BEYOND SCREENING AND PROGRESS MONITORING: AN EXAMINATION OF THE RELIABILITY AND CONCURRENT VALIDITY OF MAZE COMPREHENSION ASSESSMENTS FOR FOURTH-GRADE STUDENTS

By

Casey F. Brasher

A Dissertation Submitted in Partial Fulfillment of the Requirements for the

Doctor of Philosophy in Literacy Studies

Middle Tennessee State University

May, 2017

Dissertation Committee:

Dr. Amy Elleman, Chair

Dr. Jwa K. Kim

Dr. Aimee Holt

For you, Granny

ACKNOWLEDGEMENTS

First and foremost, I want to acknowledge my Lord and Savior, Jesus Christ. Thank you Jesus.

Second, my husband, Jason, has provided me encouragement and support throughout my very long journey. He always believed in me and supported me even when it was not easy. I would not have been able to go back to school as a mother of five working full time without him by my side. I love you, Jason.

My children have been my motivation. I want them to know that they can accomplish anything they set their mind to do! I love you Zack, Landon, Triston, Harper, and Hadley.

I am thankful for the support of my mom and dad. Thanks for encouraging me and believing in me.

I am fortunate to have the intellectual and emotional support of my dissertation committee, Dr. Jwa K. Kim and Dr. Aimee Holt, and my chair, Dr. Amy Elleman. Thank you for your thoughtful guidance, reassurance, and continued support throughout this process. I am blessed by knowing each of you. Funding for this project was possible thanks to the Sawyer-Rudler Research Scholarship.

Lastly, I want to acknowledge the administrators, teachers, and fourth-grade students in Maury County for their participation with this project.

ABSTRACT

Reading comprehension assessments often lack instructional utility because they do not accurately pinpoint why a student has difficulty. The varying formats, directions, and response requirements of comprehension assessments lead to differential measurement of underlying skills and contribute to noted amounts of unshared variance among tests. Maze is an assessment tool used to screen and monitor reading comprehension performance. This type of assessment consists of words deleted throughout the passage replaced with three options, the correct choice, and two distractors. Students are required to select the correct option during the process of reading. Maze emerged as an assessment of reading comprehension to guide teachers in selecting an independent reading level for students. However, the purpose of maze shifted to screening and monitoring reading performance rather than instructional planning. Yet, there is a pressing need for an assessment or system of assessments that can inform instruction for students with reading comprehension weaknesses. The present study examined the validity and reliability of different types of maze assessments (fixed-word deletion, word-feature deletion, and sentence deletion) and a multiple choice assessment. All passages were created from informational news stories. All four assessment conditions demonstrated acceptable to excellent levels of internal consistency. Correlations between conditions analyzed in the study and validated measures of reading comprehension varied significantly. The sentence deletion version of maze demonstrated significant correlations with two of three of the comprehension tests and had a significant correlation with a composite score for reading comprehension. Correlations to reader skills varied across types of maze. The conditions created for this study seemed to tap into a dimension of reading comprehension not measured by validated, standardized comprehension measures. Passage length and genre were suggested as possible reasons for the differences between the assessment conditions analyzed in this study and the validated comprehension tests. Further, a maze task involving sentence deletion emerged as a potential alternative to the way maze assessments are standardly created. Implications for policy and practice are discussed in terms of analyzing student performance across measures when assessing reading comprehension.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER ONE

INTRODUCTION

Reading comprehension is well agreed upon in the research community and the classroom as the goal of reading. Therefore, assessing the purported goal of reading is incredibly important in education. Although reading comprehension assessments are used to guide important educational decisions for children, there is evidence to suggest that they differentially measure underlying skills (e.g., Cutting & Scarborough, 2006; Eason, Sabatini, Goldberg, Bruce & Cutting, 2013; Keenan, Betjemann, & Olson, 2008), do not accurately distinguish children in the lowest and highest percentiles (Keenan & Meenan, 2014), and do not directly inform instruction (Cutting & Scarborough, 2006; Francis, et al., 2006; Sweet, 2005). Reading comprehension tests are broad measures that are typically not useful for addressing why a student struggles with comprehension. The available assessments tap underlying skills differently to the point that scores can vary substantially in some cases. Identifying specific reading profiles and comprehension processes is critical for targeting appropriate interventions for students (Compton, Fuchs, Fuchs, & Bryant, 2006); yet, research on the instructional usefulness of reading comprehension assessment is limited.

In addition to lack of instructional utility of the available comprehension tests, there are noted administration and response format differences among them as well. Two factors have been proposed as reasons for the differences among comprehension assessments in several studies (Cutting & Scarborugh, 2006; Keenan et al. 2008; Keenan

& Meenan, 2014; Kendeou, Papadopoulos, & Spanoudis, 2012; Nation & Snowling, 1997; Spear-Swerling, 2004). The first of the proposed factors pertains to item features, response format, and text characteristics of the comprehension assessment (Spear-Swerling, 2004) and the second refers to the underlying skills of the reader, such as decoding and fluency (Cutting & Scarborough, 2006; Keenan et al. 2008). Due to the criticisms of current reading comprehension assessments there is a need for a new approach. One such approach identified in previous research is to administer multiple assessments of comprehension in a systematic way. Multiple assessments are more likely to adequately capture the broad construct of reading comprehension (Francis et al., 2006; RAND Reading Study Group, 2002).

Theoretical Framework for Reading Comprehension

In addition to understanding how the components of the test and reader skills contribute to performance, understanding theory can be useful for explaining how the component skills are related and assessed. It has been suggested that the available comprehension tests are not driven by theory during passage construction or item development which contributed to the noted lack of consistency (Compton, Miller, Elleman, & Steacy, 2014). McMaster, Espin, and van den Broek (2014) contend cognitive theory should be linked to the definition of reading comprehension and to the development of comprehension assessments. In their view, a thorough understanding of the underlying cognitive skills tapped by components of an assessment can directly inform instruction. Existing comprehension tests have a variety of components. For example, there are many tests of comprehension that present multiple choice questions to students. Other assessments require the student to summarize or retell the events of the passage after reading. McMaster and colleagues (2014) emphasize the idea of coherence which was examined in terms of characteristics and properties of readers, texts, and instructional contexts. Readers described as struggling in the McMaster et al. (2014) study were noted to have trouble developing a clear, mental representation of text when reading. Also, they often do not engage in strategies to *fix* comprehension mistakes as good readers do. Most of the available comprehension tests measure comprehension after reading rather than during the process of reading. Assessments that provide information about the strategies and processes used during reading, rather than after, would be consistent with cognitive theory and may provide useful information for instruction.

Sociocultural theory, another perspective on how meaning is constructed when reading, focuses on the interaction between reader, text, and task. The characteristics of the text, especially cohesion, and the characteristics of the reader, such as background knowledge and ability to make inferences, are important aspects of the theory (McNamara & Kendeou, 2011). Despite the significance attributed to each factor within the theory, the interaction between the reader and text as well as social interaction surrounding the text that are emphasized. Comprehension intervention based on sociocultural theory would emphasize close reading and analysis of text as well as discussion surrounding it.

Two additional theoretical frameworks posited as important for understanding reading comprehension include the *simple view of reading* and *construction-integration (C-I) theory*. Gough and Tunmer (1986) originally proposed the *simple view of reading* as a model to describe the components necessary for successful reading comprehension. Research studies have provided support in favor of the *simple view of reading* with word recognition and language comprehension consistently emerging as the components necessary for reading comprehension (Catts, 2009; Francis et al., 2006; Keenan & Meenan, 2014). Three subgroups of children with deficits in reading can be classified in terms of the components of the *simple view of reading*: poor word recognition, poor comprehension, or weaknesses in both areas (Catts, Compton, Tomblin & Bridges, 2011; Compton, Fuchs, Fuchs, Elleman, & Gilbert, 2008; Elleman, Compton, Fuchs, Fuchs, & Bouton, 2011). Catts, Hogan, and Fey (2003) specifically suggested designing reading comprehension assessments based on the *simple view of reading*. Such assessments would identify if intervention for a student should emphasize word recognition or language comprehension skills.

In the *construction-integration model* originally proposed by Kintsch (1988) comprehension is linked to cognition. The model emphasizes background knowledge as well as information presented in the text. To clarify, when reading, all theories discussed rely on construction of meaning at different levels. The *CI model* specifically accentuates the construction of meaning at a linguistic level and conceptual level. The linguistic level requires knowledge of the word features such as how it sounds. The conceptual level requires integration of information presented in the text and background knowledge of the reader (Kintsch, 1988). Intervention within the cognitive perspective, *simple view*, or *CI model* focus on the development of underlying skills important for the development of comprehension ability and on strategies to emphasize coherence and meaning construction.

Text and Item Features of Reading Comprehension Assessments

A meta-analysis on reading comprehension assessments by Garcia and Cain (2014) revealed that the text variables and administration procedures were significant in determining the amount of variance decoding skill attributed to the comprehension measure. The text features reviewed in the meta-analysis included the genre and length of the passage as well as the format of the task. The administration procedures varied on the tests reviewed. For example, differences were found across many factors such as the response required from the student, presence of time limitations, and whether the text was read aloud or silently. The inclusion of a timing component and length of passage, in term of number of sentences, were important variables that contributed to differences in performance across students with different skills (Garcia & Cain, 2014; Keenan & Meenan, 2014; Spear-Swerling, 2004). Kendeou et al. (2012) examined the importance of a number of underlying child skills on comprehension performance. The skills examined in the study include rapid naming, phonological and orthographic processing, fluency, vocabulary, and working memory. The authors found that processing demands of the tests varied based on specific text features such as passage length and availability of text during questioning. Also, incorporation of timing into an assessment was also found to influence the processing demands posed by the test.

Cain and Oakhill (2006) suggested that different response formats (i.e. cloze, true/false sentence recognition, sentence verification task, multiple choice questions, and open-ended responses) could be one source of inconsistency in the assessment of reading comprehension. Response format refers to the way a reader is expected to demonstrate comprehension such as orally answering open-ended questions, choosing a response among multiple choice options, retelling the main points of a story, or providing a missing word in a sentence or paragraph. The cloze method in which students are required to read a passage silently and provide a missing word is a method of comprehension assessment that has been shown to correlate strongly with decoding skill (Cain & Oakhill, 2006; Francis et al., 2006). Keenan et al. (2008) found that assessments requiring retell or oral answers to questions had differing patterns of variance for predicting underlying skills than assessments using the cloze method or picture selection. Specifically, retell and verbal responses tapped listening comprehension and language skills more than cloze or picture selection which were more influenced by decoding.

Reader Skills Important for Reading Comprehension

Modest correlations among reading comprehension measures suggest that different tests have substantial amounts of unshared variance and are differentially related to skills important for comprehension such as reading fluency, vocabulary, working memory, phonological skills, and orthographic processing (Kendeou et al., 2012).

Kendeou et al. (2012) proposed a structural equation model to examine the relative importance of various skills across three different comprehension tests. All three assessments, maze, cloze, and retell were interrelated (*r* = .32 - .42) but had substantial amounts of unshared variance. For instance, reading comprehension as measured by the Woodcock-Johnson III passage comprehension subtest (WJ-III PC) which uses the cloze procedure was significantly predicted by working memory and orthographic processing, whereas reading comprehension as measured by maze was significantly predicted by reading fluency and vocabulary (Kendeou et al., 2012). In the study, performance on recall was significantly predicted by working memory, orthographic processing, and phonological skills. Findings such as these show how underlying reader skills are measured differently on reading comprehension assessments. Although reading comprehension assessments purport to measure a unitary construct, most tap the underlying skills known to be important for reading comprehension differently. For example, decoding ability and reading fluency skills have been shown to have varying impacts on performance depending on the characteristics of the test and items. A variety of component skills are necessary to complete a reading comprehension assessment; yet, performance across component skills can vary significantly for students with similar scores on the broad comprehension measure.

Many of the available comprehension assessments tap skills differently which make informing instruction and intervention needs for students problematic. To say that a student is struggling with reading comprehension based on results from a particular test does not provide useful information about why that student is struggling, and thus cannot help the teacher develop targeted lessons. For example, Keenan et al. (2008) found that the amount of variance accounted for by decoding was significantly different based on the particular reading comprehension assessment that was used. Specifically, decoding skill accounted for significantly more variance on the WJ-III PC and Peabody Individual Achievement Test than on the Gray Oral Reading Test, 3rd edition and the Qualitative Reading Inventory, 3rd edition. Betjemann, Keenan, Olson and Defries (2011) used factor analysis to analyze five comprehension measures in terms of decoding and listening comprehension ability. Although both skills were important across all five measures, there were differing patterns of influence. Specifically, some of the tests analyzed in the study were more impacted by decoding skill whereas other tests demonstrated a stronger relationship to listening comprehension providing additional evidence about the differences among tests. Also, in Garcia and Cain (2014), decoding skill seemed to have lesser impact on comprehension performance for older children than younger ones across all formats and administration procedures of the tests analyzed in the study. A possible implication from this finding is that comprehension measures are more dependent on decoding ability for younger students. In general, several studies have found decoding and word reading ability to contribute more variance to comprehension assessments that consist of the cloze procedure (Keenan et al., 2008; Keenan & Meenan, 2014; Kendeou et al., 2012). However, Nation and Snowling (1997) found that oral language skills made significant contributions on a comprehension test that used the cloze procedure. The passages used in this study were longer than those commonly used on cloze assessments, suggesting that the impact of the response required by students as well as features of the text, such as passage length, both impact the skills needed to be successful at a particular test.

Maze as a Comprehension Assessment

Cloze assessment, or fill in the blank, has been used as a comprehension assessment for many years. A very early study (see Louthan, 1965) analyzed the types of words deleted and the difficulty of text using cloze as a comprehension measure for students in seventh grade. The students in the study showed differing patterns of comprehension performance based on the type of passages they read. Some of the passages had deletions of specific word types such as nouns, verbs, modifiers, prepositions, conjunctions, or pronouns. Other passages contained deletions following every fifth word. The author noted that deleting every fifth word, or any fixed-word deletion strategy, creates a situation in which the type of word deleted will be random. In this study, students read passages with specific types of words deleted, random words deleted based on every fifth word, or intact passages. Students who read intact passages performed superior to the students in the two cloze conditions answering questions about the passage. No differences were found for students in the two cloze conditions indicating that deleting specific types of words did not impact performance more or less than deleting words randomly by type as occurs with fixed-word deletion, or every fifth word (Louthan, 1965).

Shanahan, Kamil, and Tobin (1982) devised several studies to determine if cloze assessment measured comprehension across multiple sentences or at the sentence level only. The authors created scrambled probes with sentences containing word deletion in random order and compared performance to intact probes of comparable readability. Readability was found to be the significant factor at predicting student performance rather than whether the passages were scrambled or intact. The conclusion from this study was that cloze was a measure of sentence-level comprehension rather than reflecting comprehension across sentences (Shanahan et al., 1982). Helfeldt, Henk, and Fotos (1986) described the use of cloze as an instructional tool to guide teachers in choosing texts at the appropriate level for their students. In this study, the authors found that random deletion of words was more consistent and reliable than cloze assessments based on a fixed-deletion method, such as every fifth word.

Maze emerged as an alternative to cloze in the early 1970's as a comprehension assessment. Early in the development of maze, it was described as a multiple choice type of cloze assessment but over time has emerged as a reliable and valid assessment of reading comprehension distinct from its cloze predecessor (Decker et al., 2014; Fuchs & Fuchs, 1992; Graney, Martinez, Missall, & Aricak, 2010; Guthrie, 1973; Johnson, Semmelroth, Allison, & Fritsch, 2013; Marcotte & Hintze, 2009; Shin, Deno & Espin, 2000; Williams, Ari, & Santamaria, 2011); however, similar to other reading comprehension assessments, maze does not typically provide instructional information for teachers to develop targeted interventions for students with comprehension weaknesses. Maze has specific advantages and disadvantages as a comprehension assessment. It is a quick assessment that assesses basic comprehension ability not all skills necessary to define the construct. It is cost- and time-effective in comparison to oral reading fluency assessment since group administration is an option. Oral reading fluency tests require individual administration so a student can read aloud while someone closely monitors performance typically for one minute. However, it should be noted that oral reading fluency has been shown to be more reliable and valid at predicting overall reading ability than maze in some studies. For example, Ardoin et al. (2004) found that maze did not add unique variance beyond oral reading fluency to measures of broad reading skill. The efficiency of group administration should be taken into account when the individual variances of the two measures are similar.

The noted criticisms of reading comprehension assessments highlight the need for a new approach to address the lack of instructionally useful information provided. The available assessments, including maze, are not accurate at pinpointing deficits or instructional needs. Accurate and informative comprehension assessments are essential for developing targeted, skill-specific interventions for students. In terms of general comprehension assessment recommendations, the RAND Reading Study Group (2002) suggested criteria for developing more informative reading comprehension assessments with the purpose of improving early identification of specific comprehension weaknesses in children. These guidelines include reliability and validity at the level of the individual item, theory-driven item development, and sensitivity to different levels of reading.

Specifically, regarding maze as a comprehension assessment, January and Ardoin (2012) suggested the development of probes with sentence-level deletions and better distractors to improve the instructional utility of the assessment. The idea of improving distractors is not a new one. McKenna and Miller (1980) recommended improving distractor selection to increase construct validity of maze as a reading comprehension measure. The noted recommendation in the study was to include distractors that are the same part of speech and require the reader to understand previous sentences to rule it out as incorrect. A review of maze studies completed by Parker, Hasbrouck, and Tindal (1992) also recommended increasing the difficulty of distractors as a way to address the criticism that maze assesses sentence-level comprehension only. The authors specifically noted that the most common version of maze construction (i.e., fixed-word deletion) should be revised to improve construct validity.

CHAPTER TWO

REVIEW OF LITERATURE

Despite the need for a comprehension assessment that provides instructionally appropriate information for the large percentage of students in upper-elementary and beyond who struggle with reading comprehension, the overall purpose of maze as a comprehension assessment has shifted away from providing instructional information to providing an overall evaluation of reading skill. Some of the earlier studies on maze (i.e., Guthrie, 1973; Guthrie, Seifert, Burnham, & Caplan, 1974; Pikulsi & Pikulsi, 1977) focused on the instructional information that could be gained. Yet, the overall trend has been to use maze as an assessment to screen and monitor performance. The overall purpose of the present study was to design a system of efficient comprehension measures that can be used to inform classroom comprehension instruction.

To accomplish this, a comprehensive search of the literature was completed on the use of maze as a comprehension assessment. Studies with assessment materials in a language other than English or that include maze only as an outcome measure in a specific intervention study were excluded. An initial search was performed using the 'JEWL search engine' from James E. Walker Library of Middle Tennessee State University. The search engine accesses multiple databases, including *ERIC*, Education Source, Academic OneFile, ScienceDirect, JSTOR, and PsycArticles. The initial search terms included: reading comprehension, assessment, and maze. From the initial search, twenty-two studies were found. Two of the articles were excluded because assessment materials were created in a language other than English. An additional six studies were excluded because the purpose of maze was as an outcome measure only in an intervention study. A second search was completed in the same database with the terms reading assessment and maze. An additional fifteen studies were produced that were not part of the initial search. Nine of the fifteen articles were excluded for reporting maze only as an outcome measure in an intervention study. The reference sections of the twenty studies were carefully reviewed and an additional fifteen articles were obtained that fit the inclusion criteria.

In general, several aspects about the construction, administration, and scoring of maze as a comprehension assessment have been explored in research. There are multiple terms used to describe maze in the studies reviewed such as maze fluency (Foorman & Petscher, 2010), curriculum based measurement silent reading (Brown-Chidsey, Davis, & Maya, 2003), CBM maze (Graney et al., 2010; Mercer et al., 2012), CBM-mR (Yeo, Fearrington, & Christ, 2012) and maze selection measures (Ticha, Espin, & Wayman, 2009). Interestingly, only six of the 35 studies reviewed created maze with a strategy other than fixed word-deletion (e.g., every *n*th word). Two of the studies did not report a deletion pattern. General validity and reliability of maze were adequate but varied considerably across the studies reviewed. In regard to reliability, 14 of the studies reviewed did not report reliability information. Only five of the studies reported reliability for maze above .90. Four of the studies reported inter-rater reliability only which ranged from 93-100%. The remaining studies provided a wide range of reliability coefficients between .26 and .89. Eleven studies did not report validity information. Similar to the reliability, there was a wide range of validity coefficients reported across 19 studies from .23 to .88. The studies included concurrent and predictive validity to standardized reading comprehension tests and end-of-year state level tests in reading. All of the comparisons across studies were made with only one measure of reading comprehension as the criterion. One study used teacher judgment to assess the face validity of maze.

Instructional Utility of Maze

Several of the earliest studies found on maze (Guthrie, 1973; Guthrie et al., 1974; Pikulsi & Pikulsi, 1977) focused on the instructional information provided from the assessment. For example, Guthrie (1973) examined the differences in overall performance, proficiency with different parts of speech, and whether errors were syntactic or semantic across groups of readers with varying skills. Good readers in this study were defined as reading at the grade-appropriate level on the Gates-MacGinitie Vocabulary Test and scored within the Average range on the Peabody Picture Vocabulary Test. Syntactic and semantic errors did not vary for good readers and students reading more than two grade levels below expected grade placement based on the Gates-MacGinitie Vocabulary Test. Guthrie et al. (1974) addressed instructional questions by administering maze to good and poor readers as well. In this study, good and poor readers were primarily differentiated by performance on an oral reading fluency and a maze task. Poor readers had a lower overall mean and larger standard deviation than good readers on both oral reading and maze performance compared to the good readers indicating maze is able to discriminate students with varying skill levels.

Pikulsi and Pikulsi (1977) also focused on utility by examining the accuracy of maze with assisting teachers in determining an instructional reading level and group placement. In this study, maze was not deemed as a useful instructional tool since it overestimated the students' level in reading compared to teacher ratings of comprehension performance. Validity in this study was determined solely based on teacher ratings of performance rather than validated comprehension measures. Teachers judged the comprehension of students to be lower than the level suggested by the maze assessment used in this study.

Gillingham and Garner (1992) evaluated the instructional information provided from maze by looking at proficiency across different types of maze as validated by a summarization task of the material read. Specifically, the authors constructed different types of maze by creating tasks involving deletion of words within a passage with distractors, deletion of entire sentences with sentence distractors and a correct option, and a paragraph type of maze in which students had to choose the best fitting main idea of the paragraph from three options. Gillingham and Garner (1992) hypothesized that students would reach proficiency at the word, sentence, and paragraph levels in a hierarchical pattern to reflect increasing skill. The study was conducted with students in seventh grade and high school level. At both grades, a hierarchical pattern of performance was found such that students proficient with the paragraph maze task were also proficient with the sentence and word maze task. This means that students who had trouble with the maze task involving deletion of words were likely to have trouble with the sentence and paragraph maze tasks. Also, in this study, those proficient with the paragraph maze task generated the most detailed summaries.

Gillingham and Garner (1992) found a trend with some seventh-grade students in reaching proficiency with the paragraph type of maze and not with the word or sentence type. This finding was surprising for the authors since it was expected that the paragraph maze task would be more difficult than the other maze tasks. The authors suggested the unexpected pattern of performance found for some seventh grade students was likely attributed to frequent guessing (Gillingham & Garner, 1992). Yet, the authors did not seem to link the possibility of a non-hierarchical pattern of performance to differing reading skills or deficits. For instance, students proficient at determining the most appropriate main idea for a paragraph but unable to select the appropriate word among three options may have specific trouble with reading words but demonstrate a strength with other skills important for comprehension such as listening comprehension or vocabulary.

Although the instructional information provided from an assessment of reading is recognized as important, the type of research on comprehension assessments are often limited in this regard. McKenna and Miller (1980) provided a review of research on maze as a type of assessment. A focus of the research review was to analyze the types of distractors used when constructing maze. The authors focused on deletion of content words in the passage rather than a fixed-deletion strategy within their review. Parker et al. (1992) provided another review of maze. The way words are deleted as well as the length of the passage were significant factors discussed by these authors. In fact, a common feature of both reviews was the call for more research on the usefulness of maze in assisting teachers with making instructional decisions. A link is needed between the construction and format of the maze task, as highlighted in many studies, to the comprehension skills and strategies of the reader.

