AN INVESTIGATION OF THE RELIABILITY AND VAILIDITY OF CURRICULUM-BASED MEASUREMENT MAZE PROBES: A COMPARISON OF 1-MINUTE, 2-MINUTE, AND 3-MINUTE TIME FRAMES

A Dissertation

Submitted to

the Temple University Graduate Board

in Partial Fulfillment

of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

by

Israel A. Sarasti

August, 2009

ABSTRACT

Title: An Investigation of the Reliability and Validity of Curriculum-based Measurement
Maze Probes: A Comparison of 1-minute, 2-minute, and 3-minute Time Frames
Candidate's Name: Israel A. Sarasti
Degree: Doctor of Philosophy
Temple University, 2009
Doctoral Advisory Committee Chair: Joseph DuCette, Ph.D.

Prevention science has suggested that universal screening can enhance educational and mental health outcomes in the schools (Greenberg et al., 2003). A three-tier model of prevention has been proposed by Albers, Glover, and Kratochwill (2007) and Brown-Chidsey and Steege (2005) employing universal screening assessments of basic academic skills at Tier-1. Curriculum-based measurement maze (CBM-maze) probes are universal screeners that were developed as measures of reading comprehension. They are characterized as easy to administer, time-efficient, valid, and reliable (Parker, Hasbrouck, & Tindal, 1992). CBM-maze probes are short stories consisting of 400 words where every seventh word is omitted and replaced with three answer choices. Students are given 3-minutes to read the passage silently and select a word from the answer choices that restores the meaning of the story. Maze probes have been utilized as reading comprehension assessments for universal screening (Tier 1) and progress monitoring (Tier 2 and Tier 3; Espin, Deno, Maruyama, & Cohen, 1989; D. Fuchs & Fuchs, 1992).

The current research study was conducted to further extend the research on the reliability and validity of CBM-maze probes. More specifically, it investigated if there were any differences between 1-minute, 2-minute, and 3-minute time frames, alternate form reliability, concurrent validity, and social validity of the maze probes.

Results indicated differences in correct word selections (CWS) between 1-minute, 2-minute, and 3-minute time frames with significant interaction effects noted for the 2-minute maze probe. Alternate form reliability correlation were statistically significant and moderately strong (*r* = .47 to .71). Concurrent validity correlations between the STAR Reading norm referenced test (computer adaptive reading comprehension test) and CBM-maze probes yielded statistically significant and moderate correlations (*r* = .30 to .50). Tabulations of the assessment rating scale indicated that students perceived maze probes as acceptable measures for reading comprehension. Implications for practice, cautions in interpreting the results, and future directions are discussed.

ACKNOWLEDGEMENTS

I would like to acknowledge the following individuals for contributing to my doctoral education at Temple University.

- My family, specifically my mother who, unconditionally supports me in all my efforts
- My aunt Elsa for her spiritual guidance
- The late Christine D. Nield-Capote for guiding my growth as an individual and professional
- My dissertation committee, Dr. Joseph DuCette, Dr. Catherine Schifter, and Dr. James Connell for their support through this process
- Finally, Mrs. Coccia, and the faculty/students at St. Denis school for participating in the study

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Figure Page

CHAPTER 1

INTRODUCTION

Prevention Science

Over one decade ago a national research agenda on the concept of prevention science began to emerge. This new research discipline aimed at preventing or moderating human dysfunctions was a collaborative effort between the fields of medicine, social services, psychology, criminology, education and human development (i.e., biological and social sciences; Coie et al., 1993; Reiss & Price, 1996; Weissberg, Kumpfer, & Seligman, 2003). Prevention science is a systematic effort to prevent and ameliorate maladaptive behaviors while promoting adaptive behaviors in societies across the life-span (Reiss & Price, 1996). It seeks to foster the interplay between science and practice when designing and implementing prevention and intervention programs within an ecological framework (i.e., characteristics of the person and environment and their interplay). A review of prevention research in 1995 by the National Institute of Mental Health (NIMH) suggested that prevention research should look at the developmental processes of risks and protective factors that may influence the development of mental disorders (Reiss & Price, 1996). That is, research should be aimed at “reducing risks and enhancing protective factors” (p. 1112).

Coie et al. (1993) have suggested that national programs developed for prevention research should employ “practical applications in schools, hospitals, playgrounds, homes, clinics, industries and community agencies” (p. 1020). These recommendations have been further supported by Reiss and Price (1996) who called for the many subfields of psychology (i.e., school, clinical, developmental) to contribute to the development of prevention science in partnership with other fields. While theoretical prevention frameworks vary across disciplines, a common thread among them is the three-tier model for preventing negative outcomes (Weissberg, Kumpfer, & Seligman, 2003). Caplan (1964) conceptualized a three-tier prevention model utilizing (1) primary, (2) secondary, and (3) tertiary intervention categories. Contrastingly, the model by the Institute of Medicine (IOM) proposes (a) universal preventive interventions, (b) selective preventive interventions, and (c) indicated preventive interventions (Weissberg, Kumpfer, & Seligman, 2003). Both of these models are aimed at addressing negative outcomes which include physical illness, mental disorders, violence, school failure, and poverty within a three-tier framework.

The phenomenon of resilience is an aspect has been studied within the field of prevention science research. Masten (2001) defines resiliency as "a class of phenomena characterized by good outcomes in spite of serious threats to adaptation or development (p. 228). According to Masten, a variable-focused model of resiliency concentrates on potential qualities of the individual and the environment that may impact at-risk consequences or adversity. In contrast, a person-focused approach to resiliency concentrates on individual characteristics in resilient children in comparison to other groups of children. In the field of education, prevention science research has contributed largely to a variable-focused model of resiliency providing interventions geared towards parenting, bullying, school failure, substance abuse, and adolescent pregnancy (Nation et al., 2003). In a review of effective prevention programs, Nation et al. (2003) found that programs that engaged children and their environmental contexts (i.e., home-school collaborations) were more likely to produce change. Comprehensive programs with sufficient dosage of the intervention that included a skill development focus were common principles found across effective prevention programs. Greenberg et al. (2003) have suggested that universal interventions (i.e., tier one) that do not provide sufficient dosage of the intervention may lack support for children in high-risk situations. Nation et al. (2003) also concluded that programs that were sensitive to the development of negative outcomes and appropriately implemented timed interventions using varied teaching methods evidenced highly successful outcomes.

Universal Screening

Nation et al. (2003), Masten (2001), and Greenberg et al. (2003) have all suggested that early identification combined with prevention and intervention services can enhance educational and mental health outcomes in schools. Currently, there is limited research and information on universal screening techniques for identifying children who may benefit from early intervention services (Albers, Glover, & Kratochwill, 2007). Within a three-tier prevention system, universal screening tools can be utilized as first tier initiatives to screen for children who may be at-risk for academic and behavioral difficulties (Glover & Albers, 2007; Greenberg et al., 2003; Kratochwill, Albers, & Shernoff, 2004). These screening tools can allow for early provision of evidence-based prevention programs and early intervention services delivered through a three-tier system (Albers, Glover, & Kratochwill, 2007).

Recent federal educational policies such as the *No Child Left Behind Act of 2001* have strongly suggested that early identification and prevention programs be implemented to screen for children who may be at-risk for academic difficulties (United States Department of Education, 2001). In the same vein, the United States Public Health service (2000) also promoted for the early identification of mental health problems in children across several settings (i.e., preschool, childcare, education, and health) in an effort to provide early intervention services. Special education legislation (i.e., Individuals with Disabilities Education Improvement Act of 2004; IDEIA) is also in support of early identification, prevention, and early intervention provisions for addressing children's academic and social/emotional needs (Albers, Glover, & Kratochwill, 2007). Inclusively, IDEIA allows for 15% of federal special education funds to be used for prevention activities (*e.g.*, universal screening).

Response to Intervention

Educational practices that are reactive and deliver services when children have "experienced failure and distress" (p. 114) have been termed "wait-to-fail" models (Albers, Glover, & Kratochwill, 2007). Kratochwill, Albers, and Shernoff (2004) have proposed that early identification of behavioral and academic problems in children can lead to the provisions of evidence-based prevention programs in a timely fashion. These programs are delivered within a multi-tiered intervention approach such as the prevention models described in Weissberg, Kumpfer, and Seligman (2003). Response to Intervention (RTI) is one proactive multi-tiered intervention system that utilizes an assessment-intervention model allowing schools to provide sound and effective instructional practices (Albers, Glover, & Kratochwill, 2007; Brown-Chidsey & Steege, 2005; Glover & Albers, 2007). In an RTI model, students at-risk are identified early and are provided with supplemental or specialized instruction in order to minimize the development of academic and/or emotional difficulties. Data-based decisions are made within all tiers enhancing students' instructional programs while providing a dynamic flow between tiers (Brown-Chidsey & Steege, 2005; Coyne, Kame'enui, & Simmons, 2001).

Figure 1. Three tiers of instruction.

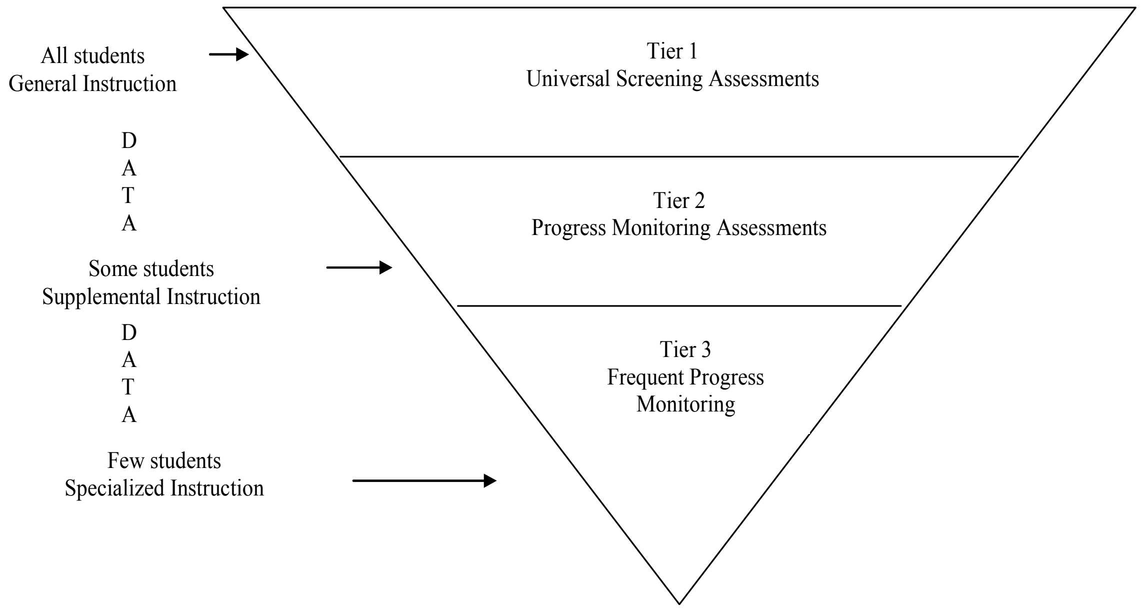


Figure 1 illustrates the three-tier RTI framework within an educational setting (Brown-Chidsey & Steege, 2005). In this model, Tier 1 represents the general student population, while Tier 2 and Tier 3 represent students identified at-risk and students with severe academic/behavioral concerns respectively. Universal screening assessment tools (i.e., curriculum-based measures in reading, math, spelling and writing) are programmed for Tier 1 and are usually administered as benchmark assessments three times a year to identify students at-risk or in need of supplemental instruction. Progress monitoring assessments in the basic skill areas are programmed for Tier 2 (usually bi-weekly) and Tier 3 (usually weekly) respectively in order to assess the effectiveness of the educational programs or interventions implemented. This process allows dynamic movement between Tier 1, Tier 2 and Tier 3 as deficits are identified or remediated (Brown-Chidsey & Steege, 2005; Haager, Klinger, & Vaughn, 2007).

Theoretical RTI models have been proposed for the elementary grades (K-5) by Coyne, Kame'enui, and Simmons (2001), Fuchs, Mock, Morgan and Young (2003), and Kratochwill, Albers, and Shernoff (2004). These models stress the importance of assessing basic skills in the areas of reading, math, spelling, and writing at Tier 1. In the early grades (K-3), the big ideas in reading are monitored using universal screening instruments as suggested by the National Reading Panel (2001). The ideas include phonemic awareness, phonics, fluency, vocabulary, and comprehension (Coyne, Kame'enui, & Simmons, 2001; Haager, Klinger, & Vaughn, 2007).

One RTI model that has been empirically evaluated the implementation of a three-tier process is the System to Enhance Educational Performance (STEEP; for a full account see VanDerHeyden, Witt, & Gilbertson, 2007). The data gathered using a multiple-baseline single-subject design across schools (N=5) from the STEEP implementation research suggested a reduction in special education referrals. Moreover, those referrals that proceeded to an evaluation for special education services were more likely to qualify for services. During the STEEP implementation, academically at-risk students received strategic early interventions versus waiting to qualify for special education services to receive those targeted interventions (VanDerHeyden, Witt, & Gilbertson, 2007). Using a problem solving model (PSM) within a three-tier RTI framework Marston, Reschly, Lau, Muyskens, and Canter (2007) discuss how a response to intervention approach was implemented in the Minneapolis Public Schools using varied research-based instructional programs (Nation et al., 2003), universal screening assessments, and progress monitoring tools (Coyne, Kame'enui, & Simmons, 2001; Deno, Espin, & Fuchs, 2004; D. Fuchs, Mock, Morgan, & Young, 2003). The researchers suggest that this type of PSM fosters the use of evidence-based interventions within each tier and increases the monitoring of student's progress in basic skills across all three tiers. The theoretical and conceptual models of prevention science (Coie et al., 1993), which encompass both the RTI and PSM frameworks of VanDerHeyden, Witt, and Gilbertson (2007), and Marston, Reschly, Lau, Muyskens, & Canter (2007) begin to explore multi-tiered prevention programs with methodological rigor utilizing broad-based universal screening tools.

Statement of the Problem

The need for preventive approaches to education, social welfare, and physical/mental health has been well documented in the literature (see Coie et al., 1993; Reiss & Price, 1996). Specifically, early deficiencies in basic academic skills such as reading which can lead to school failure (National Reading Panel, 2001), school dropout (Nation et al., 2003; National Reading Panel, 2001; Weissberg, Kumpfer, & Seligman, 2003), and the development of emotional/behavioral problems (United States Public Health Service, 2000; Weissberg, Kumpfer, & Seligman, 2003) have been cited as concerns targeted for prevention and intervention services. The proposal for prevention frameworks which include a three-tier approach (Kratochwill, Albers, & Shernoff, 2004; Weissberg, Kumpfer, & Seligman, 2003) seem to have promise in addressing the aforementioned issues. Greenberg et al. (2003) and Kratochwill, Albers, and Shernoff (2004) have suggested that application of universal screening for children in schools can be highly effective in identifying children at-risk for academic and mental health concerns.

Glover and Albers (2007) indicate universal screening tools in educational settings should identify students in need of specific instruction in a "contextually appropriate, technically sound, and usable" (p. 118) manner. Curriculum-based measures (CBM) of basic skills (i.e., reading, spelling, writing, and math) appear to fit these criteria. These measures are described by Deno (1985) as efficient, reliable, and valid indicators of academic competence that can be utilized for screening decisions and gauging individual student progress. In the area of reading, CBM reading (reading a story aloud for 1-minute) has been the approach most widely used for universal screening (Marston, Reschly, Lau, Muyskens, & Canter, 2007; VanDerHeyden, Witt, & Gilbertson, 2007; Wayman, Wallace, Wiley, Ticha, & Espin, 2007). Because reading aloud involves the complex integration of decoding and fluency skills, it can be thought of as an overall indicator of reading competence in the early grades (Deno, Mirkin, & Chiang, 1982; L. S. Fuchs, Fuchs, Hosp, & Jenkins, 2001). Over the past 30 years, the data on CBM-R have indicated this assessment tool is (a) highly efficient in detecting children at-risk for reading failure for grades K-3 (Wayman, Wallace, Wiley, Ticha, & Espin, 2007); (b) a valid and reliable measure to progress monitor reading growth (D. Fuchs & Fuchs, 1986); and (c) can serve as a tool to make eligibility decisions for special education services (D. Fuchs, Mock, Morgan, & Young, 2003; Marston, Reschly, Lau, Muyskens, & Canter, 2007; VanDerHeyden, Witt, & Gilbertson, 2007).

At the upper elementary grades (4-6) students' individual progress and performance on CBM-R have been reported to level off or reach an asymptote making them less sensitive in detecting growth (Shinn, Good, Knutson, & Tilly, 1992; Yovanoff, Duesbery, Alonzo, & Tindal, 2005). Alternate reading comprehension assessment tools such as curriculum-based measurement maze (CBM-maze) procedures have been developed to address this plateau, as well as the face validity of CBM-R as a measure of reading comprehension (D. Fuchs & Fuchs, 1992; Parker, Hasbrouck, & Tindal, 1992). More specifically, the maze procedure involves silently reading a passage or short story. The first sentence in the story is left intact and thereafter every 7th word is replaced with three multiple-choice answers. For this task the students are asked to read the story and circle/mark the correct word replacements within a specified time frame (i.e., 1 minute or 3 minutes). Of the answer choices, two are distracters, and one is the contextually correct word to complete the phrase or sentence. CBM-maze probes have been shown to be valid, reliable and efficient measures of silent reading comprehension for students in the upper grades (Marston, 1989; Shin, Deno, & Espin, 2000).

While there is compelling evidence for CBM-R as a universal screening tool at the early grades, more research needs to be conducted on developmentally appropriate measures that can serve as universal screeners for more complex reading performance. Preliminary research on CBM-maze has some documentation of the tools efficacy in detecting growth levels (Espin, Deno, Maruyama, & Cohen, 1989; Shin, Deno, & Espin, 2000) and differentiating between monolingual students and English language learners at the upper grades (Wiley & Deno, 2005). Currently though, CBM-maze standardized procedures indicate several administration times (1-minute, 2-minute, and 3-minute time frames) are acceptable (Hosp, Hosp, & Howell, 2007). Additionally, the maze probes have varied in presentation formats (i.e., computer assisted versus paper and pencil; D. Fuchs & Fuchs, 1992). These variabilities need to be investigated to see if they significantly impact student performance.

Rationale

Universal screening assessments can vary with respect to target domains (i.e., academic or social/emotional), constructs, format administration, content, recommended frequency and administration time frames (Glover & Albers, 2007). Three general aspects of universal screeners have been suggested by Glover and Albers (2007) as being especially important "(a) the appropriateness for the intended use, (b) their technical adequacy, and (c) their usability" (p. 119). With respect to reading, universal screeners should be sensitive to the developmental nature of the reading process. That is, they should be able to reliably detect growth and screen out students at-risk at every developmental stage (i.e., decoding or comprehension).

Context of the Current Study

Curriculum-based measurement maze probes were developed as an alternate form for assessing general reading comprehension outcomes (D. Fuchs & Fuchs, 1992). Reliability and validity studies for maze probes have been documented to be adequate in the research literature (see Wayman, Wallace, Wiley, Ticha, & Espin, 2007). The presentation formats and timing administration have been varied though. For example, one of the initial maze procedures was administered using an un-timed format (Guthrie, Siefert, Burnham, & Caplan, 1974), others have researched maze procedures using 1-minute time frames (Jenkins & Jewell, 1993; Wiley & Deno, 2005), while yet others have conducted studies where CBM-maze probes were administered using 3-minute time frames (Ardoin et al., 2004; Sarasti, 2007; Shin, Deno, & Espin, 2000). Given the variability in administration times, the technical adequacy of the maze probes can be suspect to yield variable outcomes as a function of the different time frames. The purpose of current investigation is to extend the research on the technical adequacy of CBM-maze procedures taking into account the developmental stages of the reading process. More specifically, it will investigate the reliability (alternate form) and validity (concurrent) of CBM-maze probes by comparing 1-minute, 2-minute, and 3-minute time frames for upper elementary grade students. At this stage students are expected to have reached a mastery level of decoding and fluency on such measures as CBM-R and are now at a stage where they have to read for meaning. An additional aspect of this investigation is capturing the student's acceptability and perceptions of CBM-maze probe as a reading comprehension assessment tool. The social validity aspect of this study supports a unitarian model of validity as proposed by Messick (1995). This investigation will further extend the research literature on CBM-maze procedures as developmentally appropriate measures for universal screening of general reading comprehension.

Research Questions

1. Are there any significant differences in correct word selections (CWS) between 1-minute, 2-minute, and 3-minute CBM-maze probes for 4th and 5th grade students?
2. What are the alternate form reliabilities between 1-minute, 2-minute, and 3-minute maze probes for 4th and 5th grade students?
3. What is the relationship between 1-minute, 2-minute, and 3-minute CBM-maze probes and the STAR Reading computer adaptive comprehension test?
4. What are students' acceptability ratings of the CBM-maze probes?

Definition of Terms

The following is a list of terms used within this experiment.

Basal reader: Textbooks used to teach reading to school aged children, usually at the elementary level. They are anthologies that contain various forms of literature such as short stories, narratives, poems, and original works.

Curriculum-based measurement (CBM): A standardized method for assessing basic skills (i.e., reading math, spelling, and writing).

Curriculum-based measurement maze (CBM-maze): A timed silent reading comprehension fluency measure utilizing a maze procedure.

Formative assessment: The use of frequent or repeated measures to monitor instruction and learning.

General outcome measure: assessment of a broad domain (i.e., reading) using a repeated sampling of performance on a task to assess change in proficiency on that task. The data collected is a dynamic indicator of the broad domain assessed that enable individuals to make predictions about improvements in performance on the task assessed (Deno, Espin, & Fuchs, 2004).

Lexile level: a numeric representation of a reader's ability or text difficulty where a reader can comprehend about 75 percent of what is being read (Lennon & Burdick, 2004).

Maze procedure: a short story or reading passage where every seventh word is replaced by multiple-choice items. One choice is correct and restores the meaning of the sentence or phrase. The other choices are incongruous distractors within the context of the story (D. Fuchs & Fuchs, 1992; Hamilton & Shinn, 2003; Parker, Hasbrouck, & Tindal, 1992).

Maze probe: The individual test developed from a maze procedure (see Appendix A) that is completed as a fluency measure within a specific time frame (D. Fuchs & Fuchs, 1992).

Oral reading fluency (ORF): The rate per minute at which an individual reads words correctly aloud (Daly, Chafouleas, & Skinner, 2005; Hosp, Hosp, & Howell, 2007); ORF is also referred to as Curriculum-based measurement in reading (CBM-R; Wayman, Wallace, Wiley, Ticha, & Espin, 2007).

