ORAL READING FLUENCY AND MAZE MEASURES AS PREDICTORS OF PERFORMANCE ON NORTH CAROLINA END-OF-GRADE ASSESSMENT OF READING COMPREHENSION

by

Tara Watkins Galloway

A dissertation submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Special Education

Charlotte

2010

Approved by:

_____
Dr. LuAnn Jordan

_____
Dr. Claudia Flowers

_____
Dr. Chris O'Brien

_____
Dr. John Beattie

_____
Dr. Susan Sell

ABSTRACT

TARA WATKINS GALLOWAY. Oral reading fluency and maze measures as predictors of performance on North Carolina end-of-grade assessment of reading comprehension. (Under direction of DR. LUANN JORDAN)

Current legislation (IDEA, 2004; NCLB, 2001) mandates all students, including students with disabilities, demonstrate progress toward the same standards. However, students continue to struggle with attainment of statewide academic standards as measured by high-stakes assessment. The purpose of the current study was to examine the degree that Dynamic Indicators of Basic Early Literacy Skills Oral Reading Fluency (DIBELS ORF) and Maze Curriculum-Based Measures (AIMSweb Maze-CBM) predict standard scores on the North Carolina End-of-Grade (EOG) Assessment of Reading Comprehension. The study also investigated differences in the relationship as a function of grade, examined the accuracy of established cutoff scores, and determined optimal cut scores. Participants included 336 students in third, fourth, and fifth grades. Results of the study were consistent with previous research, indicating the significance of fluency measures for determining the likelihood of proficiency on high-stakes assessments. Findings indicated ORF and Maze measures significantly predicted proficiency, with ORF accounting for the most variance in EOG scores. Receiver Operating Characteristic (ROC) Curves revealed statistically significant Area Under the Curve (AUC) values for ORF and Maze. Sensitivity levels were adequate for recommended cutoff values; specificity levels were less than adequate. Optimal cutoff scores to maximize sensitivity and specificity yielded slightly different cutoff points for ORF and Maze. Implications for practice, limitations, and suggestions for future research are provided.

DEDICATION

I would like to dedicate this dissertation to my family. Most importantly, this is dedicated to my husband, Dewey, who has been my source of strength throughout this entire journey. From the time I applied to the program, you have inspired me, given me hope, and motivated me to give one hundred percent. You have made it all possible, despite the many sacrifices. I appreciate you sticking by me and often times leading the way with your steadfast love and commitment. Thank you for being my rock and my salvation during this process.

I would also like to dedicate this dissertation to my children, Zane and Kailey, who have been an incredible source of inspiration for me. No matter what, you continued to encourage and motivate me. I never could have accomplished this without you. I will never forget the day you both looked up at me with your precious eyes and said, "You can do it, mom!" From that day forward, I was determined to make you proud. Thank you for helping me to make my dream come true. I hope that both of you will always hold on to your dreams and never give up.

Finally, this process and final product is dedicated to my parents. It is impossible to thank you for the many values that you have instilled in me through the years. Thank you for your constant belief in my capabilities. I have achieved goals that I never thought possible. You planted the seed and let it grow. Thank you for allowing me to bloom. I am grateful for opportunities you provided for me. Daddy, though you have gone on to eternal life in Heaven, your memory has served as a constant reminder to be persistent and to keep my focus. Your beautiful spirit lives on and will remain in the eyes of your children and grandchildren forever. I will always remember that I am a "Watkins."

ACKNOWLEDGEMENTS

I would like to thank so many people who have been instrumental in helping me through this process in pursuit of one of my lifelong dreams. First, I would like to thank my advisor and dear friend, Dr. LuAnn Jordan, for being there when I needed support, confidence, guidance, or motivation to complete each step along the way. I am so grateful that God saw fit for you to be a part of my life. You have given me personal and professional growth opportunities beyond expectation. I realize that this process has demanded a great deal of time, effort, and sacrifice. I am grateful that I had my friend and role model by my side the entire time. Second, I would like to thank the members of my dissertation committee, Dr. Claudia Flowers, Dr. John Beattie, Dr. Chris O'Brien, and Dr. Susan Sell for continuous support, valuable feedback, and expertise throughout my dissertation. I greatly appreciate the time you invested in this endeavor on my behalf. All of you have gone above and beyond the call of duty to help me to reach my full potential.

Next, I would like to say a special "thank you" to Dr. Nancy Cooke and Dr. Charlie Wood for taking me under your wings and sacrificing your time to help me to grow as a researcher and teacher throughout the program. To Dr. Diane Browder, I would like to say thank you for academic guidance and expert advice as a valued member of my portfolio committee. To Dr. Dave Test, Dr. Richard White, Dr. Wendy Wood, Dr. Fred Spooner, Dr. Gloria Campbell-Whatley, Peggy Moore, and my other professors, I am thankful for your level of knowledge and expertise that inspired me to be successful. A special thanks to Sara and the many other friends who have completed this adventure along with me. Your friendship, knowledge, and collaboration made the journey much easier. I cannot imagine going through this process without you.

To my immediate and extended family members, thank you for your optimism, love, commitment, and encouragement throughout the entire program. Even when you thought I was crazy, you supported me in every way possible. To my siblings, Randy, Tonya, and Tieka, a special thanks to you and your families for helping me to maintain my sanity. Sometimes when you did not even realize it, you helped me to persevere, especially while Dewey was overseas in Kuwait. Sincere gratitude to Tieka for the many hours spent supporting my efforts. To Dewey's family members, especially Billy, thank you for the ongoing commitment and assistance that you offered, without hesitation. Without all of you, my goal would still be just a dream. You have all supported me in so many ways. To my mom, I love you more than words can say. You made it a priority to invest time and energy into molding each of your children into independent, strong individuals. I know that Daddy is proudly smiling down from Heaven right now.