Parker, Guillemard, Goetz, and Galarza (1996) created a different type of maze assessment task with the intent to provide instructionally-relevant information using maze-like semantic maps in the area of science. These maze-like tasks included semantically linked science words with blank sections to be completed by filling in the associated words to the ones provided. Twyman and Tindal (2007) adapted traditional maze probes for middle school students by creating two concept maps, one for examples, and the other for attributes. In this study, some semantic maps consisted of examples and the category to fit the examples had to be supplied. A second type of semantic map included attributes, or features, of a word or concept and the overall concept or word had to be filled in based on the attributes. The concept maze for attributes was found to be more challenging than the traditional maze probe or the concept maze for examples. Students scored higher on the traditional maze probe but also had the most variable performance on this measure than on the semantic map examples.

In order to compare performance on maze to other comprehension tasks, Spear-Swerling (2004) compared performance on maze to a reading comprehension assessment with multiple-choice questions. The maze task analyzed in this study was from a standardized assessment knows as the Degrees of Reading Power. The researchers compared student profiles and skills to performance on each of the reading comprehension measures. The maze passages on the Degrees of Reading Power are created with expository passages and constructed to require integration of information across all sentences. Performance on maze was a significant predictor of performance on word reading, vocabulary, and listening comprehension tasks suggesting these skills are important for completing the task. Therefore, students with difficulty on this measure may benefit from instruction in these skill areas. Performance on the multiple-choice comprehension assessment in this study was predictive of vocabulary and listening comprehension. The results of this study demonstrated that assessment with maze relies more on word reading skill than assessments requiring students to answer multiple choice questions following a passage (Spear-Swerling, 2004). The understanding of how component, or underlying, skills are measured across comprehension assessments with varying formats and task requirements could provide important, instructionally useful information.

Kendeou et al. (2012) also examined the underlying skills measured by a maze task with the intent of providing information to target specific skills during intervention. In this study, performance on maze was significantly predicted by reading fluency and vocabulary. On the WJ-III PC subtest, which utilizes the cloze procedure to assess comprehension, working memory and orthographic processing were found to be significant predictors of performance which were also predictive of student performance on a recall task. Therefore, in this study, maze measured different underlying skills than other commonly used reading comprehension assessments such as cloze and retell (Kendeou et al., 2012). This difference was significant because some skills can be directly targeted for intervention such as the ones identified as important for maze. For students with poor performance on maze, reading fluency and vocabulary skill can be further analyzed and targeted for intervention.

Carlson, Seipel, and McMaster (2014) analyzed a maze task created with sentences deleted rather than words. For this task, the sixth sentence of a seven sentence, narrative text was deleted and replaced with four response options. The four options represented specific types of responses including one of two types of inferences, paraphrase, or associations. One of the primary findings of this study suggested that students with good comprehension skills as measured by performance on a standardized measure of reading comprehension were significantly more likely to pick the causally coherent inference than students with average or lower skills. The purpose of creating an assessment in this way was to identify skills to target for students based on their choices for the missing sentence.

Some studies have specifically examined the impact of reading maze aloud on student performance. For example, a study (Hale et al., 2011b) compared maze performance for students in first and second grade in terms of predicting performance on a broad reading measure when the maze passage was read aloud or read silently. There was not a significant within-subjects main effect for reading mode (silent vs. aloud) on maze found for the first and second grade students in the study. Ticha et al. (2009) also directly compared the impact of reading aloud on maze performance. In this study, students read maze aloud and their score was based on the number correct within the measured time. Performance on maze when read aloud was predictive of overall reading skill on the criterion measures used. Performance on oral reading fluency was also predictive of overall reading skill. However, only maze growth grates were significantly related to achievement increases on criterion measures. This finding suggested that measures used to monitor reading performance may need to shift as students get older and their overall reading skill increases. Maze read aloud and oral reading fluency contributed to overall reading proficiency; however, maze was more efficient to administer to large groups and more strongly related to achievement changes on comprehension criterion measures in this study.

Reading fluency and maze. There are a variety of ways to combine reading fluency and maze assessments. Hale et al. (2011) found a moderate and significant correlation between a maze task and the WJ-III PC subtest; however, a stronger correlation was exhibited with the maze accurate response rate (*r* = .64; calculated by taking the number of correct responses on the maze task and multiplying it by 60 and then dividing it by the sum of the number of seconds required to read the passage; Hale et al., 2011). During administration of the maze task in this study, passages were read aloud by students so that the time could be recorded. The authors proposed reading maze aloud as an alternate method for assessing the reading skills of students in middle and high school since the usefulness of fluency as a general measure of reading seems to decrease as students get older (i.e., Decker et al., 2014; Hale et al, 2011b). The rate measures for assessing older students' reading skills combines maze with oral reading fluency. Reading comprehension rate was specifically proposed as an alternative to oral reading fluency measures for students in middle and high school. Reading comprehension rate was calculated by obtaining percentage correct on multiple choice questions and multiplying by 60 then diving by the number of seconds required to read the passage. It was found to be the best predictor of performance on the WJ-III PC subtest than the other combined measures of fluency and comprehension. However, maze accurate response rate, the same scoring but with percentage correct on a maze task as described earlier, was also found to be a strong predictor. These two assessments require the student to read aloud. In contrast, most maze assessments reviewed required students to respond as they read silently. Another example of combining fluency with a maze measure, Hale, Skinner, Wilhoit, Ciancio, and Morrow (2012) compared performance on a timed maze task to performance answering multiple-choice factual and inferential questions and found that the reading speed correlations were stronger for maze than for answering questions.

The impact of reading fluency skills has been examined in terms of predictive validity on several comprehension assessments (Cutting & Scarborough, 2006; Keenan et al., 2008; Keenan & Meenan, 2014; Spear-Swerling, 2004). Specifically, noted in Eason et al. (2013), oral reading rate accounted for additional variance on five comprehension measures; however, the amount varied widely ranging from 7.9% to 28.1%. Specifically, in this study, poor fluency skills were more likely to impede reading comprehension performance on tasks that require timed reading, assessments with long passages, as well as completion of multiple choice questions after reading (Eason et al., 2013).

Guthrie et al. (1974) proposed using percent correct on maze and oral reading fluency assessments as a guide for teachers to determine if fluency or comprehension should be targeted for intervention. A much later study by Decker, Hixson, Shaw, and Johnson (2014) provided a similar recommendation by combining maze with oral reading fluency to make screening decisions. Combining fluency and maze could improve screening accuracy across grade levels and provide information relevant for instruction and intervention. In Decker et al. 2014, the authors indicated that multiple measures during screening would increase the accuracy of identification of students needing intervention since the levels of the two measures across grade levels varied considerably.

Screening and Progress Monitoring

Fuchs and Fuchs (1992) analyzed the criterion validity of maze and was the first to describe maze as a form of curriculum-based measurement (CBM). They suggested maze would be more efficient than oral reading fluency CBM since maze could be administered to groups of students and administered on the computer. It appeared to function similarly to oral reading fluency as a measure of growth. Concurrent validity to a standardized comprehension measure was adequate and superior to both cloze and retell. Specific maze construction and administration characteristics were described such as how words are deleted, the types of distractors that could be included, and how much time was allowed. In this study, students monitored with maze had better achievement outcomes than the students in the control group who were not routinely monitored. Also, maze was better at monitoring growth when administered regularly than retell or cloze measures. Teacher satisfaction with the maze was reported as high (Fuchs & Fuchs, 1992). The Fuchs and Fuchs (1992) study marked the use of maze as a timed measure and provided support for it as a progress monitoring assessment to document student growth by repeated administration of probes and progress represented visually on graphs. Although it was recommended in Fuchs and Fuchs (1992) that future research provide more qualitative information about maze performance for instructional utility, the purpose of maze at this point seemed to shift away from providing instructional information to an assessment used for screening, monitoring, and predicting reading comprehension performance.

Swain and Allinder (1996) analyzed maze based on sensitivity to growth during an intervention. In this study, second grade students were monitored repeatedly in the areas of oral reading fluency and maze as part of a documented reading fluency intervention. The fluency measure was determined to be more sensitive to growth than maze. The intervention, however, was more aligned with the fluency measure than the maze task. Students in the study directly practiced reading fluency skills within the context of the intervention but did not practice the maze task or focus on comprehension, vocabulary, or word reading skills also known to be important for predicting performance on maze. A practical implication from the Swain and Allinder (1996) study is that accurately identifying growth during an intervention likely depends on appropriately matching progress monitoring tools with the intervention. Ticha et al. (2009) found that the validity of the slope on progress monitoring measurement was dependent on aligning the progress monitoring tool with the intervention outcome.

Tolar et al. (2012) looked at growth rates with maze as a progress monitoring tool and found that growth rates were not predictive of end of year reading scores for sixth, seventh, and eighth grade students (r = .25 - .38) on a standardized state test in Texas. Yeo, Fearrington, and Christ (2012) found growth rates for maze scores measured over time did not provide a significant prediction for performance on the reading portion of state assessment used in Tennessee. In addition, they found that the relationship between the progress on the fluency and maze assessments was non-significant indicating that growth might be distinct and dependent on each method of assessment. The overall correlations reported in this study between maze and the state test ranged from .45 to .66 for students in grades three through eight (Yeo et al., 2012). However, the authors urged for caution in relying solely on one measure such as maze or oral reading fluency to evaluate progress since the growth scores were not significant predictors on state tests as measured in the study and did not demonstrate a significant relationship with each other.

Brown-Chidsey et al. (2003) created maze assessments from passages within grade level curriculum materials and/or basal reading programs. The study found that the maze task, which was described as curriculum based measurement (CBM) silent reading, was a valid method for differentiating comprehension skills for students across all levels of performance. Ticha et al. (2009) found that maze was more sensitive to growth than oral reading fluency as measured by an increase in correct choices each week. The results of the study suggested maze is sensitive to growth based on a significant increase in correct choices on probes collected weekly over 10-weeks. The growth reflected on the maze task was related to reading level and to pre- to post-test change on the WJ-III PC subtest. In contrast to many of the studies on maze, the measures in this study were created from newspapers. Three different timeframes were used during administration (2-, 3-, and 4-minutes; Ticha et al., 2009). Similar to the previous studies, Shin et al. (2000) found maze to be a reliable measure for estimating growth. Shin and colleagues (2000) suggested using performance on maze to make instructional changes for students if growth over time was considered inadequate. Although maze can be useful at determining when an intervention should be adjusted, it does not provide information about how the intervention should change.

Predictive Validity of Maze

Several studies have specifically analyzed maze in terms of the predictive validity to overall reading achievement as measured by individually administered, validated measures of comprehension (Ardoin et al., 2004; Williams et al., 2011) or performance on state-wide tests of reading (Graney et al., 2010; Silberglitt, Burns, Madyun, & Lail, 2006; Yeo et al., 2012). Williams et al. (2011) identified a significant correlation of .68 between a maze task and the Nelson-Denney reading test. Similarly, maze was found to be a stronger predictor of performance on the Nelson-Denney reading test than a cloze assessment task in which students had to supply a missing word. Maze in this study also correlated with performance on literal and inferential questions for average and struggling readers enrolled in a college preparatory reading course. It should be noted that struggling readers were defined in this study as students required to repeat the preparatory course in reading due to initial poor performance (Williams et al., 2011). The definitions of average versus struggling readers in this study were problematic because skill level is not the only factor that could contribute to poor performance in the course.

Several studies have shown maze to be predictive of scores on high-stakes tests. Shin et al. (2000) found maze to have a positive relationship with the *California Achievement Test* scores. Graney et al. (2010) found a significant correlation of .67 between a maze task and the Indiana Statewide Testing for Educational Progress (ISTEP). Both maze and oral reading fluency performance were found to be predictive of performance on the ISTEP; however, the authors recommended the use of maze due to the efficiency of administering maze in group format (Graney et al., 2010). A study by Marcotte and Hintz (2009) found that the maze task explained 40 percent of the variance on the Group reading Assessment and Diagnostic Evaluation (GRADE). Maze in this study also exhibited a strong and significant relationship with oral reading fluency and sentence verification technique and a lesser relationship with retell fluency and written retell (Marcotte & Hintz, 2009). It is significant to note that both the GRADE and Woodcock-Johnson III Tests of Achievement reading scores rely on a cloze task in which students supply a missing word for sentences and paragraphs. Therefore, it is possible that the relationship maze has with reading comprehension measures may be impacted by the task or format of the measure as well as by the differences in the construction and administration of maze. Across studies reviewed, there are differences among validated measures of reading comprehension as well as variations across maze tasks in terms of passage construction, administration procedures, and response format.

Johnson et al. (2013) looked at predictive and concurrent validity of maze with a state science assessment. The authors constructed maze passages with science textbooks by deleting every seventh word and creating specific distractor options consistent with the maze administration and construction described in Fuchs and Fuchs (1992). Despite a range of topics, there was high consistency across passages. Concurrent validity with the Test of Silent Reading Efficiency and Comprehension was moderate and significant (r = .65). The maze task with science passages also exhibited a predictive validity coefficient of .46 to the Idaho State Achievement Test, Science. Decker et al. (2014) identified moderately strong correlations between a maze task and the Michigan Education Assessment Program (MEAP) for seventh (*r* = .54) and eighth grade students (*r* = .58). The authors specifically sought to answer whether maze added to the prediction of MEAP after accounting for oral reading fluency. Interestingly, differing patterns were found for students in seventh and eighth grades in this study with maze being a stronger predictor for students in seventh grade. Oral reading fluency was the stronger predictor on the state test for students in eighth grade (Decker et al., 2014).

Several other studies have found less than favorable correlations and reliability for maze scores. For example, Merino and Beckman (2010) found that maze scores alone were not predictive of performance on a computer adaptive reading test, Measures of Academic Progress (MAP). In this study, maze did predict reading scores for students in fourth grade; however, oral reading fluency was a better predictor than maze at all grade levels including fourth grade. When oral reading fluency was added into the model, maze did not add to the prediction of reading scores on MAP. Ardoin et al. (2004) found that the maze test in their study did not explain significant unique variance on the WJ-III Broad Reading composite after accounting for oral reading fluency. The contrasting findings among studies on maze could be due to the inherent differences in the comprehension tests used for comparison. Further, actual administration of maze has been shown to vary considerably across studies in terms of construction and administration.

See Table 1 for a summation of the primary studies analyzed in terms of grade level students were assessed, reported reliability and validity coefficients, type and number of passages used, as well as whether timing was included and if so, the number of minutes.

Construction of Maze Assessment

Text type. The selection of text used to construct maze passages varies widely across the primary studies analyzed. The categories of text represented across the thirty-five primary studies include basal readers or other grade-level specific reading material, science texts, expository passages (content area not specified), or narrative text. Several studies reported using maze passages from Aimsweb, which contained narrative text separated into grade-level specific passages (Shinn & Shinn, 2002). One study (Brown-Chidsey et al., 2003) specified the use of history text for maze construction. Ticha et al. (2009) constructed maze passages from newspapers. Two studies reported using material from science texts to create maze passages. Six studies reviewed did not provide information about the specific types of text used to construct the maze task.

Table 1

Analysis of Primary Studies on Maze (n = 35)

Study	Grade	Reliability	Validity	Text Type	Timed	Passage
Guthrie (1973)	2-4	.90-.93	.82-.85	Basal	-	1
Guthrie et al., (1974)	2	None	None	Basal	-	1
Pikulsi & Pikulsi (1977)	5	None	Teacher Judgment	Basal	-	1
Gillingham & Garner (1992)*	7, 9-11	None	None	Science	-	1
Fuchs & Fuchs (1992)	5-8	None	.74	NS	2.5 m	1
Jenkins & Jewell (1993)	2-6	None	.65-.76	NS	1m	3
Espin & Foegen (1996)	6-8	None	.56-.62	NS	2m	1
Parker et al., (1996)	7-8	.48-.70	.23-.62	Science	30m	1
Swain & Allinder (1996)	2	93% agree	None	NS	2.5m	1
Shin, Deno & Espin (2000)	2	.66-.83	None	Basal	3m	1
Brown-Chidsey, Davis, & Maya (2003)	5-8	.26-.69	None	History	10m	1
Spear-Swerling (2004)	4	.94	.76	Expository	75m	1

Table 1 Continued

Study	Grade	Reliability	Validity	Text Type	Timed	Passage
Ardoin et al. (2004)	3	99%	.31-.62	Basal	3m	1
Silberglitt et al. (2006)	3, 5, 7-8	.79-.97	.48-.54	Aimsweb	3m	3
Marcotte & Hintz (2009)	4	.83	.67-.72	Basal	3m	3
Fore III et al. (2009)	6-8	None	.46-.88	NS	3m	1
Ticha, Espin, & Wayman (2009)	8	>.80	.80-.88	Newspaper	2, 3, 4	1
Rutherford-Becker & Vanderwood (2009)	4-5	None	None	Aimsweb	3m	3
Foorman & Petscher (2010)	3-12	.77-.90	.51-.64	NS	3m	2
Merino & Beckman (2010)	2-5	None	None	Aimsweb	1m	3
Graney et al. (2010)	4-5	.66-.89	.41-.67	Aimsweb	3m	3
Hale et al. (2011)*	6-8	100%	.31	Aimsweb	-	2
Hale et al. (2011b)	1-2	94-96%	.86-.89	Aimsweb	3m	3
Kendeou et al. (2012)	1-2	.82	.37-.62	Narrative	1m	3

Table 1 Continued

Study	Grade	Reliability	Validity	Text Type	Timed	Passage
Mercer et al., (2012)	3-5	.82-.87	None	Aimsweb	1, 2, 3m	9
Hale et al. (2012)*	6-7	>.80	.63	Aimsweb	1m	1
Yeo et al. (2012)	3-8	None	.58-.81	Aimsweb	3m	1
January & Ardoin (2012)	3-5	None	None	Aimsweb	3m	1
Tolar et al. (2012)	6-8	.74-.86	.45-.70	Aimsweb	3m	1
Johnson et al. (2013)	7	.56-.80	.63-.67	Science	3m	3
Decker et al. (2014)	7-8	None	.54-.72	Basal	2.5m	1
Carlson, Siepel, & McMaster (2014)	3-5	.60-.80	None	Narrative	-	-
Piece, McMaster, & Deno (2010)	1-12	.86-.90	.74-.79	Proact	2m	1
Brown-Chidsey et al. (2005)	5	None	None	History	2m	3
Williams, Ari, & Santamaria (2011)	13+	None	.62-.68	Narrative	3m	4-5

Note: NS = Not specified

Due to the variety of passages used to construct maze assessments, one study (Brown-Chidsey, Johnson & Fernstrom, 2005) specifically compared performance on maze created with grade-level controlled passages to ones created from children's literature. In general, the results of the study indicate that students had more correct responses during the two-minute timeframe on grade-level controlled passages than on literature-based passages. Students demonstrated growth over time on both types of passages; however, differences were significant at the fall, winter, and spring benchmarks suggesting that performance on maze can be impacted by the type of passage used to create the measure. This finding is important since a variety of text types are used across the studies reviewed on maze. Also, it suggests that completing maze with grade-level controlled passages may be easier than literature passages but not why this difference is found. Specific passage details were not provided such as number of paragraphs or sentences, Lexile range, level of cohesion, and difficulty level of the words used. The study did not address how using expository passages or content-area textbooks to create maze would impact performance. Consistency across types of passages is important for accurate measurement of growth (Brown-Chidsey et al., 2005). Further, type of text may be a factor contributing to the contrasting findings across studies in terms of validity.

Deletion ratio. Previous studies have primarily used a fixed-word deletion strategy by deleting every *n*th word while leaving the first and last sentence intact (e.g., Ardoin et al., 2004; Brown-Chidsey et al., 2003; Fuchs & Fuchs, 1992; Graney et al., 2010; Hale et al., 2011b; Johnson et al., 2013; Kendeou et al., 2012; Marcotte & Hintz, 2009; Mercer et al., 2012; Rutherford-Becker & Vanderwood, 2009; Shin et al., 2000; Silberglitt et al., 2006; Swain & Allinder, 1996; Ticha et al., 2009; Tolar et al., 2012; Williams et al., 2011; and, Yeo et al., 2011). Guthrie (1973) deleted every fifth word; however, deletions were specified based on equal amounts of four categories of words: nouns, verbs, modifiers, and function words. The most common deletion pattern in the studies reviewed was deletion of every seventh word.

In one of the studies considered to be the first on maze as a comprehension assessment, Kingston and Weaver (1970) compared the one in five deletion ratio to deletion of words based on part of speech. The authors termed deletion of words based on part of speech as a lexical deletion pattern. Nouns, verbs, and adjectives were the parts of speech targeted for deletion. In this study, no consistent differences were noted in terms of reliability or validity among the lexical or fixed deletion patterns. Interestingly, cloze was superior to both forms of maze in terms of predictive validity to the California Achievement Test. A review by Parker et al. (1992) found that fixed-word deletion ratios ranged from every fifth to every 46th word, with every seventh word being the most common across studies in their review.

January and Ardoin (2012) questioned the commonly used fixed-deletion strategy for maze and deleted words based on designation as function or content words. The authors of the study were interested in examining if the type of word deleted would influence accuracy on a maze task. Also, the authors were interested in determining if there would be differences in accuracy on scrambled versus intact probes. In this study, the mean for the intact probes was greater than the mean for the scrambled probes; however, the difference was not significant suggesting that students completed intact and scrambled probes with similar accuracy. Further, the main effect for type of word deleted was not significant indicating maze accuracy was not different based on the type of word deleted. However, the specific words deleted could not be controlled since it was based on placement in the passage. Following deletion of every seventh word, the words were categorized as function or content words. Control over the selection of the words to be deleted may result in differing results and as noted by the authors may also reflect a more accurate assessment of reading comprehension. Overall, the study by January and Ardoin (2012) provided support for using another measure of reading comprehension to supplement the results of maze when screening student's comprehension skills. Since performance did not vary significantly on the scrambled versus intact probes, the authors concluded that maze was a measure of sentence-level comprehension.

Across the studies reviewed, there were few exceptions to deleting single words from passages for maze. Gillingham and Garner (1992) created a hierarchical pattern of word, sentence, and paragraph types of mazes. One purpose for creating different types of maze tasks in this study was to determine if one of the maze types was more predictive of the ability to summarize the text. The sentence type of maze consisted of deletion of an entire sentence. The paragraph type of maze did not contain deletions within the passage yet students were required to select the main idea of the paragraph from three options. In Carlson et al. (2014), the sixth sentence of a seven sentence text was replaced with four specific response types. Each text had a title and an average of 80.5 words across seven sentences. The purpose of the sentence level deletion was to distinguish the types of responses made among students with varying levels of comprehension ability and to provide useful information about why readers have difficulty based on the type of response selected. Findings of the study indicate that good readers as defined by performance on a state reading test were more likely to choose the causally coherent inference response than poor or average readers. The average and poor readers were more likely to choose responses reflecting a paraphrase, local inferences, or lateral connections. The authors of the study suggest the response types developed for reading assessments can be used to assess underlying comprehension processes important for instructional and intervention planning. More work is needed to link carefully constructed responses to underlying skills. Parker et al. (1996) also used the sentence deletion strategy; however, the maze task was presented as semantic maps with choices to identify key word pairs related to science content. Deletion beyond individual words is rare across the studies reviewed.

Distractors. The type of distractors used when developing a maze assessment requires consideration during construction of passages. Kingston and Weaver (1970) noted that the distractors might be more important in terms of difficulty than the method used to delete words following the findings of their study in which differing deletion patterns did not significantly impact student performance. Distractor selection in many of the earlier studies on maze consisted of one option that was the same part of speech as the correct word but did not make contextual sense. The other option made contextual sense but was not the correct part of speech (Gillingham & Garner, 1992; Guthrie, 1973; Guthrie et al., 1974). For example, the word *he* could be replaced with three words including he and two distractors. One of the distractors could also be the same part of speech but not make sense with the passage (i.e, *they*) and the other distractor would be a different part of speech but fit the overall meaning of the passage or story (i.e., *kittens*). Guthrie (1973) originally prescribed distractor selection from word lists or from other words within the passage itself. In a study by McKenna and Miller (1980) different types of maze distractors were analyzed. The authors found that the difficulty level of items was most impacted by syntax. Distractors that were the same part of speech as the target word were more difficult to discriminate. Visual similarity among words did not contribute to difficulty of items. Other studies have considered distractors in terms of the placement on the page. McKenna and Miller (1980) discussed the placement of distractors within the sentence, underneath, or in a column to the right of the passage.