Reading comprehension: The cognitive process used for understanding text as a whole (Sternberg, 2003); making meaning and sense of text (Palincsar & Brown, 1984).

Response to Intervention (RTI): An assessment and intervention model that integrates high-quality teaching and assessment methods using systematic data-based activities (Brown-Chidsey & Steege, 2005).

Universal screening tools: assessment instruments used to identify individuals in need of specific instruction or services within a prevention oriented education model (Glover & Albers, 2007).

CHAPTER 2

LITERATURE REVIEW

The following literature review discusses three broad areas: developmental reading processes, reading comprehension assessments, and curriculum-based measures. The area of developmental reading is discussed with particular attention to the stages of reading development and the theoretical concepts of reading comprehension. Following, a historical perspective of reading comprehension assessments is presented which includes the varying methods that have been utilized to assess the reading comprehension construct. Next, curriculum-based measures in reading (CBM-R) are discussed within the framework of assessing of basic academic skills (i.e., reading, spelling, writing, and math). The research literature pertaining to the technical adequacy of CBM-R, a universal screening tool is briefly analyzed. Finally, a review of the relevant empirical literature on the maze procedures as a measure of reading comprehension is introduced. This section details the history of the CBM-maze procedure and its current advances. A detailed account of the studies employing varying CBM-maze administration formats (i.e., 1-minute, 2-minute, and 3-minute time frames) is the capstone of the literature review.

The literature search pertaining to CBM-maze procedures was conducted in a systematic manner using inclusionary and exclusionary criteria. First, an electronic search was performed using EBSCO*host*® an online reference system (EBSCO publishing, Ipswich, MA, www.ebscohost.com) to identify empirical studies commencing with the year 1992. This year was chosen because several important empirical studies on the maze procedure were published that year to include the technical adequacy, recent developments of the tool (Parker, Hasbrouck, & Tindal, 1992), and the maze procedure as a general outcome measure of reading comprehension (i.e., CBM-maze; D. Fuchs & Fuchs, 1992). The following terms were used in the electronic search: “maze procedure,” “maze probes,” “CBM maze,” “curriculum based measurement maze,” “CBM silent reading,” “multiple choice cloze,” “cloze technique,” and “maze comprehension”. Only the studies that used CBM-maze procedures and addressed time components were included. Lastly, the reference sections of those studies were searched for additional studies to include in the review of the literature.

Developmental Reading Processes

In a society, the process of reading is a foundational tenet which supports literacy, human capital, and the advancement of the society. In essence, the outcome of the reading process is comprehension (Potter & Wamre, 1990), but at a more complex level, the process encompasses analyzing, synthesizing, and creating new knowledge (Haskell, 2001; Pearson & Hamm, 2005). Theoretical models of the processes underlying the development of reading have been espoused by varying research disciplines such as educational psychology (Chall, 1996) and cognitive psychology (LaBerge & Samuels, 1974). These models suggest reading is composed of hierarchical components that build upon each other (Potter & Wamre, 1990). Hoover and Gough (1990) have suggested that reading is a combination of decoding and comprehension and is the byproduct of these sub-skills.

LaBerge and Samuels (1974) theorized that reading has foundational processes that are automatic. In their conceptualization, they suggest that decoding has a fluency component which is a combination of reading speed and accuracy. When a reader's attentional resources are consumed by decoding and fluency, fewer resources can be dedicated to comprehending what was read (LaBerge & Samuels, 1974). Moreover, LaBerge and Samuels suggested that the brain has limited attentional capacities, and thus, when cognitive processes (i.e., decoding with fluency) are taxed due to a lack of fluency a comprehension breakdown occurs. This breakdown can cause superficial comprehension and gaps in understanding. In contrast, when decoding abilities are mastered at a fluent level, the cognitive resources and attentional capacity can be primarily devoted to the process of comprehending.

The process of reading has been conceptualized in stages by Chall (1996) as "a scheme for studying and understanding the course of reading development from its beginning to its most advanced form" (p. 9). The conceptualization of the reading stages utilizes some of the following assumptions: (a) they have qualitative characteristics and are hierarchical in nature; (b) progression through the stages is a function of interacting with the environment; (c) in successive stages reading becomes more complex and abstract; (d) the reading process can be influenced by motivation, content, ideas, and values. A presentation of the developmental reading stages and processes can be seen in Table 1. The skills for each stage are subsumed by each successive stage. That is, in order to make thorough meaning of material read at stage 3, the decoding aspect of stage 1 has to be in place (Chall, 1989, 1996) much like LaBerge and Samuel's (1974) theory of automaticity. An analysis of reading material throughout the stages indicates text length, vocabulary, and sentence structure become more complex as one progresses through the stages (Chall, 1996). Finally moving through the stages can be thought of as problem solving where assimilations and accommodation are made to transfer and generate new knowledge (Haskell, 2001).

Table 1

Chall's (1996) Stages of Reading Development

Stages	decoding or meaning	oral or silent reading	Examples of material
Stage 0 Birth to age 6	meaning	oral	
Stage 1 Grades 1-2	decoding (meaning)	oral	"may I go?" said John
Stage 2 Grades 2-3	decoding (meaning)	oral/silent	Albert was a goldfish in a bowl.
Stage 3 Grades 4-8	meaning (decoding)	silent	Alex loved to visit his Great Aunt Heidi because she had a library filled with books. The library's shelves held books on every subject.
Stage 4 Grades 9-12	meaning	silent	Two of the most severe droughts of the millennium may have triggered the mass starvation at America's first English settlement at Jamestown. They also sealed the fate of the 120 inhabitants of the "Lost Colony" on Roanoke Island.
Stage 5 College to Adulthood	meaning	silent	Discussions and debates about inclusion frequently have revolved around philosophical issues, at times to the exclusion of the reality of day-to-day considerations, such as learning and behavioral considerations.

In a simple view of reading Gaugh and Tumner (1986) proposed that reading in its simplest form “is the product of decoding and listening comprehension or R = D x C” (p. 7). Inclusively, they state that comprehension is a larger concept of linguistic comprehension where words, sentences, and discourse is interpreted (Gough & Tunmer, 1986). Therefore, reading ability should be able to be predicted from word reading and listening comprehension. This theory asserts that failure at one component (i.e., decoding) will breakdown the reading process. Savage (2001) has investigated this model and has questioned if verbal abilities play a role in overall reading when compared to listening comprehension. The data gathered from his investigation indicated that listening comprehension predicted overall reading best followed by decoding and reading accuracy (Savage, 2001).

The theoretical models presented on developmental reading suggest that sub-skills in reading play an important role in overall reading abilities. The National Reading Panel has suggested five basic ideas in reading (phonemic awareness, phonics, fluency, vocabulary, and comprehension) that are in line with the models presented. It seems that from the simplest model proposed by Gough and Tunmer (1986), and LaBerge and Samuels (1974), to Chall’s (1996) stages of reading development, fluency plays an important part in developing efficient comprehension skills (Stahl & Heubach, 2006). Further analysis suggests having well developed abilities at one stage is a prerequisite for the next stage (Chall, 1996). It is noted though, that fluency in a sub-skill (i.e., decoding) at an early stage (1-3) may reach an asymptotic level (Yovanoff, Duesbery, Alonzo, & Tindal, 2005). Staying at this asymptotic level may hinder reading development if skills are not bridged to the successive stage (Stahl & Heubach, 2006). For example, having high levels of oral reading fluency (ORF; stage 2), a pre-requisite for other stages, and not having exposure to explicit comprehension instruction (stage 3) may develop gaps in the reading process (Pearson & Dole, 1987; Potter & Wamre, 1990). The assessment tools we use to measure reading at any of the stages should reflect sensitivity to the developmental and qualitative aspects of that stage and should capture the skills required of that stage efficiently.

Historical Perspectives on Reading Comprehension Assessments

The formal study of reading comprehension assessment has been significant in the field of education for over 80 years (Pearson & Hamm, 2005). Although it has been a phenomenon of the 20th century, reading comprehension assessment was a practice seen in classrooms within American education prior to its formal study at the turn of the century. The construct of reading comprehension has also been influenced by social and political values that have guided the development of varying assessment tools researched and utilized in education. The development of reading comprehension assessments can be categorized into three historical periods, (1) the beginning, (2) the revolution, and (3) current initiatives (Pearson & Hamm, 2005).

The Beginning

One of the early attempts to measure reading ability through comprehension was attempted by Simon Binet in France (Pearson & Hamm, 2005). In his intelligence test of 1895, Binet used simple incomplete sentences to measure reading ability by having subjects supply the missing words to the sentences (Sattler, 2001). In the United States, political and economical endeavors such as compulsory education played a role in the development of reading assessments. Due to these endeavors, teachers were faced with instructing students with varying degrees of literacy. They soon realized there was a need for “cheap and efficient screening tools” (i.e., universal screeners, CBM) “to determine students’ level of literacy” (Pearson & Hamm, 2005, p. 16). Other impetuses at this time period were the field of psychology and behaviorism which undoubtedly influenced the area of reading through research, objectivity, and the application of the scientific method to the field.

The first published reading assessment was introduced in 1914 as an oral reading test by William S. Gray (Pearson & Hamm, 2005). This assessment tool had to be administered individually and was not efficient for teachers. To address this concern within the emerging behavioral theories of the period, Kelly introduced the Kansas Silent Reading Tests in 1916. This task required students to complete a series of diverse tasks (i.e., fill-in-the blank, following directions) in a five minute time frame. Further developments in reading comprehension were launched by educational psychologist E. L. Thorndike circa 1917. He referred to the act of reading as “reasoning” suggesting reading had many factors. This was probably an early indication to the concept of metacognitve processes (Pearson & Hamm, 2005; Sternberg, 2003).

The advancement of psychometric theory has also played an important role in defining reading comprehension assessments. Frederick Davis conducted several studies between the 1930’s and the 1960’s using factor analytic techniques to investigate if reading comprehension was a unitary construct or was comprised of distinct components

(Davis, 1968; Pearson & Hamm, 2005). These studies were composed of assessments containing multiple-choice tests questions measuring distinct skills (i.e., finding details, main idea, drawing inferences). Initially, he concluded that reading comprehension consisted of two major factors “word knowledge” and “reasoning about what we read” (Davis, 1972, p. 674). In a study conducted by Davis (1972) using multiple regression techniques, he again concluded that reading comprehension was not a unitary factor. Table 1 displays Davis’ Eight Potential Factors. Davis’ research yielded that “remembering word meaning” explained 32% of the variance followed by “drawing inferences” (20% of the variance). The other factors contributed considerably less and failed to meet statistical significance.

Table 2

Davis’ Eight Potential Factors (Pearson & Hamm, 2005, p. 22)

Remembering word meaning	Drawing inferences from the content
Word meaning in context	Recognizing the author’s tone, mood and purpose
Understanding content stated explicitly	Recognizing literacy techniques
Weaving together ideas in the content	Following the structure of the content

The cloze technique developed by Wilson Taylor was introduced as an alternative to multiple-choice items that were being researched by Davis. This assessment technique was created by Taylor to reduce subjectivity (Pearson & Hamm, 2005). In this assessment, every 5th word was deleted from a passage, and the examinee was asked to fill in the omitted word in the blank. The exact replacement was scored as a correct response. Twenty years later, the movement of mastery learning introduced criterion-referenced tests (CRT). This type of assessment was heavily utilized in schools between the 1970's and 1980's. In multiple-choice CRTs mastery of one of several sub-skills (i.e., main idea) with a minimum of 80% accuracy was the goal. The concept of CRT was in line with Davis' theory that reading comprehension was composed of several constructs (Davis, 1972; Pearson & Hamm, 2005).

The Revolution

The revolutionary period in reading comprehension assessment was driven by cognitive psychology, sociolinguists, and literary theorists. The emergence of cognitive psychology and the constructs of intention, motivation, and metacognition were clashing perspectives with the previous ideas of behaviorism; the ideas quickly impacted reading assessments in the classroom (Pearson & Hamm, 2005). The new assessments reflected resources such as prior knowledge (Sternberg, 2003), and environmental cues (Dougherty Stahl & McKenna, 2006). Open-ended questions were introduced to assess comprehension, as well as the concept that there could be more than one correct response when assessing using multiple-choice formats. Pearson and Hamm (2005) cite that classroom assessment tools that were in favor during this period (1970-1980's) included the retell procedure aimed at measuring the depth of a student's understanding of text, and the use of think-aloud protocols used to understand a student's processing of the information.

The sociolinguists and the literary theorists proposed that a text does not have a true meaning and proposed the idea of "Reader Response". Inclusively, they suggested that “meaning is created in the transaction between the reader and the text” (Pearson & Hamm, 2005, p. 41). The major reading comprehension assessment tool that was derived from this concept was the California Learning Assessment System (CLAS) which emphasized response to literature formats and the social aspects of learning. This type of format required students to summarize, explain, justify, and interpret evidence in their answers. Students were also required to collaborate in groups and derive a collective response. The CLAS did not have a long life due to legislative mandates in California, but the assessment format has continued to influence reading comprehension assessment (Pearson & Hamm, 2005).

Current Initiatives

The National Assessment of Educational Progress (NAEP) was signed into legislature in the 1960’s by congress to assess the progress of American education. This was a form of monitoring student achievement by the federal government, as well as a mean for comparing how states compared with each other in educating students (Conley, 2003). The reading subtest of the NAEP was intended to have open-ended constructed responses and addressed the following domains: (a) forming initial understanding, (b) developing interpretations, (c) personal actions and response, and (d) demonstrate a critical stance. In essence, the NAEP Reading was designed to assess how American students negotiated complex interactions with the text (National Center for Educational Statistics, 2005; Pearson & Hamm, 2005).

Another important development in the last decade has been linking comprehension assessment to book reading levels (Pearson & Hamm, 2005). One such attempt is that of the Lexile scales. Basically, a Lexile places a student's comprehension score on the same scale as the reading material (Stenner, Smith, & Burdick, 1983). Additionally, the Lexile level or score can guide students to books they ought to be able to read. That is, the readability level (word frequency and sentence length) and comprehension requirement (hypothetically answer 75% of comprehension questions about the text correctly) are matched to the student's score ("The Lexile framework for reading", n.d.; Stenner, Smith, & Burdick, 1983).

The field of reading has seen diverse types of assessments to measure comprehension in the last century. Interestingly, some of the early assessment tools (i.e., oral reading fluency, oral retelling, multiple-choice) continue to re-emerge across the differing theoretical conceptualizations (i.e., behaviorism and constructivism) of reading. Pearson and Hamm (2005) indicate that reading comprehension is the ability to integrate the resources at our disposal to make sense of the text being read. This process is probably best captured by the NAEP reading test in its progressive thinking of assessment practices.

Curriculum-based Measures

Reading assessments have also been influenced by curriculum-based measures which have roots in behavioral psychology (Lovitt, 1967) and the 1980's movement to formatively assess student's academic progress using curriculum materials (Blankenship, 1985; Gickling & Thompson, 1985). CBM falls under the general domain of curriculum-based assessment (CBA) which tries to tie assessment directly to the local curriculum (Tucker, 1985). Blankenship (1985) described CBA as having the utility to place students in the curriculum, adjust instruction based on their performance, and evaluate/communicate student progress. This model of CBA can be compared to a criterion-referenced test where mastery of instructional objectives is evaluated. In defining CBAs, Gickling and Thompson (1985) proposed a model based on behavioral principles (i.e., task analysis) where students could be assessed using repeated measures. In general, CBA's can be categorized under the mastery measurement model (L. S. Fuchs & Deno, 1991; L. S. Fuchs & Fuchs, 1999) which proposes breaking down a concept into sub-skills and teaching those sub-skills to mastery (i.e., criterion of 80% correct) to gain an understanding of the overall skill (Deno, Espin, & Fuchs, 2004).

In contrast to CBA mastery measurement models, curriculum-based measurement (CBM) adds a fluency component to the assessment process. CBM utilizes the local curriculum as well, but also incorporates standardized administration and scoring procedures integrating measurement theory and classroom-based observational methodology (L. S. Fuchs & Fuchs, 2000, p. 170). Additionally, CBMs are a form of barometric reading of a student's global level of competence in a basic skill domain (i.e., reading; Deno, 1985).

Curriculum-based measurement

One of the basic tenets of curriculum-based measurement is to "decrease the separation between measurement and instruction" (Deno, 1985, p. 221). These measures seek to enhance teacher decision making based on student achievement data enabling teachers to improve student performance (Deno, 1992). Curriculum-based measures were generated from the research agenda conducted by Deno and Mirkin (i.e., Data-Based

Program Modification, 1997) at the Institute for Research on Learning Disabilities sponsored by the University of Minnesota (Marston, 1989). This research sought to find measures that met the following characteristics: (1) were tied to the students' curricula, (2) were short in duration to facilitate frequent administration, (3) had multiple forms for repeated measurement that were psychometrically sound, (4) were inexpensive to produce, (5) were easily communicated and understood by parents, teacher, and administrators (Deno, 1985), and (6) were sensitive in detecting student achievement over time (Deno, 1992; Marston, 1989). The results of this research identified several assessment tools that were valid and reliable to progress monitor student achievement in the basic skill areas of reading, spelling, writing, and math (for an extended account of these studies see Hosp, Hosp, & Howell, 2007; Marston, 1989).

Curriculum-based measurement reading. The curriculum-based measure that has been researched most widely is that of reading. Extended reviews of the efficacy studies on curriculum-based measurement reading (CBM-R) have been conducted by Madelaine & Wheldall (2004) and Wayman, Wallace, Wiley, Ticha, and Espin (2007). CBM-R was first identified and validated by Deno, Mirkin, and Chiang (1982). More specifically, the researchers identified several measures (i.e., reading aloud stories from a basal reader, reading aloud a lists of randomly selected words) that had the potential to meet the criteria for curriculum-based measures. The measure that proved to have the highest correlation coefficients with standardized norm-referenced criterion referenced tests of reading was having the student read aloud from their basal reader for 1 minute (Deno, Mirkin, & Chiang, 1982; Marston, 1989). The correlations ranged from .73 to .91 with most above .80. In terms of reliability, test-retest (.82 to .97), alternate form (.84 to .96), and interrater agreement (.99) proved to meet acceptable psychometric standards as well (Hosp, Hosp, & Howell, 2007; Wayman, Wallace, Wiley, Ticha, & Espin, 2007).

The technical adequacy (i.e., validity and reliability) of CBM-R has produced significant research to date. Studies have compared CBM-R to curriculum-based measures of vocabulary (Espin & Foegen, 1996), word identification fluency (Hosp & Fuchs, 2005), and high-stakes accountability tests (Ardoin et al., 2004; Hintze & Silberglitt, 2005; Wiley & Deno, 2005) and all have produced validity coefficients in the moderate range (.50 to .70). Others have investigated the abilities of CBM-R to differentiate between word callers (Hamilton & Shinn, 2003), potential differences among ethnicities (Hintze, Callahan, Mathews, Williams, & Tobin, 2002; Wiley & Deno, 2005), and students in regular education and special education (Deno, Fuchs, Marston, & Shin, 2001; Espin, Deno, Maruyama, & Cohen, 1989; Shinn, 1988). Inclusively, CBM-R has been compared to developmental reading models (Potter & Wamre, 1990), and its relation to decoding, fluency, and comprehension via factor analysis (Shinn, Good, Knutson, & Tilly, 1992). These studies have yielded that CBM-R is highly predictive of overall reading competence in the early grades (K-3), but has less predictive magnitude in the upper elementary grades (4-6). Overall, CBM-R (oral reading fluency) has a body of literature to support its efficacy as an assessment tool for general reading competence (L. S. Fuchs & Deno, 1991) and as a universal screening instrument to identify students at-risk for a reading disability (Deno, 2003; D. Fuchs, Mock, Morgan, & Young, 2003; Glover & Albers, 2007).

Maze Procedure

The maze procedure was developed in the early 1970's to improve on the cloze procedure that was identified roughly 20 years prior as a measure of assessing reading comprehension (Parker, Hasbrouck, & Tindal, 1992). Basically, the maze procedure is a multiple choice variation of the cloze procedure where students choose the correct answer from a choice of answers; the contrasting cloze uses a fill-in-the-blank format. Three early versions of the maze procedure were researched by Kingston and Weaver (1970), Guthrie Seifert, Burnham, and Caplan (1974), and Cranney (1972-73). These studies served as platforms for the development of this tool as a reading comprehension measure.

Several aspects in the construction of maze procedures have been studied according to Parker, Hasbrouck, and Tindal (1992). They include (a) passage selection, (b) deletion ration, (c) distractor selection, (d) number of distractors, (e) placement of options, and (f) test administration and scoring. The reading passages used to construct maze probes have come from varying sources such as student reading basals, or generic reading passages utilized as CBM-R probes (Espin, Deno, Maruyama, & Cohen, 1989; Howe & Shinn, 2002). They can range from 60 to 400 words in length for the elementary grades. Contrastingly, Cranney's (1972-73) development of the maze procedure for college students utilized 1,500 word passages for each probe with deletions approximately every 46 words. With respect to maze deletion ratios (nth word deleted from passage), a range of 1/5 to 1/46 have been cited in the literature (Parker, Hasbrouck, & Tindal, 1992). The deletion ratio of 1/7 seems to be the most widely accepted because it provides enough of a context for the reader to make meaning of the passage. The distractor type and number of distractors used to construct maze probes have included words from word lists, words deleted from the maze probe text (Kingston & Weaver, 1970), and syntactic (part of speech) and semantic (meaning) relationship of distractors (Guthrie, Siefert, Burnham, & Caplan, 1974). L. S. Fuchs, Fuchs, Hamlett, and Ferguson (1992) have suggested the following parameters for creating distractors for maze probes:

> "the first sentence is intact; thereafter every seventh word is deleted and replaced with three choices. Only one choice is semantically correct. Distractors are not auditorily or graphically similar to the correct replacement; they are either the same length or within one letter of the correct replacement."(p. 48)

The initial maze probes produced varied page layouts to include a chunk of text followed by a word bank (Kingston & Weaver, 1970), answer choices listed to the right of the maze text (Cranney, 1972-73), and embedded responses listed within the text (Guthrie, Siefert, Burnham, & Caplan, 1974). None of the researchers reported apparent confusion by the students when completing the maze probes due to placement options (for a visual representation of placement options see Parker, Hasbrouck, & Tindal, 1992). Finally, the administration procedures of the maze probes have included timed formats (Espin, Deno, Maruyama, & Cohen, 1989; Jenkins & Jewell, 1993) and un-timed formats (Guthrie, Siefert, Burnham, & Caplan, 1974). L. S. Fuchs, Fuchs, Hamlett, & Ferguson (1992) have included a time limit to control for fluency abilities and ceiling effects.