To my Ida Rankin Family, thank you all for believing in me. To Mr. Foulk, Kristi, Janet, Lakena, Laurie, Dana, Melanie, Shelly, and Leslie, I never could have done this without your consistent support. I am truly blessed to have friends who took on the extra load when I could not do everything expected of me. You encouraged me when things were difficult, and loved me unconditionally. I would like to say a very special "thank you" to my principal, Mr. Ron Foulk, for helping me to overcome my obstacles. We are all climbing our own mountain. Thank you for helping me to reach the top of mine.

Most importantly, I would like to praise God. It was only through the grace of our Lord that I was able to accept and meet this challenge. I thank God each day for giving me the opportunity to grow professionally and meet the needs of all students, especially those with special needs. Through God, everything is possible!

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AIMS	Arizona Instrument to Measure Standards
AUC	Area Under the Curve
BST-R	Basic Standards Test – Reading
CAT	California Achievement Test
CBM	Curriculum-Based Measurement
CSAP	Colorado State Assessment Program
DIBELS	Dynamic Indicators of Basic Early Literacy Skills
ECT	Elementary Cognitive Tasks
ELL	English Language Learners
FCAT – SSS	Florida Comprehensive Assessment Test – Sunshine State Standards
GMORF	Growth Modeling Oral Reading Fluency Passages
GRA+DE	Group Reading Assessment and Diagnostic Evaluation
HLM	Hierarchical Linear Modeling
ISAT	Illinois Standards Achievement Test
ISF	Initial Sound Fluency
ITBS	Iowa Test of Basic Skills
K-BIT	Kaufman Brief Intelligence Test
KTEA	Kaufman Test of Educational Achievement
LNF	Letter Naming Fluency
MAEP	Michigan Educational Assessment Program
MAT	Metropolitan Achievement Test
MCA	Minnesota Comprehensive Assessment

MSAP	Maryland School Performance Assessment
NPV	Negative Predictive Value
NCDPI	North Carolina Department of Public Instruction
NC EOG	North Carolina End-of-Grade
NWF	Nonsense Word Fluency
OAKS	Oregon Assessment of Knowledge and Skills
OPT	Ohio Fourth Grade Reading Proficiency Test
ORF	Oral Reading Fluency
OSA	Oregon Statewide Assessment
OSRA	Oregon Statewide Reading Assessment
PDE	Phoneme Decoding Efficiency
PMRN	Progress Monitoring and Reporting Network
PORA	Passage Oral Reading Assessment
PPV	Positive Predictive Value
PSF	Phonemic Segmentation Fluency
PSSA	Pennsylvania System of School Assessment
R-CBM	Reading Curriculum-Based Measure
ROC	Receiver Operating Characteristic
RTF	Retell Fluency
SAT-9	Stanford Achievement Test - 9
SAT-10	Stanford Achievement Test -10
SDRT	Stanford Diagnostic Reading Test
SEM	Structural Equation Modeling

SSRS	Social Skills Rating System
SWE	Sight Word Reading Efficiency
Terra Nova	TerraNova California Achievement Test
TNF	True Negative Fraction (specificity)
TPF	True Positive Fraction (sensitivity)
WASL	Washington Assessment of Student Learning
WCPM	Words Correct Per Minute
WDRB	Woodcock Diagnostic Reading Battery
WISC-IV	Wechsler Intelligence Scale for Children
WJ-R	Woodcock Johnson Psychoeducational Battery - Revised
WJ-III	Woodcock-Johnson III Test of Achievement
WRA	Word Recognition Assessment
WRMT-R	Woodcock Reading Mastery Tests-Revised
WUF	Word Use Fluency

CHAPTER ONE: INTRODUCTION

National Reading Crisis

Although the growing importance of ensuring that all students meet grade level standards has been recognized (NCLB, 2001), educators in schools today are faced with many issues concerning reading with all students, especially students with special needs. Despite mandates in the educational systems across the nation (IDEA, 2004; NCLB), students in classrooms continue to face barriers in learning to read; therefore, many struggle with attainment of statewide academic standards. The most recent statistics are disturbing since they reveal that only 32% of fourth grade students across the nation are able to read at the proficient level (National Center for Educational Statistics, 2007).

Compelling findings indicate that students who fail to read early fall farther behind, creating a literacy gap that widens as the students get older (Stanovich, 1986). Research suggests that students with poor early reading skills are likely to have poor reading later (Good, Simmons, & Kame'enui, 2001). In a longitudinal study, Juel (1988) found 88% probability that a child who is a poor reader in first grade will be a poor reader at the end of fourth grade. Furthermore, when students fail to meet grade level expectations by third grade, they are likely to continue struggling to catch up with the standards, as 74% of children who are poor readers in third grade remain poor readers in ninth grade (Francis, Shaywitz, Stuebing, Shaywitz, & Fletcher, 1996).

Recent legislation focuses on providing quality education to all students, with and without disabilities, and has prioritized academic achievement of students in our nation. The passage of the No Child Left Behind Act (NCLB, 2001) and Individuals with Disabilities Education Act (IDEA, 2004) resulted in increased accountability for ensuring that all students demonstrate progress toward the same standards. Based on concerns relating to the academic achievement of students and the emphasis of no child being excluded from or left behind the general curriculum, current legislation includes requirements to use scientifically-based instruction and mandates implementation of statewide systems of accountability. According to NCLB, all students must be reading on grade level in third grade by the year 2014. Unfortunately, as evidenced by Adequate Yearly Progress (AYP) results, students struggle to meet expected growth on standardized End-of-Grade (EOG) tests with only 70% of schools in the United States currently making AYP (National Center for Educational Statistics, 2007).