Fuchs and Fuchs (1992) outlined specific criteria for selecting distractors such as the following: the words had to be within one letter of the same length as the correct choice, could not make contextual sense, be a nonsense word, rhyme with the correct choice, be a unique or difficult vocabulary word, or require students to read more than 1.5 lines ahead to determine the meaning. The majority of subsequent studies followed the criteria for distractors outlined in Fuchs and Fuchs (1992). Brown-Chidsey and colleagues (2003) included distractors that were the same part of speech or a different part of speech. Only the correct choice made contextual sense. Similarly, several of the studies reviewed used distractors that were the same part of speech as the target word (e.g. Graney et al., 2010; Mercer et al., 2012; Silberglitt et al., 2006; Shinn & Shinn, 2002). Shinn and Shinn (2002) described the procedure for the development of the maze passages in Aimsweb that was used by eleven of the thirty-five studies. The distractors included in Aimsweb are words selected at random from the story.

Parker et al. (1992) specifically recommended increasing the difficulty of distractors by making them the same part of speech and meaningful within the sentence. The authors suggested that incorporating more difficult distractors would improve construct validity of the common form of maze. In most studies reviewed in Parker et al. (1992), the distractors were not meaningful within the sentence and studies were split equally in terms of distractors being the same or different part of speech. It is difficult to make accurate comparisons of maze assessments across studies due to the differences in the construction of passages and creation of distractors. Further, administration and scoring of maze varied across the studies reviewed.

Administration and Scoring of Maze Assessments

Timing. The use of timing is a major factor across studies contributing to differences in the administration of maze. Fuchs and Fuchs (1992) introduced timing as a component for maze tasks allowing 2.5 minutes to complete a maze passage presented on the computer. Of the thirty-five primary studies reviewed, fourteen of them used a 3 minute timeframe to administer maze (refer to Table 1). Two studies in addition to Fuchs and Fuchs (1992) used the 2.5 minute timeframe (i.e., Swain & Allinder, 1996; Decker et al., 2014). Another timeframe reported in three studies was 1 minute (Hale et al., 2012; Kendeou et al., 2012; Merino & Beckman, 2010). In January and Ardoin (2012), performance was marked at the 3 minute mark and the student continued reading the remainder of the passage. Brown-Chidsey et al. (2003) noted that 10 minutes for each 250-word passage was too long due to the high scores of all participants in their study. The use of timing is also important for determining the score and will be further explored in the section addressing the scoring methods for maze.

Ticha et al. (2009) specifically analyzed the impact varying amounts of time during administration of maze had on the reliability and validity of the assessment. The study looked at administration of maze in two, three, and four minute timeframes. No significant impact of time was noted on reliability or validity. Across timed conditions in the study, students made an average of seven correct choices in a 1 minute time-frame (Ticha et al., 2009). Variations across timing practices during administration of maze were present across the primary studies reviewed.

Number of passages. The ideal number of passages to use for maze has been debated and the number used varied across the studies reviewed. Mercer et al. (2012) recommended administration of more than one maze passage to increase reliability for screening purposes and high-stakes decisions such as intervention placement. Mercer et al. (2012) found that reliability was higher for students in fifth grade than for students in third or fourth with the administration of two probes (r = .87). However, for students in third and fourth grades, administration of three probes was needed to establish adequate reliability of .83 (Mercer et al., 2012). Yet, in the majority of studies reviewed, administration of one passage was most common (*n* = 20). Refer to Table 1.

Three of the studies reviewed (i.e., Hale et al., 2011b; Johnson et al., 2013; Marcotte & Hintz, 2009) administered three passages, each with a 3 minute time limit. Williams et al. (2013) administered four to five passages each for 3 minutes. A few studies used three passages each with a 1 minute time limit (i.e., Jenkins & Jewell, 1993; Kendeou et al., 2012; Merino & Beckman, 2010). Interestingly, Hale et al. (2012) was an exception among the studies reviewed in that one passage was used with a 1 minute time limit. Since Mercer et al. (2012) found that administration of three passages improved reliability for students in fourth grade, three passages will be used for each assessment condition in the present study.

Scoring rules and guidelines. Scoring differences across studies were pervasive as well. Studies that administered more than one maze passage, such as in Hale et al. (2011), used a scoring method based on a median score of the passages. In some studies, the median score could be based on the number correct within 3 minutes as in Hale et al. (2011) and Williams et al. (2011) or represent the number correct within a 1 minute timeframe (see Merino & Beckman, 2010). Mercer et al. (2012) used the more common three-minute administration timeframe but based scores on the number of correct maze choices each minute. Graney et al. (2010) had students read one passage and used the number correct in 3 minutes as the score. Similarly, Decker et al. (2014) looked at the number of correct answers on one passage with a 2.5 minute timeframe. From the review of scoring guidelines, it is apparent that comparisons of student performance across different types of maze tasks is problematic to variability in how they are scored, created, and administered.

Fore III, Boom, Burke, and Martin (2009) assessed performance on maze by subtracting the number of incorrect answers from the total number of items attempted, similar to the common method for determining words correct in a minute on oral reading fluency assessments. In Fore III et al. (2009) the score was calculated based on performance on one passage read for 3 minutes. Maze as used within Aimsweb also used the number correct as the total score. Kendeou et al. (2012) and Tolar et al. (2012) described a similar scoring strategy as the number correct minus the number incorrect. However, in Kendeou et al. (2012) students read three texts each for 1 minute, and in Tolar et al. (2012) students read a single passage for 3 minutes. Yet another strategy employed for scoring was described in Foorman and Petscher (2010) and was based on the average number correct in 3 minutes. Timing is a factor that impacts how the assessments are scored across studies. To decrease the impact that timing can have on student performance and scoring, a time limit will not be enforced for the maze and multiple choice passages in the study.

Ticha et al. (2009) presented a unique approach to scoring in an attempt to control for guessing. In this study, scoring was discontinued once a student made three consecutive errors. Also, in Brown-Chidsey et al. (2005), responses following three errors in a row were not counted. Additionally, the number of errors made prior to the three consecutive errors is divided in half and that number was subtracted from the correct word choices. Again, the widely varying scoring practices across the studies reviewed have significant implications for the comparability of student performance on maze tasks.

Pierce, McMaster, and Deno (2010) compared different maze scoring procedures in terms of how reliability, validity, and growth from fall to spring would be impacted. The scoring methods analyzed in Pierce et al. (2010) included the following: correct maze choices, correct minus incorrect maze choices, correct maze choices minus half of the incorrect maze choices, two error stop rule, and three error stop rule. Studies that employed more than one scoring adjustment were not included in the analysis. For example, in Brown-Chidsey et al. (2005) scoring was based on a three-error stop rule and correct maze choices minus half of the incorrect choices. The findings of Pierce et al. (2010) revealed that all scoring methods were similar in terms of concurrent and criterion validity as well as alternate form reliability. Also, all scoring methods analyzed in the study were able to reflect growth from fall to spring. In general, the coefficients across grade levels were lowest when correct choices were considered without a scoring adjustment (Pierce et al., 2010). Interestingly, Aimsweb calculated the maze score as the number of correct responses within a 3 minute timeframe. Since the majority of studies reviewed use the Aimsweb version of maze, number of correct responses within a 3 minute timeframe is the most common scoring method. Yet, the results from Pierce et al. (2010) suggest that an adjustment when scoring may be appropriate. For the present study, the number of correct responses will be considered; however, scoring adjustment calculations could be considered following administration in terms of two- or three-error stop rules. As stated previously, timing will not be part of the scoring since students will have as much time as needed to complete the assessments.

Purpose of the Study

Although maze originally emerged as a test to identify instructional reading levels, the shift in research and practice has been to use maze as a screening, monitoring, or predictive measure rather than an assessment tool to inform instruction and intervention. A comprehensive review of maze assessment revealed that most of the available maze passages are created by deleting every seventh word (i.e., fixed-word deletion). Text manipulations on maze comprehension assessments have included a few studies that delete words based on specific features of the word (i.e., Guthrie, 1973; Gillingham & Garner, 1992) or that delete entire sentences (Carlson et al. 2014; Gillingham and Garner, 1992).

While some studies have examined different types of maze assessments, these studies did not systematically compare performance across types of maze assessments in terms of reliability and concurrent construct validity based on multiple standardized measures of reading comprehension. Also, none of the studies reviewed have explored the differing correlations and predictions of different maze types on reader skills. The overall purpose of this research was to design a valid and reliable comprehension assessment that can ultimately be used for making instructional decisions about comprehension in the classroom. Specifically, this study analyzed the reliability, including item-level properties, as well as concurrent construct validity of four conditions (three types of maze and a multiple-choice comprehension assessment). Further, the sentence deletion and word-feature deletion conditions were analyzed to determine if they improved construct validity beyond the standard version of maze (i.e., fixed-word deletion) to the prediction of reading comprehension performance. In the present study, different types of maze tasks were validated against multiple reading comprehension tests. Multiple measures of maze and reading comprehension tests were used due to the differences in the construction, administration, and responses required from students. Specifically, the following questions were addressed:

Research Questions

1) Were the four assessment conditions in this study reliable based on a measure of internal consistency?
2) What were the correlations of the four assessment conditions to three validated measures of comprehension? Were there differences in correlations across standardized comprehension measures with varying formats, administration procedures, and response requirements? It was predicted that the correlations would vary across conditions. Specifically, the sentence-deletion and word-feature deletion conditions were predicted to be more highly correlated to the comprehension measures that relied on multiple-choice (i.e., GRMT) or open-ended responses (i.e., WIAT-III) than the fixed-word deletion condition which would be more highly correlated to the comprehension measure utilizing cloze response format (i.e., WJ-IV PC).

3) What were the correlations of the four assessment conditions to reader skills important for comprehension such as word identification, decoding, reading fluency, reading vocabulary, and morpho-syntactic knowledge? It was predicted that correlations across reader skills would vary for each assessment condition analyzed in the study. Specifically, the sentence-deletion and word-feature deletion conditions were hypothesized to be more strongly correlated with measures of reading vocabulary and morpho-syntactic knowledge than the fixed-deletion condition. The fixed-deletion condition was hypothesized to correlate more strongly with measures of word identification and decoding than the other conditions.
4) Which of the assessment conditions analyzed in the study contributed significant variance to a reading comprehension composite derived from three validated comprehension tests? It was hypothesized that the sentence-deletion and word-feature deletion conditions would improve the construct validity of fixed-word deletion maze by contributing significant variance to the composite.
5) Which of the reader skills analyzed in the study contributed significant variance to reading comprehension? It was predicted that reader skills would vary based on the comprehension test used with performance on the WJ-IV PC predicted by word reading and decoding skills more than vocabulary and syntax with opposite patterns predicted for the other two comprehension measures, GRMT-4 and WIAT-III RC.

CHAPTER THREE

METHOD

Participants

Participants were 93 fourth-grade students (53% boys) enrolled in a rural school district in the Southeastern US where the principal investigator is a school psychologist. Students from four schools participated in the study. Sixty-nine percent of the students in the sample were Caucasian, 14% Black/African American, 15% Hispanic, and 1% was Asian. Approximately 12% of students from the sample received special education services. Of those receiving special education services, three students in the sample were identified as Intellectually Gifted, three students receive special education services as a student with an Other Health Impairment, and five students receive services for a Specific Learning Disability in the area of reading. Students with a range of abilities were recruited to participate from 19 fourth-grade classrooms in four schools.

Fourth-grade students were selected for participation in this study considering late elementary often denotes the transition to increasingly complex, expository text. Also, noted amounts of students struggle specifically with poor comprehension during late elementary school (McMaster et al., 2014). The studies reviewed on maze as a comprehension assessment include students in general and special education in first through twelfth grades. Some studies included participants with a specific type of disability such as an emotional or behavioral disorder (Fore et al., 2009) or learning disability with an at-risk status in the area of reading (Parker et al., 1992; Pierce et al., 2010).

Materials

Passages. The passages used to create the four assessment conditions were selected from www.newsela.org and reflect a Lexile level appropriate for fourth-grade students (Lexile range 620 to 690). Twelve passages were taken from news stories that covered a range of topics in an effort to control for varying levels of background knowledge. Passages were chosen from topics in the areas of science, health, arts, as well as opinion articles and news specifically devoted to kids. They all contained a Flesch-Kincaid grade level between 4.4 and 5.2. The passages were numbered 1 to 12. Then, the numbers 1 to 12 were entered into a random number generator (www.random.org). The first three non-repeating numbered passages were included in the fixed-word deletion assessment condition, the next three numbered passages were put into the word-feature deletion condition, followed by three passages for the sentence deletion condition, and the last three numbered passages were put in the multiple choice condition. Details about the 12 passages are included in Table 2.

Passages with fixed-word deletion provided three options for every seventh word beginning with the second sentence. The first and last sentences of the passages in this condition were left intact. Participants selected the best-fitting word from the three options provided in place of the omitted words (included within parenthesis in the text separated by commas). The fixed-word deletion passages were created by pasting the passages into the maze passage generator (www.interventioncentral.org/test-of-reading-comprehension). Distractors included words selected randomly from the passage.

Table 2

Passage Characteristics for Assessment Conditions

Assessment Condition				
Passage	# of Words	# of Items	Lexile	Flesch-Kincaid
Fixed-Word Deletion				
Small Toys	635	88	670	4.5
Opinion	591	82	650	4.8
Wild Animals	583	79	690	5.1
Word-Feature Deletion				
Storms	629	25	680	4.9
Traditions	547	25	680	5.2
Teddy Bears	557	25	670	5.2
Sentence Deletion				
Zoo	612	10	690	5.2
Toilet	683	10	680	4.9
Support Program	578	10	670	4.4

Table 2 continued

Passage Characteristics for Each Assessment Condition

Assessment Condition

Passage	# of Words	# of Items	Lexile	Flesch-Kincaid
Multiple-Choice				
Ancient	580	10	690	4.8
View	424	10	680	5.2
Finding Dory	545	10	680	5.2

Note: # of words is based on passage length prior to manipulation into assessment conditions. # of items refers to the number of items deleted in the passage or the number of questions in the multiple choice condition.

Each distractor was checked to ensure it could not make contextual sense if it replaced the correct choice. Fifteen distractors across the three passages in this condition were replaced with new randomly generated words from the passage. Each distractor was looked at to ensure that it would not make sense in the context of the story. In two instances, the distractors were replaced because one of the words made contextual sense. In two additional instances, the correct word choice was repeated as a distractor and was replaced. The distractor was a non-word in one instance and ten distractors were replaced for using a person's first or last name from the article which appeared as a non-word.

Passages with word-feature deletion patterns contained omitted words based on the following categories: pronouns, conjunction words, and words central to the content of the passage (i.e., nouns). These types of words were chosen so that students would have to comprehend information from previous sentences to make a correct selection. The first and last sentences of the passages were left intact. Twenty-five deletions were made for each passage in this condition. Each omitted word was replaced with three choices including the correct word and two distractors. One distractor included a word that was the same part of speech as the correct word (i.e., pronoun, conjunction, or noun) but did not make sense in the passage. The other word used as a distractor was a different part of speech than the correct word. All distractors were taken from existing words in the passage similar to the fixed-word condition. Gellert and Elbro (2013) created cloze passages with pronouns, conjunctions, and lexical references deleted to address the criticism that cloze assessments primarily assess sentence level comprehension. Pronouns, conjunctions, and lexical references were chosen to be deleted since they require integration of information across sentences. When applied to cloze passages, the selection of specific words to delete that required the reader to make inferences across sentences improved the validity and sensitivity of the assessment. Specifically, the new version of cloze correlated strongly with a question-answer measure of reading comprehension and contributed significant variance above and beyond word reading and decoding skill (Gellert & Elbro, 2013). However, when applied specifically to maze passages in another study, January and Ardoin (2012) found no effect for type of word deleted on maze accuracy.

In the sentence-deletion condition, entire sentences were deleted and replaced with three options. The correct sentence and two distractor sentences were included within parentheses and separated by a forward slash. As with the other maze types described, the first and last sentence of each passage was left intact. For each passage, in this condition, 10 sentences were deleted and replaced with the correct sentence and two distractor sentences. One distractor sentence was designed to be syntactically incorrect. The words in the target sentence were scrambled so that the word order did not make sense. The other distractor sentence did not fit meaningfully within the paragraph but was an intact sentence taken from another part of the passage. For this assessment condition, participants selected the best-fitting sentence from the three options provided in place of the omitted sentence. Most of the primary studies analyzed on maze assessment used the single-word deletion strategy. Gillingham and Garner (1992) created maze passages with entire sentences deleted to address the hierarchy of comprehension at the word, sentence, and paragraph level when completing maze assessments. Carlson et al. (2014) replaced the sixth sentence of a seven sentence passage with four potential options designed to reflect the following types of responses: causally coherent, paraphrase, local inferences, and lateral connections. Good readers consistently chose the causally coherent sentence while average and poor readers were more likely to choose one of the other three alternatives (Carlson et al., 2014). Unfortunately, none of the studies using sentence level maze reported reliability.

In the multiple choice assessment condition, passages were left intact and 10 multiple choice questions were constructed. Some of the questions were adapted from the ones included with the original article on the website (newsela.org). The multiple choice questions included a clearly defined sentence stem or question. The distractors were consistent grammatically with the stem so that no tense or inconsistency could provide a clue about the correct answer. Distractors were similar in length to the correct answer. For example, after reading a passage about protecting a natural park area from oil drilling, students are presented with 10 questions such as the following:

1. Which selection BEST expresses a main idea of the article?
    a. Naylor and others want to protect the park's quiet and solitude.
    b. North Dakota is in the middle of an oil boom.
    c. The park is named after Theodore Roosevelt, U.S. president.
    d. Valerie Naylor fell in love with the park more than 40 years ago.

The position of the correct answer was varied in a random manner for all conditions. The three passages for each condition along with the all questions for the multiple choice condition are included in the Appendix. Refer to Table 2 for a summary of passage characteristics for each condition.

For the administration of the group measures, doctoral level graduate assistants read scripted directions and monitored student performance during testing. The 12 passages completed by all participants were scored by the primary investigator. All 12 passages for 10% of files were double scored and double entered by independent coders. Inter-rater agreement for scoring was 100%

Validated measures of reading. Doctoral level graduate assistants completed a training session for the individual measures administered in the study. Training sessions typically lasted 1 to 2 hours. All graduate assistants were familiar with administering educational tests and working in a school setting. They were encouraged to take the measures home to practice administering them. Scoring and administration was checked regularly by the primary investigator throughout the study. Discrepancies were addressed with graduate assistants. Administration errors noted during the fidelity checks included not obtaining an appropriate ceiling on the subtests assessing word reading and decoding. Errors were also noted with the basal on the WJ-IV PC subtest. The primary investigator corrected the errors by administering additional items to establish appropriate basals and ceilings so that the tests could be appropriately scored. All protocols were double-scored by a graduate assistant and the primary investigator.

Table 3 lists the assessment battery that was individually administered to participants. The assessment battery was intended to provide measures of underlying reader skills shown to be important for reading comprehension performance (Keenan et al. 2008). These skills included reading fluency, decoding, word recognition/identification, reading vocabulary, inferencing skill, and morpho-syntactic knowledge. The measures selected had adequate reliability and validity for assessing the reading skills of fourth-grade students. Administration of all individual measures was audio-recorded. Ten percent of audio-recordings were checked by a second examiner and each assessment was independently scored. Inter-rater agreement for scoring was

Table 3

Individual Assessment Battery for Fourth-Grade Students

	Assessment Battery
Underlying Skill Assessed	Measure
Reading Comprehension	WIAT-III Reading Comprehension subtest
	WJ-IV Passage Comprehension subtest
	GRMT-4 Reading Comprehension subtest
Word Reading and Decoding	WIAT-III Word Reading
	WIAT-III Pseudoword Decoding
Reading Fluency	Easy CBM Passage Reading Fluency – fourth grade
Vocabulary	WJ-IV Reading Vocabulary subtest
Morpho-syntactic Knowledge	Test of Morphological Awareness (Carlisle, 2000)

Note: WIAT-III = *Wechsler Individual Achievement Test – 3rd edition;* WJ-IV = *Woodcock-Johnson Tests of Achievement – 4th edition;* GRMT-4 = *Gates-MacGinitie Reading Test, 4th edition, Form S; Easy CBM =curriculum based measurement of oral reading fluency*

calculated by the following formula for each measure: Agreements/ (Agreements + Disagreements) X 100. Overall, reliability of administration and scoring of all individual measures was 97.23%. The majority of scoring disagreements were on the WIAT-III RC test in which partial scoring was an option. All discrepancies were reviewed and reconciled by the primary investigator by closely examining scoring criteria provided in the manuals for each test.

Reading comprehension. Previous research on reading comprehension assessment has demonstrated the inconsistency across measures on factors such as response format (Cain & Oakhill, 2006), text variables (Garcia & Cain, 2014), as well as procedures for administering the test (Garcia & Cain, 2014; Keenan & Meenan, 2014). Therefore, three standardized measures of reading comprehension were administered to all students in the sample. Response format is typically defined as the way students are required to demonstrate their comprehension of a passage or story on a test. The tests varied by the type of response expected from the student with one utilizing a cloze procedure in which the student supplied a missing word (i.e., WJ-IV PC) and another required students to respond verbally to open-ended questions about a story (i.e., WIAT-III RC). The third test required students to fill in a circle in response to a multiple choice question (i.e., GRMT-4). Two of the measures were individually administered (i.e., WJ-IV PC and WIAT-III RC) and one was administered in a group format (i.e., GRMT-4).

The Wechsler Individual Achievement Test-Third Edition Reading Comprehension (WIAT-III RC) subtest was administered to all students in the sample. The average reliability for all grade levels is .88 and listed specifically as .85 for fourth-grade students in the manual (Breaux, 2010). This assessment used expository and narrative text to assess comprehension. Students were asked literal and inferential questions by the examiner after reading each of three passages designated for their grade level. All students read the same three passages for fourth-grade students which covered a range of difficulty. Two of the passages were expository and one was narrative with lengths of 48, 100, and 126 words. There are between six and eight questions for each passage. There was no timing component to the test and students were told they could read the passages aloud or silently. Questions were answered verbally by students and their responses were recorded verbatim by the examiner. Points were awarded based on complete, partial, or incorrect answers with students earning 2, 1, or 0 points. The protocol provided guidelines for scoring responses. Inter-rater agreement for this measure was 94.29%. Again, scoring disagreements were resolved by closely consulting the test manual. When an answer was not found in the manual to clarify the disagreement, the original scoring was retained.

On the WJ-IV PC subtest, the median reliability was reported as .89 for students between the ages of 5 and 19 years with a reported reliability of .89 at ages 9 and 10 years as reported in the test manual (McGrew, Laforte, & Schrank, 2014). The WJ-IV PC used a cloze response format to assess reading comprehension. Students were required to provide a missing word for sentences and short paragraphs (e.g., “Woof,” said the _____, biting the hand that fed it.”). The number of sentences encountered depends on the age-appropriate start point and continued correct performance. Passages were discontinued following five consecutive incorrect responses. The items increase in difficulty as evidenced by longer sentences, more complex vocabulary, and more abstract topics. The longest passage on the WJ-IV PC contained 49 words. Each response was scored as correct or incorrect. All students in the sample started with the same item but several had to reverse to items below the start point for fourth-grade students. Inter-rater agreement for this measure was 98.32%.

The WIAT-III RC and WJ-IV PC were individually administered reading comprehension assessments in contrast to the Gates-MacGinitie Reading Test, 4th Edition (GRMT-4) which was group administered. The GRMT-4 consisted of 11 passages followed by literal and inferential multiple-choice comprehension questions. Reported reliability ranges from .87 to .92 across forms of the GRMT-4 as noted by the test's authors in the manual (MacGinitie, MacGinitie, Maria, & Dreyer, 2000). Form S was used in the present study and the 35-minute time limit was enforced. Students were required to mark their answer to a multiple-choice question on an answer sheet. Passages ranged in length from 65 to 126 words. Five of the 11 passages had 100 or more words. Standardized instructions were provided at the beginning of the assessment and examples were provided for students on marking their answer in the correct spot on the answer sheet. The GRMT-4 was administered to students in a group setting of 13 to 26 students.