Early studies on the technical adequacy of the maze procedure have yielded moderate reliability and validity coefficients. Internal consistency, test-retest, and alternate form reliability for the studies conducted by Kingston & Weaver (1970) and Guthrie (1974) resulted in coefficients ranging from .79 to .97 across both studies. Concurrent validity studies with standardized reading tests (i.e., Gates-MacGinitie, Stanford Reading Comprehension subtest, California Achievement Test reading subtests) have reported modest to high coefficients (.48 to .85) depending on the comparison measure. An investigation into the validity of the maze procedure has yielded low to moderate correlations (.19 to .53) with teacher rankings of student reading ability (Parker, Hasbrouck, & Tindal, 1992).

Recent advances on the maze procedure have produced research investigating test format, passage selection, and passage difficulty. Most notably, the research conducted by L. S. Fuchs, Fuchs, Hamlett, and Ferguson (1992) and D. Fuchs and Fuchs (1992) has addressed the feasibility of administration by incorporating technology with the maze procedure. Through the use of a computer program the researchers devised an individually administered maze probe with immediate scoring and feedback for the student and teacher. The computer program was used as a progress monitoring tool with a graphical representation of the student's progress (for an example see L. S. Fuchs, Fuchs, Hamlett, & Ferguson, 1992). For this specific maze procedure, students were allotted 2.5 minutes to complete the probe; the probe was scored by counting correct word replacements.

Brown-Chidsey, Davis, and Maya (2003) have researched the sensitivity of maze probes to differentiate among developmental reading levels (i.e., grade levels) and students receiving special education. In their study, the students (grades 5-8) had 10 minutes to complete a 250 word passage. The results of the repeated measures analysis of variance (ANOVA) yielded that all the passages were able to discriminate students based on grade level and special education status. Some interesting details of this study were (a) the time factor for a short passage, (b) the change in term from CBM maze probe to CBM silent reading, and (c) the test format where the multiple choice answers were placed below a blank underlined space which did not conform to the format suggested by L. S. Fuchs, Fuchs, Hamlett, & Ferguson (1992).

One of the most recent studies on the maze procedure was conducted by Twyman and Tindal (2007). The researchers adapted a traditional maze probe to accommodate for higher order comprehension and to address the need for a more complex general outcome measure of reading comprehension at the middle school level. Three maze probes were developed: (1) a traditional maze with four one-word distractors, (2) a concept maze (examples) with four one-word or text phrases as distractors varying in difficulty, and (3) a concept maze (attributes) with three multiple choice selection based on inferences from the text (Twyman & Tindal, 2007). The sample consisted of 240 middle school students and the test was administered in an un-timed format. The results yielded that the concept maze (attributes) was more challenging and stable than the other two maze probes (traditional and examples). Interestingly, students scored higher and with more variability on the traditional maze probe. The authors note that traditional maze probes (i.e., L. S. Fuchs, Fuchs, Hamlett, & Ferguson, 1992) may be valid indicators for general reading comprehension skill, but may not tap into more advanced construction of knowledge and conceptual learning required of students in middle school (Chall, 1996; Twyman & Tindal, 2007).

Curriculum-based measurement maze probes (CBM-maze)

While some of the early versions of the maze procedure were un-timed, the literature base on the maze has supported a timed approach such as a maze probe (i.e., 1-minute, 3-minute time frames). The timed fluency measure reduces distribution skewness, increases validity coefficients, and allows for better discrimination among student performance (L. S. Fuchs, Fuchs, Hamlett, & Ferguson, 1992; Parker, Hasbrouck, & Tindal, 1992). Hosp, Hosp, and Howell (2007) have suggested that either the 1-minute or 3-minute probes appear to be technically adequate measures, but the research investigating the differences in time frames is limited. Table 2 provides a list of empirical studies utilizing CBM-maze probes as outcome measures with varying time frames.

1-minute CBM-maze probes. The Basic Academic Skills Sample (BASS) was developed as a group administered screening tool of basic skills (Deno, Maruyama, Espin, & Cohen, 1989). It included math computation probes, spelling lists, and maze probes. The maze probes (approximately at a second grade reading level) were administered 3 times using 1-minute time frames much like CBM-R procedures. A report on the technical adequacy of the BASS maze probes yielded correlations ranging from .77 to .86 with CBM-R measures for a random sample of students in grades 3, 4, and 5 (Espin, Deno, Maruyama, & Cohen, 1989). Inclusively, the data reported by the researchers indicated a stable pattern of growth in maze scores from grades 1 to 6 between winter and spring administrations (Wayman, Wallace, Wiley, Ticha, & Espin, 2007).

Table 3

List of Varying Time Frames Used for CBM-maze probes

	Sample			Technical Adequacy		
Study	N	Grade(s)	Time (min)	Validity	Reliability	Growth
Espin, Deno, Maruyama, & Cohen (1989)	2,604	1-6	1	✓		✓
Jenkins & Jewell (1993)	335	2-6	1	✓		✓
Wiley & Deno (2005)	36/33	3/5	1	✓		
Fuchs & Fuchs (1992)	63	4 5	2.5			✓
Fuchs, Fuchs, Hamlett, & Ferguson (1992)	63	2-8	2.5			✓
Espin & Foegen (1996)	176	6-8	2	✓		
Markell & Deno (1997)	42	3	2	✓		
Allinder, Dunse, Brunken, Obermiller-Krolikowski (2001)	50	7	2.5	✓		✓
Hamilton & Shinn (2003)	66	3	2	✓		
Brown-Chidsey, Johnson, & Fernstrom (2005)	21	5	2	✓		✓
Shin, Deno, & Espin (2000)	43	2	3		✓	✓
Ardoin et al. (2004)	75	3	3	✓	✓	

The use of “formative teaching, or adapting instruction to students’ current knowledge base, represents one of the most basic tenets of effective instruction” (Jenkins & Jewell, 1993, p. 421). Repeated measures of CBM-R and CBM-maze probes were collected and compared to standardized reading achievement measures (*Gates-MacGinitie Reading Test* and *Metropolitan Achievement* Tests; MAT) to assess for criterion validity. A sample of 335 students in grades 2-6 were administered three CBM-R and three maze probes. The results for the maze probe comparison with the Gates-MacGinitie and MAT comprehension measures yielded correlation coefficients between .65 to .76, and .60 to .74 respectively (Jenkins & Jewell, 1993). Additionally, correlations between CBM-R and the standardized reading comprehension measures ranged from .60 to .88 across the varying reading subtests. The researchers reported that the correlations in grades 2 through 4 tended to be stronger with CBM-R, and dropped in magnitude for grades 5 and 6. An analysis of the maze probe correlation with the standardized comprehension tests depicted consistency in growth pattern and magnitude across grade levels.

Wiley and Deno (2005) have investigated the predictive validity of both CBM-R and maze probes with state mandated accountability measures (i.e., *Minnesota Comprehensive Assessme*nt; MCA). A variation to their research included 33 third graders and 36 fifth graders of whom 15 and 14 were English language learners (ELL) respectively. Pearson correlation coefficients for grades 3 and 5 between the maze probe and the MCA were significant and at.70 or above for native English speakers, but in the .50s for ELLs. In contrast, CBM-R showed a decline in magnitude for both grades and populations (Wiley & Deno, 2005). Additional analysis of the data employing multiple regression techniques provided support for CBM-R and maze probes as predictive measures on the MCA. More specifically, the results yielded that the maze probe appeared to be a better predictor of the MCA at grade 5 than at grade 3 for native English speakers while having significant contributions to performance on the MCA were evidenced for both grade levels. The results for ELLs revealed patterns consistent with the research in lower grades where CBM-R seems to have a stronger magnitude with overall reading competence. This may be due to the fact that the students are in the process of acquiring basic reading skills (i.e., decoding and fluency) and may not be able to integrate vocabulary and prior knowledge into the reading process efficiently (Pearson & Hamm, 2005; Stahl & Fairbanks, 2004).

The use of 1-minute time frames in the administration of maze probes by Espin, Deno, Maruyama, and Cohen (1989) , Jenkins and Jewell (1993) and Wiley and Deno (2005) reflect the standardized administration procedures described by Deno (1985, 1992) and mirror that of the more established general outcome measure of CBM-R. Interestingly, the collective results of these studies support developmental reading models that suggest decoding, fluency, and comprehension are separate abilities which become more complex as a function of grade level (L. S. Fuchs, Fuchs, Hosp, & Jenkins, 2001; Hoover & Gough, 1990; Shinn, Good, Knutson, & Tilly, 1992).

2-minute CBM-maze probe variations. As indicated by Table 2 the majority of studies employing CBM-maze probes utilized a variant of the 2 minute time frame (2 or 2.5 minutes). D. Fuchs & Fuchs (1992) investigated four reading comprehension measures: (1) question answering tests, (2) recall procedures, (3) cloze techniques, and (4) maze probes which had the potential to be used as progress monitoring tools. In this investigation, the researchers employed a computerized version of the maze probe to monitor reading comprehension growth. The participants were 33 special education teachers of which 22 were in the CBM group and 11 were in the control group. Each teacher chose two students to progress monitor with a final sample consisting of 63 students due to attrition (D. Fuchs & Fuchs, 1992). The students were monitored twice weekly using the computerized CBM-maze probe for 18 weeks. Students had 2.5 minutes to complete the standardized 400 word generic passages. The results indicated that the CBM-maze probes were more sensitive to growth (i.e., slope) than cloze, question answering, and retell procedures. The teachers rated the maze probes with high acceptability and thus, were more apt to design effective instruction based on the maze results (D. Fuchs & Fuchs, 1992).

As part of the same research agenda to investigate the feasibility of technology to progress monitor student achievement, L. S. Fuchs, Fuchs, Hamlett, and Ferguson (1992) conducted a similar study utilizing the computerized CBM-maze probes as one of the dependent measures. The sample consisted of 33 special educators and 63 students that were randomly assigned to either CBM expert system consultation (the computer program indicated when instructional changes needed to be made based on students' performance on the maze probes), CBM with no consultation (no feedback on instructional changes), and a control group. Students were monitored twice weekly for 17 weeks. For both of the CBM groups, student achievement was higher than the control group and the teachers in the expert consultation group made more relevant instructional program modifications (D. Fuchs & Fuchs, 1992). The students' performance on the maze probes was higher for both CBM groups generating an effect size of .16. Other researchers have also employed the computerized CBM-maze probes developed by Fuchs & Fuchs (1992) to measure reading comprehension growth. In a sample of 50 seventh graders enrolled in three remedial reading classes, Allinder, Dunse, Brunken, and Obermiller-Krolikowski (2001) found that instruction on specific reading strategies aimed to increase oral reading fluency yielded higher student performance on maze probes. The researchers found significant differences between the intervention group and the control group (no specific strategy). Specifically, rate of growth (i.e., slope) was significantly higher for the strategy group (.44) than the control group (.09). These studies have utilized CBM-maze probes as progress monitoring tools. The overall results provide evidence that CBM-maze probes are able to detect growth dynamically as a general outcome measure of reading comprehension (Allinder, Dunse, Brunken, & Obermiller-Krolikowski, 2001; D. Fuchs & Fuchs, 1992; L. S. Fuchs, Fuchs, Hamlett, & Ferguson, 1992).

The issue of word-callers has been cited in the literature as concern for teachers (Pearson & Hamm, 2005). Students who are able to read aloud with accuracy but have limited comprehension abilities have been termed "word callers". Hamilton and Shinn (2003) investigated this phenomenon with two groups of third graders, word-callers rated by their teachers and similarly fluent peers. Student performance on varying reading measures (i.e., CBM-R, maze probes, question answering) yielded significantly different and consistently lower scores on CBM-R and comprehension measures for the word-callers. With respect to CBM-maze probes, an analysis of the results provided by Hamilton and Shinn (2003) support the tool as a viable measure in discriminating between similarly fluent peers and those students who may be at-risk for reading failure (word-callers). Some interesting findings noted by the researchers were that the teachers over estimated student performance on all the measures, but conversely were able to correctly identifying students who were word-callers from those who were performing at grade level.

CBM–maze probes have been used as measures to experimentally manipulate text difficulty (Markell & Deno, 1997). In a study involving 42 third graders, Markell and Deno (1997) utilized reading passages from a basal series to investigate the effects of passage difficulty. Three second, forth, and sixth grade level passages were developed as outcome measures according to standardized procedures (Deno, 1985; Deno & Fuchs, 1987). The students (1) read the three passages out loud (CBM-R); (2) completed maze probes composed of the same passages, and immediately after, (3) answered 8 orally presented questions at the three respective grade levels. The results revealed higher performance on CBM-R paralleling higher scores on the maze probes and question answering tasks. Further analysis of the data using discriminant analysis revealed that CBM-R could reliably discriminate comprehenders from non-comprehenders using grade level material (versus below or above grade level). It was noted though, that large increases of CBM-R (10-15 words) were needed to predict higher performance on the comprehension tasks with certainty (Markell & Deno, 1997). The manipulation of CBM-maze passages was also investigated by Brown-Chidsey, Johnson, and Fernstrom (2005) using controlled passages (generic stories) and literature based passages (taken from library books). With a sample of 21 randomly selected fifth grade students, the researchers investigated if there were any differences in correct selections on maze probe passages administered the during Fall, Winter, and Spring. Correlations between passages were reported at .80, .74, and .92 respectively, and significant differences were found between passages (generic were higher) for all three administrations (Brown-Chidsey, Johnson, & Fernstrom, 2005).

Much of the research on CBM measures for general outcome measurement has been conducted at the elementary level. At the middle and high school levels studies investigating measures to progress monitor student growth on content-area tasks has begun to emerge in the research literature (*e.g.*, Espin & Foegen, 1996). The relations between three variables CBM-R, maze probes, and vocabulary matching were compared to question answering, daily tests, and posttests by Espin & Foegen (1996). The sample for this study consisted of 184 urban middle school students of which 13 were identified with mild disabilities. Data analysis conducted with correlational techniques and multiple regressions revealed that vocabulary was the most efficient at predicting student performance on content area tasks (i.e., question answering; Espin & Foegen, 1996). While the regression analysis revealed that vocabulary matching accounted for the largest portion of the variance, the correlations with question answering, daily tests, and posttests were all within a close range of each other (.52 to .65). This suggests that at the middle school level all three general outcome measures (i.e., CBM-R, maze probes, vocabulary matching) appear to be valid predictors of general reading comprehension abilities, but vocabulary matching appears to predict specific content area knowledge with better precision.

3-minute CBM-maze probes. The technical adequacy of CBM-maze probes has been studied and documented with varying populations and different curriculum materials (e.g., Wayman, Wallace, Wiley, Ticha, & Espin, 2007). The rate of growth was initially studied by D. Fuchs and Fuchs (1992) which found that CBM-maze probes have the ability to detect stable growth over time. In an extension of this study, Shin, Deno, & Espin (2000) assessed the growth over one year's time of 43 second graders. The reliability (alternate form), validity, and sensitivity to growth was investigated using correlation techniques and hierarchical linear modeling (HLM). Data were collected monthly using the 3-minute maze probes across the school year. Regarding alternate form reliability, all coefficients were significant at the .01 level and ranged from .69 to .91 with an average correlation of .81 (Shin, Deno, & Espin, 2000). Overall growth rates and individual growth rates proved to be reliable based on Hierarchical Linear Modeling (HLM) analyses. Finally, the relation between growth rates (slope) and scores on the California Achievement Tests reading subtests "revealed a significant positive relation" (p. 168). Contrastingly, the difference between growth rates for regular education student and special education students was not significant.

Ardoin et al. (2004) investigated CBM-R and maze probes as potential universal screening measures. The researchers sought to investigate the predictive validity of one single CBM-R probe versus the median of three CBM-R probes with the *Woodcock-*

Johnson-III (WJ III) Broad Reading Cluster (individually administered) and the Iowa Test of Basic Skills (ITBS; group administered). Additionally, the incremental validity of adding the administration of a CBM-maze probe with the CBM-R probes was investigated (Ardoin et al., 2004). Seventy-seven third grade students were administered the CBM-R probes, maze probes, WJ III, and ITBS measures. Correlations between all measures were statistically significant at the .01 level. The correlations between CBM-R and the WJ III subtests were significantly higher than those for CBM-maze probes and the WJ III. Regression analyses indicated that CBM-R accounted for most of the variance in predicting performance on the WJ III subtests. The researchers concluded that administering one CBM-R probe would be as efficient as administering three probes as suggested in standard CBM procedures by Shinn (1989). Moreover, results yielded that concurrently administering a CBM-maze probe with the CBM-R probes did not significantly add to the regression equation for the population sampled. When these results are framed within developmental reading theories such as Chall's (1996), the higher correlations with CBM-R are not surprising in that third graders are still building fluency competence versus developing the more complex skill of comprehension as measured by maze probes (Paris, 2006).

Integrative Summary

Both developmental reading theories and socio-political developments have impacted the field of reading and in turn the measurement tools used to assess the reading process. While the research seems to suggest that distinct sub-skill in reading (Pearson & Hamm, 2005; Shinn, Good, Knutson, & Tilly, 1992) contribute to the overall reading process, fluency abilities, vocabulary, motivation, life experiences, and cognition can play a role in the breadth and depth of comprehension. The integration of cognitive models (*e.g*., LaBerge & Samuels, 1974) with developmental reading models (Chall, 1996) suggest that fluency at specific stages such as stage 2 (decoding; learning to read) can foster the development of overall reading comprehension (stage 5). Gough and Tunmner (1986) have suggested that fluency on its own can not constitute reading comprehension (i.e., word-callers).

Curriculum-based measurement in reading can be described as an oral reading fluency measure (L. S. Fuchs, Fuchs, Hosp, & Jenkins, 2001). The research investigating the psychometric properties on CBM-R have shown to be strong and valid in measuring students' early reading development (stage 2, learning to read). CBM-R appears to be less dynamic in assessing general reading competence at the upper grade levels (Wiley & Deno, 2005). Additionally, the face validity of CBM-R as a measure of reading comprehension has been questioned by teachers (Parker, Hasbrouck, & Tindal, 1992). The development of the CBM-maze probes as fluency measures of silent reading comprehension appear to be a promising alternative for assessing the more complex skills of the reading process (D. Fuchs & Fuchs, 1992). Further validation of the psychometric properties of CBM-maze probes specifically with respect to administration procedures and time frames is warranted.

CHAPTER 3

METHODS

Research Questions

This investigation addressed the following research questions:

1. Are there any significant differences in correct word selections (CWS) between 1-minute, 2-minute, and 3-minute CBM-maze probes for 4th and 5th grade students?
2. What are the alternate form reliabilities between 1-minute, 2-minute, and 3-minute maze probes for 4th and 5th grade students?
3. What is the relationship between 1-minute, 2-minute, and 3-minute CBM-maze probes and the STAR Reading computer adaptive comprehension test?
4. What are students' acceptability ratings of the CBM-maze probes?

Participants and Setting

A total of 85 students enrolled in 4th and 5th grades attending one elementary school in the metropolitan Philadelphia, PA area served as participants for this investigation. A power analysis using G*Power 3 (Faul, Erdfelder, Lang, & Buchner, 2007) was conducted and indicated that a sample size of 80 students would provide adequate power (.80) for main and interaction effects using a repeated measures analysis of variance (ANOVA). A medium effect size (.25) and an alpha level of .05 was assumed (Hintze, Christ, & Keller, 2002). Permission was obtained from the school principal to conduct the study. Every student in both grades was given a parent informed consent form to take home to their parents, as well as a child assent form. All of the students (N = 85) returned the permission forms signed by their parents giving permission to participate in the study. Each student also provided assent to participate. Thus, every student participated in the study.

Overall, the catholic parochial school served 362 students in pre-kindergarten through 8th grade. The student population was predominantly suburban, middle to upper-middle class and had a racial make-up of 94% Caucasian, 4% African American, 2% Hispanic, and 2% Asian with no students identified as English language learners (ELL) or receiving special education services. Remedial reading and math services were offered and funded through Title I and administered by the county intermediate unit to 5% of the student population in kindergarten through 4th grade. All of the teachers in the school were certified by the state department of education and had an average of 15 years of teaching experience.

The sample for this study consisted of the entire 4th and 5th grades. These grade levels were chosen because students in the 4th and 5th grades are entering Chall's Stage 3 of reading development "Reading for Learning the New" (Chall, 1996), and decoding/fluency abilities (i.e., oral reading fluency) begin to level off at these grade levels (Hintze & Shapiro, 1997; Wayman, Wallace, Wiley, Ticha, & Espin, 2007). Each grade level had two sections and two teachers with approximately 20 students per class. A summary of the participants is provided in Table 4 (Brown-Chidsey, Davis, & Maya, 2003).

Table 4

Participant Information

	N	Percent
Total	85	100
Male	43	50.6
Female	42	49.4
4th grade	46	54.1
Class 1	23	
Class 2	23	
5th grade	39	45.9
Class 1	19	
Class 2	20	
Receiving general education instruction only	82	96.5
Receiving Title 1 remedial reading services	3	3.5

Inferential statistics were performed to investigate the relationship between grade level (4th and 5th), gender (male and female) and type of instruction (general education and remedial). The analysis suggested no significant relationship between grade and gender, χ^2 (1) = .57, p = .45; or grade and type of instruction, χ^2 (1) = 3.67, p = .06. Additionally, a between-groups (class 1, 2, 3, 4) one-way ANOVA revealed no significant differences in semester reading grades across classes F (3, 81) = 1.54, p = .21. Mean reading grades for class 1-4 were 92.26 (3.11), 92.78 (5.39), 95.11 (3.64), 92.20 (6.67) respectively.

Materials

CBM-maze probes

For this investigation, commercially available curriculum-based measurement maze (CBM-maze) probes were utilized (www.AIMSWeb.com). The probes are original stories consisting of 300-word passages for grades one through eight (see appendix A). At each grade level, there are 3 benchmark probes and 20-30 progress monitoring probes. The reliability and validity of the AIMSWeb reading passages was conducted using the oral reading fluency measures passages. The measures were then constructed into CBM-maze probes using standardized procedures (D. Fuchs & Fuchs, 1992). The technical manual reports alternate form reliability ranging from .83 to .90 for the oral reading fluency general outcome measures (GOM; Howe & Shinn, 2002). Criterion and concurrent validity for CBM-maze probes with standardized reading comprehension tests are acceptable (.60 to .83) based on supporting literature (D. Fuchs & Fuchs, 1992; Wayman, Wallace, Wiley, Ticha, & Espin, 2007). The readability levels for the AIMSWeb reading passages were calculated using Lexile levels (for a detailed description see Stenner, Smith, & Burdick, 1983). Briefly, a Lexile provides a common metric for measuring text difficulty and student reading ability where the student can read with moderate success (i.e., 75% comprehension). It has a semantic component (word difficulty) and a syntactic component (length of sentence in a text) which together produce a single Lexile measure for a text ranging from 200L to 1700L (Lennon & Burdick, 2004). Howe and Shinn (2002) report correlations between Lexile levels and other readability formulas (i.e., Dale-Chall, Flesch, Powers-Sumner-Kearl, SMOG, and Spache) ranging from .78 to .98 across grades for CBM-R probes.