Initiatives to Improve Student Achievement in Reading

A report released by the U.S. Department of Education National Commission on Excellence in Education, *A Nation at Risk* (1983), indicated 23 million adults in America were unable to complete the simple tests of everyday reading, writing, and comprehension. This publication began a wave of reform initiatives aimed at raising standards and outcomes for students. Since then, progress made in understanding how children learn to read (Adams, 1990; Snow, Burns, & Griffin, 1998; NRP, 2000) has prompted changes in beginning reading instruction (Cowen, 2003). Ongoing research efforts have demonstrated the importance of responsive instructional supports to accelerate reading progress (Chard et al., 2008) and established what works to help

students become proficient readers. Consensus reports have documented the critical components of reading instruction and emphasized the importance of including these components in daily instruction. For example, Anderson, Hiebert, Scott, and Wilkinson (1985) emphasized the value of automatic word recognition in a report entitled *Becoming a Nation of Readers: The Report of the Commission on Reading.*

Following a research synthesis report entitled *Preventing Reading Difficulties in Young Children* (Snow et al., 1998), congress mandated the largest, evidence-based review ever conducted on how children learn to read. The National Reading Panel (NRP, 2000) was developed in an attempt to raise student outcomes. Members of the Panel were charged with reviewing more than 100,000 research studies (Armbruster, Lehr, & Osborn, 2001). Using rigorous research standards, the Panel conducted an assessment of effective approaches to teach children to read and provided information about reading development. Based on findings, the panel identified the empirically validated foundational skills referred to as the "big ideas" in reading (Good et al., 2001). These "big ideas" of reading were found as the skills necessary to include when teaching reading, including (a) phonemic awareness, (b) phonics, (c) vocabulary, (d) text comprehension, and (e) fluency. Children need to gain these important skills in order to become independent readers. With these findings, fluency was recognized as "one of several critical factors necessary for reading comprehension" (NRP, 2000, p. 11).

With knowledge of critical components in reading, there is considerable evidence that student achievement in reading is alterable (Coleman, Buysse, & Neitzel, 2006; Denton, Fletcher, Anthony, & Francis, 2006; National Joint Committee on Learning Disabilities, 2005). Educators face increased accountability for student performance, as

measured by high-stakes assessments. Students with disabilities are expected to demonstrate the same knowledge of standards and reach the same level of achievement as students without disabilities. For students who are at risk for not meeting minimal acquisition of skills on statewide high-stakes testing, educators need effective tools to gauge student progress toward expected state curriculum goals and make effective instructional decisions to change learning trajectories.

Assessment of Student Performance

With federal mandates that all students will make progress (IDEA, 2004; NCLB, 2001), it is vital for educators to recognize that reading assessment can help to improve learning if it occurs at stages other than at the end of the learning cycle (Kennedy, Chan, Fok, & Yu, 2008). There has been growing concern with summative assessments that measure student knowledge at one point in time, primarily because these types of assessments do not influence student learning (Kennedy et al.). However, federal mandates such as NCLB (2001) require summative assessment for all students for accountability purposes. This form of assessment restricts the amount of feedback and practice (Fuchs, Fuchs, Compton, Bouton, Caffrey, & Hill, 2007) because it is not acceptable for testing administrators to interact with the student during testing (Caffrey, Fuchs, & Fuchs, 2008). Actually, it is believed by some that summative assessment practices often influence student learning in negative ways (Biggs, 1998). There is also concern that summative assessment practices actually cause some students to give up (Stiggins, 2002).

Testing is the centerpiece of current education policy and is at the heart of NCLB; however, standardized testing cannot be used to tailor instructional decision making

(Pressley, 2006). In contrast, using formative assessment data gathered on a regular basis, educators can make future educational decisions and guide the course of instruction when changes are needed. When classroom measures are reliable predictors of progress toward achieving grade-level reading skills (Schilling, Carlisle, Scott, & Zeng, 2007), data collected can provide valuable information to track student progress toward valued learning outcomes.

Curriculum-Based Measurement – Oral Reading Fluency (CBM-ORF). Curriculum-based measurement (CBM) procedures are brief, repeatable fluency measures that assess a broad range of academic skills reflecting end-of-year goals. Reading curriculum-based measurement (R-CBM: Shinn, 1989) is a widely accepted, empirically valid and reliable index of reading and has been identified as a strategy for monitoring yearly progress (U.S. Department of Education, 2002). Oral reading fluency (ORF) rate has been found useful as a method for monitoring overall reading growth (Fuchs, Fuchs, & Maxwell, 1988). Since ORF correlates highly with reading comprehension, reading progress can be monitored using curriculum-based measurement oral reading fluency (CBM-ORF) measures (Deno, 1985; Fuchs & Fuchs, 1992; Parker, Hasbrouck & Tindal, 1992).

Of all curriculum-based measures to assess skills, measures of oral reading fluency have the most theoretical and empirical support (Marston, 1989). With oral passage reading measures, students read a passage of approximately 250 words under timed conditions, while examiners score words correct and errors per minute. The CBM practice of timed oral reading is an effective tool for educators to monitor growth

(Marston & Magnusson, 1985) and to adjust instruction in a system of formative assessment (Fuchs, Tindal, & Deno, 1981).