Word reading and decoding. The Wechsler Individual Achievement Test-Third edition (WIAT-III) Word Reading and Pseudoword Decoding subtests were administered as measures of basic word reading and decoding. On the WIAT-III Word Reading subtest, students were required to read aloud from a list of increasingly complex words. As noted in the test manual, the average reliability across all grade levels of the WIAT-III for this subtest was .97 and the reported reliability for fourth grade was .98 (Breaux, 2010). The Pseudoword Decoding subtest from the WIAT-III required students to read aloud from a list of increasingly complex nonsense words. For this subtest, the average reliability was .98 for all grade levels and for fourth grade. Inter-rater reliability for WIAT-III Word Reading and Pseudoword Decoding subtests was 97.19% and 94.38%, respectively.

Reading fluency. For a measure of oral reading fluency, students read a passage at a fourth-grade level as determined by easycbm.com for 1 minute. Words read correctly in 1 minute as well as number of errors were recorded. Standardized instructions were read verbatim from the assessor form. Students were told they would have one minute to read as much of the passage as they could and were encouraged to do their best reading. If a student hesitated or sounded out a word longer than 3 seconds, the correct word was supplied. Oral reading fluency measured by correct words per minute has been shown to be a strong indicator of overall reading skill (see Fuchs, Fuchs, Hosp & Jenkins, 2001). Specifically, for the easycbm.com passages, test-retest reliability for fourth grade ranges from .86 to .96 with a median of .95. Further, predictive and concurrent validity ranged from .55 to .69 when compared to a state test for students in third through eighth grades (Tindal, Nese, & Alonzo, 2009). Inter-rater agreement for this sample was 98.75%.

Reading vocabulary. The WJ-IV Reading Vocabulary subtest was administered to all participants and is considered a measure of general vocabulary and verbal knowledge. There were two parts to this subtest. The student was required to read a word and provide a synonym for the word in the first part. The second part required the student to provide an antonym for the target word. Students must be able to read the words correctly and were not provided help with decoding the words. As reported in the manual, median reliability for ages 5 to 19 was .88 (.79 to .92 range) with .89 reported as the reliability for students ages 9 and 10 years. Inter-rater agreement for this measure was 98.23%.

Morpho-syntactic knowledge. To assess knowledge of morphology and syntax, an experimental measure (Carlisle, 2000) was administered to all students in the sample. This measure contained a sentence completion task in which students are required to manipulate the morphological structure of word to fit a sentence while also preserving appropriate syntax. For example, the student may be provided the word humor and then asked to complete the following sentence: The story is quite _________. Students provided verbal responses to the sentences which were read aloud to them by the examiner. Responses were recorded by the examiner and scored as correct or incorrect. The measure contained 28-items. Inter-rater reliability was 99.64%.

Design

A within-subjects design was used to analyze the following four reading comprehension assessment conditions as independent variables: (a) maze passages with fixed-word deletion (every seventh word); (b) maze passages with word-feature deletion (pronouns, conjunctions, nouns); (c) maze passages with sentences deleted; and (d) intact passages followed by multiple choice questions. Each of the four assessment conditions were analyzed in regards to the reliability and concurrent construct validity.

Sampling Procedure

District permission was sought initially to work in four schools in a rural district in the southeastern US. The assistant director of schools and building level principals signed their approval for the project. The primary investigator set a briefing session at each of the four schools to discuss the project with the fourth-grade teachers. All 19 teachers agreed to allow the students in their class to send parental consent forms home. The primary investigator met with each class and invited all students to participate in the study. Students were told that the purpose of the study was to develop a new test to measure reading comprehension. Parental consent forms were given to students by the primary investigator. Parents were informed that the purpose of the study was to obtain general information about how students perform on different types of reading comprehension tests. Students with signed informed consent forms were pulled from class for group testing sessions to complete the 12 passages representing the four assessment conditions analyzed in this study. At the first session, students were read a student assent script. It was emphasized that participation was voluntary and their performance would not be reflected in a class grade. All students agreed to participate and signed student assent. Refer to the Appendix E for the parental consent and Appendix F for the student assent forms used in this study.

Testing Procedure

The assessments were originally divided into three sessions consisting of four passages each. Assessment session one included four passages (two fixed-word deletion; two content-word deletion) as well as the group-administered reading comprehension measure, Gates-MacGinitie Reading Test, 4th edition (GRMT-4). Assessment session two included four passages (one fixed-word deletion; one content-word deletion; two multiple-choice). Assessment session three included four passages consisting of three sentence deletion maze passages and one multiple choice passage. Standardized directions were presented at each session relevant to the measures used. For example, sentence deletion required different instructions and examples than the other two maze assessments; therefore, all three passages with sentence deletion were administered in one setting. The order of passages within each assessment condition was counterbalanced to control for passage effects as well as fatigue. The GRMT-4 followed administration of maze passages in session one for some groups and preceded it in other groups due to the limited number of testing booklets.

It should be noted that following the administration of the sessions at the first school, the sessions were extended from three to five due to the length of time it was taking students to finish. Student fatigue with the number and length of passages was problematic. Further, students were confused by the directions when different assessment conditions were included in one session (i.e., sentence deletion and multiple choice passages in one session). Group administration was altered to include passages from each condition together (e.g., fixed-word deletion, sentence deletion) and the GRMT-4 was administered in a separate group session. An example was added to the instructions for the sentence deletion condition due to student confusion with the task.

Following completion of all group measures (12 passages, across 4 conditions, and the GRMT-4), the individual battery was administered to students in one session. When time constraints at the school shortened the testing session, the order of tests was preserved and testing continued on subsequent days in the same order. The two individually administered reading comprehension measures, the WIAT-III RC and WJ-IV PC, were administered first and were counterbalanced across students. Following administration of the two reading comprehension measures, a measure of word reading (WIAT-III word reading subtest), a measure of decoding (WIAT-III Pseudoword decoding subtest), a measure of reading fluency (easycbm.com fourth-grade passage reading fluency probe), and a measure of morpho-syntactic knowledge (Test of Morphological Awareness) were administered to all 93 students. All students were assessed by the primary investigator, a licensed school psychologist, or by trained doctoral level graduate assistants. Group measures were typically completed in classrooms. At one of the schools, the majority of group measures were completed in the library. The individual testing sessions took place in a quiet area outside the classroom either in a conference room or office.

CHAPTER FOUR
RESULTS

Analysis

In the current study, reliability was evaluated first using α to determine internal consistency of the items across the four conditions. The reliability at the condition and passage level was calculated. α measures how closely items are related in that as the correlations among items increase, α increases. Acceptable levels of α for informal tests is .7 to .8. For standardized measures, α levels of .8 or .9 are considered acceptable.

Concurrent construct validity of all four assessment conditions to multiple, validated measures of reading comprehension was explored with correlations, factor analysis, and hierarchical regression. Concurrent validity explores the relationship between two measures taken during the same time period. The relationship between measures represents the degree to which the measure taps the same skill. Correlations between each of the reading comprehension assessment conditions and the three validated measures of reading comprehension were conducted. Next, correlations between each of the assessment conditions and validated measures of word reading, decoding, reading fluency, inferencing skill, and morpho-syntactic knowledge were calculated. The correlations to reader skills is intended to determine if the different comprehension assessment conditions are measuring skills important for comprehension similarly or not.

The relative strength of each of the four assessment conditions were examined with factor analysis and hierarchical regression to specifically address the fourth research question addressing which assessment condition analyzed in this study contributed significant variance to a reading comprehension composite. Lastly, the final question addressed how reader skills contributed variance to different comprehension tests. Further regression analyses were conducted to determine how reader skills contributed variance to the assessment conditions analyzed in this study.

All score distributions were approximately normal with acceptable values of skewness (+/- 2). The WJ-IV PC subtest demonstrated a moderately skewed distribution to the left indicating a build-up of scores below the mean. The distribution of scores for the GRMT-4 and the WIAT-III RC subtest was slightly skewed to the left. These trends may reflect lower comprehension skills for the sample of participants. Descriptive statistics for all measures of reader skills in the study are reported in Table 4. Standard scores are reported for the measures in the study for which they were available. The GRMT-4 has a standard growth score reported but it was on a different metric than the other two comprehension tests analyzed in the study; therefore, the raw score average was reported. In general, the mean scores reflected average skills with standard scores between the ranges of 85 to 115 in all areas. However, vocabulary and passage comprehension as measured by the WJ-IV subtests reflected the lowest mean scores. The WJ-IV was the most recently revised test of the individual battery. It should be noted that the vocabulary measure used in this study assessed reading vocabulary rather than listening vocabulary. Students were required to correctly read the word in order to provide a synonym or antonym. The mean words correct in a minute for the fluency measure (*M* = 133 words correct) was at the 75th percentile for fourth-grade benchmark expectations in the fall. The mean score represented the 50th percentile for expectations for the winter and spring of fourth grade.

Table 4

Means of Reader Skills (N= 93)

Reader Characteristics	Measure Range	*M*	*SD*	Student Range
Comprehension (GRMT-4) [b]	0-48	25.13	10.72	2-45
Comprehension (WJ-IV PC) [a]	40-160	92.86	11.76	42-119
Comprehension (WIAT-III RC) [a]	40-160	100.2	12.43	60-141
Word Reading (WIAT-III) [a]	40-160	102.09	14.14	68-136
Decoding (WIAT-III) [a]	40-160	102.49	14.20	71-139
Fluency (cwpm) [b]	0-250	133.13	37.05	22-216
Vocab (WJ-IV) [a]	40-160	96.73	13.84	59-120
Syntax (Carlisle) [b]	0-28	13.98	5.34	1-24

Note: WIAT-III RC= *Wechsler Individual Achievement Test – 3rd edition Reading Comprehension subtest;* WJ-IV PC = *Woodcock-Johnson Tests of Achievement – 4th edition passage comprehension subtest;* GRMT-4 = *Gates-MacGinitie Reading Test, 4th edition, Form S.*

a = standard score b = raw score

Exploratory Factor Analyses

To investigate the construct validity of the assessment conditions, an exploratory factor analysis using principal components analysis was completed on the items in each condition and on the total scores for each condition. Further, the three validated measures of comprehension used in this study (i.e., GRMT-4; WIAT-III RC; WJ-IV PC) were also entered into a factor analysis. The rotation used was promax because of the high correlations among variables. This rotation computes factor loadings with the assumption that factors are highly correlated. First, the items for each assessment condition were entered. Kaiser's Rule to retain all eigenvalues above 1.0 and an examination of scree plots were used to determine the number of factors.

Fixed-word deletion. The Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy was .560 which is just above the minimum criterion. Bartlett's test was significant (*p* = .000) indicating data were suitable for factor analysis. A one-factor solution was the best fit for the 249 items entered for this condition. Although Kaiser's Rule suggested a 54-factor solution, examination of the scree plot indicated a flattening at factor one which accounted for 29.96% of the variance of the measure. The remaining factors identified by Kaiser's Rule contribute 5% or less of the variance to the total measure.

Sentence deletion. The KMO measure of sampling adequacy was .754 and Bartlett's test was significant (*p* = .000) for this condition both indicating the data were suitable for factor analysis. A one-factor solution emerged as the best fit for the 30 items entered for this condition. Kaiser's Rule suggested an eight factor solution; however, the scree plot flattens significantly at factor one which accounted for 25.28% of the variance of the measure. The second factor identified by Kaiser's rule contributed 8.6% of the variance to the measure with the remaining factors contributing 5% or less of the variance to the measure.

Word-feature deletion. For this condition, the KMO measure of sampling adequacy was .560 and Bartlett's test was significant (p = .000). Similar to the other conditions, a one factor solution emerged after examining the scree plot. Kaiser's Rule suggested a 23-factor solution. The first factor accounts for 23.75% of the variance of the measure with each subsequent factor identified by Kaiser's Rule as contributing 5% or less of the variance to the measure.

Multiple choice. For the multiple choice condition, KMO measure of sampling adequacy of .606 is adequate and Bartlett's test is significant (*p* = .000) indicating the items are appropriate for factor analysis. Based on Kaiser's Rule, a 12-factor solution emerged for the original 30 items in this condition. For this condition, the scree plot flattens significantly after the first factor which accounts for 16.1% of the variance. The remaining factors identified by Kaiser's Rule contribute 6% or less of the variance to the total measure.

Assessment condition totals. When the total scores for fixed-word deletion, sentence deletion, word-feature deletion, and multiple choice conditions were entered using principle component analysis and promax rotation, a one-factor solution emerged based on Kaiser's Rule and examination of the scree plot accounting for 74.22% of the variance for total scores across conditions.

All measures. All of the individual measures and the assessment condition totals for fixed-word deletion, sentence deletion, word-feature deletion, and multiple choice were entered into an exploratory factor analysis. Again, principle components analysis was used with promax rotation. The KMO measure of sampling adequacy was .845 which is a strong value and Bartlett's test was significant (p = .000). When all measures were entered, a two-factor structure emerges based on Kaiser's rule and examination of the scree plot. The four assessment conditions created for this study load on one of the two factors and the remaining measures utilized load on the other factor together. See Table 5.

Reliability

Refer to Table 6 for α statistics for individual passages and each of the four assessment conditions. The fixed-word deletion condition demonstrated the highest α levels at the condition and passage level followed closely by the word-feature deletion condition. Since the number of items were not consistent across conditions, the larger number of items on these two conditions compared to the sentence deletion and multiple choice conditions contributed to the higher levels of α which is impacted by the number of items. The reliability coefficients for the sentence deletion and multiple choice

Table 5

Two-Factor Structure of Measures

Test/Measure	Component One	Component Two
GRMT-4	.685	
WJ-IV PC	.845	
WIAT-III RC	.796	
Fixed-word Maze		.919
Word-feature Maze		.949
Sentence Maze		.793
Multiple Choice		.842

Note: WIAT-III RC= *Wechsler Individual Achievement Test – 3rd edition Reading Comprehension subtest;* WJ-IV PC = *Woodcock-Johnson Tests of Achievement – 4th edition passage comprehension subtest;* GRMT-4 = *Gates-MacGinitie Reading Test, 4th edition, Form S.*

Table 6

Reliability of Conditions and Passages

Assessment Condition *Passage*	# of items	*M*	*SD*	% Correct	α
Fixed-Word Deletion	**249**				**.990**
Opinion	82	54.77	22.71	66.79%	.978
Small Toys	88	67.73	21.48	76.96%	.977
Wild Animals	79	55.21	20.39	69.88%	.974
Word-Feature Deletion	**75**				**.953**
Storms	25	18.75	5.73	75%	.902
Teddy Bears	25	17.47	5.91	69.88%	.883
Traditions	25	17.83	5.80	71.32%	.887
Sentence Deletion	**30**				**.885**
Support Program	10	6.24	2.74	62.4%	.783
Toilet	10	5.26	2.46	52.6%	.685
Zoo	10	5.54	2.42	55.4%	.670

Table 6 continued.

Reliability of Conditions and Passages

Assessment Condition *Passage*	# of items	*M*	*SD*	% Correct	α
Multiple Choice	**30**				**.781**
Ancient	10	4.26	2.01	42.6%	.430
Finding Dory	10	5.31	2.28	53.1%	.625
Park's View	10	4.73	2.30	47.3%	.592

conditions were more comparable since each condition has 30 items. The sentence deletion condition demonstrated higher internal consistency than the items in the multiple choice condition. All four assessment conditions analyzed in this study demonstrated acceptable to excellent levels of internal consistency. However, at the passage level, only seven out of 12 demonstrated reliability at an acceptable level or higher. All three passages in the fixed-word deletion and word-feature deletion condition demonstrated acceptable levels of α. As noted previously, these conditions have the inherent advantage for reliability analysis by including more items. Only one of the passages in the sentence deletion condition demonstrated acceptable reliability whereas none of the passages in the multiple choice condition had an acceptable α level.

Since the number of items vary considerably across conditions, the correlations between items and item means were examined. Corrected item-total correlations indicate the degree to which a response to the items was predictive of the total score. Corrected item-total correlations below 0.3 indicate that the item did not predict the total score. By investigating corrected item-total correlation values and changes in reliability when an item was deleted from the scale, there were items within each condition that were problematic. Specifically, 15 items were deleted in the fixed-word deletion condition, six from the word-feature condition, four from the sentence deletion condition, and 18 items were deleted from the multiple choice condition. All deleted items had corrected item-total correlations below 0.3 and a corresponding increase in α level if the item was deleted. Reliability for each condition was reevaluated following the deletion of items. Overall, reliability for the sentence deletion and multiple choice conditions were most improved by the deletion of targeted items. The α level of the multiple choice condition increased from .781 to .790 which may seem like a modest increase but 18 items were deleted from the condition total and reliability still improved. Reliability for the sentence deletion condition improved from .885 to .894 with the deletion of four items. Little to no change occurred with the level of α for the other two conditions.

In addition to examining overall reliability following removal of the targeted items, the corrected item-total correlations were reexamined for the remaining items in each condition. For the multiple choice condition, corrected item-total correlations for the 12 remaining items ranged from .30 to .51. For the sentence deletion condition, the corrected item-total correlations ranged from .25 to .68. Two items continue to fall below the 0.3 criterion used for deletion; therefore, two more items could be considered for deletion and may improve reliability closer to the .9 range. In the word-feature deletion condition, a .001 increase in reliability occurred after removing 6 of 30 items. Corrected item-total correlations for the remaining 24 items ranged from .29 to .62. Although one item has a correlation below 0.3, overall reliability would not increase for the condition by deleting the item. In the fixed-word deletion condition, corrected item-total correlations ranged from .29 to .78; however, only one correlation was below 0.3 on the scale and deletion of the item would not increase reliability for the condition.

Concurrent Validity

To examine the concurrent validity of the assessment conditions, correlations between measures of comprehension and reader skills were closely examined. First, correlations between the assessment conditions and the three validated measures of comprehension were explored. As predicted, there were varying correlations among the conditions and validated comprehension measures which have varying response formats and administration procedures. The fixed-word deletion condition was correlated significantly with the WJ-IV PC subtest as predicted, $r = .275$, $p = 001$. Yet, there was no significant relationship between this condition and the GRMT-4 ($r = .185$, $p = .079$) or the WIAT-III RC subtest ($r = .144$, $p = .172$). The WJ-IV PC subtest which utilizes a cloze response format exhibited the strongest correlation with the fixed-word deletion condition yet this comprehension measure correlated significantly with the other three assessment conditions as well. The WIAT-III RC subtest did not correlate significantly with any of the four assessment conditions examined in this study. On the WIAT-III RC subtest, students were required to provide open-ended answers to questions posed about the passage by the examiner. The GRMT-4 demonstrated a significant correlation with the sentence deletion condition only, r = .309, p = .003. There was no significant relationship between the GRMT-4 and the remaining assessment conditions. The administration of the GRMT-4 required students to respond to multiple choice questions by filling in a circle on a separate answer sheet.

The three comprehension assessments (WJ-IV PC, WIAT-III RC, and GRMT-4) correlated significantly with one another. However, given that all three purport to measure the same construct and are used interchangeably to describe a student's reading comprehension level, the correlations indicate that the assessments are measuring the construct differently. Correlations range from .438 to .691 and are all significant ($p <$ .001). As expected the two individually-administered comprehension assessments, the WIAT-III RC and WJ-IV PC, shared the strongest correlation, r = .691, p = .000. Administration of these two assessments were counterbalanced during individual testing. On average, students demonstrated higher performance on the WIAT-III RC subtest (M = 100, SE = 1.31) than on the WJ-IV PC subtest (M = 93, SE = 1.23). This difference, 7.32, BCa 95% CI [5.35, 9.11], was significant t (90) = 7.31, p = .000, d = .56. The three comprehension tests correlated significantly with all reader skills analyzed in the study. Although all correlations were significant (*p* < .001) for reader skills and the GRMT-4, fluency and word reading demonstrated the largest correlations to this measure. In contrast, the WJ-IV PC demonstrated the strongest correlations with word reading and syntax. Whereas, the strongest correlation for the WIAT-III RC among reader skills was vocabulary. Fluency and syntax were tied with the second strongest correlation to the WIAT-III RC.

Next, correlations were examined between the four assessment conditions and reader skills. For the fixed-word deletion condition, the correlation with vocabulary was the strongest (*r* = .32, *p* = .002), followed by syntax (*r* = .25, *p* = .015), fluency (*r* = .228, *p* = .030), and decoding (*r* = .21, *p* = .045). The correlation between the fixed-word deletion condition and word reading was not significant, *r* = .16, *p* = .118. For the word-feature deletion condition, the correlation with decoding (*r* = .29, *p* = .005) was strongest followed by vocabulary (*r* = .26, *p* = .010), syntax (*r* = .24, *p* = .021), fluency (*r* = .23, *p* = .027), and word reading (*r* = .21, *p* = .039). In the sentence deletion condition, decoding and fluency skills evidenced the strongest correlations (*r* = .273, *p* = .009), followed by word reading (*r* = .25, *p* = .016), and vocabulary *(r* = .24, *p* = .017). The relationship between the sentence deletion condition and syntax was not significant, *r* = .17, *p* = .099). Lastly, correlations for the multiple choice condition were strongest for vocabulary (*r* = .37, p = .001), followed by syntax (*r* = .264, p = .011), decoding (*r* = .24, *p* = .020), and word reading (*r* = .23, *p* = .024). The correlation between the multiple choice condition and fluency was not significant, *r* = .19, *p* = .070.

In sum, the correlations across reader skills varied for each of the assessment conditions but not in the way that was predicted. For example, it was predicted that the fixed-word deletion condition would correlate more strongly with measures of word identification and decoding than the other conditions. Yet, the strongest and most significant correlation for this condition was with vocabulary. Surprisingly, word reading skills were not significantly correlated with this condition. It should be noted that the reading vocabulary measure used required students to read the word correctly prior to providing a synonym or antonym. Therefore, the vocabulary measure used was dependent on the students' ability to read the word. Despite predicting that vocabulary and syntax would correlate stronger with the sentence deletion and word-feature deletion conditions than decoding, decoding emerged as the strongest correlation for both conditions. Syntax was not correlated significantly with the sentence condition. Syntax was predicted to be important for the sentence deletion condition since one of the distractors included a sentence with the words scrambled. Another interesting and unexpected finding is that fluency correlated significantly with all assessment conditions except the multiple choice condition. See Table 7 for correlations between all variables.

To address the fourth research question, a two-stage hierarchical regression analysis was completed to address the amount of significant variance the sentence deletion contributed to reading comprehension after controlling for performance on the

Table 7

Correlations of Assessment Conditions to Validated Measures of Reading

									Assessment Condition			
	Gates	WJ-P	W-RC	WR	PD	Fluency	Vocab	Syntax	Fixed	Word	Sent	MC
Gates	-											
WJ-P	**.470****	-										
W-RC	**.438****	**.691****	-									
WR	**.544****	**.704****	**.523****	-								
PD	**.434****	**.521****	**.320***	**.813****	-							
Fluency	**.598****	**.643****	**.590****	**.734****	**.573****	-						
Vocab	**.484****	**.679****	**.637****	**.670****	**.502****	**.561****	-					
Syntax	**.480****	**.701****	**.590****	**.661****	**.534****	**.616****	**.691****	-				
Fixed	.185	**.275****	.144	.165	**.211***	**.228***	**.326****	**.254***	-			
Word	.133	**.261***	.138	**.217***	**.292****	**.232***	**.267***	**.242***	**.897****	-		
Sent	**.309****	**.247***	.062	**.251***	**.273****	**.273****	**.249***	.174	**.580****	**.615****	-	
MC	.144	**.254***	.163	**.237***	**.243***	.191	**.378****	**.264***	**.644****	**.704****	**.565****	-

Note: Correlations in bold are significant at p = .05, $^{*}p$ = .001**; WJ-P: Woodcock-Johnson Tests of Achievement, Passage Comprehension; W-RC: WIAT-III Reading Comprehension; WR: WIAT-III Word Reading; PD: WIAT-III Pseudoword Decoding; Fluency: easycbm 4th grade fluency passage; Vocab: Woodcock-Johnson Tests of Achievement, Reading Vocabulary; Syntax: Carlisle Test of Morphological Awareness; Fixed: Fixed-word deletion maze passages; Word: Word-feature deletion maze passages; Sent: Sentence deletion maze passages; MC: multiple-choice passage fixed-word deletion, or standard, maze. Prior to conducting the regression analyses, all three validated measures of reading comprehension were normalized and averaged into one score for a reading comprehension composite. The composite score was then used in subsequent analysis. The composite variable was created to account for the three different comprehension tests which had differing correlations with underlying skills and differing passages and response formats. The tests varied in terms of administration and one test had a different scoring metric than the other two. All four assessment conditions were significantly and similarly correlated with the reading comprehension composite ($p < .05$). Specifically, the sentence deletion condition had the strongest correlation with the reading comprehension composite ($r = .24$, $p = .018$) followed closely by the fixed-word deletion ($r = .24$, $p = .021$), then, the multiple choice condition ($r = .23$, $p = .032$), and the word-feature deletion condition ($r = .21$, $p = .042$).