For this investigation, 3 probes were chosen at each grade level (4th and 5th) from the 30 progress monitoring probes. The probes selected reflect Lexile reading levels corresponding to the end of the year curriculum; this was done because this investigation was conducted during the last month of the school year. For example, the Lexile levels in 4th grade range from 650L to 850L and correspond to the beginning and end of year expected reading levels respectively. In the 5th grade, the Lexile levels range from 750L to 950L. Table 5 provides the titles and Lexile reading levels for the probes chosen.

Table 5

CBM-maze Probes Chosen for the Investigation

Grade	Title	Lexile Reading Level
4th		
	# 8 It was a fine winter…	750L
	# 16 Mr. Le Sung…	750L
	# 30 A young polar bear…	760L
5th		
	# 12 Maria made beautiful…	870L
	# 20 It was an especially…	880L
	# 8 First street school…	930L

STAR Reading

The STAR reading test is a computer adaptive norm-referenced reading comprehension/reading achievement measure for grades 1-12. It is a web-based program that adapts to the students level of proficiency. The reliability for STAR reading was calculated using test-retest, split-half, alternate form, and the estimation of generic reliability correlations and ranged from .82 to .95 (Renaissance Learning, 2006). Concurrent validity with other reading achievement measures (i.e., Gates-MacGinitie, Stanford Achievement Test 9, Iowa Tests of Basic Skill, and Terra Nova) yielded correlations ranging from .69 to .91 across all grade levels. The test consists of 27 multiple choice cloze items with 4 answer choices. The items include 20 vocabulary in-context questions and 7 authentic text passage questions (see appendix B). Students have an average of one minute to complete each item and are given a 10 second warning before STAR reading automatically moves to the next question. Various scores are generated based on the student's reading ability and include standard scores ranging from 0-1400, normal curve equivalents, grade equivalents, and instructional reading levels respectively (Renaissance Learning, 2006). STAR reading is categorized as a valid assessment by the U.S. Department of Education's National Center of Student Progress Monitoring (www.studentprogress.org).

Procedures

Data Collection

CBM-maze probes. The investigation was conducted over a one week period during the last month of the school year. The data was collected by two school psychologists (the researcher and a colleague from the local intermediate unit) and the school principal. All three data collectors had graduate training in administering and scoring curriculum-based measures (i.e., CBM-maze probes) and had previous experience using this type of assessment procedure. The day prior to data collection of the CBM-maze probes, a brief refresher training and scoring session was conducted by the researcher using the AIMSWeb training workbook (Shinn & Shinn, 2002). A score of 95% or better on the implementation integrity checklist provided in the workbook was required prior to initiation of the study (Hintze, Christ, & Keller, 2002).

One week before data collection, the school principal and the researcher choose a date to conduct the investigation based on availability and previously scheduled activities on the school calendar. The 4th and 5th grade classroom teachers were given that date one week in advance, and informed the students that the investigation would take place the following week on the specified day. Data collection for the CBM-maze probes took place in the morning during the students' language arts block.

Table 6

CBM-maze Probe Randomization Matrix

Group (N)	Counterbalanced CBM-maze Probes		
1 (15)	1-minute	2-minute	3-minute
2 (14)	1-minute	3-minute	2-minute
3 (13)	2-minute	1-minute	3-minute
4 (15)	2-minute	3-minute	1-minute
5 (15)	3-minute	2-minute	1-minute
6 (13)	3-minute	1-minute	2-minute

The morning of the investigation, the students were randomly assigned to one of six groups (see Table 6) by having them count off 1, 2, 3, 4, 5, and 6. These groups represent the six possible combinations of counterbalancing the three CBM-maze probes (1-minute, 2-minute, 3-minute). The administration of the CBM-maze probes was conducted in group format in an empty classroom in the school building. The school principal picked up one group at a time from their class (i.e., group 1; 4th and 5th grades), escorted them to the testing room, and walked the students back when the testing was complete. The researcher and other school psychologist were in the testing room and conducted the CBM-maze probe assessments for each of the six groups. Meanwhile, the classroom teachers remained in their classroom with the students not being tested and followed their daily instructional plans.

In the testing rooms, each student was given a packet containing a maze cover sheet (see Appendix B) and the three CBM-maze probes in the specific order listed in Table 2. For example, all 4th graders received a packet with the Maze cover sheet and probe #8, #16 and #30 in that specific order. The 5th graders also received a packet with the cover sheet and their corresponding grade level CBM-maze probes. The students were then asked to write their name and assigned group number on the maze cover sheet. The CBM-maze practice test and probes were then administered using standardized directions as indicated in the administration manual (Shinn & Shinn, 2002, p. 14).

> When I say ‘Begin’ turn to the first story and start reading silently. When you come to a group of three words, circle the 1 word that makes the most sense. Work as quickly as you can without making mistakes until I say ‘Stop’ or you are all done. Do you have any questions?

The students were instructed to complete only one CBM-maze probe at a time within the different time frames (1-minute, 2-minute, 3-minute). The administrator adhered to each group’s time frame order as indicated in Table 3 by following a treatment integrity script developed for each group (see Appendix D). Additionally, both the researcher and the school psychologist walked around the room monitoring students and making sure the circled only 1 word.

Scoring was conducted on the same day by the two school psychologists after all 6 groups had been administered the three CBM-maze probes. The scoring procedures outlined in the training manual were followed (Shinn & Shinn, 2002). The number of correct word selections (CWS) on each CBM-maze probe served as the datum reported for each student. If a student scored three or more incorrect selections consecutively, scoring was discontinued and only the correct selections before the three incorrect were tallied (D. Fuchs & Fuchs, 1992; Hosp, Hosp, & Howell, 2007).

STAR Reading. The STAR Reading test was administered in the school computer lab during the same week the CBM-maze probes were administered. Each class took the computer-adaptive reading comprehension test during their weekly scheduled computer lab time. The computer lab teacher served as the test administrator for the STAR reading test for both 4th and 5th grade classes. She followed the standardized directions in the software manual/pretest instructions (Renaissance Learning, 2006). The students completed a brief practice test embedded within the software prior to completing the actual reading comprehension test. Each testing session took approximately 25 minutes of computer lab time per class. At the end of each class period (1 hour), the computer lab teacher printed out a "Summary Report" for each class and a "Parent Report" for each student to take home. For the purpose of data analysis, the datum recorded for each student was the standard score achieved.

Interscorer. Interscorer agreement data was calculated for 20% of the CBM-maze probes administered during the investigation. An equal number of probes for each grade and condition (i.e., 1-minute, 2-minute, 3-minute) was be checked for agreement by an independent scorer (school principal). If any scoring discrepancies occurred, the datum recorded by the researcher was used in the analyses. The percentage agreement was calculated by a percentage agreement formula [(agreements/ agreements +

disagreements) x 100] as suggested in House, House & Campbell (1981). Percentage agreement averaged 98% for the CBM-maze probes.

Procedural integrity. To ensure the fidelity of each test administration, a procedural integrity checklist was completed by the test administrator for each of the CBM-maze probe administrations and STAR Reading administrations (Kennedy, 2005). The checklist consisted of the steps to be followed for the CBM-maze administration (see Appendix D). A mark (1 = completed; 0 = not completed) was placed after each step completed or not completed. The number of step completed was calculated per administration. The school principal completed a separate procedural integrity checklist for 33% of the administrations for the CBM-maze probe (2 out of 6 six groups) to cross validate test administration procedures. A separate integrity checklist was completed for 1 session of the STAR Reading test administration (25%) by the school principal as well. Point-by-point agreement was conducted for the steps on both of the checklists and yielded procedural integrity scores of 100%.

Data Analysis

The initial step in data analysis was to screen the data set for outliers, normality and meeting parametric assumptions as they pertain to repeated measures analysis of variance and bivariate correlations (Meyers, Gamst, & Guarino, 2006; Wells & Hintze, 2007). To address research question 1, a repeated measures analysis of variance (ANOVA) was employed to investigate if there were any differences in scores between the different time frames on the CBM-maze probes (Christ, Johnson-Gros, & Hintze, 2005; Ellis, 1999; Hintze, Christ, & Keller, 2002; Hintze, Owen, Shapiro, & Daly, 2000).

The within-subject variables were the 1-minute, 2-minute and 3-minute maze probes, and the between-subject grouping variables were grade level (i.e., 4th and 5th).

Research question 2 was investigated by examining the relationship between the 1-minute, 2-minute and 3-minute CBM-maze probes. The correlations between forms provided alternate-form reliability estimates across time frames (Brennan, 2001; Fishman & Galguera, 2003). A Fisher transformation was utilized to test for significant differences among the correlations (Meng, Rosenthal, & Rubin, 1992).

To address research question 3, which investigated the concurrent validity of the CBM-maze probes, bivariate correlational analyses were conducted between the CBM-maze probes and the STAR reading test. The standard scores on the STAR Reading test and CWS on the maze probes were correlated to identify the magnitude of the relationships between the various time frames of the maze probes to the norm-referenced reading test. Additionally, the correlations attained were also analyzed for statistical significance using a Fisher transformation (Silberglitt, Burns, Madyun, & Lail, 2006).

Social Validity

The social validity (research question 4) of the CBM-maze probes as reading comprehension measures was investigated through an assessment rating scale (see Appendix E) to identify the students' acceptability and perception of the maze probes (Cooper, Heron, & Heward, 1987; Kennedy, 2005). Descriptive statistics were used to analyze the ratings on the assessment rating scale. In order to do this, the "No, Maybe, and Yes" ratings were converted to likert type scores ranging from "1, 2, 3", and the tallies of these scores were utilized to calculate the descriptive statistics (Daly, Chafouleas, & Skinner, 2005; Sarasti, 2007).

CHAPTER 4

RESULTS

Introduction

The purpose of this study was to investigate the technical adequacy of CBM-maze probes using three different time frames for a group of 4th and 5th grade students. Four questions were utilized to explore the reliability and validity of commercially available generic CBM-maze probes (i.e., AIMSWeb probes). Three different aspects in relation to time frames (1-minute, 2-minute, and 3-minute) were investigated: alternate form reliability, concurrent validity, and social validity. This chapter presents the results of the study using each research question as a guiding framework. First, descriptive statistics and preliminary data analysis are presented. Then, each research question and its corresponding results are offered.

Descriptive Statistics

Several steps were implemented to ensure that the original data were accurate and problem free before running any data analyses. Initial data entry into the Statistical Package for the Social Sciences (SPSS) Graduate Pack 16.0 for windows was conducted by the primary researcher. Additionally, the data was also entered into SPSS by the second school psychologist who assisted in data collection and scoring. Code and value cleaning was then implemented by both individuals to check for major errors including out of range scores on both categorical and continuous variables and the legitimacy of the data (Meyers, Gamst, & Guarino, 2006). No missing data were found for either the CBM-maze probes or the STAR reading test, thus yielding a total of 85 students for inclusion in the sample.

Table 7

Descriptive Statistics for Dependent Measures: Maze Probes and STAR Reading Test

	N	Range	CWS Means (*SD*)	Skewness (*SE*)	Kurtosis (*SE*)
4th Grade	46				
1-minute maze		2-12	8.15 (2.60)	-.49 (.35)	-.41 (.69)
2-minute maze		6-22	12.96 (3.85)	.16 (.35)	-.66 (.69)
3-minute maze		9-33	21.22 (5.63)	.02 (.35)	-.69 (.69)
5th Grade	39				
1-minute maze		4-11	7.28 (2.25)	.16 (.38)	-1.18 (.74)
2-minute maze		9-23	15.77 (3.86)	-.09 (.38)	-.80 (.74)
3-minute maze		8-33	22.10 (6.55)	.-.24 (.38)	-.52 (.74)
Combined 4th/5th Grade	85				
1-minute maze		2-12	7.75 (2.46)	-.18 (.26)	-.85 (.52)
2-minute maze		6-23	14.25 (4.08)	.04 (.26)	-.74 (.52)
3-minute maze		8-33	21.62 (6.05)	-.09 (.26)	-.61 (.52)
STAR Reading test	85	431-1218	731 (209)	.75 (.26)	-.42 (.52)

Note. CWS = correct word selections.

Table 7 displays the descriptive statistics of each CBM-maze probe by grade and as a combined sample. Forty-six, 4th graders and thirty-nine, 5th graders make up each grade level respectively. An analysis of the descriptive statistics suggested that distributions of CWS means were overall normally distributed (+/- 1.00 for skewness and kurtosis) across 1-minute, 2-minute, and 3-minute time frames for both grade and the combined sample. One exception was noted for the 5th grade 1-minute maze probe where kurtosis was -1.18; a histogram indicated a rather platykurtic distribution. A visual inspection of Normal Q-Q Plots and Detrend Normal Q-Q Plots by grade and combined sample suggested normal distributions for the three CBM-maze probes as well (Meyers, Gamst, & Guarino, 2006). The data for the STAR Reading test appeared normally distributed when inspected for skewness, kurtosis and plots.

To inspect for univariate outliers, the combined sample for the three maze probes was converted to Z-scores. Any score exceeding +/- 2.58 $p = .01$ was considered for deletion. Z-scores ranged between -2.33 and 2.15 and did not exceed the stipulated cut score. Assessment of multivariate outliers for the three maze probes and STAR reading was computed using the Mahalanobis distance statistic for each case (Meyers, Gamst, & Guarino, 2006; Stevens, 2002). A stringent $p < .001$ criterion indicated no significant multivariate outliers for any of the dependent variables.

A further visual inspection of CWS means and standard deviations suggested a parallel and steady increase in scores for 4th grade, 5th grade, and the combined sample across the three time frames. The means for the 2-minute and 3-minute maze probes generally tended to increase by 2-times as much and 3-times as much their respective 1-

minute probe score (see Table 7). For example, students' means for the combined sample yielded 7 CWS for the 1-minute probe, 14 CWS for the 2-minute probe, and 21 CWS for the 3-minute probe indicating a pattern of x, 2x, and 3x. The same approximate pattern was noted for 5th grade. While the multiplicative pattern deviates to some extent in the 4th grade students' maze probe scores, it is apparent that CWS still increased by an approximate pattern of x, 2x, and 3x.

Based on the analysis of the descriptive statistics, scores for all the dependent measures appeared to be normally distributed and met all necessary parametric assumptions for both repeated measures analysis of variance (ANOVA) and Pearson Product-moment correlations. Scores were all within normal limits requiring no listwise or pairwise deletions. Therefore, for subsequent data analyses, all cases (*N* = 85) for the 1-minute, 2-minute, and 3minute maze probes and their corresponding STAR Reading scores were included and considered acceptable data.

Research Question 1

Are there any significant differences in correct word selections (CWS) between 1-minute, 2-minute, and 3-minute CBM-maze probes for 4th and 5th grade students?

In order to investigate whether there were any significant differences between maze probes time frames, three separate repeated measures ANOVA were conducted. For all cases the three time frames (1-minute, 2-minute, 3-minute) served as the within-subject factor while grade (4th and 5th) and counterbalanced groups (groups 1 – 6) served as between-subject factors. The first analysis employed a 3 (1-minute, 2-minute, 3-minute) x 6 (counterbalanced groups) repeated measures ANOVA on the first factor (time frames). The between subjects main effect for group was not found to be significant *F* (5, 79) = 1.29 *p* > .05. Contrastingly, the interaction effect between maze probes (time frames) x group was found to be statistically significant *F* (8.7, 137.40) = 2.72 *p* < .007, η^2 = .15. The omnibus test of main effects yielded significant differences in mean CWS between 1-minute, 2-minute, and 3-minute time frames *F* (1.74, 137.40) = 436.97 *p* < .001 with partial η^2 = .85. A pairwise comparison using a Bonferroni adjusted multiple comparison test (*p* < .01) indicated significant differences in estimated mean CWS between 1-minute (*M* = 7.76, *SE* = .38) and 2-minute (*M* = 14.23, *SE* = .43) maze probes, 1-minute and 3-minute (*M* = 21.68, *SE* = .64) maze probes, and 2-minute and 3-minute maze probes.

Figure 2. CBM-maze probe CWS means by group.

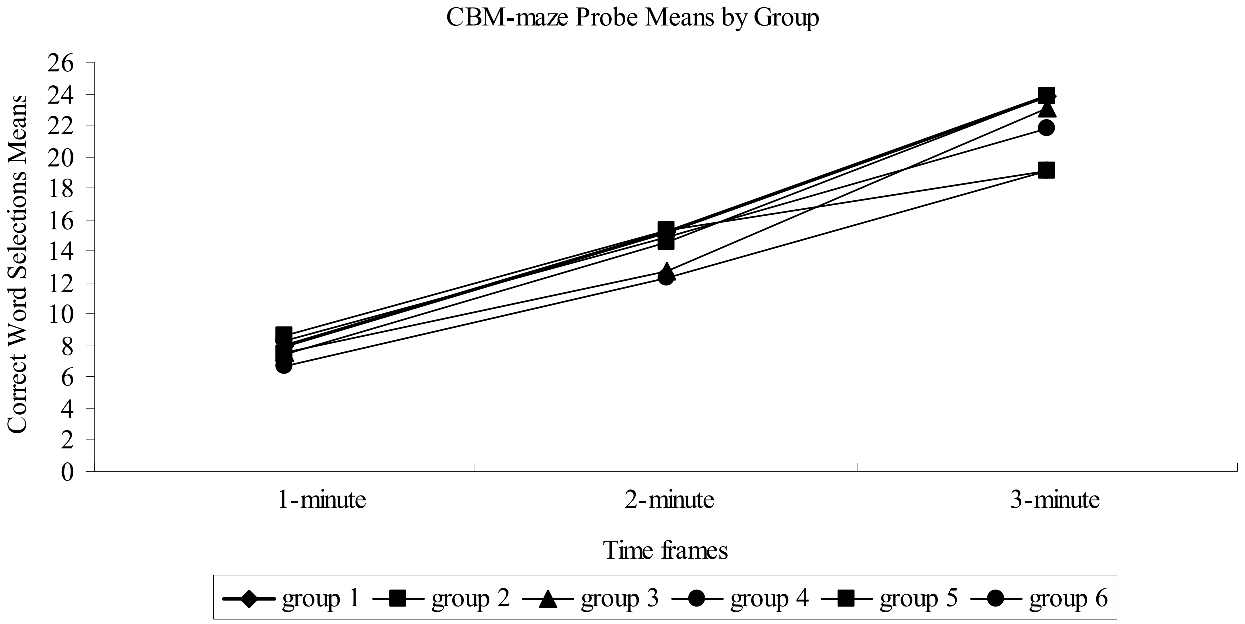


Figure 2 displays the group means across the three time frames. A visual inspection indicates a steady increase in level (CWS) and trend (slope) for most groups across all three maze probes. A noticeable interaction between groups is seen for the 2-minute maze probe and the 3-minute maze probe. Additionally, the means for group 4 are lower for all three maze probes. Group 3 mirrors the same trend for the 1-minute maze probes and 2-minute maze probe as group 4, but then shows a sharp increase in level and trend for the 3-minute maze probe. Finally, group 5 shows a steady increase in level and trend for the first two maze probes and then a decrease in level and trend.

The second repeated measures ANOVA employed a 3 (1-minute, 2-minute, 3-minute) x 2 (4th and 5th grade) repeated measures ANOVA on the first factor (time frames). The between subjects main effect for grade was not found to be significant F (1, 83) = 1.47 $p > .05$. The interaction effect between maze probes (time frames) x grade was found to be statistically significant F (1.52, 126.08) = 7.44 $p < .002$, $\eta^2 = .08$. The omnibus test of main effects yielded significant differences in mean CWS between 1-minute, 2-minute, and 3-minute time frames F (1.52, 126.08) = 426.62 $p < .001$ with partial $\eta^2 = .84$. A pairwise comparison using a Bonferroni adjusted multiple comparison test ($p < .01$) indicated significant differences in estimated mean CWS between 1-minute ($M = 7.17$, $SE = .27$) and 2-minute ($M = 14.36$, $SE = .42$) maze probes, 1-minute and 3-minute ($M = 21.66$, $SE = .66$) maze probes, and 2-minute and 3-minute maze probes.

Figure 3. CBM-maze probe CWS means by grade and combined sample.