Affirming the significance of oral reading fluency (ORF) measures, Deno (1985) stated ORF can be "used as a 'vital sign' of reading achievement in much the same sense that heart rate or body temperature is used as a vital sign of physical health" (p. 224). Evidence on reliability of CBM-ORF is positive (Mehrens & Clarizio, 1993) with strong relations between CBM-ORF and comprehension (Deno, Mirkin, & Chaing, 1982; Fuchs, et al., 1988; Yovanoff, Duesbery, Alonzo, & Tindal, 2005) and high reliability for oral reading proficiency of students (Hintze, Owen, Shapiro, & Daly, 20000; Marston, 1989; Shinn, Good, Knutson, & Tilly, 1992). Additionally, some studies have ruled out general cognitive ability or processing speed and efficiency (Kranzler, Brownell, & Miller, 1998) as well as bias with respect to ethnicity or socioeconomic status (Hintze, Callahan, Matthews, Williams, & Tobin, 2002) as factors in oral reading ability.

Theoretical frameworks for understanding the process of reading provide a basis for conceptualizing ORF as an indicator of overall reading competence (Fuchs, Fuchs, Hosp, & Jenkins, 2001). The automaticity model of reading by LaBerge and Samuels (1974) prompted a theory that if competent skills are developed in a short time frame, it allows attention to be reallocated to more complex comprehension functions. In this model, components of reading are executed automatically and higher processes wait for lower ones to develop (LaBerge & Samuels). This theory is based on the assumption that reading development requires the ability to recognize words efficiently at a lower level, which frees attention needed to process text for comprehension (LaBerge & Samuels).

This model of reading validates fluency as a valid indicator of word skills and comprehension of text.

Among the first studies conducted to determine the relationship among measures of reading performance and achievement, Deno et al. (1982) found usefulness in formative measures for continuous evaluation of student growth. In the same year, Fuchs, Fuchs, and Deno (1982) validated CBM in a study that examined the technical adequacy of three informal reading inventory procedures. Also, in one of the largest studies to compare CBM-ORF with standardized assessment, data from United States Department of Education Office of Educational Research and Improvement revealed that fourth grade students with higher fluency rates had higher reading proficiency on the 1992 National Assessment of Educational Progress (NAEP) integrated reading performance record (Pinnell, Pikulski, Wixson, Campbell, Gough, & Beatty, 1995).

The correlational relationship between CBM-ORF and statewide, standardized, high-stakes assessments has been investigated by a number of researchers in various states (Buck & Torgesen, 2003; Crawford, Tindal, & Stieber, 2001; Hintze & Silberglitt, 2005; McGlinchey & Hixon, 2004; Shapiro, Keller, Lutz, Santoro, & Hintze, 2006; Sibley, Biwer, & Hesch, 2001; Silberglitt & Hintze, 2005; Stage & Jacobsen, 2001). For example, Crawford et al. (2001) used scores on CBM-ORF to predict statewide test performance in Oregon for reading and math. Also, in Washington, Stage and Jacobsen (2001) found a .44 correlation between ORF and the reading section of the standardized state assessment in Oregon. Sibley et al. (2001) reported CBM data identified those who did not meet established standards on the Illinois State Assessment. Buck and Torgesen (2003) found ORF predicted whether students attained proficiency on Florida

Comprehensive Assessment Test. In Minnesota, Hintze, and Silberglitt (2005) found performance on CBM as an accurate predictor of students who are likely to pass the reading portion of the state assessment. In another study in Minnesota, Silberglitt and Hintze (2005) set CBM-R cut scores to determine whether students were on track to pass third grade achievement tests. Finally, Shapiro et al. (2006) found moderate to strong correlations with mid-year ORF assessment in reading and Pennsylvania high-stakes test.

Dynamic Indicators of Basic Early Literacy Skills (DIBELS). Dynamic Indicators of Basic Early Literacy Skills (DIBELS; Kaminski & Good, 1996) are curriculum-based assessments developed by researchers at the University of Oregon which employ fluency measures. The DIBELS system meets the stringent criteria for endorsement of Student Progress Monitoring. DIBELS includes measures identified as critical to early development in reading and provides benchmarks in order to determine whether students are making progress toward grade level reading goals. Appropriate levels of reliability and validity for screening, monitoring progress, and evaluating the outcomes of instructional programs have been established for DIBELS (Good, Gruba, & Kaminski, 2002). Several reports of research have shown significant relationships between DIBELS Oral Reading Fluency (ORF) and year-end reading comprehension assessments (e.g., Barger, 2003; Good et al., 2001).

By examining correlations, Shaw and Shaw (2002) investigated the relationship between DIBELS ORF and Colorado State Assessment Program (CSAP) and found 90% of the students who scored at the benchmark goal on spring ORF scored proficient on CSAP. Similarly, in Oregon, Good et al. (2001) found that 96% of students in third grade who scored above 110 correct words per minute on ORF met expectations on the Oregon

Statewide Assessment. Vander Meer, Lentz, and Stollar (2005) conducted a study in Ohio with fourth grade students and reported that 72% of the students who met the DIBELS ORF benchmark in third grade passed the Ohio Reading Proficiency test. Likewise, Wilson (2005) used DIBELS ORF to identify students likely and unlikely to meet proficiency on the Arizona Instrument to Measure Standards (AIMS). Wood (2006) used hierarchical linear modeling to investigate the relationship between DIBELS ORF and used efficiency statistics with cut scores to predict pass/fail on the Colorado statewide assessments. In Michigan, Schilling et al. (2007) examined predictive validity of DIBELS and found DIBELS significantly predicted year end achievement on Iowa Test of Basic Skills (ITBS).

Recently, Baker et al. (2008) found slope on DIBELS ORF added to the accuracy of predicting performance above level of performance. In Florida, Roehrig, Petscher, Nettles, Hudson, and Torgesen (2008) examined predictive validity and found the third administration of DIBELS ORF was the strongest correlation with performance on Florida state assessment (FCAT-SSS) and SAT-10. Shapiro, Solari, and Petscher (2008) added a reading comprehension measure (4Sight Benchmark) to DIBELS and found the combination of the two measures was the best predictor of performance on the Pennsylvania state assessment. Catts, Petscher, Schatschneider, Bridges, and Mendoza (2009) found second grade ORF level predicted outcomes on third grade SAT-10. Most recently, Goffreda, Diperna, and Pedersen (2009) found DIBELS ORF predicted later reading proficiency.