The sentence deletion condition was further examined in comparison to the fixed-word deletion, which is the most common form of maze. Specifically, the amount of variance the sentence deletion condition accounts for in reading comprehension after considering the variance explained by the fixed-word deletion condition was explored (see Table 8).

Table 8

Hierarchical Regression: Sentence Deletion

		B	*β*	*t*	*p*	*Adj. R^2* of Model
Model 1	Constant	-0.585		-2.23	.028	.048
	Fixed-Word Deletion	0.003	.243	2.35	.021	
Model 2						
	Constant	-0.615		-2.34	.022	.051
	Fixed-Word Deletion	0.002	.139	1.01	.311	
	Sentence Deletion	0.020	.156	1.14	.258	

Note: n = 93

The fixed-word deletion condition accounted for 4.8% of the variance in performance on the reading comprehension composite which was calculated from normalizing and averaging three, validated comprehension measures into one score. This indicates that the fixed-word deletion condition was significant at predicting reading comprehension performance, *b* = .003, *p* = .021. However, after controlling for performance on the fixed-word deletion condition, performance on the sentence deletion condition did not contribute significant variance to reading comprehension. When the sentence deletion condition was entered first into the model, sentence deletion was predictive of overall performance contributing 3.7% of the variance on the reading comprehension composite. Fixed-word deletion entered second did not improve the prediction of the model significantly. See Table 9.

Table 9

Hierarchical Regression: Fixed-Word Deletion

		B	*β*	*t*	*p*	*Adj. R*^{2} of Model
Model 1	Constant	-0.439		-1.96	.053	.037
	Sentence Deletion	0.026	.219	2.12	.037	
Model 2						
	Constant	-0.645		-2.34	.021	.044
	Sentence Deletion	0.014	.115	0.87	.386	
	Fixed-Word Deletion	0.002	.168	1.27	.204	

Note: n = 93

Again, using the reading comprehension composite as the dependent variable and entering fixed-word deletion, sentence deletion, and word-feature deletion conditions into the model after controlling for word reading skills, the fixed-word deletion condition is the only one that emerged as a significant predictor (*b* = .005, *p* = .026). See Table 10.

Table 10

Hierarchical Regression: All Conditions

		B	*β*	*t*	*p*	*Adj. R^2* of Model
Model 1	Constant	-4.257		-9.46	.000	.050
	Word Reading	0.042	.712	2.35	.021	
Model 2						
	Constant	-4.372		-9.46	.000	.051
	Sentence Deletion	0.001	.007	0.06	.946	
	Fixed-Word Deletion	0.005	.364	2.26	.026	
	Word-Feature Deletion	-0.014	-.271	-1.58	.116	

Note: *n = 93*

When entering word reading, vocabulary, syntax, and fluency into the model at the same time, fluency emerged as the strongest predictor of performance on a derived comprehension composite contributing 35.7% of the variance of the model. Vocabulary emerged behind fluency as the skill most predictive of the overall reading comprehension skill accounting for 30.8% of the model. Syntax also emerged as a significant predictor. Surprisingly, word reading skill did not. When all four reader skills are entered into the model, 70.9% of the variance in the reading comprehension composite is explained. See Table 11.

Table 11

Regression Analysis: Reader Skills

		B	*β*	*t*	*p*	*Adj. R^2* of Model
Model 1	Constant	-3.948		-8.82	.000	.709
	Word Reading	0.006	.105	1.07	.287	
	Vocabulary	0.018	.308	3.52	.001	
	Syntax	0.033	.209	2.35	.021	
	Fluency	0.008	.357	4.02	.000	

Note: n = 93

A general linear model was conducted to control for Type I SS by entering sentence deletion and fixed word deletion into the prediction of the reading comprehension composite. Under this model, neither condition nor the interaction between them contributed significant variance to the reading comprehension composite analyzed in the study.

Lastly, reader skills measured in the study were further analyzed to determine which of the skills contribute significant variance to each of the three validated comprehension measures. The GRMT-4 was only significantly predicted by fluency skills and did not demonstrate a significant relationship with word reading, vocabulary, or syntax. The WJ-IV PC subtest was significantly predicted by word reading, vocabulary, and syntax and a non-significant relationship emerged for fluency. The WIAT-III RC subtest was significantly predicted by vocabulary and fluency skills. A non-significant relationship emerged between the WIAT-III RC subtest and word reading and syntax. See Tables 12, 13, and 14. As noted, using the reading comprehension composite as the dependent variable and entering all reader skills at the same level, including word reading, reading vocabulary, syntax, and fluency, all skills except word reading emerged as a significant predictor (Table 11).

Table 12

Regression Analysis: GRMT-4

		B	*β*	*t*	*p*	*Adj.* R^2 of Model
Model 1	Constant	345.204		11.50	.000	.370
	Word Reading	0.331	.122	0.86	.392	
	Vocabulary	0.390	.140	1.11	.270	
	Fluency	0.412	.393	3.08	.003	
	Syntax	0.447	.061	0.47	.634	

Note: n = 93

Table 13

Regression Analysis: WJ-IV PC

		B	*β*	*t*	*p*	*Adj. R^2* of Model
Model 1	Constant	37.287		5.187	.000	.612
	Word Reading	0.207	.092	2.24	.028	
	Vocabulary	0.196	.230	2.32	.022	
	Fluency	0.051	.160	1.60	.113	
	Syntax	0.624	.224	2.78	.007	

Note: n = 93

Table 14

Regression Analysis: WIAT-III RC

		B	*β*	*t*	*p*	*Adj. R^2* of Model
Model 1	Constant	54.191		6.12	.000	.477
	Word Reading	-0.09	-.108	-.83	.406	
	Vocabulary	0.359	.397	3.46	.001	
	Fluency	0.113	.039	2.88	.005	
	Syntax	0.427	.276	1.54	.125	

Note: n = 93

CHAPTER FIVE

DISCUSSION

The present study sought to examine the reliability and validity of four types of comprehension assessments as well as promote an improved understanding of the underlying skills tapped by reading comprehension assessments. Maze is unique in that it measures comprehension during the process of reading. Most assessments, such as question-answer, retell, and multiple-choice questions, measure comprehension after reading. Available comprehension assessments, including maze, have been criticized for having little to no instructional usefulness for teachers. Differential measurement of reader skills across reading comprehension assessments limits the instructional usefulness of these types of tests. They are insufficient for pinpointing the specific skills that need to be targeted for intervention.

The assessment conditions in this study were found to have acceptable to excellent levels of reliability with the fixed-word deletion and word feature deletion conditions demonstrating the highest levels. Following careful consideration and deletion of specific items, reliability for the multiple choice and sentence deletion conditions improved. Validity was more difficult to establish. Correlations suggest that each of the assessment conditions are tapping several underlying skills known to be important for reading comprehension. Correlations among standardized, validated measures of comprehension varied considerably across conditions. When considering the amount of variance each of the assessment conditions contribute to a reading comprehension composite calculated from normalizing and averaging performance across the three validated measures used in this study, the fixed-word deletion condition was the only one to contribute significant variance to the model. Although the relationship was determined significant, the amount of variance was minimal. Specifically, fixed-word deletion condition accounted for 4.8% of the variance which was likely not meaningful in a practical way. The sentence deletion version of maze did not contribute significant variance on reading comprehension when controlling for performance on the fixed-word maze condition. Yet, all four assessment conditions correlated significantly with a reading comprehension composite created from normalizing and averaging scores from the three validated measures of reading comprehension. Sentence deletion was the only condition to correlate with two of three validated comprehension tests. Also, based on a factor analysis, all four conditions are measuring the same factor.

Text and Item Features of Reading Comprehension Assessments

Text variables and administration procedures vary across assessments of reading comprehension. Cain and Oakhill (2006) identified response format as one source of inconsistency in the measurement of reading comprehension. In this study, several response formats were used. Specifically, on the standardized measures, response format included open-ended verbal responses to questions posed by the examiner, filling in a circle for a multiple choice response, and providing a missing word in the context of a passage. The reading comprehension assessment conditions for this study also incorporated different response formats such as circling one of three words, circling one of three entire sentences to fit the passage, or responding to multiple-choice questions after reading intact passages. Words were deleted from text at different rates and both items and types of distractors varied across the assessment conditions studied.

As predicted, correlations were found to vary across comprehension tests with different formats. Specifically, the hypothesis that the sentence deletion condition would be more highly correlated to the GRMT-4 which utilizes multiple-choice questions than the WJ-IV PC which uses a cloze response format was supported. The sentence deletion task was the only assessment condition in this study to correlate significantly with two of three comprehension tests. Specifically, the sentence deletion maze condition correlated significantly with the GRMT-4 and the WJ-IV PC subtest. All four assessment conditions demonstrated a significant correlation to the WJ-IV PC suggesting that the cloze format may be measuring comprehension similarly to the maze and multiple choice response formats used in this study. The lack of significant correlation to the WIAT-III RC subtest may indicate that this test was measuring comprehension differently than the conditions in this study. The WIAT-III RC is the only measure that required students to respond verbally to questions posed by the examiner. Similarly, and unexpectedly, the multiple choice assessment condition in this study exhibited a significant relationship with the WJ-IV PC subtest but was not significantly correlated to the other two comprehension measures, the GRMT-4 and WIAT-III RC. The lack of correlation between the multiple choice condition and the GRMT-4 is particularly surprising since both rely on responses to multiple choice questions.

Again, it is worth noting that the WIAT-III RC measure did not correlate significantly with any of the assessment conditions analyzed in this study. This finding may suggest that responding to open-ended questions verbally tapped a different set of underlying skills than the other response formats. As noted, the responses expected from students varied considerably across conditions and measures. However, the WIAT-III RC was the only measure that required students to respond verbally to open-ended questions. Although the varying correlations previously described suggest that tests and conditions are measuring underlying reader skills differently, the WIAT-III RC, in particular, may be tapping underlying skills that the other measures are not.

As noted in the literature review, text variables and administration procedures vary across assessments of reading comprehension. Specifically, length of passage has been found to contribute to differences in performance for students with varying skills (Garcia & Cain, 2014; Keenan & Meenan, 2014; and, Spear-Swerling, 2004). The passages used to construct the assessment conditions in this study were pulled from expository texts covering a variety of topics. The passages ranged in length from 424 to 683 words which is considerably longer than the passages used in the standardized measures. The WJ-IV PC, GRMT-4, and WIAT-III RC all share a common characteristic of presenting short passages for students to read. Specifically, the longest passage on the WJ-IV PC contains only 49 words. The three passages on the WIAT-III RC contain 100, 48, and 126 words. The 11 passages on the GRMT-4 ranged in length from 65 to 126 words. Passage length may have been a significant factor in this study contributing to decreased performance on the assessment conditions. Also, passage length could be contributing to the different factor scores for the assessment conditions in this study and the validated comprehension measures.

As noted, group sessions were shortened following administration of the conditions in the first school due to student fatigue. The assessment battery was reorganized to decrease the number of passages students had to read in one setting. Even after the reorganization to shorten the sessions, students were taking 45 minutes to just over an hour to complete each session. It is possible that length of the passages in the assessment conditions inadvertently tapped into other skills such as reading stamina and motivation. If so, passage length could have decreased the validity of the measures. In particular, the participants did not seem to possess much reading stamina. Complaints from students about the length of the passages were persistent across schools and classrooms. At times, students seemed to rush through the passages rather than put forth their best effort. A study by Nation and Snowling (1997) used longer passages to create maze assessments than many other studies such as Keenan et al., 2008. The longer passages may have been the reason Nation and Snowling found stronger correlations to oral language than word reading and decoding skills. Oral language was not assessed in the present study. The measures used, even the vocabulary measure, required reading. Interestingly, none of the assessment conditions correlated with the WIAT-III RC test in which lengthy oral responses were required. Again, passage length may have been a significant confound in this study and may be linked to oral language skills which were not assessed.

Another potential confound with the passages constructed for this study was the lack of cohesion across sentences. The expository nature of the passages and low cohesion may have increased the difficulty level for students in the sample. Passages with low cohesion generally require background knowledge on the topic. Also, the sentences were short and lacked syntactic complexity which could have been problematic for the sentence deletion condition. In addition to passage length, other features of the passage could be better controlled to increase the validity of the conditions. One passage could be constructed to represent different conditions to control for passage features and effects. For example, the same passage could have been used to construct each maze type studied and multiple choice questions.

Further, the specific items chosen for deletion were reviewed within each of the assessment conditions to determine if there were important qualitative characteristics of the item that may explain why it did not contribute to performance on the total test. As noted in the multiple choice condition, 18 items were removed which was more than half of the items in the original condition. Of the 18 items removed, eight of them had a phrase in the question referring the reader back to the passage. For example, the question stem began with the phrase *According to the article* or *Based on information presented in the article*. Three of the deleted items required students to make an inference about the passage by asking questions such as *What does he mean?* or *How does he feel?* Three additional deleted items asked students to select the best fitting sentence or choose the best main idea for the passage. Similarly, one deleted item contained the word most in the question (i.e., *which detail is most important?*). The three remaining deleted items involved identifying the true statement among choices, defining a word in the passage, and finding facts such as dates and numbers from the story.

In the fixed-word deletion category, 15 items were deleted. Upon review of the deleted items, six of them contained a distractor that could fit meaningfully within the passage. Another item involved deletion of the last name of a character in the article. Four items targeted for deletion had the omitted word as the first word of a sentence. Four of the deleted items did not seem to have a shared characteristic. The word-feature deletion items were divided into the part of speech of the word that was omitted. Three of the items deleted had a pronoun as the target word, two had conjunctions as the target word, and one had a noun. Further, for one of the items deleted, a typo was present on the correct word choice (i.e., *an* instead of *and*). Two of the items deleted had the target word and distractors divided across two lines. In the sentence deletion condition, two of the four items had distractor sentences that could make sense within the passage. The other two items did not seem to have a shared characteristic. Placement of the omitted word and distractors may be an important consideration for maze test development.

Reader Skills

Prior research has documented that comprehension tests have substantial amounts of unshared variance and are related to reader skills such as fluency, vocabulary, as well as decoding and word level skills differently (Kendeou et al., 2012). In prior research, the WJ-IV PC subtest has been shown to be significantly predicted by working memory and orthographic processing whereas maze with fixed-word deletion was predicted by reading fluency and vocabulary. The results of the present study support the assumption that varying assessment conditions can have differing relationships to underlying skills deemed important for reading comprehension. Specifically, in the present study, the maze assessment with fixed word deletion had significant correlations with vocabulary, syntax, fluency, and decoding. Similar and significant correlations were found between underlying reader skills and the other assessment conditions analyzed in this study. However, of particular interest, is that the fixed-word deletion maze was the only condition studied that did not correlate significantly with word reading. Surprisingly, of all four assessment conditions, the sentence-deletion maze shared the highest correlation with word reading.

Also, of interest, was the lack of significant correlation between the syntax measure used in this study and the sentence deletion maze. This finding was surprising because it was predicted that syntax would be an important skill for choosing the correct sentence because one of the distractors was the correct sentence with the word order scrambled. However, this finding could be explained by the syntax measure used in this study. Specifically, the syntax measure relied heavily on morphological knowledge of words rather than word order within sentences. Measuring syntax by having students choose or construct a sentence with appropriate word order may demonstrate a different relationship to the sentence deletion condition. The highest correlations (p < .001) for the sentence deletion maze were with decoding and fluency, followed by word reading, and then vocabulary. Fluency was correlated to all three maze conditions analyzed in this study but was not significantly correlated to the multiple choice condition. Timing was not implemented for any of the assessment conditions in this study. Students had as much time as needed to complete the assessments. Eason et al. 2013 found fluency to be a significant predictor for completion of multiple choice questions and for assessments with long passages whether timing was implemented or not.

The results of this study support the findings from previous studies that assessments with varying response formats and administration requirements measure underlying skills known to be important for comprehension differently. Consistent with previous research (i.e., Kendeou et al., 2012), reading fluency and vocabulary skills were significantly correlated with performance on a maze assessment with fixed-word deletion. However, inconsistent with several previous studies (i.e., Keenan et al., 2008; Keenan & Meenan, 2014; and, Kendeou et al., 2012), the results of the present study did not show a significant correlation between word reading and performance on the fixed-word deletion condition. Although the fixed-word deletion condition had a significant correlation with decoding, it was lower than the correlations for vocabulary, syntax, and fluency. The passages used for all conditions were constructed from news articles which was not common among studies reviewed. The expository nature of the passages may have impacted the correlations with underlying skills, particularly increased correlations with vocabulary and decreased correlations with word reading.

Maze

A thorough review of maze as a comprehension assessment revealed varying levels of reliability and validity across studies. For the most part, instructional information provided by maze is limited and the trend has been to use maze as a screening, monitoring, or predictive measure. Several factors varied in terms of the construction, administration, and scoring of maze assessments in the studies reviewed. In terms of construction, text type, deletion ratio, as well as number and type of distractors were important distinctions. Administration factors impacting maze included incorporation of timing as well as type, length, and number of passages. Scoring guidelines for maze varied as well across studies.

For the present study, maze was constructed from news articles. In addition to a potential confound with passage length mentioned previously, the expository nature of the passages may have been an important variable contributing to the findings for the manipulated versions of maze (i.e., sentence deletion and word feature deletion). Prior research has shown that expository texts are often more difficult for students than narrative because of varying text structures, longer and more technical words, as well as higher demands on prior knowledge (Saenz & Fuchs, 2002). Prior knowledge was not assessed in this study. An attempt to control for prior knowledge was made by including expository passages about a variety of topics. Previous research has demonstrated that students often perform better on comprehension tasks derived from narrative rather than expository text (Saenz & Fuchs, 2002). In the studies reviewed on maze, the dominant trend has been to use narrative text to construct maze passages. Only five of the 35 studies reviewed used expository passages such as a science or history text. One study created maze passages from newspapers (Ticha et al., 2009).

Also, in the present study, student fatigue with the testing sessions was problematic and may have contributed to decreased performance. Across sessions and conditions, completion time varied significantly. It was apparent that some students were finishing too quickly and not putting the same level of effort into their performance as others. Further, students were aware that their performance on the tasks would not reflect their grades and may have had limited motivation to put forth their best effort. In general, mean percentage correct across most passages was below 70% despite the fact that the mean scores on the validated comprehension measures were average indicating their scores on the conditions analyzed in the study may reflect an underestimate of their skills.

Limitations and Future Directions

Due to the nature of collecting data in the schools, several factors were not well controlled in the study. For example, the number of sessions were increased due to student fatigue after data was collected at the first school. Further, group size varied from 13 to 26 based on the availability of testers and space. In one session, school lock down procedures were initiated during testing. Fire drills interrupted more than one session.

Text features can significantly impact comprehension as noted in previous research (Garcia & Cain, 2014; Keenan & Meenan, 2014; Spear-Swerling, 2004). Although attempts were made to control for Lexile and readability levels, there were other features that were not well controlled in the present study. For example, prior knowledge was not assessed. A variety of topics were included across passages in an attempt to control for varying levels of prior knowledge with one particular topic. All of the passages were derived from news articles written specifically for kids. The majority of studies reviewed used a pencil-paper version of the maze task with only four studies specifically noting computer administration of passages (i.e., Foorman & Petscher, 2010; Fuchs & Fuchs, 1992; Gillingham & Garner, 1992; Swain & Allinder, 1996). No comparisons have been made in the research between pencil-paper and computer administration of maze. Computer administration has a noted advantage in terms of saving paper and time spent scoring responses. The computer allows for time-efficient administration and scoring (Fuchs & Fuchs, 1992).

An evaluation of the distractors used in the assessment conditions could provide important information and was not analyzed in the present study. Distractor quality is important for influencing performance. In particular, the distractors may have been problematic for the multiple choice condition rather than the questions. Further analysis of distractors would also provide useful information for the sentence deletion and word-feature conditions. Distractor analysis provides information about incorrect responses. For example, when students chose an incorrect sentence on the sentence deletion maze, were they more likely to choose the one with inappropriate syntax? As noted, an unanticipated finding was that the sentence deletion condition did not have a significant relationship with the measure of syntax used in this study. In the word-feature condition, incorrect responses could be analyzed to determine if students were more likely to choose the word that was the incorrect or correct part of speech. The percentages of students who chose each type of distractor could be analyzed and compared across high ability and low ability groups based on total scores. Analyzing correct and incorrect responses may contribute to important instructional information as well as identify distractors that did not provide useful information. Also, scoring adjustments could be applied to the conditions rather than just counting the number correct. Pierce et al. (2010) found that applying a scoring adjustment improved validity of maze. Scores could be recalculated following a two- or three-error stop rule. After two or three incorrect responses in a row, correct responses are no longer counted toward the total. Such analyses would be an important next step in analyzing the assessment conditions of interest; however, it was beyond the scope of the current study.

Further, late elementary often denotes the transition to increasingly complex, expository text, such texts were used to create the passages for each of the assessment conditions. The validated measures of reading comprehension all relied on a combination of narrative and expository passages. The expository nature of the passages could have been a factor impacting student performance. Future studies could create comparable assessments with narrative and expository texts and note differences in performance and correlations with reader skills. Attempts were made to ensure the passages used in the study across conditions were similar in terms of length, Lexile, and Flesch-Kincaid. Results from the Coh-Metrix analysis indicated that the passages across conditions were high in syntactic simplicity which means the sentence structures were short and simple. Passages were also marked by low referential cohesion indicating little overlap between words and ideas in sentences. It has been suggested that passages with low cohesion require more inferences and may be more dependent on one's prior knowledge of the topic (Graesser, McNamara, Louwerse, & Cai, 2004). The passages in the sentence deletion condition were designed to improve the way maze is standardly created by requiring integration of information across sentences to get a correct response. The low cohesion within passages may have made sentence deletion too difficult.

In addition to controlling for the cohesiveness of the tests used, each of the conditions (i.e., fixed-word deletion, multiple choice, sentence deletion, and word feature deletion) could be reconstructed with shorter passages. With shorter passages, the number of items across conditions could be held constant so that reliability analysis would be more comparable across conditions. Reliability can increase with fewer items when the items retained are high quality items. As noted in this study, A increased for the multiple choice condition in this study when 18 out of 30 items were removed.

Implications for Research, Policy, and Practice

Combined findings of the study indicate that the maze assessment conditions have acceptable reliability and validity. The manipulated versions of maze created for this study, sentence deletion and word-feature deletion, did not improve the validity of the fixed-word deletion, or standard, maze assessment. However, the assessment conditions created for this study seemed to tap into a dimension of reading comprehension not measured by validated, standardized comprehension measures. Passage length and genre were suggested as possible reasons for the differences. Further, a maze task involving sentence deletion emerged as a potential alternative to the way maze assessments are standardly created. The sentence deletion condition demonstrated significant correlations to two of the three comprehension tests used in the study and correlated significant with the reading comprehension composite. A system of reading comprehension assessments incorporating multiple response formats as well as varying tasks would likely be beneficial teachers to pinpoint specific skills that could be targeted during instruction. A system of assessments would allow analysis of student performance across measures when assessing reading comprehension.