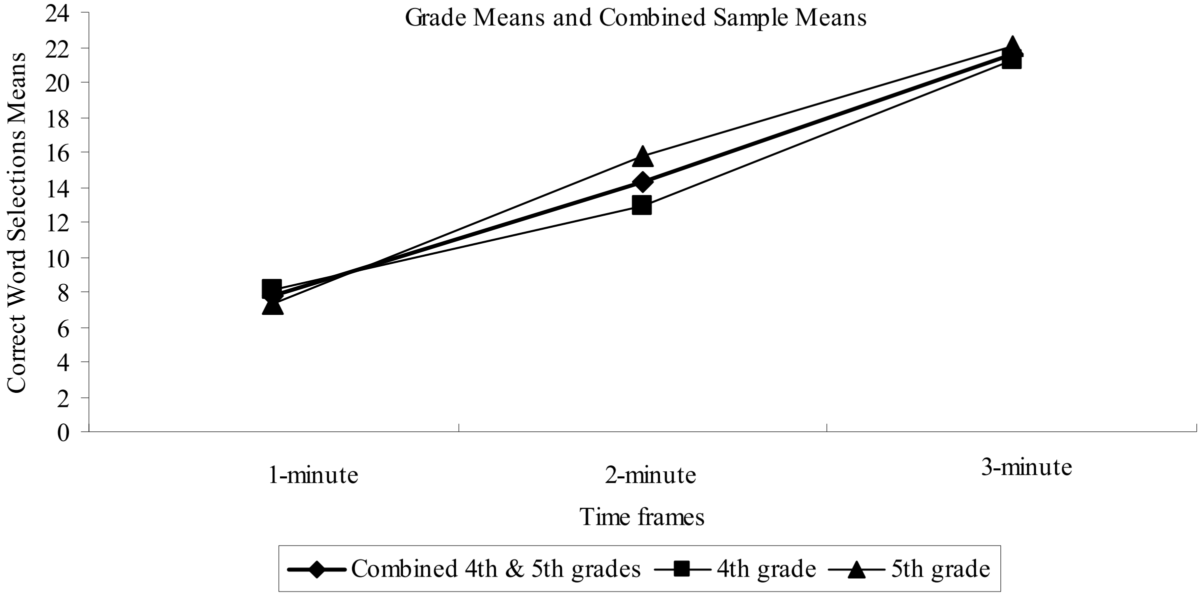
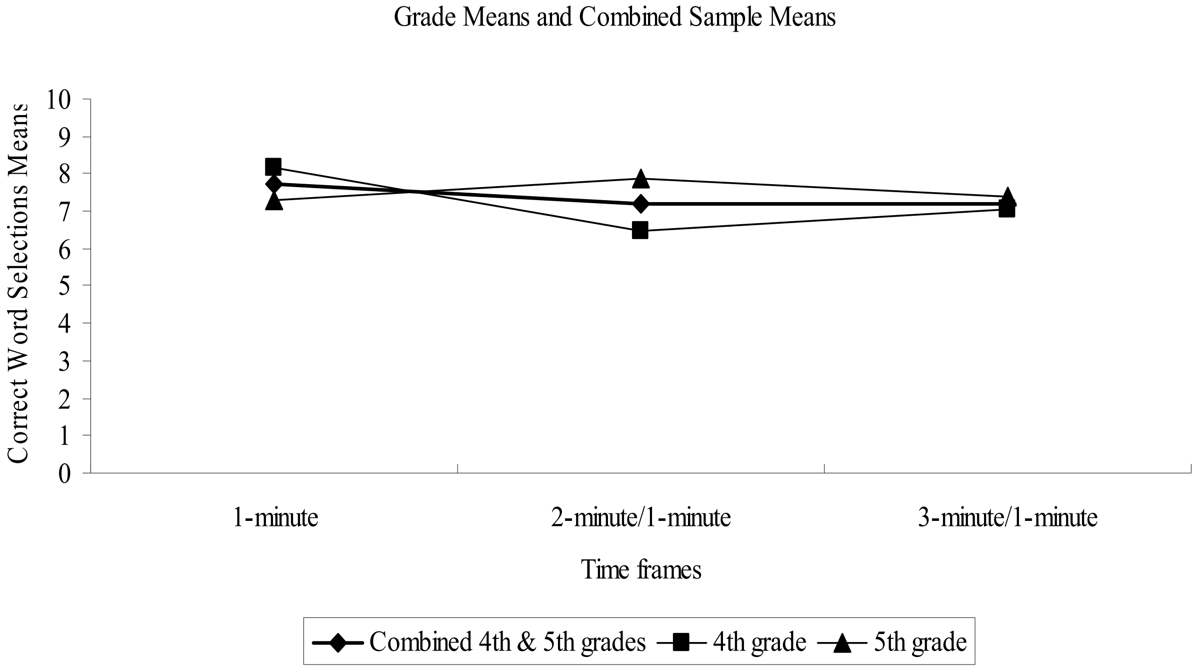


Figure 3 displays the means by grade and the combined sample across the three time frames. A visual inspection indicates a parallel and steady increase in level (CWS) and trend (slope) for grades and the combined sample across all three maze probe time frames. A noticeable interaction is seen between 4th (*M* = 12.96; *SD* = 3.85) and 5th (*M* = 15.77; *SD* = 3.85) grade for the 2-minute maze probe.

A third repeated measures ANOVA was conducted to investigate if there were any significant differences in CWS across maze probes when the time frames were held constant. This was based on the multiplicative pattern (x, 2x, 3x) found in mean CWS displayed in Table 7. In essence, the CWS datum for each student on the 2-minute time frame was converted to 1-minute by dividing each score by 2. The same process was used for the 3-minute time frame but with each score being divided by 3 to arrive at a 1-minute conversion as well. The resulting data yielded three 1-minute maze probes administered at six counterbalanced times. As with the previous repeated measures ANOVA conducted, the within factor was CBM-maze probe time frames (all 1-minute) and the between factor was grade (4th and 5th). The between subjects main effect for grade was not found to be significant F (1, 83) = .49 $p > .05$. The interaction effect between maze probes (time frames) x grade was found to be statistically significant F (1.86, 154.45) = 7.44 $p < .001$, $\eta^2 = .15$. The omnibus test of main effects yielded significant differences in CWS between the three converted 1-minute time frames F (1.86, 154.45) = 426.62 $p < .05$ with partial $\eta^2 = .05$. A pairwise comparison using a Bonferroni adjusted multiple comparison test ($p < .05$) indicated significant differences between 1-minute (*M* = 7.71, *SE* = .27) and the 2-minute/1-minute conversion (*M* = 7.18, *SE* = .21) maze probes. No significant differences were observed between 1-minute and 3-minute/1-minute conversion (*M* = 7.22, *SE* = .22) maze probes, and 2-minute/1-minute and 3-minute/1-minute maze probes.

Figure 4 displays the means by grade and the combined sample across the three 1-minute converted time frames. A visual inspection of the1-minute converted CWS means depicts some variability in 4th grade means (*SD*) 8.15 (.36), 6.48 (.28), and 7.07 (.30) respectively. Contrastingly, the 5th grade means appeared more stable 7.28 (.39), 7.88 (.30), and 7.36 (.32). A noticeable interaction was evidenced within the 2-minute/1-minute time frame conversion where an ascending trend in 4th grade and a descending trend in 5th grade were noted. The combined sample means for the converted CBM-maze probe time frames depict a stable trend ranging at about 7 CWS per minute and supporting the x, 2x, and 3x multiplicative pattern.

Figure 4. CBM-maze probe 1-minute CWS converted means by grade and combined sample.



Across all three repeated measures ANOVAs significant differences were observed for main effects and interaction effects for the within factor CBM-maze probe time frames. Interaction effects accounted for a small portion of the variance ranging from 8% to 15% for the first two analyses and 0% for the 1-minute maze conversion. This interaction effect was specifically evident for the 2-minute maze probe. The main effect analyses for time frame by group and time frame by grade accounted for 85%, 84% and 10% of the variance respectively. On the other hand, while the means for the 1-

minute maze conversion were significantly different from each other, they only accounted for 5% of the variance. This data suggests that as students have more time to complete the CBM-maze probes their CWS increases, but when time is held constant the magnitude of the difference may not be as meaningful.

Research Question 2

What are the alternate form reliabilities between 1-minute, 2-minute, and 3-minute maze probes for 4th and 5th grade students?

To investigate alternate form reliability of the CBM-maze probe measures, Pearson product-moment correlations were conducted between the 1-minute, 2-minute, and 3-minute probes (see Table 8). Bivariate correlations were conducted by grade level (4th and 5th) and for the combined sample. All correlations for the varying combinations of CBM-maze probes for grade levels and the combined sample were statistically significant at the *p*. < .01 level. There was a moderately strong relationship (.47 to .71) between 1-minute, 2-minute, and 3-minute time frames with the highest correlations noted in 5th grade. Each correlation between 4th and 5th grade was examined to see if there were any significant differences in relationship magnitude for 1-minute, 2-minute, and 3-minute CBM-maze probes. A Fisher transformation was utilized to compare the coefficients. An alpha level of .01 was set as the criterion to demonstrate significance (Silberglitt, Burns, Madyun, & Lail, 2006). Results found no significant differences between 4th and 5th grade coefficients. Therefore, the correlation coefficients for the combined sample appear to represent both 4th and 5th grade adequately.

Table 8

Alternate Form Reliability Bivariate Correlations

		CBM-maze Probe Time Frames		
	N	r_{m1m2}	r_{m1m3}	r_{m2m3}
4th grade	46	$.67^{**}$	$.47^{**}$	$.54^{**}$
5th grade	39	$.62^{**}$	$.65^{**}$	$.71^{**}$
Combined 4th/5th Grade	85	$.53^{**}$	$.52^{**}$	$.61^{**}$
Fisher's Z		.38	1.17	1.25

Note. r_{m1m2} = correlation between 1-minute and 2-minute maze probes; r_{m1m3} = correlation between 1-minute and 3-minute maze probes; r_{m2m3} = correlation between 2-minute and 3-minute maze probes
***p*. < .01 (2-tailed)
Z-critical is 2.58 for p < .01

Research Question 3

What is the relationship between 1-minute, 2-minute, and 3-minute CBM-maze probes and the STAR Reading computer adaptive comprehension test?

Bivariate correlations were run between CBM-maze probe time frames and the STAR Reading test to identify the magnitude of relationship between the measures. The results are broken down by grade level and the combined sample and are listed in Table 9. A Fisher transformation was utilized to compare the coefficients between grades. An alpha level of .01 was set as the criterion to demonstrate significance (Silberglitt, Burns, Madyun, & Lail, 2006). Results found no significant differences between 4th and 5th grade coefficients. For the combined sample, moderate positive correlations between the STAR Reading test and the 1-minute, 2-minute, and 3-minute maze probes were evidenced. All three correlations were statistically significant at the *p*. < .01 level and explained roughly 9%, 25%, and 13% of the variance. The relationship between the 2-minute maze probe and the STAR Reading test was approximately 2 times stronger than the 3-minute maze probe and a little less than 3 times stronger than the 1-minute maze probe.

Table 9

Concurrent Validity Bivariate Correlations

		Time Frames		
	N	1-minute maze	2-minute maze	3-minute maze
STAR 4th grade	46	.36*	.46**	.37*
STAR 5th grade	39	.38*	.43**	.35*
STAR Combined 4th /5th grade	85	.30**	.50**	.36**
Fisher's Z		.10	.17	.10

*p. < .05 (2-tailed)
**p. < .01 (2-tailed)
Z-critical is 2.58 for p < .01

A Stieger's Z-test for correlated correlations (Meng, Rosenthal, & Rubin, 1992) was computed to examine if one or more of the correlation between CBM-maze probe time frames and the STAR Reading test were significantly stronger than another. An alpha level of .01 (+/- 2.58) was set as the criterion to demonstrate significance. The results yielded no significant difference between the correlations for the 1-minute and 2-minute maze and the STAR Reading test Z = -2.09, *p* > .01; between the 1-minute and 3-

minute maze and the STAR Reading test Z = .60, $p > .01$; and between the 2-minute and 3-minute maze and the STAR Reading test Z = 1.63, $p > .01$.

Research Question 4

What are students' acceptability ratings of the CBM-maze probes?

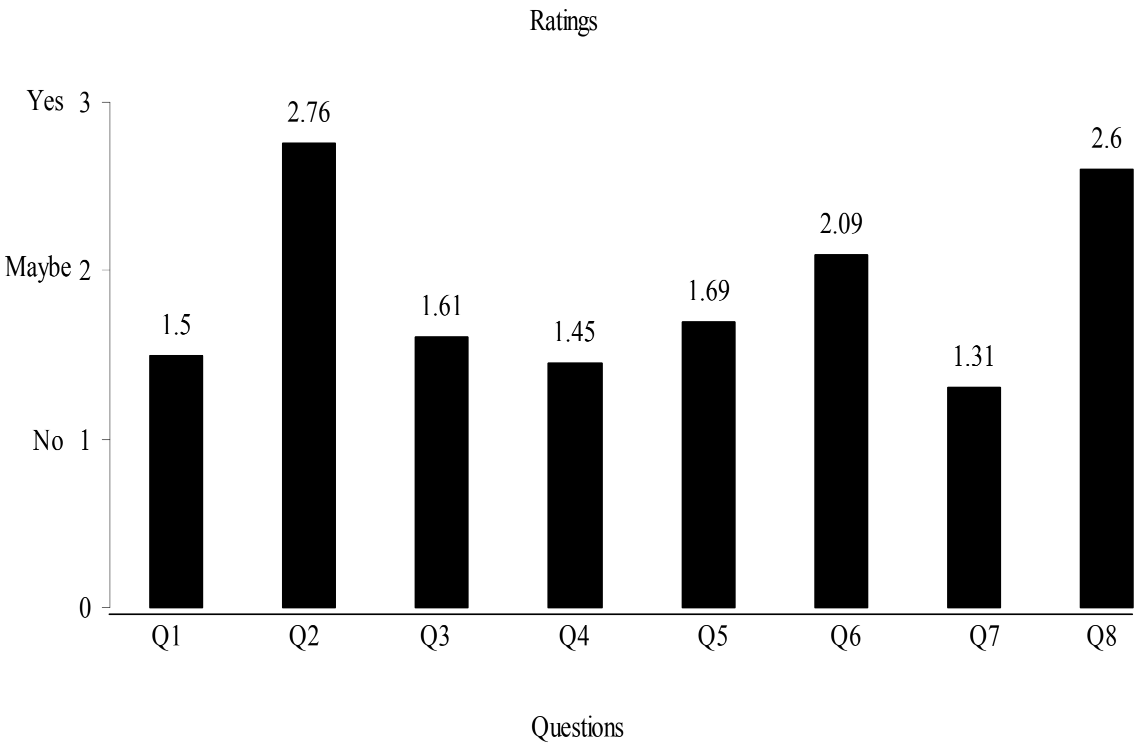
Table 10

Children's Assessment Acceptability Rating Scale Tabulations

	Student Rating Frequencies			
	No (1)	Maybe (2)	Yes (3)	Mean (*SD*)
1. I like taking CBM-maze probes.	39%	47%	14%	1.5 (.69)
2. CBM-maze probes are a good way to see how much I understand what I read.	3%	17%	80%	2.76 (.50)
3. My friends would like to take CBM-maze probes.	39%	61%	0	1.61 (.49)
4. I liked the 1st story the best.	69%	15%	16%	1.45 (.74)
5. I liked the 2nd story the best.	58%	15%	27%	1.69 (.87)
6. I liked the 3rd story the best	36%	18%	46%	2.09 (.91)
7. The stories were hard to understand.	76%	15%	9%	1.31 (.62)
8. The stories were easy to understand.	9%	21%	70%	2.60 (.66)

The social validity of the CBM-maze probes as reading comprehension measures was investigated by administering each student an assessment acceptability rating scale consisting of 8 items (see Table 10). The items were rated as “No,” “Maybe,” and “Yes” by students. For the purpose of calculating descriptive statistics and analyses, these categorical ratings were converted into likert scale ratings of 1 (No), 2 (Maybe), and 3 (Yes). The tabulations of the means for each question are visually displayed below.

Figure 5. Students’ mean ratings on the Children’s Assessment Acceptability Rating Scale.

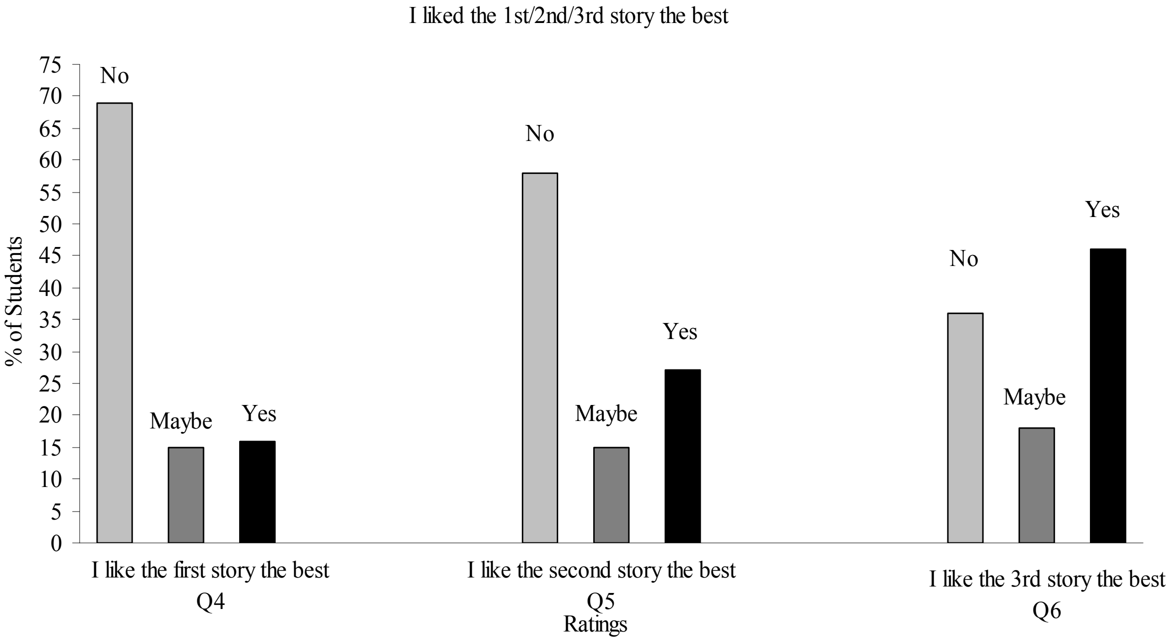


Questions 2, 7, and 8 were related specifically to the CBM-maze probes as comprehension measures. The highest rating of the scale was observed for question 2 (*M*

= 2.76; *SD* = .50) indicating that students perceived that CBM-maze probes were acceptable measures for reading comprehension. Roughly 70% of students indicated that the stories were easy to understand (questions 7 and 8). With respect to taking CBM-maze probes as comprehension measures, approximately 40% of students did not like taking them (question 1; *M* = 1.15, *SD* = .69) and would not recommend them to their peers (question 3; *M* = 1.61, *SD* = .49).

Finally, questions 4, 5, and 6 queried students on which story they liked the best, 1st story, 2nd story, or 3rd story. Due to the counterbalancing procedures implemented in this investigation, the ratings for these questions suggest they are more reflective of possibly story content and readability rather than actual time to complete each CBM-maze probe (time frames). That is, since not all groups completed the 1-minute maze probe first, the students were not rating the same time frames at the same times. A trend across probes was seen where students gradually endorsed ratings of "Yes" from the first probe to the third probe, 16%, 27%, and 46% respectively (see figure 5). The converse was noted for the rating of "No" where endorsed ratings gradually declined 69%, 58%, and 36%. Interestingly, a flat trend was observed for the "Maybe" rating 15%, 15%, and 18%. A slow increase in trend was also observed for Q4, Q5, and Q6 (see Figure 4) yielding the following means and standard deviations: 1.45 (.74), 1.69 (.87), and 2.09 (.91).

Figure 6. Percentage of ratings for questions 4, 5, and 6.



A repeated measures ANOVA was conducted to investigate if there were any significant differences in mean ratings across Q4, Q5, and Q6. The mixed design ANOVA was employed utilizing (Q4, Q5, Q6) as the within subject factor and 4th and 5th grades as the between subjects factor. The between subjects main effect for grade was not found to be significant F (1, 83) = .007 $p > .05$. The interaction effect between maze probes (time frames) x grade was not found to be statistically significant either F (2, 166) = .002 $p > .05$. The omnibus test of main effects yielded significant differences in mean ratings between Q4, Q5, and Q6 F (2, 166) = 9.216 $p < .001$ with partial $\eta^2 = .10$. A pairwise comparison using a Bonferroni adjusted multiple comparison test ($p < .01$) indicated no significant differences between Q4 (M = 1.46, SE = .08) and Q5 (M = 1.69, *SE* = .10), but significant differences were observed between Q4 and Q6 (*M* = 2.10, *SE* = .10) and Q5 and Q6.

The results of the descriptive statistics and repeated measures ANOVA suggests that as the students had practice taking CBM-maze probes, their liking (“Yes”) endorsements increased. Students’ mean ratings were significantly higher for Q6 than Q4 and Q5 supporting the hypothesis that they liked the third maze probe the best. Moreover, students’ ratings of disliking (“No”) decreased by approximately 50% from Q4 (69%) to Q6 (36%). Overall, it seems the students felt more confident in taking the CBM-maze probes as they gained experiences with the test format.

Summary

This chapter presented the results of the reliability and validity of CBM-maze probes with a sample of 4th and 5th grade students. More specifically, the results of the technical adequacy of 1-minute, 2-minute, and 3-minute time frames were presented using each research question as a guiding framework. In the next chapter, Chapter 5, a discussion of the results will be offered comparing previous research on the technical adequacy of CBM-maze probes to the current results. Finally, implications for practice within a 3-tier approach to prevention and intervention will be addressed.

CHAPTER 5

DISCUSSION

Introduction

The primary purpose of this study was to investigate the reliability and validity of 1-minute, 2-minute, and 3-minute curriculum-based measurement maze (CBM-maze) probe reading comprehension measures. Alternate form reliability, concurrent validity, and social validity of the maze probes were compared for the three time frames for a sample of 4th and 5th grade students. Descriptive and inferential statistics were utilized to analyze correct word selections (CWS) means for the CBM-maze probe time frames by grade level and as a combined sample.

This chapter provides a summary and discussion of the results for each research question addressed in this study. The technical adequacy (i.e., reliability and validity) of CBM-maze probes will be discussed in relation to current and previous research. Following, the implications for practice to include the use of CBM-maze probes as tools for universal screening and progress monitoring in the area of reading comprehension are discussed. Finally, cautions for interpreting the results of this study and future directions for research are presented.

Summary of Research

Four research questions were addressed in this study. These questions and a summary of their results follow.

1. Are there any significant differences in correct word selections (CWS) between 1-minute, 2-minute, and 3-minute CBM-maze probes for 4th and 5th grade students?

Three separate repeated measures analysis of variance (ANOVA) were employed to identify if there were any significant differences in CWS between the three different time frames. For the first two repeated measures ANOVA, the omnibus tests yielded significant differences (*p* < .001) between 1-minute, 2-minute and 3-minute time frames with groups (1-6) and grade levels (4th and 5th) as between factors. Adjusted Bonferroni pairwise comparisons indicated significant differences between the three time frames, as well as significant interaction effects for the 2-minute maze probe for both repeated measures analyses. Contrastingly, no significant differences were noted for the between subjects main effects for either groups or grades. The third repeated measures ANOVA utilized 1-minute converted CWS means. Again, the omnibus test yielded significant differences for the three maze probes (all 1-minute). Adjusted Bonferroni comparisons indicated significant differences in CWS means between the first (1-minute) and second maze (2-minute) probe only. Likewise significant interaction effects were noted for the 2-minute maze probe with no significant main effect difference on the between subject factor (grades).

2. What are the alternate form reliabilities between 1-minute, 2-minute, and 3-minute maze probes for 4th and 5th grade students?

Pearson product-moment correlations between 4th and 5th grade and the combined sample yielded overall moderate correlations (*r* = .47 to.71) between the three CBM-maze probes (time frames). A Fisher Z-transformation analysis of the 4th and 5th grade correlations indicated no significant differences in CWS means between grades. Thus, the combined sample ($4^{th}/5^{th}$ grade) correlations appear to be a viable representation for alternate form reliability analyses.

3. What is the relationship between 1-minute, 2-minute, and 3-minute CBM-maze probes and the STAR Reading computer adaptive reading comprehension test?

Results of the correlations between the STAR Reading test and the 1-minute (r = .30), 2-minute (r = .50), and 3-minute (r = .36) maze probe time frames for the combined sample yielded significant and moderate correlations. A further analysis of the correlations utilizing Stieger's Z-test indicated no significant differences between the three correlation coefficients at the .01 level.