Maze Curriculum-Based Measurement (Maze-CBM). Another curriculum-based assessment currently used in schools is *AIMSweb Maze Curriculum-Based Measurement*

(Maze-CBM; Edformation, 2009) using *Edformation's Standard Reading Maze Passages.* AIMSweb Maze-CBM has been proven to be a reliable and valid measure of student's reading comprehension skills. Maze is a timed, multiple-choice cloze task in which the student completes the passage by choosing the correct word from three choices given in parenthesis. Maze is administered using standardized directions. Recently, Maze-CBM was evaluated by members of the National Technical Review Committee (TRC) of the National Center on Response to Intervention (NCRTI, 2009). The Maze measure fully met all seven standards giving it the highest rating possible for predictive validity and reliability.

Despite strong predictive validity and reliability, few researchers have studied the addition of Maze-CBM to ORF for prediction of outcomes on statewide assessments (Ardoin et al., 2004; Silberglitt, Burns, Madyun, & Lail, 2006; Wiley & Deno, 2005). However, results are promising for adding a Maze measure of comprehension for prediction of performance on statewide assessments. Ardoin et al. examined the contribution of Maze in addition to CBM using 77 students in third grade. Using hierarchical multiple regression and simultaneous regression, researchers found CBM and Maze had high correlations with reading achievement and comprehension. However, CBM was a better predictor at overall reading achievement than Maze.

In another study to determine whether the Maze procedure adds to the predictive power of general outcome measures of oral reading on state assessments for ELL students, Wiley and Deno (2005) found adding the measure of comprehension aided in predictive performance of ORF with non-ELL students on the Minnesota Comprehensive Assessment in Reading. In fact, results indicated the Maze task was a better predictor

than oral reading for non-ELL students in fifth grade, while oral reading was slightly better than the Maze task for EL students in third and fifth grade. Findings suggested CBM and Maze procedures can be used in assessment of English language learners reading proficiency. Researchers suggested future research on the potential of Maze and oral reading measures for identifying students who are at risk for failing state assessments.

In Minnesota, Silberglitt et al. (2006) analyzed the relationship between curriculum-based measures (CBM and Maze) and state accountability tests as a function of grade level. Data for 5,472 students in third, fifth, seventh, and eighth grades from five rural or suburban districts in Minnesota were correlated to test scores on the Minnesota Comprehensive Assessments-Reading (MCA-R). Results indicated coefficients for all grade levels met or exceeded .50 and R-CBM and Maze were both significantly related to state accountability test scores. Maze accounted for 24% to 29% of variance between CBM and state test scores on the MCA, with the overall value of prediction diminishing significantly as grade level increased. Researchers suggested further empirical investigation to explore the decline in predictive power of CBM in later grades and careful consideration of establishing target scores or introduction of additional assessment tools.

Significance of Study

Many studies exist examining the correlation of ORF and statewide high-stakes tests. Studies have been conducted in many states related to ORF scores as a predictor of performance on standardized testing. However, there is limited research related to using DIBELS Oral Reading Fluency (ORF) measures as a predictor of achievement on high-

stakes, standardized tests and identification of optimal cut scores to predict a proficient score. Additionally, no research was found related to using a combination of DIBELS (ORF) and Maze measures as predictors of performance of End-of-Grade Reading Comprehension.

Since there is wide-scale use of these measures in schools across the nation, there is a need to specifically address which measure or combination of measures most accurately predict proficiency on high-stakes assessments. There is also a need to determine the optimal cut score to use in predicting which students are at risk for not meeting proficiency by the end of the grade. With this information, educators can gauge student progress toward standards of proficiency by using DIBELS ORF and Maze-CBM probes as formative assessments to facilitate more accurate instructional decisions.

Educators are held accountable for student performance on high-stakes testing in reading. Current legislation mandates that each child, with or without a disability, demonstrate progress toward meeting the same standards using a statewide system of accountability. As a result of this widespread adoption of statewide tests and the importance placed on test results, academic progress of students needs to be closely monitored through the use of other measurement systems that are available more frequently (i.e., formative assessment). Research on the utility of formative curriculum-based assessments to predict performance on the high-stakes test of reading comprehension is necessary for educators to improve instructional programs (Crawford et al., 2001) by making instructional decisions based on data gathered from assessments that occur at stages other than the end of the learning cycle (Kennedy et al., 2008).

Therefore, this study examined the relationship between a measure of oral reading fluency and an additional measure of reading comprehension to the statewide, high-stakes reading assessment in North Carolina. Specifically, this study investigated the degree that Dynamic Indicators of Basic Early Literacy Skills (DIBELS) Oral Reading Fluency (ORF) scores and AIMSweb Maze curriculum-based measurement (Maze-CBM) comprehension scores predicted standard scale scores on the North Carolina End-of-Grade (EOG) Test of Reading Comprehension. The study also examined whether grade level differences existed in the magnitude of the relationship, investigated the accuracy of established DIBELS benchmark cutoff scores, and determined optimal cut scores to predict proficiency on the statewide assessment of reading comprehension for third, fourth, and fifth grades.

Research Questions

1. Using DIBELS Oral Reading Fluency (ORF) and AIMSweb Maze-CBM universal screening scores, which measure or combination of measures are the best predictors of standard scale scores on a state developed reading accountability measure for third, fourth, and fifth grade students?
2. Is there a difference in the magnitude of the relationship between EOG and ORF and Maze among third, fourth, and fifth grade?
3. How accurate are published DIBELS ORF risk level cutoff scores for ORF and AIMSweb Maze scores for identifying third, fourth, and fifth grade students who will or will not be proficient as measured by the statewide grade level NC EOG Reading Comprehension test?