REFERENCES

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

Betjemann, R. S., Keenan, J. M., Olson, R. K., & DeFries, J. C. (2011). Choice of reading comprehension test influences the outcomes of genetic analysis. *Scientific Studies of Reading, 15*, 363-382.

Breaux, K. C. (2010). *Wechsler Individual Achievement Test- 3rd edition (WIAT-III) technical manual with adult norms*. NCS Person, Inc.

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

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

Cain, K., & Oakhill, J. (2006). Assessment matters: Issues in the measurement of reading comprehension. *British Journal of Educational Psychology, 76*, 683-696.

Carlisle, J. F. (2000). Awareness of the structure and meaning of morphologically complex words: Impact on reading. *Reading and Writing*, *12*, 169-190.

Carlson, S. E., Seipel, B., & McMaster, K. (2014). Development of a new reading comprehension assessment: Identifying comprehension differences among readers. *Learning and Individual Differences*, *32*, 40-53.

Catts, H. W. (2009). The narrow view of reading promotes a broad view of comprehension. *Language, Speech, and Hearing Services in Schools, 40*, 178-183.

Catts, H. W., Compton, D., Tomblin, B. J., & Bridges, M.S. (2011). Prevalence and nature of late-emerging poor readers. *Journal of Educational Psychology, 104*, 166-181.

Catts, H. W., Hogan, T. P., & Fey, M. E. (2003). Subgrouping poor readers on the basis of individual differences in reading-related abilities. *Journal of Learning Disabilities*, *36*, 151-164.

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

Compton, D. L., Miller, A. C., Elleman, A. M., & Steacy, L. M. (2014). Have we forsaken reading theory in the name of 'quick fix' interventions for children with reading disability? *Scientific Studies of Reading, 18*, 55-73. doi: 10.1080/10888438.2013.836200

Cutting, L. E., & Scarborough, H. S. (2006). Prediction of reading comprehension: Relative contributions of word recognition, language proficiency, and other cognitive skills can depend on how comprehension is measured. *Scientific Studies of Reading, 10*, 277-299.

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

Eason, S. H., Sabatini, J., Goldberg, L., Bruce, K., & Cutting, L. E. (2013). Examining the relationship between Word Reading Efficiency and Oral Reading Rate in predicting comprehension among different types of readers. *Scientific Studies of Reading, 17*, 199-223.

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

Foorman, B. R., & Petscher, Y. (2010). Development of spelling and differential relations to text reading in grades 3-12. *Assessment for Effective Intervention*, *36*, 7-20.

Fore III, C., Boom, R. T., Burke, M. D., & Martin, C. (2009). Validating curriculum-based measurement for students with emotional and behavioral disorders in middle school. *Assessment for Effective Intervention*, *34*, 67-73. doi: 10.1177/1534508407313234.

Francis, D. J., Snow, C. E., August, D., Carlson, C. D., Miller, J., & Iglesias, A. (2006). Measures of reading comprehension: A latent variable analysis of the diagnostic assessment of reading comprehension. *Scientific Studies of Reading, 10,* 301-322.

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

Fuchs, L.S., Fuchs, D., Hosp, M. K., & Jenkins, J. R. (2001). Oral reading fluency as an indicator of reading competence: A theoretical, empirical, and historical analysis. *Scientific Studies of Reading, 53*, 239-256.

Garcia, J. R., & Cain, K. (2014). Decoding and reading comprehension: A meta-analysis to identify which reader and assessment characteristics influence the strength of the relationship in English. *Review of Educational Research*, *84*, 74-111. doi:10.3102/003465313499616.

Gellert, A.S., & Elbro, C. (2013). Cloze tests may be quick, but are they dirty? Development and preliminary validation of a cloze test of reading comprehension. *Journal of Psychoeducational Assessment, 31*(1), 16-28.

Gillingham, M.G., & Garner, R. (1992). Readers' comprehension of mazes embedded in expository texts. *Journal of Educational Research*, *85*, 234-241.

Gough, P. B., & Tunmer, W. E. (1986). Decoding, reading, and reading disability. *Remedial and Special Education, 7*(1), 6-10.

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

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers*, *36*, 193-202. doi: 10.3758/BF03195564

Guthrie, J.T. (1973). Reading comprehension and syntactic responses in good and poor readers. *Journal of Educational Psychology*, *65*, 294-299.

Guthrie, J.T., Seifert, M., Burnham, N.A., & Caplan, R. I. (1974). The maze technique to assess, monitor reading comprehension. *The Reading Teacher*, *28*, 161-168.

Hale, A. D., Hawkins, R. O., Sheeley, W., Reynolds, J. R., Jenkins, S., Schmitt, A. J., & Martin, D. A. (2011). An investigation of silent versus aloud reading comprehension of elementary students using maze assessment procedures. *Psychology in the Schools*, *48*, 4-13. doi: 10.1002/pits.20543.

Hale, A. D., Henning, J. B., Hawkins, R. O., Sheeley, W., Showmaker, L., Reynolds, J., & Moch, C. (2011b). Reading assessment methods for middle-school students: An investigation of reading comprehension rate and maze accurate response rate. *Psychology in the Schools*, *48*, 28-36. doi: 10.1002/pits.20544.

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

Helfeldt, J. P., Henk, W. A., & Fotos, A. (1986). A test of the alternative cloze test formats at the sixth-grade level. *The Journal of Educational Research, 79*, 216-221.

January, S. A., & Ardoin, S. P. (2012). The impact of context and word type on students' maze task accuracy. *School Psychology Review*, *41*, 262-271.

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

Johnson, E. S., Semmelroth, C., Allison, J., & Fritsch, T. (2013). The technical properties of science content maze passages for middle school students. *Assessment for Effective Intervention, 38*, 214-223. doi: 10.1177/1534508413489337

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

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

Kendeou, P., Papadopoulos, T. C., & Spanoudis, G. (2012). Processing demands of reading comprehension tests in young readers. *Learning and Instruction, 22*, 354-367. doi:10.1016/j.learninstruc.2012.02.001.

Kingston, A.J., & Weaver, W.W. (1970). Feasibility of cloze techniques for teaching and evaluating culturally disadvantaged beginning readers. *The Journal of Social Psychology*, *82*, 205-214.

Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review*, *95*(2), 163.

Louthan, V. (1965). Some systematic grammatical deletions and their effects on reading comprehension. *The English Journal, 54*(4), 295-299.

MacGinitie, W.H., MacGinitie, R. K., Maria, K., & Dreyer, L. G. (2000). *Gates-MacGinitie Reading Tests Form S & T, Fourth Edition*. Itasca, IL: Riverside.

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

McGrew, K.S., LaForte, E.M., & Schrank, F.A. (2014). Technical Manual. *Woodcock-Johnson IV Tests of Achievement*. Rolling Meadows, IL: Riverside.

McKenna, M.C., & Miller, J. W. (1980). The effects of age and distractor type on maze performance. In M.L. Kamil (Ed.), *Perspectives on reading research and instruction: 29th yearbook of the National Reading Conference* (pp. 288-292). Washington, D.C.: National Reading Conference.

McMaster, K.L., Espin, C.A., & van den Broek, P. (2014). Making connections: Linking cognitive psychology and intervention research to improve comprehension of struggling readers. *Learning Disabilities Research & Practice*, *29*, 17-24.

McNamara, D.S., & Kendeou, P. (2011). Translating advances in reading comprehension research to educational practice. *International Electronic Journal of Elementary Education, 4* (1), 33.

Mercer, S. H., Dufrene, B. A., Zoder-Martell, K., Harpole, L. L., Mitchell, R. R., & Blaze, J. T. (2012). Generalizability theory analysis of CBM maze reliability in third- through fifth-grade students. *Assessment for Effective Intervention*, *37*, 183-190.

Merino, K., & Beckman, T. O. (2010). Using reading curriculum-based measurements as predictors for the measure academic progress (MAP) standardized test in Nebraska. *International Journal of Psychology: A Biopsychosocial Approach*, *6*, 85-98.

Nation, K., & Snowling, M. (1997). Assessing reading difficulties: The validity and utility of current measures of reading skill. *British Journal of Educational Psychology, 67*, 359-370.

Parker, R., Guillemard, L., Goetz, E., & Galarza, A. (1996). Using semantic map tests to assess subject matter comprehension. *Diagnostique*, *22*, 39-62.

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

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

Pikulski, J. J., & Pikulski, E.C. (1977). Cloze, maze, and teacher judgment. *The Reading Teacher, 30*, 766-770.

RAND Reading Study Group. (2002). *Reading for understanding: Toward an R & D program in reading comprehension*. Washington, D. C.: RAND Education.

Rutherford-Becker, K. J., & Vanderwood, M. L. (2009). Evaluation of the relationship between literacy and mathematics skills as assessed by curriculum-based measures. *The California School Psychologist*, *14*, 23-34.

Saenz, L. M., & Fuchs, L. S. (2002). Examining the reading difficulty of secondary students with learning disabilities expository versus narrative text. *Remedial and Special Education*, *23*, 31-41.

Shanahan, T., Kamil, M. L., & Tobin, A. W. (1982). Close as a measure of intersentential comprehension. *Reading Research Quarterly*, *1,* 229-255.

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

Shinn, M. R., & Shinn, M. M. (2002). *AIMSWEB training workbook: Administration and scoring of reading maze for use in general outcome measurement*. Retrieved November 30, 2015, from http://www.cnyric.org/tfiles/folder1052/Administrationandscoringofreadingmaze.pdf

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

Spear-Swerling, L. (2004). Fourth graders' performance on a state-mandated assessment involving two different measures of reading comprehension. *Reading Psychology*, *25*, 121-148. doi:10.1080/02702710490435727.

Swain, K.D., & Allinder, R. M. (1996). The effects of repeated reading on two types of CBM: Computer maze and oral reading with second-grade students with learning disabilities. *Diagnostique*, *21*, 51-66.

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

Tindal, G., Nese, J. F., & Alonzo, J. (2009). Criterion-related evidence using easycbm reading measures and student demographics to predict state test performance in grades 3-8 (Technical Report 0910) Eugene, OR: *Behavioral Research and Teaching*.

Tolar, T. D., Barth, A. E., Francis, D. J., Fletcher, J. M., Stuebing, K. K., & Vaughn, S. (2012). Psychometric properties of maze tasks in middle school students. *Assessment for Effective Intervention*, *37*, 131-146. doi: 10.1177/1534508411413913

Twyman, T., & Tindal, G. (2007). Extending curriculum-based measurement into middle/secondary schools: The technical adequacy of concept maze. *Journal of Applied School Psychology, 24*, 49-67.

Williams, R. S., Ari, O., & Santamaria, C. N. (2011). Measuring college students' reading comprehension ability using cloze tests. *Journal of Research in Reading*, *34*, 215-231.

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*, 147-158. doi: 10.1177?1534508411420129.

APPENDICES

APPENDIX A

Fixed-Word Deletion

Small toys can be a big danger

Last month, Christin Rivas was playing **(shows, said, with)** a couple of small magnets at **(her, growing, silver-ball)** school. The 14-year-old student needed both **(hands, machine, type)** to grab something, so she put **(the, are, one)** mini-magnets in her mouth. Someone made **(her, could, your)** laugh, and…gulp. She swallowed them. **(What, Another, Five)** days later, Christin was in the **(hospital, who, buckyballs)** having the magnets taken out. Along **(bulletin, teens, with)** the magnets, doctors took out a **(panic, small, also)** part of her colon. The colon **(coiled, magnets, is)** the last part of the body's **(pooped, digestive, toys)** system. Christin used to love magnets, **(the, perform, in)** kind you can form into shapes **(and, hole, was)** use to perform magic tricks. But **(Orlando, she, risk)** found out the hard way what **(a, body's, doctors)** already know. It's not just little **(five, kind, kids)** who get into trouble with magnets. **(Tweens, Tracked, Cause)** and teens are also at risk. **(Eventually, It, Warnings)** are on many toys. But that **(think, has, swallow)** not stopped a growing number of **(no, life, kids)** from putting the little magnets in **(their, work, life)** mouths or noses. The magnets can **(get, something, and)** stuck in their bodies. There, they **(of, one, can)** cause serious harm. "Kids swallow a **(online, sent, lot)** of objects", said Dr. Tejas Mehta. **(He, The, Come)** treated Christin at the hospital. "Magnets **(cause, their, large)** more damage than anything else."

Magnets **(that, work, old)** very hard to find each other. **(Study, Their, System)** force can cause intestines to twist **(candy, she, and)** become blocked. The intestines are the **(doctors, long,**

stuck), coiled tubes connected to the stomach. **(Each, They, Form)** help the body break down food **(and, removed, large)** absorb nutrition. Eventually, the magnets can **(make, you, showed)** a hole in the intestines. This **(a, be, can)** cause an infection, Mehta said. Four **(home, connected, out)** of five kids who swallow magnets **(find, right, will)** need an operation to remove them, **(products, out, said)** Mehta. Some kids need a serious **(like, know, type)** of operation. In it, all or **(pins, help, part)** of the large intestine is removed.

(From, Doctor, Intestine) 2002 to 2011, the number of kids getting **(sick, mini-magnets, used)** from magnets jumped. A study showed 22,500 **(needed, kids, toddlers)** under 21 went to the hospital because **(mouth, taken, of)** magnets. Usually the kids swallowed the **(putting, magnets, will)**. But one in four kids put **(become, got, the)** magnets up their noses, said Julie **(this, Brown, normally)**. She is a doctor at a **(hospital, pass, their)** in Seattle. She helped write the **(study, hard, set)**.

The problem shows no sign of **(slowing, been, found)** down, she said. “There is something **(twist, very, takes)** tempting about magnets”, Brown said. “You **(want, four, Dr.)** to put them in your mouth. **(You, Having, Tempting)** want to try to separate the **(safety, use, magnets)** with your teeth. To toddlers, they **(five, from, look)** like the little silver-ball sprinkles on **(small, cupcakes, other)**”, she said.

Buckyballs is a set (shapes, hands, of) 216 magnets. People can stick them together **(mouths, usually, to)** make things. Last year, a product **(Seattle, about, safety)** group said Buckyballs should not be **(example, sold, went)**. But you can still find them **(online, eight, little)**. And Brown said that more than 3 **(million, them, together)** magnet sets have been sold. Other

(products, about, nutrition), with magnets are also a problem, **(she, my, put)** said. One example is bulletin boards **(than, no, with)** little magnet pins. There also are **(refrigerator, there, not)** magnets that look like candy.

Some **(remove, hit, doctors)** don't know magnets can be **(anything, serious, along)** problem. When Christin's mom took her **(other, daughter, advice)** to the hospital, a doctor just **(sent, magnets, pooped)** Christin and her mother home. The **(doctor, up, tubes)** said the magnets would pass out **(noses, should, of)** her digestive system when she pooped. **(So, Stick, That)** advice didn't sound right to Rivas. **(Teeth, Sold, She)** looked online. "That's when I hit **(very, the, with)** panic button", she said. Her mother **(took, that, separate)** Christin to another hospital. There, doctors **(last, tracked, it's)** the magnets with an X-ray machine.

(Decided, Together, Normally), anything put in the mouth takes **(need, many, six)** to eight hours to come out **(way, done, of)** the body. But Christin's magnets got **(stuck, someone, warnings)** for 24 hours. So, the doctors decided **(she, magnets, to)** operate. The doctors safely removed the **(magnets, serious, a)**. But the operation could cause her **(swallowed, operation, to)** have a blocked intestine in late **(life, intestines, blocked)**. A blocked intestine can be very **(dangerous, digestive, bodies)**. "Don't even think about touching them **(or, still, mother)** buying magnets", Christin advised. "I messed **(up, there, but)** my intestines. I worry about that **(stopped, things, down)** the road."

Opinion: Why do some 'winners' later become cheaters?

People compete against each other in many ways. They play against each other in **(and, sports, mind)**. They try to get the top **(grades, what, wanted)** in school. They try to outshine **(questions, figure, others)** at work. Competition can drive people **(to, talking, rolled)** do great things. However, it also **(released, there, makes)** people behave badly. Some figure the **(easiest, might, computer)** way to win is to cheat. **(They, Are, More)** act unfairly or lie.

What about **(to, people, winning)** itself, though? Does winning make people **(lot, behave, makes)** differently? Could it make them more **(or, likely, dice)** to cheat later on? Amos Schurr **(shows, has, and)** Llana Ritov decided to try to **(answer, through, results)** those questions. Schurr is a business **(with, professor, successful)**, and Ritov is a scientist who **(told, flashing, studies)** the human mind. Both live in **(Israel, game, shekels)**. The two released a report of **(teacher, their, does)** finding earlier this week.

Schurr has **(by, successful, always)** wondered about people with a lot **(those, of, it)** success. Some start out honest and **(easiest, wanting, form)** to do good things, but end **(people, ways, up)** lying, cheating, and stealing. Others stay **(the, fine, honest)**. What makes these two types of **(answer, successful, actually)** people turn out differently? Could it **(into, have, us)** something to do with the nature **(split, become, of)** their success?

Schurr and Ritov say **(there, won, really)** are two kinds of success. People **(can, when, report)** succeed by beating others in some **(those, want, form)** of competition. Or, they can succeed **(by, should, across)** doing a good job at something. **(Cheat, Competition, Schurr)** and Ritov say their study shows **(that, had, human)** the first kind of success changes **(people, two, at)**. When people beat others in a **(competition, behave, studies)** they start to think differently. They **(feel, beat, become)** more likely to behave badly.

Schurr **(get, and, was)** Ritov performed several experiments to find **(a, out, other)** what happens when people win competitions. **(If, Grade, They)** wanted to see if winning changes **(says, people, first)**. First, they had groups of students **(their, compete, cheated)** against each other. The students were **(could, groups, told)** to say how many signs were **(each, ended, flashing)** across a computer screen. Those who **(were, the act)** closest would be declared the winners. **(Badly, Performed, They)** would be given a prize. The **(winners, take, decide)** were actually picked by chance. They **(findings, had, starts)** not really beaten other students, but **(types, to, they)** believed they had. The students were **(then, can, also)** given something else to do. First, **(turn, remaining, they)** were split into pairs. One student **(scientist, well, was)** given two dice and a cup **(they, with, students)** a hole in the bottom. The **(outshine, other, against)** was told just to watch. The **(lie, there, pair)** then played a game. Each had **(signs, a roll)** chance to win a certain amount **(students, beating, of)** money. The money could be up **(that, can, to)** 12 Israeli shekels, roughly the same as 12 **(quarters, make, stealing)**.

The first student was told to **(based, shake, some)** the dice and keep the cup **(of, later, over)** them so that only he **(could, always, outcome)** see the results of the roll. **(His, Live, Keep)** outcome would be between 2 and 12. It **(that, honest, would)** decide how many shekels he could **(changes, however, take)**. The other student would receive the **(likely, could, remaining)** amount. The students who believed they **(several, right, had)** won the first competition cheated. They **(said, think, chance)** they had rolled a higher number **(beaten, than, top)** they actually had. They ended up **(we, play, taking)** more money than they deserved. The **(students, experiments, else)** who had not won the earlier **(decided, competition, for)** did not cheat.

What is it **(about, stealing, did)** winning that makes people behave badly? **(About, When, Good)** we win a competition, it makes **(number, wanting, us)** feel better than other people, says **(given, cup, scientist)** Dacher Keltner. It makes us feel **(over, wondered, like)** we have more power than others. **(We, Receive, Success)** begin to think we have the **(one, right, business)** to behave however we want. Cheating **(starts, thought, drive)** to seem just fine.

Schurr and **(we, deserved, Ritov's)** study may have a lesson for **(power, try, us)** all. Perhaps success should not be **(based, kind, happens)** on beating other people. Instead, it **(however, say, should)** be based on doing something well. **(There, People, Succeed)** might be less lying and cheating **(if, quarters, competition)** more people thought that way.

Wild Animals Losing Their Fear as Largest Predators are Dying Out

Large animal hunters like wolves or tigers are known as apex predators. They are at the top of **(than, predators, the)** food chain. No other animal except **(would, sounds, man)** hunts them. Apex predators are very **(important, islands, over)**. They help keep nature in balance. **(When, Elk, For)** example, wolves hunt elk. By killing **(some, fear, example)** elk, they keep elk herds from **(getting, with, nowhere)** too large. When there are too **(many, killing, smaller)** elk, trees cannot grow well. Soon, **(there, bold, birds)** have nowhere to live. One problem **(leads, however, be)** to another.

Scientists know apex predators **(help, else, for)** keep nature balanced. However, most only **(look, like, can)** at the way they kill other **(bears, manage, animals)**. Scientist Justin Suraci thinks there is **(recordings, crab, something)** else that is important about apex **(fish, predators, Canada)**. He said it is that other **(acted, animals, now)** are frightened of them. Suraci says **(other, man, listening)** animals spend much of their time **(watching, eating, areas)** and listening for apex predators. Even **(time, if, realize)** they are never caught, that fear **(is, important, frightened)** part of their lives. It makes **(brought, a, started)** big difference in the way they **(another, behave, area)**.

Fearful animals spend time worrying about **(open, he, getting)** eaten, Suraci says. They hide, they **(bring, run, take)**, they carefully look and listen. In **(other, turn, his)** that means they spend less time **(are, in, eating)**. More of the even smaller animals **(they, have, test)** eat manage to stay alive. Suraci **(says, they, stay)** all that has changed. Apex predators **(happen, bringing, are)** quickly disappearing all over the world. **(At, Near, Start)** the same time, fear is disappearing

(must, too, tigers). Smaller animals are no longer afraid **(hunters, well, now)** that the big hunters are gone. **(Others, Caught, They)** are behaving differently and in ways **(fearful, hide, that)** are harming nature.

Suraci decided to **(again, them, run)** an experiment to test his ideas. **(Soon, Do, His)** plan was to reintroduce fear to **(an, seem, there)** area with no apex predators left. **(Top, World, He)** wanted to see what might change. **(Look, Suraci, Watching)** traveled to the Gulf Islands in **(fewer, it, British)** Columbia, Canada. He picked the raccoons **(make, living, danger)** there to study. The raccoons were **(raccoons, bother, once)** hunted by big cats and bears. **(Now, Apex, Thinks)** those predators are all gone, killed **(beach, scientist, off)** by humans. The raccoons have gotten **(smaller, very, did)** bold, and no longer seem scared **(picked, at, hunt)** all. They flock to the beach **(carefully, in, most)** search of fish and crab. They **(was, do, those)** not even bother to look around **(shore, British, them)**, or to hide.

To make the **(one, something, raccoons)** fearful again, Suraci set up speakers **(of, alive, near)** the shore. The speakers played the **(said, sounds, plants)** of dogs barking. The raccoons quickly **(traveled, grow, started)** to act differently. They spent far **(listen, to, less)** time in open areas than before. **(Quickly, When, To)** they did appear in the open, **(large, few, they)** spent much more time looking around. **(Means, Overall, His)** the raccoons spent around two-thirds less **(some, time, harming)** looking for food than before. The **(far, left, difference)** in the way the raccoons acted **(soon, after, more)** spread to other animals.

Those raccoons **(Columbia, eat, by)** crabs and fish. Because they were **(is, even, now)** eating fewer of those animals, there **(much, were, problem)** soon more crabs and fish around. **(There,**

Speakers, Says) were also fewer periwinkle snails than **(before, flock, two-thirds)**. Crabs eat periwinkle snails. The new **(and, turn, fear)** brought back balance, Suraci says. Balance **(balance, reintroduce, was)** lost after humans killed the apex **(makes, predators, periwinkle)**.

When the balance of nature is **(upset, way, plan)** all sorts of problems happen. There **(real, because, can)** be too many of some animals, **(apex, and, killed)** too few of others. Plants and **(trees, birds, barking)** can start to die out. Diseases **(can, ideas, wanted)** start to spread. Of course, recordings **(decided, behave, of)** barking dogs is not the answer. **(Has, It, Overall)** would not take long for animals **(an, to, problems)** realize the danger was not real. **(Changed, Long, Instead)**, we must start trying to bring **(apex, except, all)** predators back. Bringing them back is **(they, gotten, the)** only way to make nature healthy **(leads, again, scientists)**, Suraci says.