4. What are students' acceptability ratings of the CBM-maze probes?

Each student completed the Children's Assessment Acceptability Rating Scale (CAARS; see Figure 5 and Appendix E) in order to measure their perceptions of the CBM-maze probes as comprehension measures. Student's perceptions of CBM-maze probes as measures of reading comprehension were relatively high (Q2; *M* = 2.76, *SD* = .50) and the students found the stories easy to read and understand (Q7 and Q8). Interestingly, relatively low ratings were endorsed by the students when queried if they liked taking CBM-maze procedures (Q1) or would recommend them to a peer (Q3) *M* = 1.5 and 1.61, respectively. Students also endorsed liking the 3^{rd} CBM-maze probe the best in comparison to the first and second maze probe. A repeated measures ANOVA indicated significant differences in main effects for the within subjects factor (Q4, Q5, Q6) with Q6 having significantly higher ratings than Q4 and Q5.

CBM-maze Probe Technical Adequacy

The technical adequacy of CBM-maze probes has been documented by several researchers utilizing various test formats (i.e., computerized, paper and pencil) and varying time frames (Parker, Hasbrouck, & Tindal, 1992; Wayman, Wallace, Wiley, Ticha, & Espin, 2007). Specifically, alternate form reliability has been investigated by Shin, Deno and Espin (2000), Brown-Chidsey, Davis and Maya (2003), and Seo (2005). Concurrent and predictive validity for CBM-maze probes has been established with reading aloud procedures (Espin, Deno, Maruyama, & Cohen, 1989; Markell & Deno, 1997), norm referenced reading comprehension assessments (Ardoin et al., 2004; Jenkins & Jewell, 1993), and high-stakes state accountability measures as well (Fore, Boon, & Martin, 2007; Wiley & Deno, 2005). The results obtained in this study extend the research base on the technical adequacy of CBM-maze probes.

Reliability. A broad definition of reliability involves quantifying the consistencies and/or inconsistencies of an examinees scores (Brennan, 2001). That is, reliability measures the degree of consistency over replications of measurement procedures for an individual's scores (Fishman & Galguera, 2003). The results for the alternate form reliability investigation for this study yielded consistent moderately high correlations across 4th and 5th grades and the combined sample. While the correlations noted between the three CBM-maze probe time frames did not reach acceptable reliability standards (.80 and above) indicated by the *Standards for Educational and Psychological Testing* (American Educational Research Association (AERA), American Psychological Association (APA), National Council on Measurement in Education (NCME), 1999), they were all significant at the .01 level with a sample size of 85 students. Inclusively, students' performance on the varying times frames was consistent over the three replications and not significantly different from each other.

The lower reliability coefficients may have been impacted by several mediating factors, (a) the variation in time frames (1-minute, 2-minute, and 3-minute), (b) the readability levels of each different form, (c) the students' experience with the CBM-maze probe procedures and test format, and (d) range restriction of CWS in the 4th grade. The results of the repeated measures ANOVA revealed that with each successive time frame, students' CWS scores were significantly higher suggesting that each individual time frame was independent of each other. That is, the 3-minute probe was able to discriminate from the 2-minute and 1-minute probes. The multiplicative pattern noted between the time frames (x, 2x, 3x) was evidenced for 5th grade and the combined sample, but deviated to some extent in the 4th grade. It appears then, that this deviation and range restriction of scores (see Table 7) may have contributed to the lower reliability coefficients observed.

Some of the correlation coefficients obtained in this investigation were similar to the correlation coefficients reported by Seo (2005). In that study, two generic CBM-maze probes were developed from newspaper articles and utilized as the dependent measures for reading comprehension. Correlations between the two forms used in that study ranged from .69 to .85 (p. 62). One difference in the Seo (2005) study was that the research design allowed at least 1-week between administration times. This was not the case in the current investigation where students completed each of the three alternate form CBM-maze probes one after the other in a counterbalanced order. Additionally, Seo (2005) allowed students a total of five minutes to complete each maze probe and then had students mark/circle off after 1, 2, 3, 4, and 5 minutes. The data for each time frame was derived from the 5-minute sample on the two administrations. Contrastingly, this investigation utilized three different CBM-maze probes for each of the three different time frames. An inspection of the CWS means reported by Seo (2005) indicated a similar multiplicative pattern (x, 2x, 3x) for the first three time frame, but then some leveling off was noted with lower CWS means for the 4-minute and 5-minute time frames (e.g., 7, 14, 21, 27, 32; p. 59).

Another finding related to reliability in this investigation was that the repeated measures ANOVA yielded no significant differences between grade levels for the three forms. At each grade level, the forms chosen for this investigation were calibrated using the Lexile Framework ("The Lexile framework for reading", n.d.; Stenner, Smith, & Burdick, 1983) to approximate students' end-of-year Lexile reading levels. In essence, this meant controlling for readability at both grade levels. These results indicate that utilizing the Lexile Framework as a measure of student readability levels was efficient at controlling for grade level readability to some extent. On the other hand, when CWS means for each individual form (i.e., 2-minute) by grade level was inspected, some variability was noted within the 4th grade 2-minute probe suggesting other factors such as story content, vocabulary, or prior knowledge may have played a role in producing lower reliability coefficients. This variability may also have been attributed to individual student performance within developmental reading stages (Ehri, 1995). For instance, the observed variability was found in the 4th grade sub-sample which is considered a transition or bridge between decoding and fluency and general comprehension by some developmental reading theorists (Chall, 1996; Yovanoff, Duesbery, Alonzo, & Tindal, 2005). Hence, at this stage one may expect higher variability in student performance which were findings also observed by Brown-Chidsey, Davis, and Maya (2003) for one of the three maze probes used in their study.

Finally, the test format and experience with the actual maze procedure may have also impacted reliability estimates. For example, on the CAARS questions relating to which CBM-maze probe students liked the most (see Figure 5, Q4, Q5, and Q6), the mean rating for the third probe was significantly higher than the ratings for the first and second probes. Therefore, one can assume that as students' practiced the maze procedure and gained familiarity with the test format, their acceptability of the maze procedure increased. While the CBM-maze probe test format is technically composed of multiple choice items (see Figure 7, #1), this format may not have been as familiar to the students as a typical fill in the blank item with three answer choices (see Figure 7, #2). Hence, a possibility exists that this variation could have impacted the students' consistency in performance across the three CBM-maze probes while they gained familiarity with the procedure. Nonetheless, CWS means and standard deviations for the combined sample depicted stable growth across all three times frames (see Table 7).

Figure 7. Test format examples.

1. The dog (**apple, broke, ran**) after the cat.
2. The dog ___________ after the cat.
   (a) apple
   (b) broke
   (c) ran

Concurrent validity. The significant and moderate correlations obtained for the concurrent validity investigation between the CBM-maze probes and the STAR Reading computer adaptive test indicated that the 2-minute maze probe had the stronger relationship over the 1-minute and 3-minute maze probes. Inclusively, the magnitude of the relationship for the 2-minute maze probe was about twice as strong as the 3-minute maze probe, and three times as strong as the 1-minute maze probe. This is a noteworthy finding in that concurrent validity has not yet been established between curriculum-based measures and computer adaptive reading tests. Inclusively, in comparison to other research that has utilized AIMSWeb generic CBM-maze probes, the results obtained here fall within the same moderate range (Fore, Boon, & Martin, 2007; Ardoin et al., 2004). A review of the literature on the concurrent and predictive validity of CBM-maze probes with norm referenced tests (i.e., Woodcock-Johnson Psychoeducational Battery III, SAT-9, CAT) and criterion referenced measures (i.e., high stakes state accountability tests) using traditional paper-pencil tests indicate most correlations reported were in the moderate to high moderate range .31 to .73 as well (Wayman, Wallace, Wiley, Ticha, & Espin, 2007). Interestingly, the correlations attained between CBM-maze probes and the

STAR Reading test (computer adaptive test) in this investigation fall within the same range as those found using traditional paper-pencil formats. This suggests that CBM-maze probes may be adequate measures of general comprehension across varying testing formats.

*Social vali*dity. The findings of the CAARS are not surprising. The lower ratings endorsed by the students for engaging in and recommending the maze procedure to other peers (see Table 10; Q1 and Q3) appeared to be a function of participating in the study during the last month of school, possibly having to complete four comprehension tests in one day, or not being familiar with the maze procedure itself. Regardless, a majority of the students felt that CBM-maze probes were a good source to measure their comprehension abilities suggesting value implications and relevance utility (Messick, 1995). Another interesting finding regarding students' perceptions of CBM-maze probes was the higher acceptability ratings across the three replications suggesting the students felt more confident with the procedure with each successive exposure. In contrast though, a positive relation between the increases in CWS scores for students and the corresponding time frames could not be made; the counterbalancing procedures implemented in this investigation presented a confounding variable in establishing that relation.

Cautions in Interpreting the Results

This investigation has produced some preliminary results for the use of generic CBM-maze probes. These results have several limitations and should be interpreted within the parameters of those limitations. First and foremost, the sample size was not very diverse and representative of current census data, but was a convenience sample. As such, generalization or comparison of the results should be done with a similar population. Inclusively, the students who participated in this study were all at or above grade level expectations in reading according to their classroom teachers and school reading grades. Basically then, discriminative power was reduced since the number of students at-risk for reading failure or receiving special education services were minimal to none for this sample.

Second, this investigation utilized commercially available generic CBM-maze probes (i.e., AIMSWeb). Most of the other studies have developed their own maze probes for their studies (Brown-Chidsey, Davis, & Maya, 2003; Espin & Foegen, 1996), provided varying standardized directions (Deno, Maruyama, Espin, & Cohen, 1989; Shinn & Shinn, 2002), used varying time frames (Parker, Hasbrouck, & Tindal, 1992), and have used varying readability formulas to calculate readability levels (Brown-Chidsey, Johnson, & Fernstrom, 2005; Hamilton & Shinn, 2003). All of these variations, including the ones in this study make it difficult to generalize the results across studies. For that reason, this investigation specifically sought to utilize the most common source of CBM-maze probes commercially available; those appear to be the ones produced by AIMSWeb. The reliability and validity for the AIMSWeb reading passages reported by the publishers were conducted on the oral reading fluency measures (Howe & Shinn, 2002), not the actual maze probes with the multiple choice items embedded in the reading passage. This also poses a further limitation in interpreting the results because one does not know if the embedded multiple choice items (see Figure 7, #1) affect the readability levels of the maze probes.

A final limitation for this investigation was that the students in the sample had never been exposed to CBM-maze procedures according to the school principal and the classroom teachers. Therefore, the results obtained in this investigation may be an underestimate of their true performance on the measures. A qualitative analysis of the STAR Reading tests data indicated that the average student for the combined sample was performing around the 75th percentile on the norm-referenced reading comprehension measure. On the other hand, a comparison of CWS means for students on this study to the norms provided on the AIMSWeb website indicated that CWS means for 4th grade, 5th grade, and the combined sample were just above the 50th percentile on the 3-minute maze probe (AIMSWeb, 2008; Hosp, Hosp, & Howell, 2007). This discrepancy suggests the students may need more familiarity with the probes to perform optimally on the measures.

Implications for Practice

Reading comprehension is a multidimensional construct and at the least can be conceptualized as a product of vocabulary and fluency (Gough & Tunmer, 1986; Hoover & Gough, 1990; Twyman & Tindal, 2007). Other researchers contend that reading comprehension and its measurement should be assessed using multiple domains (Pearson & Hamm, 2005; Sarroub & Pearson, 1998). These domains include apprehending literal information, as well as inferring information and transferring or constructing new knowledge (Haskell, 2001). The maze procedure was developed to be a simple and efficient general comprehension measure based on curriculum based assessment practices (Gickling & Thompson, 1985; Parker, Hasbrouck, & Tindal, 1992). While CBM-maze probes are comprehension measures that mostly combine fluency, vocabulary, and literal comprehension, other researchers have begun to extend the procedure to include inferential types of comprehension (for a full description see Twyman & Tindal, 2007).

Universal Screening

The premise of universal screening is to identify individuals at-risk for academic, behavioral, and/or social/emotional difficulties in order to provide evidence-based prevention and intervention programs (Albers, Glover, & Kratochwill, 2007). This investigation sought to extend the research on the reliability and validity of CBM-maze probes used by some school systems as universal screeners for global reading comprehension. More specifically, the results of this research have yielded implications for applied practice in elementary schools in two broad areas (a) the technical adequacy of CBM-maze probes as comprehension measures, and (b) standardizing the administration procedures of CBM-maze probes as universal screeners for reading comprehension within a 3-tiered prevention and intervention service delivery model (i.e., response-to-intervention; Elliot, Huai, & Roach, 2007; Glover & DiPerna, 2007; Greenberg et al., 2003).

The three alternate forms utilized in this study show some promise regarding their reliability, but are only approaching acceptable psychometric standards when making relative instructional decisions. As practitioners we should use caution in interpreting the results of this type of measure as a student's absolute performance of reading comprehension. Moreover, when utilizing the data to screen students for supplementary services or to gauge student growth over the school year, multiple sources of data should be evaluated to either certify a problem or support estimates of growth (Shinn, 1989). For instance, a teacher would not recommend a student for either a remedial program or gifted/enrichment program based on a one-time administration of a CBM-maze probe (fall, winter, or spring administrations). Best practices would dictate to monitor the student over a brief period (i.e., a week or two) with maze probes in conjunction with class work data and administering other time-efficient tests such as the STAR Reading test. In this manner, the maze probe would be one piece of information in screening for appropriate instruction.

The results regarding the validity of CBM-maze probes attained in this investigation were only in the moderate range with the STAR Reading test. These findings are not surprising because maze probes do not tap into a broad array of reading comprehension abilities. That is, up to this point, most CBM-maze forms are primarily loaded with literal comprehension and less complex vocabulary. Practitioners using maze probes to assess comprehension abilities should take note of this finding and use maze probes as general screeners versus a diagnostic tool.

The procedures for administering CBM-maze probes vary in the literature in comparison to oral reading fluency. In oral reading fluency or CBM-reading the standard time frame is a 1-minute administration. The only variation to this procedure is that some researchers suggest administering three 1-minute probes and using the median score as the student's datum to control for variability (Shinn, 1989), while other researchers have suggested that a one 1-minute sampling is sufficient (Hintze, Owen, Shapiro, & Daly, 2000). In either case, 1-minute is the standard for oral reading fluency. For CBM-maze probes, administration time frames have varied significantly from 1-minute to 10-minutes, to providing unlimited time to complete the probe. The results of this investigation indicated the 2-minute probe had the stronger validity coefficients suggesting this time frame maybe be sufficient as a universal screening tool. Regarding the number of probes administered, practitioners may want to just administer one probe during universal screenings (fall, winter, spring) as this research and the research by Seo (2005) found a relatively stable x, 2x, 3x multiplicative pattern. Overall, whatever procedure and time frame is chosen, practitioners may want to implement a one year cycle with those procedures and then evaluate the effectiveness, utility, and feasibility of CBM-maze probes within a 3-tier model of effective instruction such as response-to-intervention (RTI).

Progress Monitoring

In a 3-tier prevention/intervention RTI model, students' progress is more frequently monitored. This can occur at a minimum once a month within Tier 2 or up to 2 times a week when intensive specialized instruction is being implemented in Tier 3. CBM-maze probes have been validated to be sensitive in detecting growth over time when reading interventions were implemented (Allinder, Dunse, Brunken, & Obermiller-Krolikowski, 2001; L. S. Fuchs & Fuchs, 1999; L. S. Fuchs, Fuchs, Hamlett, & Ferguson, 1992). Therefore, since maze probes are easily administered via a group format, and multiple forms are available, maze probes may be a useful way for practitioners to establish growth or deterioration in the spirit of evidence-based practices. Additionally, by progress monitoring students, programmatic instruction can be adjusted accordingly (Deno & Mirkin, 1977).

Future Directions

Glover and Albers (2007) have suggested that universal screening tools (i.e., CBM-maze probes) have three important evaluative aspects (1) appropriateness for the intended use, (2) technical adequacy, and (3) usability. The following future directions stemming from this study are offered incorporating these three aspects.

Appropriateness and Intended Use of CBM-maze Probes

Future research should continue to investigate what time frame is most appropriate and efficient for CBM-maze probes as a universal screening tool and progress monitoring tool for general reading comprehension. Inclusively, researchers may want to further investigate possible differences in student performance between maze probes administered via a paper-pencil format, such as the ones utilized in this investigation, and maze probes administered via computer assisted technology (i.e., Shin, Deno, & Espin, 2000). Thirdly, the development and refinement of maze procedures that tap into the multidimensional construct of reading comprehension such as the concept maze developed by Twyman and Tindal (2007) for use in upper grades and content areas (i.e., social studies) may be of interest to researchers. Studies should also investigate the use of maze probes for low-stakes (i.e., supplemental instruction) and high stakes (i.e., special education eligibility) decisions while identifying the sensitivity, specificity and hit rate for these decisions.

Technical Adequacy

Research in this domain needs to continue to refine the construction of CBM-maze probes addressing the current limitations in reliability and validity. For example, item response theory (IRT) may be an avenue to investigate differential item functioning, as well assisting in developing more challenging distractors. In this same vein, the answer choices developed for each multiple choice item may be investigated to include challenging vocabulary. Replicating the reliability studies using Generalizability Theory should also be considered to investigate the variances contributed by varying facets (i.e., person, grade, probe etc.).

Usability

The acceptability of CBM-maze probes by teachers, administrators, and students needs further evaluation. Teachers and administrators need to feel comfortable administering, scoring, and interpreting the results of maze probes. Therefore research could look at the benefits associated with administering maze probes at all three tiers of an RTI model versus just at tier one as universal screeners. Future studies in the area of usability should incorporate the above mentioned aspects, as well as the development of infrastructures to support the use of maze procedures in schools.

Conclusions

The maze procedure was developed over 30 years ago as a tool to measure reading comprehension (Parker, Hasbrouck, & Tindal, 1992). Researchers such as Jenkins & Jewell (1993) and D. Fuchs and Fuchs (1992) have studied the maze procedure as a form of curriculum-based measurement (CBM). Both researchers found promising results for CBM-maze probes as measures of reading comprehension. In their critical review of the maze procedure, Parker, Hasbrouck and Tindal (1992) suggested that the administration time frames utilized for maze procedures were highly variable and further validity evidence needed to be collected to consider maze probes acceptable measures of reading comprehension. More recently, in a review of universal screening tools (a form of CBM's), Glover and Albers (2007) indicated similar results based on their review of the literature. The current study investigated three common administration time frames cited in the literature for CBM-maze probes, as well examined the concurrent validity of the measures with a norm-referenced computer adaptive reading test (STAR Reading test). While the results obtained in this study were not definitive, they indicate that the 2-minute maze probe yielded higher concurrent validity coefficients when compared to 1-minute and 3-minute administrations. Moreover, reliability coefficients across all three time frames were consistent, but only in the moderate range. Thus, it appears that a 2-minute administration of a maze probe may be a valid and reliable universal screening assessment as a piece of data make instructional decisions.

REFERENCES

AIMSWeb. (2008). Retrieved June 24, 2008, from http://www.aimsweb.com

Albers, C. A., Glover, T. A., & Kratochwill, T. R. (2007). Introduction to the special issue: How can universal screening enhance educational and mental health outcomes? *Journal of School Psychology, 45*, 113-116.

Allinder, R. M., Dunse, L., Brunken, C. D., & Obermiller-Krolikowski, H. J. (2001). Improving fluency in at-risk readers and students with learning disabilities. *Remedial and Special Education, 22*, 48-54.

Ardoin, S. P., Witt, J. C., Suldo, A. M., Connell, J. E., Koenig, J. L., Resetar, J. L., et al. (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.

Blankenship, C. A. (1985). Using curriculum-based assessment data to make instructional decisions. *Exceptional Children, 52*, 233-238.

Brennan, R. L. (2001). An essay on the history and future of reliability from the perspective of replications. *Journal of Educational Measurement, 38*, 295-317.

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 passages. *School Psychology Review, 34*, 387-394.

Brown-Chidsey, R., & Steege, M., W. (2005). *Response to intervention: Principles and strategies for effective practice.* New York: The Guilford Press.

Chall, J. S. (1989). Learning to read: The great debate 20-years later--A response to "Debunking the great phonics myth". *Phi Delta Kappan, 70*, 521-538.

Chall, J. S. (1996). *Stages of Reading Development* (2nd ed.). New York: McGraw-Hill.

Christ, T. C., Johnson-Gros, K. N., & Hintze, J. M. (2005). An examination of alternate assessment durations when assessing multiple-skill computational fluency: The generalizability and dependability of curriculum-based outcomes within the context of educational decisions. *Psychology in the Schools, 42*, 615-622.

Coie, J. D., Watt, N. F., West, S. G., Hawkins, J. D., Asarnow, J. R., Markman, H. J., et al. (1993). The science of prevention: A conceptual framework and some directions for a national research program. *American Psychologist, 48*, 1013-1022.

Conley, D. T. (2003). *Who governs our schools? Changing roles and responsibilities*. New York: Teachers College Press.

Cooper, J. O., Heron, T. E., & Heward, W. L. (1987). *Applied Behavior Analysis*. Upper Saddle River, NJ: Prentice-Hall, Inc.

Coyne, M. D., Kame'enui, E. J., & Simmons, D. C. (2001). Prevention and intervention in beginning reading: Two complex systems. *Learning Disabilities Research & Practice, 16*(2), 62-73.

Cranney, A. G. (1972-73). The construction of two types of cloze reading tests for college students. *Journal of Reading Behavior, 5*(1), 60-64.

Daly, E. J., III, Chafouleas, S., & Skinner, C. H. (2005). *Interventions for reading problems: Designing and evaluating effective strategies*. New York: The Guildford Press.

Davis, F. B. (1968). Research in reading comprehension. *Reading Research Quarterly, 3*, 499-545.

Davis, F. B. (1972). Psychometric research on comprehension in reading. *Reading Research Quarterly, 7*, 628-678.

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

Deno, S. L. (1992). The nature and development of curriculum-based measures. *Preventing School Failure, 36*, 5-10.

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

Deno, S. L., Espin, C., & Fuchs, L. S. (2004). Evaluation strategies for preventing and remediating basic skill deficits. In *Interventions for Academic and Behavior Problems II: Preventive and Remedial Approaches*. Bethesda, MD: National Association of School Psychologists.

Deno, S. L., & Fuchs, L. S. (1987). Developing curriculum-based measurement systems for data-based special education problem solving. *Focus on Exceptional Children, 19*(8), 1-15.

Deno, S. L., Fuchs, L. S., Marston, D., & Shin, J. (2001). Using curriculum-based measurement to establish growth standards for students with learning disabilities. *School Psychology Review, 30*(507-524).

Deno, S. L., Maruyama, G., Espin, C., & Cohen, C. (1989). *The basic academic skills samples (BASS)*. Minneapolis, MN: University of Minnesota.