4. What are the optimal DIBELS ORF and AIMSweb Maze-CBM cut scores to use when attempting to predict satisfactory reading comprehension by the end of third, fourth, and fifth grade level as measured by EOG performance?

Definitions

This section includes terms used through the study and their definitions. The terms are critical for understanding the procedures and generalizing the results of study.

Automaticity - The ability to perform a task while devoting little attention to the task. Examples of typical "automatic" behaviors include driving, typing, and reading (LaBerge & Samuels, 1974).

Bottom-Up Model - Early cognitive model of reading which depicts cognitive processing of information proceeding from lower to higher order stages. In a "bottom-up" model, the progression of reading would be identifying letters, followed by attaching sounds to letters, then identifying words, followed by processing the word meaning, and finally understanding the meaning of the sentence (Tracey & Morrow, 2006).

Comprehension - Comprehension was identified by the NRP as one of the five components of reading. Comprehension refers to an active process of reading and understanding through the interaction between the reader and the text (NRP, 2000).

Curriculum-Based Measurement - Curriculum-based measurement is simple, standardized, short-duration fluency measures of reading, spelling, written expression, and mathematics computation (Deno, 1985).

Curriculum-Based Assessment - Curriculum-based assessment (CBA) is any set of measurement procedures that use direct observation and recording of a student's

performance in the local curriculum as the basis for gathering information to make instructional decisions (Deno, 1985).

Fluency - Fluency was identified by the NRP as one of the five components or "big ideas" of reading. Fluency is referred to as one of several critical factors needed to improve reading comprehension. Fluency measures a student's speed, accuracy, and expression when reading (NRP, 2000).

Formative Assessment - Formative assessment is conducted to enable learning and is "carried out during the instructional process for the purpose of improving teaching or learning" (Shepard, 2006, [p. 627]).

General Outcome Measures (GOM) - Assessment tools that can function as an index of student progress through the curriculum over time. GOMs are standardized measures that provide educators with information to guide instruction, based on student performance (Deno, 2003).

National Reading Panel (NRP) - The panel identified five components of reading instruction: (a) phonemic awareness, (b) phonics, (c) fluency, (d) vocabulary, and (e) comprehension. The panel was comprised of 14 individuals selected to search for effective early reading strategies found in scientific-based research. The individuals consisted of "leading scientists in reading research, representatives of colleges of education, reading teachers, educational administrators, and parents" (NRP, 2000 [p.1]).

Progress Monitoring - Progress monitoring is a method of keeping track of student's academic development using technically adequate measures. Progress monitoring requires frequent collection of data, interpretation of the data, and changes to instruction based on the results (Speece, 2007).

Prosody - The phonological subsystem that encompasses the tempo, rhythm, and stress of language (Whalley & Hansen, 2006). Reading with appropriate grouping of words, pause in appropriate places, use appropriate intonation, and express text theme and coherence with few mispronunciations, hesitations, and false starts.

Sensitivity - The proportion of true positives correctly identified as positives (Swets, 1988).

Specificity - The proportion of true negatives correctly identified as negatives (Swets, 1988).

Summative Assessment - This type of assessment is conducted after learning in order to document achievement and mastery. Summative assessment is "used to verify attainment of important milestones in students' developing competence" (Shepard, 2006, [p. 636]).

Top-Down Model - Cognitive model of reading which emphasizes the importance of the reader's background knowledge in the reading process. In a "top-down" model, a reader is assumed to use knowledge about the topic, text structure, sentence structure, word meaning, and letter-sound correspondences to make predictions and confirm hypotheses during the reading process (Tracey & Morrow, 2006).

Vocabulary - Vocabulary was identified by the NRP as one of the five components or "big ideas" of reading. Vocabulary refers to understanding of used words (NRP, 2000). Vocabulary is important for reading comprehension.

Delimitations of the Study

The study is delimited by use of archival data that was collected in only one elementary school in a suburban school district located in the Southwestern region of North Carolina. The school was participating in a reform program with the introduction

of universal screening and progress monitoring; therefore, data were readily available for the researcher. In addition, data from only one school were analyzed in this study because it was the only elementary school in the district using Dynamic Indicators of Basic Early Literacy Skills (DIBELS) and AIMSweb Maze-CBM measures. Participants included all third, fourth, and fifth grade students in the school.

Summary

In summary, research is needed to examine the relationship of a measure of oral reading fluency and an additional measure of reading comprehension to the statewide reading assessment in North Carolina. The intent of this study was to investigate the relationship between outcomes from a state's large-scale reading comprehension assessment and scores on the DIBELS Oral Reading Fluency (ORF) and AIMSweb Maze-CBM comprehension measures. The degree that DIBELS and AIMSweb predict the comprehension measures on NC EOG Reading Comprehension test scores was examined and optimal cut scores to predict proficiency on the statewide assessment for third, fourth, and fifth grades were determined. Since these measures are widely used, results of this study have implications for educators of students with and without disabilities in all elementary school reading programs. Chapter 2 provides a review of related literature important to this study. A description of the methodology used is described in Chapter 3. A summary of the results is presented in Chapter 4. Finally, a discussion, including implications of this study, limitations, and areas of future research is presented in Chapter 5.

CHAPTER 2: REVIEW OF THE LITERATURE

In the past 30 years, there has been a paradigm shift in education as statewide systems of accountability have been mandated to ensure that all students demonstrate adequate yearly progress toward meeting state performance standards (NCLB, 2001). The purpose of this chapter is to examine, review, and synthesize the literature on the relationship between standardized assessments of reading comprehension and reading curriculum-based measures. Specific areas relevant to the topic include (a) reading fluency, (b) formative and summative reading assessment, (c) curriculum-based measurement, and (d) prediction of student performance on overall reading and high-stakes assessments.