APPENDIX B

Word-Feature Deletion

Many Warned in Time to Get to Safety as Severe Storms Hit Midwest

A group of thunderstorms marched across the Midwest bringing rare, November tornadoes whirling into town. Luckily, weather forecasters were able to find where the worst of the storms would go. **(His, Their, Water)** predictions of where the tornadoes were headed combined with television and radio warnings almost certainly saved lives. City officials also used text-messaging alerts and storm sirens. The storms blew through 12 **(states, winds, heard)**. The strong winds flattened whole neighborhoods in minutes. Only eight people were killed. The storms could have hurt **(many, it, town)** more people. By Monday, another reason the low number of deaths came out. In the hardest hit town, most families were in church.

“I don’t think we had one church damaged”, said Gary Manier, mayor of Washington, Illinois. Washington was hit by the storms. The **(beeping, warning, tornado)** cut a path bigger than two football fields in Washington. It damaged **(nor, get, or)** destroyed as many as 500 homes. The weather also hit parts of the Midwest, from Michigan to western New York. Daniel Bennett was about to give a sermon at his Washington church. There were 600 to 700 people waiting to listen to him when he heard a beeping sound, then another and another. “About 24 phones started going off in service, **(he, it, leg)** said, “and everybody started looking down”. The beeping was a text message from the National Weather Service. **(She, Form, It)** warned that a twister, or tornado, was near them. Bennett stopped service and told everyone to get to a safe place. Many townspeople said those text **(satellites, clear, messages)** helped keep people safe. It's got to be connected to how they could get the warnings so quickly, Bennett said.

In Indiana, Taylor Grenna heard emergency sirens go off. **(He, Few, City)** got a message on his cell. A friend also called to warn him the storm was close. Glenna went outside **(or, and, hurt)** saw icy hail falling from the sky. Then, he heard a loud boom. **(They, He, Damage)** ran to his basement just in time. On Monday, Glenna looked at the damage on crutches. He hurt his leg when the wind knocked his home off **(who, storm, its)** base. "I would say we had pretty good **(warning, scientists, combine)**", Glenna said. "We just didn't listen to it".

Forecasting has gotten better. Now, there are faster computers. Scientists can guess what the **(chief, weather, told)** will do using math. Weather forecasters tell when large storms will form. In the past 10 years, **(she, they, math)** have doubled how far in the future they can go. Bill Bunting is the chief forecaster at the National Weather Service. He said it wasn't clear until Saturday that a large **(storm, cut, number)** could hit. That's when the information from weather stations, weather balloons **(or, go, and)** satellites came together. It suggested there was more than enough **(water, boom, called)** in the air for a big storm. Water is fuel for storms.

The storms started because of unusually warm, wet air from Louisiana to Michigan. **(He, Stand, It)** was then hit by an upper-level cold front. The crash of hot **(but, and, start)** cold, dry and wet, is what makes tornados. The storm moved fast. That meant the storm hit more places before **(they, it, field)** died down. It touched more people. **(Or, Give, But)** in places where it hit, the system may have been slightly less. About 90 minutes after the tornado plowed through Washington, rain and high winds hit downtown Chicago. Officials at Soldier Field, a football field, emptied the stands. **(They, She, Front)** ordered the NFL teams, the Bears and the Baltimore Ravens, off the field. Fans were allowed back to their seats shortly after 2 p.m. The **(game, storm, moved)** started again after about a two-hour delay. Still, this has been an easy year for twisters in the U.S. The number of **(twisters, games, meant)** is low compared to previous years.

Teaching Traditions to Save Young Lives

In a small white building, a group of Native Alaskan young people gathered on a cold February night. They came to dance, sing and drum together. **(They, He, Music)** also came to save each other. The students call themselves the Native Survivors. They are trying to stop the sadness felt by too many Native Alaskan **(seniors, dropped, youths)**. They learn their people's traditional music and crafts. They find **(destruction, peace, happening)** through sewing. They make friends through dance.

Many Native Alaskan Eskimos are poor **(and, but, talks)** have no job. Many struggle with alcohol. **(Some, None, Gathered)** even lose hope and end their own lives. Since the Native Survivors began, though, something amazing has happened. No one has ended their life in Hooper Bay. "Because of us", said Wilma Bell-Joe, who began the **(Alaska, brings, group)** in 2013. As she spoke, the building echoed with music and the beat of drums. "These kids are inspired and **(discouraged, treat, encouraged)**", said Bell-Joe, age 35. Bell-Joe herself drank alcohol in the past. **(They, She, And)** now uses her experience to show the teens the trouble it brings. Native Survivors is a program of Americorps. The group sends volunteers to work in **(communities, buildings, through)** across the United States. In Alaska, some of the AmeriCorps workers help the environment through recycling **(and, yet, interested)** gardening. Others, like Bell-Joe, are interested in health. In the town of Huslia, young people learned to race dogs and care for **(them, sees, her)**. Young people and adults learn more than dog care or recycling. They develop important skills **(grow, or, and)** become confident.

Bell-Joe grew up in Hooper Bay, one of 13 children. **(Her, Many, Discovered)** family lived in a two-bedroom home, her parents sleeping in one room and all the children in the other. She dropped out of school more than once. Native **(Survivors, Americans, they)** started with just two students. It now has 53 members and the group is growing, Bell-Joe said. Not even 5 feet tall, Bell-Joe looks almost like a teenager **(themselves, herself, skills)**. Sometimes she hears of a problem at Hooper Bay School. She dresses up like a teen **(for, and, to)** slips inside to see for herself. She said she talks to the teachers about what she sees. She reminds them that **(who, they, safety)** are there to teach and encourage, not treat children unfairly. “They are not there to talk down on the kids”, she said.

The modern world can seem far away from Hooper **(Mall, Is, Bay)**. Bell-Joe just learned about the selling website Amazon last month. **(You, She, Reminds)** discovered Facebook only a year ago. Learning traditional skills help give the kids a feeling of safety, she said. Bell-Joe’s parents teach at the little **(building, drum, sew)**. Her father is the mayor of Hooper Bay. **(Her, It, Workers)** mother teaches Yup’ik, the native language, at the town’s preschool. Other elders teach how to make harpoons and beading. They teach kids how to make dance fans **(and, nor, discover)** sew qasperet, the Yup’ik word for cloth pullovers. These traditional **(plays, so, skills)** teach patience, Bell-Joe said. They also help kids work through problems rather than blow up in anger.

On the cold February night, her father led the drumming. **(Her, Whose, Reached)** mother led the dancers. The Native Survivors danced into the night. They acted out the story of a big family celebration, a loud, happy time. Then they reached out **(and, blow, or)** danced their way to an imaginary crying baby. They gave comfort through their dance.

Animal that was Model for Teddy Bear is not off Endangered Species List

Teddy bears were named after real bears. The stuffed toys got their name from Louisiana black bears. President Theodore Roosevelt was invited to hunt Louisiana black **(bears, called, tree)** over 100 years ago. The hunt was in Mississippi. Mississippi is in the southern part of the **(were, story, country)** on the Gulf of Mexico. The 26th president did not find a bear. The people who invited him caught a bear **(an, or, store)** tied it to a tree. They thought that it would be easy for the president to kill the animal. **(Roosevelt, Mississippi, selling)**, however, would not shoot the bear.

A newspaper told the story with a cartoon, and a candy store owner in New York say **(them, it, people)**. The store owner put two toy bears in his shop window. Then **(he, you, hunt)** asked the president if he could call them "Teddy's bears". Teddy is short for Theodore. The toy **(nor, plants, and)** the story became popular. Soon the candy store owner was selling many of Teddy's bears. Later, the bears were simply called teddy bears.

Many **(years, states, worked)** ago, real Louisiana black bears were dying out. They once lived in the states of Texas, Mississippi, and Louisiana. Then, the bears were only left in Louisiana. **(Species, Caught, Louisiana)** is next to Texas and Mississippi on the Gulf of Mexico. Now, leaders say that the bears are healthy. **(He, Toys, They)** do not have to be protected by the Endangered Species Act anymore. **(It, Job, Who)** is a law that protects plants and animals that are in danger of disappearing forever.

Sally Jewell is a United States official. **(Her, It, Taking)** department looks after plants and animals. She shared the news about the bears at the Tensas River National Wildlife Refuge in

Louisiana. A **(lives, owner, refuge)** is a safe place for animals. Most of the Louisiana's black bears live there. Taking the bears off the list does not mean that nobody will take care of them. The state of Louisiana will take over the **(cartoon, related, job)**. Part of the job will be planting more hardwood forests, where the bears like to live. Hardwoods are trees that lose **(their, his, should)** leaves in the fall. Jewell said, "The work's really just beginning".

Some people have worked for many reasons to protect the bears. **(They, We, Fuller)** do not think that the Louisiana black bear should be taken off the list. Michael J. Robinson works for a group that helps to protect wild animals. **(They, He, Blocks)** said that some of the bears may not be Louisiana black bears at all. They might be related to black bears that were brought from the **(link, state, disagree)** of Minnesota. Deborah Fuller is a scientist **(parents, she, who)** disagrees with Robinson. She said that scientists found no link to the Minnesota bears. **(However, Nor, Children)**, she said the bears may share some genes. Genes are the building blocks for everything in the human body. **(They, His, Bears)** are passed from parents to children.

One part of Louisiana may be home to almost 700 bears, Fuller said. In another **(number, part, think)** of the state, there are between 350 and 600 bears. She said that the numbers are important **(or, but, take)** so is whether the bears can do well. Scientists think that **(she, they, group)** can. Harold Schoeffler does not agree with Fuller. He works for a group that protects nature. **(Who, He, Home)** helped to get the bears on the protected list. Schoeffler said that there are not enough Louisiana black bears yet to take them off the list.

APPENDIX C

Sentence Deletion

A Support Program is Helping Foster Kids be Confident Adults

When parents do not make safe homes for their children, the government finds another palace for the kids to stay. Some live with family members. **(Some stay with their parents/ Some are taken in by other families/ Some taken other families are by in)**. This is called foster care. When Michael McKernan was growing up, his father was in prison. When he was a teen, his mother abandoned him and his brothers and sisters. They all went into foster care. His younger siblings ended up in a different foster home than him. **(McKernan also lived with different family members for a few years/ McKernan different family members few years for a also lived with/ McKernan also finds another place for kids to stay)**. Without a stable home life, McKernan dropped out of high school. He started working overnights at Wal-Mart. When he turned 18, he still had a lot to learn. McKernan was living with his grandmother. He did not have parents to help guide him. Luckily, he met a social worker named William Childress.

Social workers help people like McKernan who need support. Childress works with a program called YVLifeSet. **(The program gave his name registered for a test/ The program who are just extra help becoming adults to foster kids/ The program offers extra help to foster kids who are just becoming adults)**. Childress helped push McKernan toward a brighter future. Childress encouraged him to join YVLifeSet, and McKernan accepted. Childress became McKernan’s counselor. He nudged him, the way a father would. **(He regularly and gave in with McKernan**

and checked encouragement and advice him stead/ He abandoned him regularly by giving him thick practice books to take a test/ He checked in with McKernan regularly and gave him steady encouragement and advice.)

McKernan set his sights on college. He wanted to go further than his parents. Because he had been in foster care, the government might help him pay for college. **(First, GED to pass the he would have to/ First, he would have to pass the GED/ First, he would stay up late to finish)**. The GED is a test students can take to make up for not finishing high school. He studied hard, staying up late with thick practice books. The test was set to take place on Dec. 11. His heart was pounding as he walked up to the sign-in table and gave his name. The woman checked her list and told him that she could not find his name. He was not registered. **(McKernan was sure but registered he had there was nothing for the test he could do/ McKernan took the GED test and passed after registering for it again/ McKernan was sure he had registered for the test, but there was nothing he could do)**. He went home, feeling hopeless. He could take the test in a year, but by then it would be too late to get help paying for college. He could not pay for it on his own.

Childress happened to be in town that same day. When he stopped by McKernan's grandmother's house to check in, he saw the gloomy look on the teenager's face. **(He knew something was wrong/ He knew everything went well/ He was wrong something knew)**. McKernan told him what had happened. "Now wait," Childress told him, "Let's see what we can do". Childress made some calls and learned about an exam the following week in another area.

(McKernan would not take the test after all/ McKernan after all take the test would get to/ McKernan would get to take the test after all). McKernan took the GED exam and passed.

Since then, things have been going well for him. **(He started college and is enjoying his classes/ He is studying hard hoping to get into college/ He college and his classes is enjoyed started)**. He now dreams of studying in another country. McKernan is not alone. YVLifeSet and other programs like it have helped thousands of foster kids. Erin Valentine is an expert in foster care. She measured how useful YVLifeSet is for foster kids by doing a story of them. **(She discovered that finding homes YVLifeSet increased kids' changes of as grownups and good jobs foster/ She discovered that YVLifeSet increased foster kids' changes of finding homes and good jobs as grownups/ She discovered that YVLifeSet had hurt foster kids and hard to find good homes for them as grownups)**. She said that the social workers at YVLifeSet act almost like parents. Having that support can make a big difference.

A Brand New Toilet for the World's Poor

University students in California have teamed up with a company called Kohler that makes things for bathrooms. Their goal is to build a brand new kind of toilet. This might make you laugh. Don't. It's not funny. Water made dirty by poop can cause deadly diseases in poor communities. **(Many people there don't have toilets like we do/ There don't many people like we do have toilets/ Most people there have toilets like we do)**. For example, the United Nations says that more than 600 million people in India don't use toilets. They don't use pits, called latrines, either. They just go on the ground. And, that's just one country. Big improvements have occurred in the past 20 years. But, about 1 billion people worldwide still go out in the open. Another 700 million use dirty types of bathrooms. Some people use 'hanging latrines' that dump directly into streams. **(It can get when the waste is in that disposed into waterway/ When the waste is disposed in that way, it can get into the water/ The water is safe from waste being disposed that way)**. If people drink that water it can cause diarrhea. It causes people to go the bathroom many times a day. If it goes on too long, the body loses the water and salts it needs to survive.

Diarrhea kills as many as 1.5 million people each year. **(Most of those water made dirty by are from deaths human waste/ The toilet has improved over the next couple hundred years/ Most of those deaths are from water made dirty by human waste)**. Most of the victims are under 5 years old. Can Kohler help? The company is mostly known for producing expensive sinks and faucets. It's not known for making toilets for the world's poor. **(Plumbing products of makers world's biggest is but Kohler one the/ Kohler is a brand new kind of toilet/ But Kohler is one of**

the world's biggest makers of plumbing products). In 2011, the Bill and Melinda Gates Foundation announced its 'Reinvent the Toilet Challenge'. The foundation was started by Microsoft billionaire Bill Gates. It gives the most money of any foundation. That made Kohler take notice.

It's been a long time since the toilet changed much. The modern flush toilet isn't much different than the one invented in the 1500s by Sir John Harington. He installed one of his toilets in a palace for Queen Elizabeth I. The toilet was improved over the next couple hundred years. It has worked pretty well. It helps lower disease – when it's well-connected. It needs to be attached to a system to treat the waste it swirls away. **(Most people in the world have modern pipes or waste treatment plants/ In much of the world, there are no pipes or waste treatment plants/ In much of the world, waste treatment no proper or there are plants)**. They cost too much to build for poor countries. "That's very, very expensive", said Doulaye Kone, an engineer who grew up in a village in Africa's Ivory Coast. He works at the Gates Foundation. "It will work in very few (developing world) cities". **(So the foundation asked for simple designs/ So, the foundation asked for the most expensive designs/ So simple designs for the asked foundation)**. They don't need to be hooked up to water, sewer or electrical lines. It gave money to eight universities. One of those getting money was the California Institute of Technology (Caltech).

Caltech's plan was to develop a toilet and waste treatment system all in one. The toilet would be powered by a solar panel. **(It would be hooked up to electrical lines/ It from the get energy would sun/ It would get energy from the sun)**. And it would store it for use at night. Kohler gave Caltech parts for the first design. The Caltech toilet went on to win the Gates challenge in 2012. Caltech will test the toilet in India later this year. **(One cost with the problem Caltech is toilet/ One problem with the Caltech toilet is cost/ One problem with the Caltech toilet is the simple design)**. It will take $1,500 to $2,000 to build at first. And it's complicated since it has a solar panel and battery. **(But, the Gates Foundation is giving money to other toilet makers/ But, the Gates Foundation has stopped giving money for toilets and moved on to solar panels/ But, the Gates Foundation makers to other toilet money is giving)**. American Standard Brands also received money.

American Standard has taken a low-tech approach. The company developed a plastic toilet pan with a trap door and water seal. It's designed to close off the holes under latrines from the open air to stop flies from entering. **(Last year, American Standard donated more than 500,000 of the devices to Bangladesh, the company said/ Last year, American Standard did not donate devices to Bangladesh or Kenya, the company said/ Last year, American Standard 500,000 donated of the devices more than the company said to Bangladesh)**. It now is working in Kenya to develop a simple device for areas that have little water.

Zoo Recording Tiger Voices to Gather Information, Help Animals in the Wild

Strannik is a huge tiger who weighs 422 pounds. He lives at the Milwaukee County Zoo. Not long ago, Strannik got ready to eat his breakfast. The zookeeper passed some beef through the bars of his cage. Strannik laid his ears back and made a chaffing sound. The sounds were all caught on a small recorder pointed toward Strannik. Chuffs, roars, growls, and whines all mean something in Strannik's tiger vocabulary. **(They are like the words he speaks in tiger language/ He speaks in tiger language like the words they are/ It will help pick out tigers in the wild)**. The Milwaukee zoo has three female tigers named Amba, Tula, and Nuri. They make important sound, too. The Milwaukee zoo is recording the tiger sounds for a special group called the Prusten Project. **(Count how many in the wild tigers are living this information scientists will help/ Recorders cost $600 to $900 each/ This information will help scientists count how many tigers are living in the wild)**.

Amanda Ista is a zookeeper at the Milwaukee zoo. "As zookeepers there's not a lot we can do for the animals in the wild", she said. **(She added, "tigers in this project wild we can through help"/ "We cannot help tigers in the wild", she added/ "We can help tigers in the wild with this project, though", she added)**. Other American zoos are also helping the Prusten Project. They are recording their tigers, too. Scientists will listen to the recordings. They hope to figure out each animal's vocal fingerprint. Think of it like a fingerprint but instead a voice print. No two voices are exactly the same. They hope to use the information to build a computer program. It will help pick out specific tigers in the wild. **(The sounds were caught on a small recorder pointed toward Strannik/ That way, scientists will be able to accurately count the**

beautiful but often hard-to-find animals/ Beautiful but often hard-to-find animals will be able to that way, scientists to accurately count the).

Scientists plan to place the same recorders in the wild, too. They will travel to countries like India and Sumatra. **(Listen on tigers are living in the wild use the recorders they will to eavesdrop/ It turns sounds into pictures and recordings to build a computer program/ Then, they will use the recorders to eavesdrop or listen on tigers)**. It will help scientists figure out how many tigers still live in the wild.

Courtney Dunn is in charge of the Prusten Project. She said there are only about 3,200 tigers left in the wild. **(They are safe since many are left/ They an tiger animal are endangered/ They are an endangered animal)**. Dunn used to work at the National Tiger Sanctuary in Missouri. She met a very talkative tiger known for her strange sounds. Dunn wondered if tiger sounds could help save the tigers. She remembered the Whalesong Project. It has helped whales. Dunn studies how tiger vocalizations can be used to pick out certain animals. Dunn uses a software program. It turns sounds into visual spectrograms or pictures. **(It shows how tiger's vocal cords vibrate, or move/ It shows how two voices are exactly the same/ It shows or move a tiger's vocal cords how to vibrate)**. Dunn can tell whether a tiger is male or female with this information. She also studies tigers' long calls. They have a very deep roar. It can carry 3 miles and is mainly used for mating and marking their spot.

Recorders cost $600 to $900 each. Prusten Project is a nonprofit group. It relies on money donated, or given, by animal lovers. Dunn plans to travel to Sumatra and set up recorders. **(In the meantime, Dunn and her helpers are listening to recordings/ In the meantime, Dunn has**

given up her work to pick out certain animals/ In the meantime, recordings listening to are her and Dunn helpers). Zookeepers are recording 50 tigers at 15 zoos in the United States.

Strannik's name means pilgrim or wanderer in Russian. He arrived two months ago at the Milwaukee zoo. **(Zookeepers are helping the Prusten Project in Milwaukee/ Zookeepers are slowly introducing Strannik to his fellow tigers in Milwaukee/ Strannick to his fellow tigers zookeepers Milwaukee introducing slowly are)**. The female tigers heard him before they saw him. Strannik was roaring from the other side of the zoo. "He's very talkative. He's actually very laid back as tigers go", said Ista. She sprinkled Strannik's favorite scent in his area. It smelled like apple pie. Ista held her hand up and called 'open'. **(Strannik opened his mouth and gulped down his breakfast/ Strannik roared to his fellow tigers from the other side of the cage/ Strannik his breakfast opened and down mouth gulped)**. The tiger's big pink tongue curled around the ground beef. He made a chuffing so

APPENDIX D

Multiple Choice

Ancient Skull from Kabul’s Great Wall Twists the Legend of the Cruel King

Almost everyone here knows the legend of Kabul's cruel king. Kabulis recite it when they see Shew Darwaza Mountain. Look closely, they tell visitors. You will see an ancient wall running along the mountain's edge. It looks like the teeth of a saw. This is the Great Wall of Kabul. The wall holds terrible secrets. Kabul is the capital of Afghanistan in Asia.

The legend takes place about 1,500 years ago. This was around the year 500 A.D. Kabul's king forced his male subjects to build the Great Wall to protect the city from invaders. Anyone who refused to work on the wall was buried inside it. "Maybe something like that happened", said Aziz Ahmed Panjshiri, a historian. "We have many legends about the cruel king."

A few years ago, history got some help from science. In April 2013, heavy rains caused part of the wall to fall down. Something smooth and pale was found in the damp dirt. It was a human skull. "This skull shows that the stories were true", said Abdul Ahad Abassy. He is in charge of Afghanistan's Department of Historical Monuments. But Science added another chapter to the legend of Kabul's cruel king. Experts in Germany tested the skull. They found that it was not, in fact, 1,500 years old. Instead, it's just about 500 years old. They sent it back and said, "This is not old", Panjshiri said.

Afghanistan is an ancient country. Finding a 500-year-old skull is not a big deal. The city of Balkh, in northern Afghanistan, is almost 5,700 years old, Panjshiri said proudly. Balkh was once one of the greatest cities on Earth. It was the center of a great empire. It stretched from Greece in the west to India in the east.

Some scholars say the Great Wall was built about 1,500 years ago. Most of them believe it was built about 200 years later. Panjshiri claims it is much older. In one popular story, the wall was built 1,500 years ago by Zamburak Shah. The king was so cruel that his subjects killed him and buried him inside his own wall. Sometimes, the tale says, it was a beautiful slave girl who tricked the king.

Most people say the king buried his subjects in the wall. As is common in Afghan tales, this story has no happy ending. Three years ago, rain uncovered the skull and other bones. In April 2013, the mayor of Kabul sent Panjshiri and a photographer to investigate.

Word spread about Kabul that the legend was not true. The German scientists dated the skull to about 1,550 or 1,000 years after Zamburak Shah died. Now the story has changed. A group of popular kings who ruled Kabul 500 years ago have been blamed. The greatest of these kings, known as Babur, was a warrior poet. He loved Kabul so much that he was buried there.

"When skeletons or skulls come up, it's easy to weave stories around them", said Thomas Barfield. He is a professor at Boston University. Barfield added that stories can tell much about how people think and feel in the present. Sometimes archaeologists figure out that the story does not quite fit the facts, Barfield said with a laugh. Archaeologists study ancient things to find out about the past. "Once again, science ruins a really good story".

Answer the following multiple choice questions from reading the passage.

1. Which detail from the article is MOST important to include in its summary?
   a. The Great Wall of Kabul might hold terrible secrets.
   b. A photographer was sent to the wall to take pictures of the skull.
   c. The skull was found in the Great Wall of Kabul.
   d. The city of Baikh is almost 5,700 years old.
2. Which selection from the article BEST expresses the main idea?
   a. This is the Great Wall of Kabul. The wall holds terrible secrets. Kabul is the capital of Afghanistan in Asia.
   b. “The skull shows that the stories were true”, said Abdul Ahad Abassy. He is in charge of Afghanistan’s Department of Historical Monuments.
   c. Experts in Germany tested the skull. They found that it was not, in fact, 1,500 years old. Instead, it was about 500 years old.

   d. Most people say the king buried his subjects in the wall. As is common in Afghan tales, this story has no happy ending.