Deno, S. L., & Mirkin, P. K. (1977). *Data based program modification: A manual.* Arlington, VA: Council for Exceptional Children.

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

Dougherty Stahl, K. A., & McKenna, M. C. (2006). *Reading research at work: Foundations of Effective Practice*. New York: The Guilford Press.

Ehri, L. C. (1995). Stages of learning development in learning to read words by sight. *Journal of Research in Reading, 18*, 116-125.

Elliot, S. N., Huai, N., & Roach, A. T. (2007). Universal and early screening for educational difficulties: Current and future approaches. *Journal of School Psychology, 45*, 137-161.

Ellis, M. V. (1999). Repeated measures designs. *The Counseling Psychologist, 27*, 552-578.

Espin, C., Deno, S. L., Maruyama, G., & Cohen, C. (1989). *Basic academic skills samples (BASS).* Paper presented at the American Educational Research Association, San Francisco.

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

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*, 175-191.

Fishman, J. A., & Galguera, T. (2003). *Introduction to test construction in the social and behavioral sciences: A practical guide*. New York: Rowman & Littlefield Publishers, Inc.

Fore, C., Boon, R. T., & Martin, C. (2007). Concurrent and predictive validity of curriculum-based measurement for students with emotional and behavioral disorders *International Journal of Special Education, 22*, 24-31.

Fuchs, D., & Fuchs, L. S. (1986). Effects of systematic formative evaluation: A meta-analysis. *Exceptional Children, 53*, 199-208.

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

Fuchs, D., Mock, D., Morgan, P. L., & Young, C. L. (2003). Responsiveness-to-intervention: Definitions, evidence, and implications for the learning disabilities construct. *Learning Disabilities Research & Practice, 18*, 157-171.

Fuchs, L. S., & Deno, S. L. (1991). Paradigmatic distinctions between instructionally relevant measurement models. *Exceptional children, 57*, 488-501.

Fuchs, L. S., & Fuchs, D. (1999). Monitoring student progress toward the development of reading competence: A review of three forms of classroom-based assessments. *School Psychology Review, 28*, 659-671.

Fuchs, L. S., & Fuchs, D. (2000). Analogue assessment of academic skills: Curriculum-based measurement and performance assessment. In E. S. Shapiro & T. R. Kratochwill (Eds.), *Behavioral assessment in schools: Theory, research, and clinical foundations* (2nd ed.). New York: The Guildford Press.

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

Fuchs, L. S., Fuchs, D., 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, 5*, 239-256.

Gickling, E. E., & Thompson, V. P. (1985). A personal view of curriculum-based assessment. *Exceptional Children, 52*, 205-218.

Glover, T. A., & Albers, C. A. (2007). Considerations for evaluating universal screening assessments. *Journal of School Psychology, 45*, 117-135.

Glover, T. A., & DiPerna, J. C. (2007). Service delivery for response to intervention: Core components and directions for future research. *School Psychology Review, 36*, 526-540.

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

Greenberg, M. T., Weisssberg, R. P., O'Brien, M. U., Zins, J. E., Fredericks, L., Resnik, H., et al. (2003). Enhancing school-based prevention and youth development through coordinated social, emotional, and academic learning. *American Psychologist, 58*, 466-474.

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

Haager, D., Klinger, J., & Vaughn, S. (Eds.). (2007). *Evidence-based reading practices for response to intervention*. Baltimore, MD: Paul H. Brookes Publishing Co.

Hamilton, C., & Shinn, M. R. (2003). Characteristics of word callers: An investigation of the accuracy of teacher's judgments of reading comprehension and oral reading skills. *School Psychology Review, 32*, 223-235.

Haskell, R. E. (2001). *Transfer of learning: Cognition, instruction and reasoning*. New York: Academic Press.

Hintze, J. M., Callahan, J. E., Mathews, W. J., Williams, A. S., & Tobin, K. G. (2002). Oral reading fluency and prediction of reading comprehension in African American and Caucasian elementary school children. *School Psychology Review, 31*, 540-553.

Hintze, J. M., Christ, T. C., & Keller, L. A. (2002). The generalizability of CBM survey-level mathematics assessments: Just how many samples do we need? *School Psychology Review, 31*, 514-528.

Hintze, J. M., Owen, S. V., Shapiro, E. S., & Daly, E. J. (2000). Generalizability of oral reading fluency measures: Application of G theory to curriculum-based measurement. *School Psychology Quarterly, 15*, 52-68.

Hintze, J. M., & Shapiro, E. S. (1997). Curriculum-based measurement and literature-based reading: Is curriculum-based measurement meeting the needs of changing reading curricula? *Journal of School Psychology, 35*, 351-375.

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

Hoover, W. A., & Gough, P. B. (1990). The simple view of reading. *Reading & Writing, 2*, 127-160.

Hosp, M. K., & Fuchs, L. S. (2005). Using CBM as an indicator of decoding, word reading and comprehension: Do the relations change with grade? *School Psychology Review, 34*, 9-26.

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

Howe, K. B., & Shinn, M. M. (2002). *Standard reading assessment passages (RAPs) for use in general outcome measurement: A manual describing development and technical features*. Eden, Prairie, MN: Ed Formation, Inc.

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

Kennedy, C. H. (2005). *Single case designs for educational research*. Boston, MA: Pearson: Allyn and Bacon.

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.

Kratochwill, T. R., Albers, C. A., & Shernoff, E. (2004). School-based interventions. *Child and adolescent psychiatric clinics of North America, 13*, 885-903.

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

Lennon, C., & Burdick, H. (2004). The Lexile framework as an approach for reading measurement and success. Retrieved February 23, 2008, from http://www.lexile.com/uploads/White%20Papers/Lexile-Reading-Measurement-and-Success-0504.pdf

The Lexile framework for reading. (n.d.). Retrieved June 26, 2008, from http://www.lexile.com/DesktopDefault.aspx?view=ed&tabindex=2&tabid=16&tabpageid=335

Lovitt, T. (1967). Assessment of children with learning disabilities. *Exceptional Children, 34*, 233-239.

Markell, M. A., & Deno, S. L. (1997). Effects of increasing oral reading: Generalization across reading tasks. *The Journal of Special Education, 31*, 233-250.

Marston, D. (1989). A curriculum-based measurement approach to assessing academic performance: What it is and why do it. In *Curriculum-based Measurement: Assessing Special Children*. New York: The Guilford Press.

Marston, D., Reschly, A. L., Lau, M. Y., Muyskens, P., & Canter, A. (2007). Historical perspectives and current trends in problem solving: The Minneapolis story. In D. Haager, J. Klinger & S. Vaughn (Eds.), *Evidence-based Reading Practices for*

Response to Intervention (pp. 265-286). Baltimore, MD: Paul H. Brookes Publishing Co.

Meng, X., Rosenthal, R., & Rubin, D. B. (1992). Comparing correlated correlations. *Psychological Bulletin, 111*, 172-175.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*, 741-749.

Meyers, L. S., Gamst, G., & Guarino, A. J. (2006). *Applied multivariate research: Design and interpretation*. Thousand Oaks, CA: Sage Publications, Ltd.

Nation, M., Crusto, C., Wandersman, A., Kumpfer, K. L., Seybolt, D., Morrissey-Kane, E., et al. (2003). What works in prevention: Principles of effective prevention programs. *American Psychologist, 58*, 449-456.

National Center for Educational Statistics. (2005). NAEP Question Tool. Retrieved February 4, 2008, from http://nces.ed.gov/nationsreportcard/itmrls/

National Reading Panel. (2001). What the public told us. *In NRP Progress Report* Retrieved February 3, 2008, from www.nationalreadingpanel.org/Publications/Interim_Report/section4.htm

Paris, S. G. (2006). Connecting scientific and practical approaches to reading assessment. In K. A. Doughrty-Stahl & M. C. McKenna (Eds.), *Reading Research at Work: Foundations for Effective Practice* (pp. 363-372). New York: The Guilford Press.

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.

Pearson, P. D., & Dole, J. A. (1987). Explicit comprehension instruction: A review of research and a new conceptualization of instruction. *The Elementary School Journal, 88*, 151-165.

Pearson, P. D., & Hamm, D. N. (2005). The assessment of reading comprehension: A review of practices-past, present and future. In *Children's Reading Comprehension and Assessment*. Mahwah, NJ: Lawrence Erlbaum Associates.

Potter, M. L., & Wamre, H. D. (1990). Curriculum-based measurement and developmental reading models: Opportunities for cross-validation. *Exceptional Children, 57*, 16-25.

Reiss, D., & Price, R. H. (1996). National research agenda for prevention research: The National Institute of Mental Health report. *American Psychologist, 51*, 1109-1115.

Renaissance Learning. (2006). STAR reading computer-adaptive reading test and database: Technical manual. Retrieved February 20, 2007, from www.renlearn.com

Sarasti, I. A. (2007). *The effects of reciprocal teaching comprehension-monitoring strategy on 3rd grade students' reading comprehension.* Unpublished doctoral dissertation, University of North Texas.

Sarroub, L., & Pearson, P. D. (1998). Two steps forward, three steps back: The stormy history of reading comprehension assessment. *The Clearing House, 72*, 97-105.

Sattler, J. M. (2001). *Assessment of children: Cognitive applications* (4th ed.). San Diego, CA: Jerome M. Sattler, Publisher, Inc.

Savage, R. (2001). The 'Simple View' of reading: Some evidence and possible implications. *Educational Psychology in Practice, 17*, 17-33.

Shin, J., Deno, S. L., & Espin, C. (2000). Technical adequacy of Maze task for curriculum-based measurement of reading growth. *Journal of Special Education, 34*, 164-173.

Shinn, M. R. (1988). Development of curriculum-based local norms for use in special education decision-making. *School Psychology Review, 17*, 61-80.

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

Shinn, M. R., Good, R. H., Knutson, N., & Tilly, W. D. (1992). Curriculum-based measurement of oral reading fluency: A confirmatory factor analysis of its realtion to reading. *School Psychology Review, 21*, 459-479.

Shinn, M. R., & Shinn, M. M. (2002). *AIMSWeb training workbook: Administration and scoring for reading maze for use in general outcome measurement*. Eden Prairie, MN: Edformation, Inc.

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

Stahl, S. A., & Fairbanks, M. M. (2004). The effects of vocabulary instruction; A model-based meta-analysis. In *Reading Research at Work: Foundations of Effective Practice*. New York: The Guilford Press.

Stahl, S. A., & Heubach, K. (2006). Fluency-oriented reading instruction. In K. A. Doughrty-Stahl & M. C. McKenna (Eds.), *Reading research at work: Foundations for effective practice* (pp. 177-204). New York: The Guilford Press.

Stenner, A. J., Smith, M., & Burdick, D. S. (1983). Toward a theory of construct definition. *Journal of Educational Measurement, 20*, 305-315.

Sternberg, R. J. (2003). *Cognitive Psychology* (3rd ed.). Belmont, CA: Wadsworth/ Thomas Learning.

Stevens, J. P. (2002). *Applied multivariate statistics for the social sciences* (4th ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Tucker, J. A. (1985). Curriculum-based assessment: An introduction. *Exceptional Children, 52*, 199-204.

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.

United States Department of Education. (2001). *No Child Left Behind*. Retrieved March 10, 2008, from http://www.ed.gov.inits/nclb/titlepage.html

United States Public Health Service. (2000). *Report of the Surgeon General's Conference on Children's Mental Health: A national action agenda*. Washington, DC: Department of Health and Human Services.

VanDerHeyden, A. M., Witt, J. C., & Gilbertson, D. (2007). A multi-year evaluation of the effects of a response to intervention (RTI) model on identification of children for special education. *Journal of School Psychology, 45*, 225-256.

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

Weissberg, R. P., Kumpfer, K. L., & Seligman, M. E. (2003). Prevention that works for children and youth: An introduction. *American Psychologist, 58*, 425-432.

Wells, C. S., & Hintze, J. M. (2007). Dealing with assumptions underlying statistical tests. *Psychology in the Schools, 44*, 495-502.

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

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

APPENDIX A:

SAMPLE CBM-MAZE PROBE

Jason and Max picked next Friday to carry out their special mission. Friday was a week away. They (**agreed, had, branches**) so many things to accomplish. In (**plan, order, at**) to reach their final goal, the (**next, branches, boys**) made a plan for each day (to, of, each) the week. They had to work (**hard, creek, big**) every day to finish each task. (**Pile, Could, Had**) they do it all?

On Monday, (**creek, big, they**) agreed to meet and put plan (**near, wood, A**) into action. Plan A was to (**gather, work, day**) as many fallen branches as they (**could, on, had**) carry. They hauled the wood from (**neat, a, the**) edge of the cornfield and stacked (**agree, it, they**) in a big pile at the (**plan, edge, hauled**) of the forest.

On Tuesday, the (**rocks, by, boys**) met near the lazy creek and (**put, climb, wood**) plan B into motion. They dug (**up, near, the**) rocks the size of footballs from (**and, night, the**) creek's bottom. By dusk, they had (**rode, arranged, to**) the rocks in a neat circle (**a, next, up**) to the pile of branches they (**their, found, had**) hauled the night before.

On Wednesday, (**plan, the, work**) C was to climb into the (**attic, umbrellas, they**) above Jason's garage. They searched around (**Max, in, with**) flashlights and both found backpacks. They (**spoke, under, wore**) their packs as they rode their (**without, bikes, garage**) to the edge of the forest (**to, end, for**) complete the day's work.

On Thursday (**they, it, work**) rained. They had to drop the (**up, plan, forest**) for the day. Still, Jason and (**went, backpack, Max**) met at the end of their (**bikes, driveways, on**) under umbrellas. They quietly spoke. They (**rained, decided, tent**) their mission would work without plan (**0, fire, was**). When the sun went down on (**only, Friday, evening**), they met at the edge of (**the, out, and**) forest. There sat their tent. They'd (**stacked, tasks, set**) it up on Wednesday evening. The (**circle, special, wood**) was ready to go into their (**campfire, many, night**) ring. Their next step was to (**big, build, climb**) a warm fire.

Note. Adapted from http://www.aimsweb.com/uploaded/files/sample_maze.pdf.

APPENDIX B:

MAZE COVER SHEET AND STANDARDIZED DIRECTION

Maze Cover Sheet

Name:_________________________________

Grade:________ Group:________

Practice Test

1. The dog (**apple, broke, ran**) after the cat. The cat ran (**fast, green, for**) up the hill. The dog barked (**in, at, is**) the cat.

Pass maze task out to students. Have students write their names on the Cover Sheet. Make sure they do not turn page until you tell them to do so.

Say to the students:

"*When I say 'Begin" I want you to read some stories. You will have 1, 2, or 3 minutes to read the story and complete the task. Listen carefully to the directions. Some if the words in the story are replaced with a group of three words. Your job is to circle the 1 word that makes the most sense in the story. Only one word is correct.*"

2. **Go over practice test:**

Let's practice one together. Look at your first page. Read the first sentence silently while I read it aloud: 'The dog apple, broke, ran after the cat.' The three choices are apple, broke, ran. 'The dog apple after the cat.' That sentence does not make sense. 'The dog broke after the cat.' That sentence does not make sense. 'The dog ran after the cat.' That sentence does make sense, so circle the word ran."

(Make sure students circle the word ran)

Let's go to the next sentence. Read it silently while I read it aloud. The cat ran fast, green, for up the hill. The three choices are fast, green, for. Which word is the correct word for the sentence?

Students answer fast

"Yes, 'The cat ran fast up the hill.' is correct, so circle the correct word fast."

(Make sure students circle fast)

"Silently, read the next sentence and raise your hand when you think you know the answer."

(Make sure students know the correct word. Read the sentence with the correct answer)

"That's right, 'The dog barked at the cat.' is correct. Now what do you do when you choose the correct word?"

(Students answer "Circle it." Make sure students understand the task)

"That's correct, you circle it. I think you're ready to work on the stories on your own."

Start testing by saying…..

> *"When I say "Begin", turn to the first story and start reading silently. When you come to a group of three words, circle the 1 word that makes the most sense. Work as quickly as you can without making mistakes. If you finish the first side turn to the back of that page and keep working until I say 'Stop'.*

3. **Then say,** *"Begin"* **start your stopwatch.**

4. **Monitor students to make sure they understand that they are to circle only 1 word.**

5. **If a student finishes before the time limit, record the time on the page.**

6. **At the end of either 1, 2, or 3 minutes say:** *"Stop. Put your pencils down."*

7. **Then say**, "*Now you will try the next passage. Remember, when you come to a group of three words, circle the 1 word that makes the most sense. Work as quickly as you can without making mistakes. Turn to the next story and pick up your pencils."* **Clear your stop watch.** "*Ready? Begin"* **Start your stopwatch**.

8. **Monitor students to make sure they understand that they are to circle only 1 word.**

9. **If a student finishes before the time limit, record the time on the page.**

10. **At the end of either 1, 2, or 3 minutes say:** *"Stop. Put your pencils down."*

11. **Then say,** "*Now you will do the last passage. Remember, when you come to a group of three words, circle the 1 word that makes the most sense. Work as quickly as you can without making mistakes. Turn to the last story and Pick up your pencils."* **Clear your stop watch.** "*Ready? Begin"* **Start your stopwatch**.

12. If a student finishes before the time limit, record the time on the page.

13. **At the end of either 1, 2, or 3 minutes say:** *"Stop. Put your pencils down and turn your booklets face down."*

14. Repeat steps 3-14 for remaining groups.

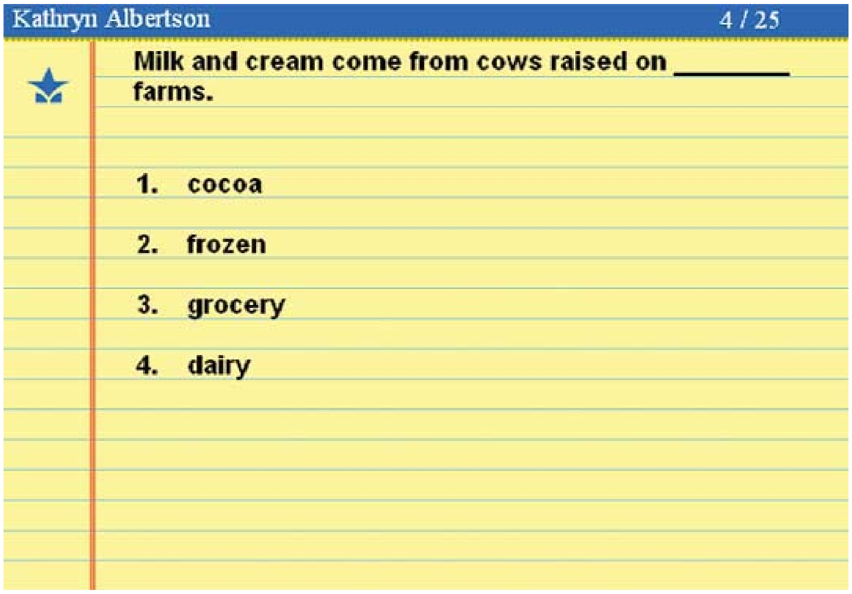
15. Collect maze task.

Group	Counterbalanced CBM-maze Probes		
1	1-minute	2-minute	3-minute
2	1-minute	3-minute	2-minute
3	2-minute	1-minute	3-minute
4	2-minute	3-minute	1-minute
5	3-minute	2-minute	1-minute
6	3-minute	1-minute	2-minute

Note. Adapted from Deno, Maruyama, Espin, & Cohen, 1989.

APPENDIX C:

SAMPLE STAR READING TEST

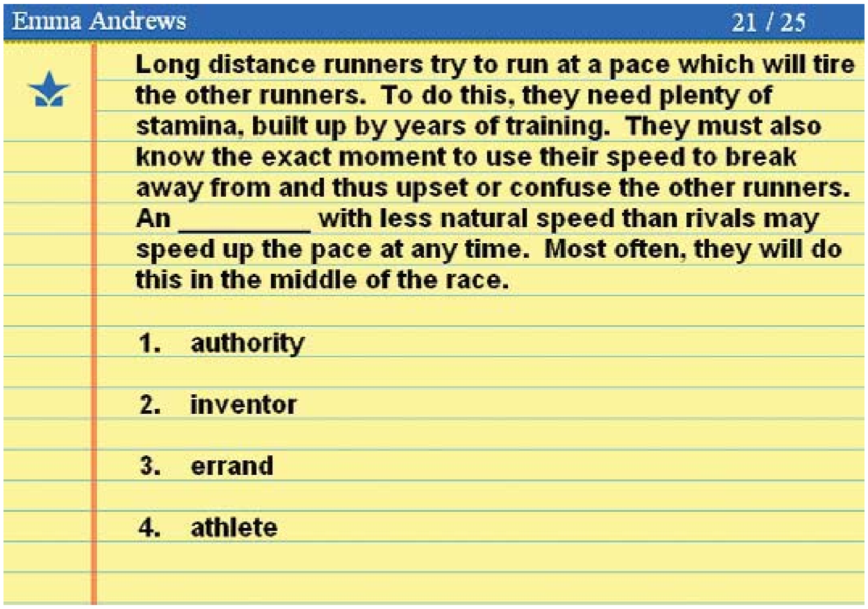


Kathryn Albertson 4 / 25

Milk and cream come from cows raised on ________ farms.

1. cocoa
2. frozen
3. grocery
4. dairy

Vocabulary-in-context Questions



Emma Andrews 21 / 25

Long distance runners try to run at a pace which will tire the other runners. To do this, they need plenty of stamina, built up by years of training. They must also know the exact moment to use their speed to break away from and thus upset or confuse the other runners. An ________ with less natural speed than rivals may speed up the pace at any time. Most often, they will do this in the middle of the race.

1. authority
2. inventor
3. errand
4. athlete

Authentic Text Questions

Note. Adapted from http://kmnet.renlearn.com/Library/R003863402GG92A3.pdf

APPENDIX D:

CBM-MAZE PROCEDURAL INTEGRITY CHECKLIST

Procedural Integrity Checklist

Group 1

1-minute	2-minute	3-minute

Place a 1 after each step completed and a 0 after each step not completed.

Maze probe assessment procedures

1-minute
__________ makes sure maze probes are in front of students
__________ Set timer to zero
__________ Read maze probe assessment instructions
__________ Starts timer
__________ Monitors for circling and completion
__________ Stop timer by saying “time’s up” after 1 minute
__________ says “turn to the next page”

2-minute

__________ reads directions
__________ Set timer to zero
__________ Says “Ready, begin”
__________ Starts timer
__________ Monitors for circling and completion
__________ Stop timer by saying “Stop” after 2 minutes

3-minute

__________ reads directions
__________ Set timer to zero
__________ Says “Ready, begin”
__________ Starts timer
__________ Monitors for circling and completion
__________ Stop timer by saying “Stop” after 3 minutes
__________ Collects packet

Procedural Integrity Checklist

Group 2	1-minute	3-minute	2-minute

Place a 1 after each step completed and a 0 after each step not completed.