The chapter is divided into seven sections. The first section of this chapter focuses on the history and theoretical foundations of reading fluency. The second section includes research relevant to formative and summative reading assessment. The third section of this chapter focuses on the history, development, reliability, validity, and technical adequacy of curriculum-based measurement (CBM). The fourth area reviews research on the relationship between scores on various reading curriculum-based measures and student outcomes on assessments of overall reading achievement. The fifth section includes research on the relationship between scores on various curriculum-based measures and student performance on state-mandated, high-stakes assessments. The sixth section presents research on the relationship between scores on CBM-Maze and student

outcomes on state-mandated, high- stakes assessments. Finally, the seventh section reviews research on the relationship between scores on Dynamic Indicators of Basic Early Literacy Skills (DIBELS) and student outcomes on state mandated high-stakes assessments.

Criteria for Selection of Relevant Literature

Studies were considered for inclusion in the literature review if they met the following criteria: (a) the study was published in peer-reviewed journal between 1980 and 2009, (b) the study examined the relationship between reading CBM and an outcome measure of overall reading achievement or high-stakes reading assessment, (c) the study was empirical and published in a peer-reviewed journal, and (d) the study included participants in elementary school. In addition, important correlational studies reported in technical reports were included in the review due to the relevance of data and frequent references to the studies within existing literature.

Electronic databases used in the search included ERIC, Academic Search Premier, Masterfile Premier, PsychInfo, PsychArticles, and Education Research Complete. Search terms included the following keywords: *oral reading fluency, fluency, curriculum-based measurement, curriculum-based measures, curriculum-based assessment, Dynamic Indicators of Basic Early Literacy Skills (DIBELS), high-stakes testing, accountability, predict, statewide assessment, formative assessment, formative measures, summative assessment, summative measures, Maze, and AIMSweb*. The process of identifying articles for inclusion included screening titles and abstracts to confirm that they related to CBM and screening the method section to verify that the study was an empirical study. Additional procedures to identify studies included examining references from identified

studies and hand searching recent 2009 publications of Exceptional Children and Learning Disabilities Research to Practice to obtain most recent literature that may not be on the database. Excluded from review were doctoral dissertations on Dissertation Abstracts International.

A total of 37 studies were located that met criteria for inclusion in the comprehensive review of literature. Eleven studies investigated the use of curriculum-based measurement (CBM) and the relation to student performance on standardized tests of overall reading achievement. An additional 10 studies were located that investigated the use of curriculum-based measurement - oral reading fluency (CBM-ORF) to predict performance on statewide, high-stakes assessment used for accountability, including one technical report and one paper presented at an annual conference frequently referenced in the literature. Three studies were included that examined the addition of CBM-Maze to predict performance on statewide assessments. Finally, 13 studies were located that examined the utility of DIBELS in predicting student achievement on high-stakes assessment, including four technical reports frequently referenced in the literature.

Reading Fluency

Fluent oral reading has been considered a significant factor in the development of reading and overall reading ability (Strecker, Roser, & Martinez, 1998). Defined as the ability to read text with speed, accuracy, and proper expression (NRP, 2000), fluency has been under the spotlight since being identified as one of the five “big ideas” of reading (NRP, Good & Kaminski, 2002). Children who are fluent readers can (a) recognize words automatically, (b) group words quickly to help them, (c) gain meaning from what they read, and (d) read aloud effortlessly and with expression (Armbruster et al., 2001).

The theoretical foundation for the construct of reading fluency can be traced to the LaBerge and Samuels model of information processing (Pikulski & Chard, 2005). In this "bottom-up" theory, the Automatic Information Processing Model, LaBerge and Samuels (1974) argued that humans are only able to do one thing at a time, and each task must be learned so well that it is automatic. According to this theory, processing in reading is a series of stages in which visual information is processed and transformed through phonological and episodic memory systems until comprehension occurs in the semantic system. In this model, good reading comprehension not only depends on accuracy, but on automaticity in decoding (Samuels, 1976).

In extensions of his earlier work, Samuels explained the theory of automatic information processing in reading (Samuels, 1979; 1997), brought attention to the importance of fluency based on the automaticity theory (Samuels, 1987), and described practical applications of the automaticity theory (Samuels, 1994). In one explanation of *automaticity*, Samuels (1994) stated,

> One way to think of automaticity is that it represents the ability to perform a task with little attention. The critical test of automaticity is that the task, which at the beginning stage of learning could be performed only by itself, now can be performed along with one or more other tasks. (p. 1130)

Based on the automaticity theory, the criteria in evaluation of learning are accuracy and automaticity. In order to process at the accuracy level, attention is necessary; however, at the automatic level, no attention is required. In this model, it is necessary to build reading skills toward an automatic level to develop higher level cognitive tasks, such as comprehension.

To apply the concept of building automaticity to reading, many simultaneous activities are required in order to read successfully (Tracey & Morrow, 2006). When a student struggles with decoding, the task of decoding the words demands all of the attention leaving no processing ability available to construct the meaning of the text. With attention at the heart of the model (Samuels, 1994), beginning readers must split attention between decoding the text and processing the meaning. On the other hand, a fluent reader reads in an effortless, flowing manner. Fluent readers are able to read words and decode text automatically; therefore, they are able to comprehend text better because more attention is freed for comprehension.

