among middle grade students. *Learning and Individual Differences*, *30*, 46–57. doi:10.1016/j.lindif.2013.11.001

Tolar, T. D., Barth, A. E., Francis, D. J., Fletcher, J. M., Stuebing, K. K., & Vaughn, S. (2012). Psychometric properties of maze tasks in middle school students. *Assessment for Effective Intervention, 37*(3), 131–146.

Torgesen, J. K., & Miller, D. H. (2009). *Assessments to guide adolescent literacy instruction*. Portsmouth, NH: RMC Research Corporation, Center on Instruction.

U.S. Department of Education, Office of Elementary and Secondary Education. (2002). Guidance for the Reading First Program. Washington, DC.

Van der Kleij, F. M., Vermeulen, J. A., Schildkamp, K., & Eggen, T. J. H. M. (2015). Integrating data-based decision making, assessment for learning, and diagnostic testing in formative assessment. *Assessment in Education: Principles, Policy & Practice*, *22*(3), 324–343.

Vaughn, S., & Fletcher, J. M. (2012). Response to intervention with secondary school students with reading difficulties. *Journal of Learning Disabilities*, *45*(3), 244–256.

Wagner, R. C., Torgesen, J. K., Rashotte, C. A., & Pearson, N. A. (2010). *Test of silent reading efficiency and comprehension.* Austin, TX: Pro Ed.

Wayman, M. M., Wallace, T., Wiley, H. I., Tichá, R., & Espin, C. A. (2007). Literature synthesis on curriculum-based measurement in reading. *Journal of Special Education*, *41*(2), 85–120.

Wiley, H. I., & 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.

Yeo, S., Fearrington, J. Y., & Christ, T. J. (2012). Relation between CBM-R and CBM-mR slopes: An application of latent growth modeling. *Assessment for Effective Intervention*, *37*(3), 147–158.

Yoo, M. S. (2015). The influence of genre understanding on strategy use and comprehension. *Journal of Adolescent & Adult Literacy*, *59*(1), 83–93.

Yovanoff, P., Duesbery, L., Alonzo, J., & Tindal, G. (2005). Grade-level invariance of a theoretical causal structure predicting reading comprehension with vocabulary and oral reading fluency. *Educational Measurement: Issues and Practice*, *24*(3), 4–12.

Zhang, X., Patel, P., & Ewing, M. (2014). *AP® potential predicted by PSAT/NMSQT® scores using logistic regression* (No. 2014-1). College Board Statistical Report.