3. According to the article, what happened when experts tested the skull?

   a. They found it was not as old as people expected.

   b. They found it was about 1,500 years old.

   c. They found that it was the skull of a young girl.

   d. They found that it belonged to the cruel king himself.

4. How old is the city of Balkh in northern Afghanistan?

   a. 1,500 years old

   b. 500 years old

   c. 5,700 years old

   d. 100 years old

5. According to the article, what happened after the age of the skull was discovered?

   a. Scientists realized that the original story was true.

   b. People were angry that the story of the cruel king was not true.

   c. Scientists visited the wall and searched for more skulls.

   d. People made a new story fit the skull that was found.

6. Based on information presented in the passage, archeologists study
    a. Ancient skulls to determine their age
    b. Ancient things to find out about the past
    c. Ancient monuments in museums
    d. Ancient leaders in Afghanistan
7. What does Barfield mean when he says that the stories people tell reflects how they feel in the present?
    a. The way people feel does not change from the past to the present.
    b. Current thoughts and feelings do not shape how we view the past.
    c. The stories people tell almost always fit what archeologists find.
    d. Current thoughts and feelings are important at shaping how we view the past.
8. Why did Abdul Ahad Abassy want to test the skull?
    a. To prove that Afghanistan is an old country.
    b. To find out if it was the skull of the cruel king.
    c. To see if the legend of the Great Wall was real.
    d. To discover when the wall was built.

9. Which of these statements is TRUE based on the article?
    a. The king who built the wall was killed by his subjects.
    b. There are many different stories about the Great Wall.
    c. A cruel king buried people from his kingdom inside the wall.
    d. Most stories about the wall have been proven by science.
10. Based on the article, what inference did experts make once the skull was examined?
    a. The skull showed that the victim was young and probably a girl.
    b. The skull confirmed that the stories of Babur's burial are true.
    c. The condition of the skull prove that the person was buried there as punishment.
    d. The age of the skull showed that the story of the cruel king was probably not factual.

Keeping a Park's Beautiful View

Theodore Roosevelt National Park is named after Theodore Roosevelt, U.S. president from 1901 to 1909. Roosevelt was a war hero, boxer, cowboy and hunter. He was also famous for working to protect natural areas.

Valerie Naylor fell in love with Theodore Roosevelt National Park more than 40 years ago. She was a teenager, visiting from Oregon with her parents. She promised that she would come back, and she did. Naylor volunteered and did research in the park for many years. Today, Valerie Naylor is in charge of the entire park. The national park covers 70,000 acres in North Dakota.

Visitors come to Theodore Roosevelt National Park for the natural sounds, fresh air and of course the beautiful views. The park is named after Theodore Roosevelt. As president, he worked to protect natural areas. He visited North Dakota in the 1880s to hunt buffalo. He also lived in the area as a rancher for a time.

Valerie Naylor is standing at Oxbow Overlook, looking out over hills, trees, and the river shimmering below her. The park is "a very, very special place," she said. But it's in danger. "It's so vulnerable." The lands around Theodore Roosevelt National Park have a lot of oil.

North Dakota is in the middle of an oil boom — companies drilling oil in the state are making a lot of money. The state and federal government are making money too. Many people are becoming rich from drilling for oil. Today, North Dakota has more than 11,000 oil wells. About 60,000 more could be built if people are allowed to drill around Theodore Roosevelt National Park. The drilling will continue for at least 20 years.

It's Naylor's job to make sure the oil drilling doesn't hurt her park. Many projects have been coming closer. "There's so many things going on so quickly. It might be a pipeline, power line, oil well, or new road," Naylor said. Naylor drives her car through a beautiful area. She points to a fence. It's the border of the park. A company wanted to build an oil pump near the fence. However, Naylor wrote a letter to the company, asking it to move the project. Eventually, after months of paperwork and discussion, the company agreed to move the pump. Naylor and others want to protect the park's quiet and solitude. They want the sound of blowing wind and singing birds, not the hum of engines. They want the darkness of the night sky, not the flares of oil drilling.

Deb Hornfeldt and Debbie Virnig were visiting the park from Minnesota. Hornfeldt said spending time in nature helps her relax and feel less stressed. "It's really important to have places like this," she said. "Natural spaces are getting smaller and smaller," Virnig added. Both Hornfeldt and Virnig agree with Naylor about the importance of the park's quiet and solitude.

Answer the following multiple choice questions from reading the passage.

1. Based on information in the article, how many oil wells does North Dakota have presently?
    a. 11,000
    b. 19,000
    c. 60,000
    d. 75,000
2. According to the article, Theodore Roosevelt was famous for all of the following EXCEPT?
    a. Boxing
    b. Hunting buffalo
    c. Protecting natural areas
    d. Building the National Park
3. Which selection from the article BEST expresses a main idea of the article?
    a. Naylor and others want to protect the park’s quiet and solitude.
    b. North Dakota is in the middle of an oil boom.
    c. The park is named after Theodore Roosevelt, U.S. president.
    d. Valerie Naylor fell in love with the park more than 40 years ago.

4. According to the article, why are visitors attracted to the Theodore Roosevelt National Park?

    a. Oil wells

    b. Narrow road

    c. Natural beauty

    d. Cowboys

5. Based on the information from the article, what can you conclude about Valerie Naylor’s profession?

    a. She is a researcher.

    b. She is a forest ranger.

    c. She is an environmental activist.

    d. She is a park superintendent.

6. Which statement shows the problems caused due to the drilling of oil wells?

    a. The lands around Theodore Roosevelt National Park have a lot of oil.

    b. The drilling will continue for at least 20 years.

    c. “It might be a pipeline, powerline, oil well, or new road”, Naylor said.

    d. Companies drilling oil in the state are making a lot of money.

7. According to the article, which of the following is correct?
    a. North Dakota has nearly 60,000 oil wells.
    b. The North Dakota park was built by Theodore Roosevelt.
    c. Valerie Naylor has been working at the park 40 years.
    d. The park covers 70,000 acres in North Dakota.
8. Which detail from the article is MOST important to include in its summary?
    a. Hornfeldt said spending time in nature helps her relax and feel less stressed.
    b. It's Naylor's job to make sure the oil drilling does not hurt her park.
    c. A company wanted to build an oil pump near the fence.
    d. Roosevelt was a war hero, boxer, cowboy, and hunter.
9. How did Naylor convince a company to move the plans for an oil well away from the park fence?
    a. She wrote letters to the company, asking it to move the project.
    b. She asked the company to use new drilling techniques.
    c. She told the company about the natural beauty of the park.
    d. She drove her car through the park to find other projects.

10. Theodore Roosevelt was famous for his conservation efforts. Conservation in this sentence means

    a. Protection.

    b. Neglect.

    c. Destruction.

    d. Vulnerable.

Finding Dory in Her Own Movie

Michael Stocker is a master animator. He can bring drawings or pictures of puppets to life. Moving pictures are known as animations. Animators are artists. They also must study how to create motion. When Stocker makes an animation, he first creates a very large set of pictures that are slightly different from one another. When the pictures are shown rapidly in order, it looks like the drawn image is moving. Animation can be made with hand drawings. It can also be done with photos of puppets or pictures made on a computer. It is used to make everything from cartoons to animated movies.

Stocker has worked for a company called Pixar for 13 years. The company makes animated movies using computers. It created the popular movie, “Finding Nemo”. Stocker’s latest project was “Finding Dory”. The movie continues the story of “Finding Nemo”. It has a different main character. The hero is Dory, a forgetful blue tang fish. For the last 3 ½ years, Stocker has led a team of 70 animators. Together they brought the story of “Finding Dory” to life. It was a long and difficult job.

The team began by making a model for each character in the movie. Computers were used to make the models, which are three-dimensional. Instead of being flat, the models are like puppets. They can be turned from side to side on the computer screen. Stocker already knew what most of the characters should look like. Dory and her friends Marlin and Nemo are well-known from swimming in “Finding Nemo”. The earlier movie came out in 2003. However, the team could not simply reuse old models, Stocker said. They had to build new characters that looked like the old ones. The voice actors recorded their parts early on. Stocker and his team had to make what they did fit the words. The voice of Dory is done by Ellen DeGeneres. She also played the part in “Finding Nemo”.

One of the hardest characters to make was Hank, the grumpy octopus. Hank the Octopus helps Dory escape from a California aquarium. “We wanted to do a realistic octopus”, Stocker said. At first it was very difficult to figure out how an octopus moves. All those legs moving at once is a tricky thing for an animator to copy. Stocker and his team spent some time at the Monterey Bay Aquarium to study the way the octopus moves around. They also held some of the animals, who turned out to be surprisingly friendly. Stocker said, “They purr like a cat!” The visits were helpful, but Hank still took two years to build.

For Stocker, there was something especially exciting about working on a follow-up to “Finding Nemo”. The earlier movie was the first thing he worked on at Pixar. Stocker is a big fan of “Nemo”. It was important for his family too. “For my kids, it was their first movie”, he said. “We’re finding out how many people love that first movie. There’s something about that world people want to live in for an hour and a half”, he said. “It’s beautiful”. Stocker said he worked hard to make sure the new movie was just as special. He wanted it to have the same magical feeling as the original.

Answer the following multiple choice questions from reading the passage.

1. Based on information in the article, how many years did it take to build the model of Hank the octopus?
   a. 3 ½ years
   b. 2 years
   c. 3 years
   d. 13 years
2. According to the article, which of the following is correct?
   a. "Finding Nemo" came out in 2001.
   b. It was easy to rebuild Dory from the original movie.
   c. Stocker worked for Pixar for 13 years.
   d. The voice actors did their parts last.
3. Which selection from the article BEST expresses the main idea?
   a. The animators of "Finding Dory" worked hard to make the movie look realistic.
   b. Animated cartoons are easier to make than movies with actors.
   c. Making movies like "Finding Dory" is too difficult and not worth it.
   d. The animated movie "Finding Dory" is going to be better than "Finding Nemo".

4. How does Stocker make Dory move?
    a. Stocker worked with animators to make Dory look real.
    b. Stocker worked for 3 years to make Dory move like a real fish.
    c. Stocker draws one picture and a computer program makes it move.
    d. Stocker puts lots of pictures of Dory in order and moves them quickly.
5. According to the last part of the passage, why was working on "Finding Dory" especially exciting for Stocker?
    a. "Finding Nemo" was the first movie Stocker had worked on at Pixar.
    b. "Finding Dory" was the first movie Stocker had ever worked on as an animator.
    c. Stocker and his team spent time at the Monterey Bay Aquarium.
    d. They had to build new characters to look like the old ones.
6. Which of the following statements from the passage BEST demonstrates how difficult it was for the animators to make the characters in "Finding Dory"?
    a. Computers are used to make the models, which are three-dimensional.
    b. The team began by making a model for each character in the movie.
    c. Hank still took two years to build.
    d. Stocker is a big fan of "Nemo". It was important for his family too.

7. According to the passage, animators are artists but they must also study what?
    a. How to make three-dimensional models
    b. How to be voice actors
    c. How to create motion
    d. How to work on a computer program
8. What surprising fact did Stocker learn about an octopus during the visits to the aquarium to study their movement?
    a. An octopus was one of the hardest characters to make.
    b. "They purr like a cat!"
    c. All those legs moving at once is a tricky thing to copy.
    d. An octopus can move all eight legs at the same time.
9. Based on the information presented in the passage, how long did it take Stocker and his team of 70 animators to make "Finding Dory"?
    a. 13 years
    b. 3 ½ years
    c. 7 ½ years
    d. 2 years

10. The main character Dory is described as a forgetful blue

    a. Octopus

    b. Clown fish

    c. Star fish

    d. Tang fish

APPENDIX E

Parent Consent

RESEARCH WITH MINORS – PARENTAL PERMISSION

A. PARENTAL PERMISSION (Parents' Copy)

Primary Investigator(s)	Casey Brasher, Ed.S. School Psychologist, MTSU		**Student** ☒
Contact information	931-486-2291 ext. 2268 or cbrasher1@mauryk12.org		
Department Institution	MTSU Department of Elementary and Special Education		
Faculty Advisor	Amy Elleman, Ph.D., MTSU	Department	Literacy Studies
Study Title	Beyond Screening and Progress Monitoring: An Examination of the Concurrent Validity and Instructional Utility of Three Types of Maze Comprehension Assessments for Fourth-Grade Students		
IRB ID	**17-2007**	**Expiration**	**09/30/2017**

Child's Name (Age <12) (type or print)

The following information is provided to you because your child may qualify to participate in the above identified research study. Please read this disclosure document carefully and feel free to ask any questions before you agree to enroll your child. The researcher must adequately answer all of your questions before your child can be enrolled. The researcher MUST NOT enroll your child without an active consent from you. Also, a copy of this consent document, duly signed by the investigator, must be provided to you for future reference.

Your child's participation in this research study is absolutely voluntary. You or your child can withdraw from this study at any time. In the event new information becomes available that may affect the risks or benefits associated with this research or your willingness to participate in it, you will be notified so that you can make an informed decision whether or not to continue your participation in this study.

For additional information about giving consent or your rights as a participant in this study, please feel free to contact the MTSU Office of Compliance (Tel 615-494-8918 or send your emails to irb_information@mtsu.edu. Please visit www.mtsu.edu/irb for general information and visit http://www.mtsu.edu/irb/FAQ/WorkinWithMinors.php for information on MTSU's policies on research with children

Please read this section and sign Section C if you wish to enroll your child. The researcher will not enroll your child without your physical signature.

1. **Purpose of the study:**
   Your child is being asked to participate in a research study because reading comprehension assessments typically tell teachers which students are struggling with comprehnding what they read but not why. Therefore, the purpose of this study is to obtain general information about how students perform on different types of reading comprehension assessments with the goal of determining which type of assessment provides the most useful instructional information for teachers.

2. **General description of procedures to be followed and approximate duration of the study:**
   The MTSU's classification of this study is

   ☒ ***Educational Tests –*** Study involves either standard or novel education practices which consists educational testing and such studies expose the minors to lower than minimal risk

   ☐ ***Psychological and/or Behavioral Evaluation –*** Although the study may or may not involve educational tests, the specific aim is to probe the child's behavioral ability.

   ☐ ***Physical Evaluation –*** The children will be asked to perform or part-take in physical activities or procedures. Examples of such studies simple physical exercises, medical or clinical intervention, pharmaceutical testing and etc. Due to the nature of these studies, your child may be exposed to more than minimal risk.

   Students with permission to participate will complete three group-administered testing sessions lasting 35 to 50 minutes each. Then, each student will also complete one individual testing session lasting 35 to 45 minutes. The testing sessions will be scheduled with classroom teachers to minimize loss to instructional time and will occur within in a timeframe of one to four weeks.

3. **What are we planning to do to your child in this study?**
   Once parent permission is obtained, students will be asked to participate in the study. Every student will be exposed to three group-administered sessions and one individual session. Multiple reading comprehension assessments will be administered along with tests assessing vocabulary, word reading, decoding, syntax, inference generation, and reading fluency. Group administration will

likely occur in classrooms. Measures that are individually administered will be completed outside the students' classrooms in a quiet room in the school building. Administration of group and individual measures will be audio-recorded to ensure appropriate fidelity and inter-rater agreement on scoring. All examiners are graduate students at Middle Tennessee State University trained to administer reading assessments and experienced working with students in elementary school settings.

4. **What will your child be asked to do in this study?**
   Your child's involvement in this study will not lead to loss of any benefits to which he or she would otherwise be entitled. Students will be asked to put forth their best effort to complete the tests. They will be required to read expository and narrative passages of varying lengths. Students will either circle a word that best fits, answer multiple choice questions, or respond verbally to read words or answer questions. The tasks students will be asked to do are similar to typical reading comprehension instruction and assessment occuring in the classroom.

5. **What are we planning to do with the data collected using your child?**
   All material will be collected each day by the principal investigator and kept safely for confidentiality. No one will have access to the materials except for authorized personnel, if necessary. Such personnel would include teachers, approved graduate assistants from the university, or administrators at the school. Sharing information with school staff would only be to provide useful information to further plan instruction and/or intervention for students in the area of reading comprehension. Further, testing materials will go through a process of de-identification. This means that a code will be assigned to each student. Student names will be taken off of testing materials and replaced with an assigned number.

6. **What are your expected costs, effort and time commitment:**
   There is no cost to you or your child for participating in this study. The time commitment for students will consist of completing the reading comprehension assessments during the course of a typical school day. The activities and tasks resemble what occurs during regular instruction for fourth-grade reading comprehension.

7. **What are the potential discomforts, inconveniences, and/or possible risks that can be reasonably expected as a result of participation in this study:**

For the Child: There are no potential risks from participating in the study. Potential discomfort for some students could result from completing reading comprehension tests; however, the students will be told that their participation is voluntary and that the results of the tests are for reseach purposes and not for assigning them grades. Loss to instructional time will be minimized by working closely with classroom teachers and school administrators.

For you the Parent: There are no potential risks, discomforts, or inconveniences for parents to allow their their child to participate in this study.

8. **How will you or your child be compensated for enrolling in this study?**
There is no compensation for participation in the study.

9. **What are the anticipated benefits from this study?**
The goal of this study is to determine which type of reading comprehension assessment provides the most useful instructional information for teachers. Despite the growing emphasis on comprehension skills, particularly beginning in fourth-grade, the available assessments are not useful for targeting specific skills for instruction. A short-term benefit would be the ability to provide instructional recommendations for fourth-grade teachers based on the information collected.

10. **Are there any alternatives to this study such that you or/and your child could receive the same benefits?**
Participation in testing is necessary to validate new assessments. Multiple assessments are given to better analyze how underlying skills are impacted by various comprehension response formats. No alternative to administering multiple assessments is known for addressing the usefulness of reading comprehension assessments for planning instruction.

11. **Will you or/and your child be compensated for study-related injuries?**
There is no compensation in case of study related injury.

12. **Circumstances under which the Principal Investigator may withdraw your child from study participation:**
There are no known circumstances under which the Principal Investigator would withdraw your child. A range of reading skills is expected.

13. **What happens if you choose to withdraw from study participation?**
If your child does not take part in the study, he or she will continue with regular classroom activities and instruction. Your child's involvement in this study will not lead to any loss of benefits to which he or she is otherwise entitled.

14. **Can you or/and your child stop the participation any time after initially agreeing to give consent/assent?**
If you and your child agree to participate, you or your child are free to end participation at any time.

15. **Contact Information.** If you should have any questions about this research study or possibly injury, please feel free to contact Casey Brasher, Ed.S. by telephone 931-486-2291 ext 2268 or by email cbrasher1@mauryk12.org OR my faculty advisor, Amy Elleman, Ph.D., at amyelleman@mtsu.edu or 615-898-5688.

16. **Confidentiality.** All efforts, within reason, will be made to keep the personal information in your child's research record private but total privacy cannot be promised. Your information may be shared with MTSU or the government, such as the Middle Tennessee State University Institutional Review Board, Federal Government Office for Human Research Protections, *if* you or someone else is in danger or if we are required to do so by law.

Consent obtained by:

____________________ ____

Date Researcher's Signature

____

Researcher's Name and Title

B. Signature Section
(Researchers' Copy)

Primary Investigator(s)	Casey Brasher, Ed.S. School Psychologist, MTSU		**Student** ☒
Contact information	931-486-2291 ext 2268 or cbrasher1@mauryk12.org		
Department Institution	MTSU Department of Elementary and Special Education		
Faculty Advisor	Amy Elleman, Ph.D., MTSU	Department	Literacy Studies
Study Title	Beyond Screening and Progress Monitoring: An Examination of the Concurrent Validity and Instructional Utility of Three Types of Maze Comprehension Assessments for Fourth-Grade Students		
IRB ID	**17-2007**	**Expiration**	**09/30/2017**

Child's Name (Age <12) (type or print)

PARENT SECTION

☐No ☐Yes I have read this informed consent document pertaining to the above identified research

☐No ☐Yes The research procedures to be conducted have been explained to me verbally

☐No ☐Yes I understand each part of the interventions and all my questions have been answered

☐No ☐Yes I am aware of the potential risks of the study

By signing below, I give permission for my child, whose name is identified above, to participate in this study. I understand I can withdraw my child from this study at any time without facing any consequences.

Date Signature of the Parent

Parental Consent obtained by:

Date PI’s Signature PI’s Name & Title

Faculty Verification if the PI is a student:

Date Faculty Signature Print Name & Title

APPENDIX F

Child Assent

RESEARCH WITH MINORS – CHILD ASSENT

Primary Investigator(s)	Casey Brasher, Ed.S., School Psychologist, MTSU		**Student** ☒
Contact information	931-486-2291 ext. 2268 or cbrasher1@mauryk12.org		
Department Institution	MTSU Department of Elementary and Special Education		
Faculty Advisor	Amy Elleman, Ph.D., MTSU	Department	Literacy Studies
Study Title	Beyond Screening and Progress Monitoring: An Examination of the Concurrent Validity and Instructional Utility of Three Types of Maze Comprehension Assessments for Fourth-Grade Students.		
IRB ID	**17-2007**	**Expiration**	**09/30/2017**

Child's Name (Age <12) (type or print)

(The PI or his/her IRB-approved representative must read the following disclosures to the child)

This information is provided to you because your parents/guardians have enrolled you to participate in a research study. Please read this carefully and feel free to ask any questions before you agree to enroll. The person who is speaking to you now must all of your questions before you participate. He/she must also give you a signed copy of this sheet. Your participation in this research study is absolutely voluntary. You can decline anytime and no one will inform your parents. You can withdraw at any time. Please visit http://www.mtsu.edu/irb/FAQ/WorkinWithMinors.php or email irb_information@mtsu.edu for more information.

(The PI or his/her IRB-approved representative must read the following disclosures to the child)

1. **Why are you doing this research?**
   I'm working on a research study to help determine the type of information we can learn from reading tests so that teachers can use that information to improve the way they teach.

2. **What will the researcher do and how long will it take?**
   If you help me with the study, you will take a few tests in class. You will be required to read a few passages and answer questions by circling responses in the group sessions. In the individual session, you will be required to read passages, lists of words, and short sentences and respond verbally.

3. **Do I have to be in this research study and can I stop if I want to?**
   You can stop at any time you want to.

4. **Will anyone know that I am in this research study?**
   We will not tell anyone unless we believe that you or someone may be in danger or we are required by law.

5. **How will this research help me or/and other people?**
   Reading comprehension is a very important skill for students in fourth grade but the tests we have now don't tell us much about what we need to teach to improve comprehension.

6. **Can I do something else instead of this research?**
   If you choose not to participate in this research study, your teacher will have assignments and tasks from the general curriculum.

7. **Who do I talk to if I have questions?**
   If you have questions, you can talk to me or your teacher.

____

____

Date Researcher's Signature Print Name and Title of the Researcher

C. Signature Section
(Researchers' Copy)

Primary Investigator(s)	Casey Brasher, Ed.S., School Psychologist, MTSU		**Student** ☒
Contact information	931-486-2291 ext. 2268 or cbrasher1@mauryk12.org		
Department Institution	MTSU Department of Elementary and Special Education		
Faculty Advisor	Amy Elleman	Department	Literacy Studies
Study Title	Beyond Screening and Progress Monitoring: An Examination of the Concurrent Validity and Instructional Utility of Three Types of Maze Comprehension Assessments for Fourth-Grade Students.		
IRB ID	**17-2007**	**Expiration**	**09/30/2017**

Child's Name (Age <12) (type or print)

CHILD SECTION (7-12 years)

□No □Yes I have read the information sheet

□No □Yes I received a signed copy from the researcher

□No □Yes I understand what I read and what I was told

□No □Yes The researcher answered all my questions

□No □Yes I am aware I can withdraw at any time

______________ ______________________________

Date Signature of the Child (waived if less than 7 years age)

Child Assent Administered by:

______________   ____

Date   PI's Signature   Print Name & Title

Faculty Verification if the PI is a student:

______________   ____

Date   Faculty Signature   Print Name & Title