Maze probe assessment procedures

1-minute
__________ makes sure maze probes are in front of students
__________ Set timer to zero
__________ Read maze probe assessment instructions
__________ Starts timer
__________ Monitors for circling and completion
__________ Stop timer by saying "time's up" after 1 minute
__________ says "turn to the next page"

3-minute

__________ reads directions
__________ Set timer to zero
__________ Says "Ready, begin"
__________ Starts timer
__________ Monitors for circling and completion
__________ Stop timer by saying "Stop" after 3 minutes

2-minute

__________ reads directions
__________ Set timer to zero
__________ Says "Ready, begin"
__________ Starts timer
__________ Monitors for circling and completion
__________ Stop timer by saying "Stop" after 2 minutes
__________ Collects packet

Procedural Integrity Checklist

Group 3	2-minute	1-minute	3-minute

Place a 1 after each step completed and a 0 after each step not completed.

Maze probe assessment procedures

2-minute
__________ makes sure maze probes are in front of students
__________ Set timer to zero
__________ Read maze probe assessment instructions
__________ Starts timer
__________ Monitors for circling and completion
__________ Stop timer by saying “time’s up” after 2 minutes
__________ says “turn to the next page”

1-minute

__________ reads directions
__________ Set timer to zero
__________ Says “Ready, begin”
__________ Starts timer
__________ Monitors for circling and completion
__________ Stop timer by saying “Stop” after 1 minute

3-minute

__________ reads directions
__________ Set timer to zero
__________ Says “Ready, begin”
__________ Starts timer
__________ Monitors for circling and completion
__________ Stop timer by saying “Stop” after 3 minutes
__________ Collects packet

Procedural Integrity Checklist

Group 4

2-minute	3-minute	1-minute

Place a 1 after each step completed and a 0 after each step not completed.

Maze probe assessment procedures

2-minute

__________ makes sure maze probes are in front of students
__________ Set timer to zero
__________ Read maze probe assessment instructions
__________ Starts timer
__________ Monitors for circling and completion
__________ Stop timer by saying “time’s up” after 2 minutes
__________ says “turn to the next page”

3-minute

__________ reads directions
__________ Set timer to zero
__________ Says “Ready, begin”
__________ Starts timer
__________ Monitors for circling and completion
__________ Stop timer by saying “Stop” after 3 minutes

1-minute

__________ reads directions
__________ Set timer to zero
__________ Says “Ready, begin”
__________ Starts timer
__________ Monitors for circling and completion
__________ Stop timer by saying “Stop” after 1 minute
__________ Collects packet

Procedural Integrity Checklist

Group 5

3-minute	2-minute	1-minute

Place a 1 after each step completed and a 0 after each step not completed.

Maze probe assessment procedures

3-minute
__________ makes sure maze probes are in front of students
__________ Set timer to zero
__________ Read maze probe assessment instructions
__________ Starts timer
__________ Monitors for circling and completion
__________ Stop timer by saying "Stop" after 3 minute

2-minute

__________ reads directions
__________ Set timer to zero
__________ Says "Ready, begin"
__________ Starts timer
__________ Monitors for circling and completion
__________ Stop timer by saying "Stop" after 2 minute

1-minute

__________ reads directions
__________ Set timer to zero
__________ Says "Ready, begin"
__________ Starts timer
__________ Monitors for circling and completion
__________ Stop timer by saying "Stop" after 1 minute
__________ Collects packet

Procedural Integrity Checklist

Group 6

3-minute	1-minute	2-minute

Place a 1 after each step completed and a 0 after each step not completed.

Maze probe assessment procedures

3-minute
__________ makes sure maze probes are in front of students
__________ Set timer to zero
__________ Read maze probe assessment instructions
__________ Starts timer
__________ Monitors for circling and completion
__________ Stop timer by saying "Stop" after 3 minute

1-minute

__________ reads directions
__________ Set timer to zero
__________ Says "Ready, begin"
__________ Starts timer
__________ Monitors for circling and completion
__________ Stop timer by saying "Stop" after 1 minute

2-minute

__________ reads directions
__________ Set timer to zero
__________ Says "Ready, begin"
__________ Starts timer
__________ Monitors for circling and completion
__________ Stop timer by saying "Stop" after 2 minutes
__________ Collects packet

APPENDIX E:

CHILDREN’S ASSESSMENT ACCEPTABILITY RATING SCALE

Children's Assessment Acceptability Rating Scale

	No	Maybe	Yes
1. I like taking CBM-maze probes.	No	Maybe	Yes
2. CBM-maze probes are a good way to see how much I understand what I read.	No	Maybe	Yes
3. My friends would like to take CBM-maze probes.	No	Maybe	Yes
4. I liked the 1st story the best.	No	Maybe	Yes
5. I liked the 2nd story the best.	No	Maybe	Yes
6. I liked the 3rd story the best	No	Maybe	Yes
7. The stories were hard to understand.	No	Maybe	Yes
8. The stories were easy to understand.	No	Maybe	Yes

APPENDIX F:

TABLE OF MEANS BY GROUP AND GRADE LEVEL

Between groups (1-6) and within time frames (1m, 2m, 3m)

means and *interaction effects*

	1-minute means	2-minute means	3-minute means
group 1	7.93	15.60	23.80
group 2	7.43	14.57	23.21
group 3	7.54	12.69	23.08
group 4	6.73	12.27	19.07
group 5	8.60	15.33	19.13
group 6	8.31	14.92	21.77
Total	7.75	14.25	21.62

Between grades (4-5) and within time frames (1m, 2m, 3m)

means and *interaction effects*

	1-minute means	2-minute means	3-minute means
4th grade	8.15	*12.96*	21.22
5th grade	7.28	*15.77*	22.10
Total	7.72	14.36	21.66

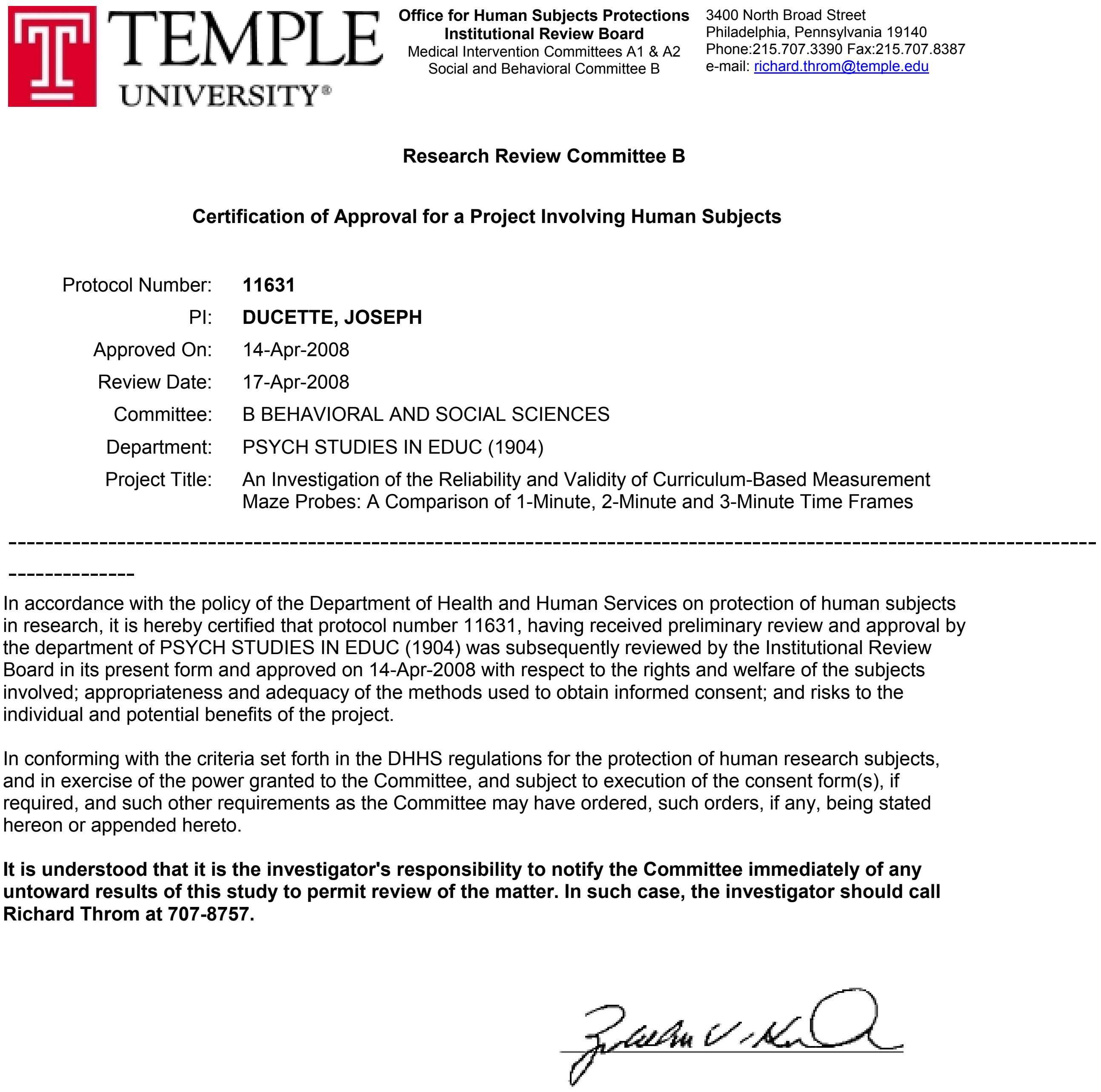
Between grades (4-5) and within time frames (1m, 2m, 3m conversion)

means and *interaction effects*

	1-minute means	2-minute means	3-minute means
4th grade	8.15	*6.48*	7.07
5th grade	7.25	*7.88*	7.37
Total	7.75	7.12	7.21

APPENDIX G:

INTERNAL REVIEW BOARD (IRB) APPROVALS



TEMPLE UNIVERSITY

Office for Human Subjects Protections
Institutional Review Board
Medical Intervention Committees A1 & A2
Social and Behavioral Committee B

3400 North Broad Street
Philadelphia, Pennsylvania 19140
Phone:215.707.3390 Fax:215.707.8387
e-mail: richard.throm@temple.edu

Research Review Committee B

Certification of Approval for a Project Involving Human Subjects

Protocol Number:	**11631**
PI:	**DUCETTE, JOSEPH**
Approved On:	14-Apr-2008
Review Date:	17-Apr-2008
Committee:	B BEHAVIORAL AND SOCIAL SCIENCES
Department:	PSYCH STUDIES IN EDUC (1904)
Project Title:	An Investigation of the Reliability and Validity of Curriculum-Based Measurement Maze Probes: A Comparison of 1-Minute, 2-Minute and 3-Minute Time Frames

---

In accordance with the policy of the Department of Health and Human Services on protection of human subjects in research, it is hereby certified that protocol number 11631, having received preliminary review and approval by the department of PSYCH STUDIES IN EDUC (1904) was subsequently reviewed by the Institutional Review Board in its present form and approved on 14-Apr-2008 with respect to the rights and welfare of the subjects involved; appropriateness and adequacy of the methods used to obtain informed consent; and risks to the individual and potential benefits of the project.

In conforming with the criteria set forth in the DHHS regulations for the protection of human research subjects, and in exercise of the power granted to the Committee, and subject to execution of the consent form(s), if required, and such other requirements as the Committee may have ordered, such orders, if any, being stated hereon or appended hereto.

It is understood that it is the investigator's responsibility to notify the Committee immediately of any untoward results of this study to permit review of the matter. In such case, the investigator should call Richard Throm at 707-8757.

ZEBULON KENDRICK, Ph.D.
CHAIRMAN, IRB

TEMPLE UNIVERSITY®

Office for Human Subjects Protections
Institutional Review Board
Medical Intervention Committees A1 & A2
Social and Behavioral Committee B

3400 North Broad Street
Philadelphia, Pennsylvania 19140
Phone:215.707.3390 Fax:215.707.8387
e-mail: richard.throm@temple.edu

MEMORANDUM

To: **DUCETTE, JOSEPH**
PSYCH STUDIES IN EDUC (1904)

From: Richard C. Throm
Institutional Review Board

Date: 17-Apr-2008

Re: Expedited Request Status for IRB Protocol:
11631: An Investigation of the Reliability and Validity of Curriculum-Based Measurement Maze Probes: A Comparison of 1-Minute, 2-Minute and 3-Minute Time Frames

----

This addendum is to be affixed to the IRB Approval Certificate

45 CFR 46 Protection of Human Subjects.

Expedited review is a type of review that can be conducted by the IRB Chair, other IRB members designated by the Chair, or a subcommittee of the IRB. A major criterion for research that can initially (initial review) reviewed through expedited process is that it must involve no more that minimal risk. The DHHS regulations and FDA regulations define minimal risk to mean that “the probability and magnitude of harm or discomfort anticipated in the research are not greater in and of themselves than those ordinarily encountered in the daily life or during performance of routine physical or psychological examinations or tests.”

This research protocol was reviewed under the following Expedited Review Category:

Expedited Category #7: Research on group characteristics or behavior (including, but not limited to, research on perception, cognition, motivation, identity, language, communication, cultural beliefs or practices, and social behavior) or research employing survey, interview, oral history, focus group, program evaluation, human factors evaluation, or quality assurance methodologies.

TEMPLE UNIVERSITY®

Office for Human Subjects Protections
Institutional Review Board
Medical Intervention Committees A1 & A2
Social and Behavioral Committee B

3400 North Broad Street
Philadelphia, Pennsylvania 19140
Phone:215.707.3390 Fax:215.707.8387
e-mail: richard.throm@temple.edu

MEMORANDUM

To: **DUCETTE, JOSEPH**
PSYCH STUDIES IN EDUC (1904)

From: Richard C. Throm
Institutional Review Board

Date: 17-Apr-2008

Re: Subpart D Status for IRB Protocol:
11631: An Investigation of the Reliability and Validity of Curriculum-Based Measurement Maze Probes: A Comparison of 1-Minute, 2-Minute and 3-Minute Time Frames

----

Addendum to the IRB Approval Certificate - append to certificate
CFR Subpart D: Additional Protections for Children Involved as Subjects in Research

Federal regulations classify permissible research involving minors into four categories, based on degree of risk and type of prospective benefit. These categories are described in relation to "minimal risk".

Minimal risk is defined as "the probability and magnitude of harm or discomfort anticipated in the **research** are not greater in and of themselves from those ordinarily encountered in daily life of during the performance of routine physical or psychological examination or tests.

Greater than minimal risk is a term used in defining Category 2 [45 CFR 46.405] and Category 3 [35 CFR 46.406]. The regulations do not provide any further definition or clarification of this term except for specifying "a minor increase over minimal risk" in regards to Category 3 only. Therefore, the protocol application should clearly describe the study risks so the IRB, in consultation with the investigator's assessment, make an appropriate determination for category of approval.

Operative Definitions.
(a) Children are persons who have not attained the legal age for consent to treatments or procedures involved in the research, under the applicable law of the jurisdiction in which the research will be conducted.
(b) Assent means a child's affirmative agreement to participate in research. Mere failure to object should not, absent affirmative agreement, be construed as assent.
(c) Permission means the agreement of parent(s) or guardian to the participation of their child or ward in research.
(d) Parent means a child's biological or adoptive parent.
(e) Guardian means an individual who is authorized under applicable State or local law to consent on behalf of a child to general medical care.

This protocol was reviewed and approved under category:
Research Category 1: Not greater than minimal risk
Requires:
1. Permission from ONE parent/legal guardian
2. Assent of minor (if child is 7 years of age or older (18)

APPENDIX H:

CONSENT AND ASSENT FORMS

CONSENT FORM

An investigation of the reliability and validity of curriculum-based measurement maze probes: A comparison of 1-minute, 2-minute, and 3-minute time frames.

Student Investigator, Israel A. Sarasti, Ed.D., Temple University Department of
Psychological Studies in Education, 610.938.9000 x2271
Principal Investigator, Joseph DuCette, Ph.D., Temple University Department of
Psychological Studies in Education, 215.207.7926

Purpose of Study

You are being asked to allow your child to participate in a study about reading comprehension. The study will test three brief multiple choice reading quizzes. Each quiz will be given in 1-minute, 2-minute, and 3-minute time frames. Each student will also take a computer-based norm referenced reading test. The results of the quizzes and reading test will be compared. Previous research has shown that that short reading quizzes can measure general reading comprehension as well as the norm referenced reading tests.

Subject Selection

Students in the 4th and 5th grades at St. Denis will be asked to participate. All the students can participate if they wish. Verbal permission to conduct the study has been given by Mrs. Coccia the school principal.

Procedures

Students will be administered the three reading quizzes during their integrated language arts (ILA) class time. Fifteen students will be tested at one time by Dr. Sarasti. The other students will remain in their class receiving their regularly planned instruction. The process will take approximately 15 minutes of instructional time per group.

The computer-based test will be given during the student's computer lab time. It will take approximately 15 minutes to read and answer 25 multiple choice questions. This test will be given the same week of the short quizzes.

At the end of the week, the students will fill out an eight question survey about the three reading quizzes.

Possible Risks

There are no foreseeable risks involved in this study.

Benefits

The expected benefits of the study are the following:

- Additional practice in reading comprehension skills
- Building reading comprehension fluency
- Additional practice taking standardized reading tests (i.e., Terra Nova reading section)

Confidentiality/Anonymity

Your child's individual information will be kept strictly confidential according to federal, state, and local laws/regulations. The child's name or any traceable identifying information will not used. Instead, each student will be coded in random order such as student 1 (S1), student 2 (S2). Then, they will be entered in the data base as S1, S2, S3, etc. The signed consent/assent forms, the brief quizzes, and computer reading test results will be kept in a locked filing cabinet in Dr. Sarasti's office.

The file and information from the study may be reviewed by the University's Institutional Review Board or federal agencies to make sure investigators are doing the study properly and following federal regulations. The results of this study may be presented or published.

This is independent research separate from St. Denis School. Your child's performance in the study will not affect his grades in any way. The results may be shared with Ms. Coccia, the school principal under coded format only. This will preserve student and teacher anonymity.

Disclaimer/Withdrawal

I am free to decide if my child participates in this study. Even if I give consent, my child is free to decide to participate. Not participating in the study or dropping out will not be held against me in any way by the investigator or by Temple University.

Institutional Contacts

If there are any further questions or want further information about my child's rights as a participant in the study, you can contact Mr. Richard Throm, Institutional Review Board Manager & Coordinator in the Office of the Vice President for Research of Temple University, 3400 N. Broad Street, Philadelphia, PA 19140, (215)-707-8757.

Final Statement and Signatures

An investigation of the reliability and validity of curriculum-based measurement maze probes: A comparison of 1-minute, 2-minute, and 3-minute time frames.

Your signature below indicates that you have read or have had read to you all of the above and that you confirm all of the following:

- **Israel A. Sarasti, Ed.D.** has explained the study to you and answered all of your questions. You have been told the possible benefits and the potential risks and/or discomforts of the study.
- You understand that you do not have to allow your child to take part in this study, and your refusal to allow your child to participate or your decision to withdraw him/her from the study will involve no penalty or loss of rights or benefits. The study personnel may choose to stop your child's participation at any time.
- You understand why the study is being conducted and how it will be performed.
- You understand your rights as the parent/guardian of a research participant and you voluntarily consent to your child's participation in this study.
- You have been told you will receive a copy of this form.

Child's name (please print)　　Date

Parent/Guardian's name (please print)　　Signature　　Date

Principal Investigator's name (please print)　　Signature　　Date

ASSENT FORM

An investigation of the reliability and validity of curriculum-based measurement maze probes: A comparison of 1-minute, 2-minute, and 3-minute time frames.

You are being asked to be part of a research project being done by the Temple University Department of Psychological Studies in Education.

This study involves reading comprehension. Dr. Sarasti wants to find out how well three quizzes compare to a computer reading test. The quizzes will be 1-minute, 2-minutes, or 3-mimutes long. The quizzes are short 1 page stories with fill in the blanks. The comparison reading test is taken on the computer. It has 25 multiple choice questions. Dr. Sarasti wants to see how well you understand what you read on the quizzes and computer reading test.

You will be asked to take the 3 quizzes during your ILA time. You will take the quizzes in groups of 15 students. Dr. Sarasti will randomly choose the groups. The quizzes will only take 15 minutes of your ILA time. If you are not taking the quizzes in the testing room, you will be in class with your teacher doing your regular class work for that day. All of the groups will be tested on the same day.

The computer reading test will be given during your computer lab time. It will take about 15 minutes for the 25 questions. The reading test will be given the same week as the 3 quizzes. At the end of the week, you will fill out a survey with 8 questions. This will tell me how you feel about the quizzes.

If you want to be part of this study, please remember you can stop doing it any time you want to. If you would like to do the study, please sign your name below.

Child's Statement and Signature of Consent
I understand what I am being asked to do for this study.
I can ask to stop at any time if I want.
I agree to participate in this study.

____
Child's name (please print)

____
Child's signature

____
Today's Date

VITA

Israel Antonio Sarasti was born in Miami, Florida on November 13, 1969. He was raised in Miami and is a product of the Miami-Dade County Public Schools. He attended New World School of the Arts (NWSA) high school music program under the direction of John de Lancie. Upon high school graduation (1988), he continued his undergraduate studies in music performance in the conservatory division of NWSA with John de Lancie and Christine D. Nield as his teachers and mentors. A Bachelors of Music degree in performance (flute and opera) was received in 1994. He worked as a chorister for the Florida Grand Opera from 1992 through 2004. From 1996 to 2004 he was a teacher with the Miami-Dade County Public School teaching elementary music and special education. In 1997, he entered graduate school at Florida International University College of Education and went on to receive a Masters of Science degree in Special Education (2000). Subsequently, he enrolled in a Specialist in School Psychology degree program at Capella University. The Psy.S. in School Psychology was completed and school psychology certification was attained in 2005. From 2004-2006 he worked for Dallas Independent School District as a licensed specialist in school psychology. While living in the Dallas, TX area enrolled in a Doctor of Education program at University of North Texas. In 2007, he completed that program and attained the Doctor of Education degree with concentrations in Curriculum and Instruction, Applied Behavior Analysis and Gifted Education. In 2006, he was accepted and began studies for the Doctor of Philosophy degree in School Psychology at Temple University in Philadelphia, PA. He currently resides in Philadelphia and works as a bilingual school psychologist for the Delaware

County Intermediate Unit. He will complete a doctoral internship at the LSU Human Development Center LAS*PIC program for the 2009/2010 school year in order to fulfill the requirements for the PhD degree in School Psychology at Temple University (2010).