A large-scale data analysis was conducted by the National Assessment of Educational Progress (NAEP), which examined the relationship between fluency and other aspects of reading ability (Pinnell et al., 1995). A representative sample of 1,136 fourth grade students participated in the study. Student performance on oral reading sessions was linked to NAEP reading assessment data to examine the role of accuracy and rate (fluency) as it relates to reading proficiency. Fluency was rated level 1, 2, 3, or 4 based on phrasing, syntax, and expressive interpretation in order to determine the relationship. Results indicated fluency and comprehension were interconnected, as oral reading fluency "demonstrated a significant relationship with reading comprehension" (p.2). Students who read more fluently were more accurate and read at a substantially faster rate. Findings confirmed higher average reading proficiency was associated with higher levels of fluency.

In a review of fluency research, Strecker et al. (1998) examined empirical evidence on the relationship between fluency and comprehension. In their review of

literature, they discussed various aspects of fluency and offered explanations for fluency development. Fluency developed when readers were able to read text with ease. All models of reading consistently suggested that "children read in qualitatively different ways over time" (p.298). Evidence throughout the literature suggested that there is a period of reading development when the reader's focus shifts from features of print to meaning; as a result, fluency is developed. Their review of research on fluency clearly supported extensive practice and modeling. Evidence also suggested that wide reading, leveled reading, and fluency training helped students to become proficient readers. Different assumptions for causes of students struggling with fluency were discussed without consensus. Researchers suggested further investigation of fluency to inform instruction and development of a fluency assessment to measure rate, accuracy, and phrasing.

Improving fluency has also been emphasized in the research on reading. Kuhn and Stahl (2003) reviewed research and theories relating to fluency. In their review, they surveyed definitions for fluency and examined studies to improve fluency. The review included a total of 71 studies, with 58 studies designed to improve fluency using assisted reading, repeated reading, classroom interventions, segmented text, and isolated word recognition fluency. Results indicated fluency instruction improved reading fluency rates and comprehension in comparison to traditional instruction. Findings indicated that the strongest results for improving reading achievement occur at a certain point in the development of reading, which is between a late preprimer level and late second grade level. For the facilitation of reading rate and accuracy, both assisted and unassisted methods were found to be effective. Based on the findings from the review, the

integration of techniques to improve fluency (i.e., paired reading, re-reading, assisted reading, echo reading, partner reading, etc.) is warranted to improve achievement in reading.

Recently, teacher knowledge of why reading fluency is important has been found to be a significant factor in students' growth in reading. Lane et al. (2009) examined the role of teacher knowledge about reading fluency and vocabulary as predictors of students' fluency growth. The shape of the students' growth was evaluated using latent growth models (LGM) and the effect of teachers' knowledge of reading fluency was measured using multilevel latent growth models (MLGM). Results indicated that students with teachers who knew more about fluency demonstrated more growth on measures of decoding in first grade, with teacher knowledge explaining 25% of the variance in growth of decoding fluency and 11% of growth on reading fluency. Teacher knowledge also yielded greater increases in reading rate and accuracy for students in second grade, explaining 59% of growth in decoding fluency and 86% of growth on reading fluency. Fluency growth leveled off in third grade with no variance in fluency explained by teacher knowledge, despite the greater amount of teacher knowledge. Knowledge of effective practices for reading fluency assessment and instruction was also evident in second grade. The combined model identified the significant predictors as (a) knowledge of why reading fluency was important, (b) knowledge of skills children need, and (c) knowledge of effective instructional methods. This study added to the literature by providing evidence that teachers with more knowledge about reading fluency had students who read more quickly and accurately, which demonstrated the importance of the overall depth of teacher knowledge in reading fluency.

Formative and Summative Reading Assessment

Two types of assessment have traditionally been considered in the process of measuring progress of students in reading. First, summative assessment is given for the purpose of documenting achievement (Shepard, 2006). Summative assessments generally occur at the end of a phase of learning and are typically given at the end of chapters, units, courses, or grade levels in order to assign grades or evaluate progress. In contrast, a second type of assessment called formative assessment is typically given in order to improve or enable learning "during the instructional process" (Shepard, p. 627). Formative assessments are typically conducted often and include informal measures that allow teachers to use ongoing results to help them enhance instruction and increase student achievement (Bloom, et al., 1971).

Despite the labels, the terms "formative" and "summative" actually apply to the function of the assessment rather than the tests themselves (William & Black, 1996). For example, curriculum-based measurement (CBM) has the capability to provide both summative and formative data (Shinn & Bamonto, 1998), with summative data representing the student's level at a specific point in time and formative data representing repeated measurements of learning over time. The difference in summative and formative assessment is the purpose or goal. In the classroom, summative assessments are generally associated with grades, while formative assessments provide information to improve instruction.

There are significant concerns with the exclusive use or overuse of summative assessment. NCLB (2001) requires teachers to use annual statewide, high-stakes testing for all students in third through eighth grade to assess ability at the end of the year. With

student performance used to pinpoint schools that need special assistance and cash incentives offered for improvements in performance, teachers face the temptation of teaching to the test in order to meet Adequate Yearly Progress goals. Nichols and Berliner (2007) claim that high-stakes testing is destructive stating that the "unintended outcomes of the high-stakes testing policy were detrimental to the education process" (p. xv). Zimmerman and DiBenedetto (2008) assert that the primary issues related to complaints of high-stakes testing are (a) that the curriculum is not reflected in the test and (b) feedback about performance has little relation to instructional decisions. Kennedy et al. (2008) add that another major concern of summative assessment is that teachers are required to test at the end of the learning cycle; therefore, high-stakes accountability tests fail to provide feedback designed to improve learning.

Concerns also exist for the exclusive use of formative assessment, especially since high-stakes testing is grounded in the assumption that it raises the standards of learning (Gorlewski, 2008). Dorn (2007) stated that in order for educators to logistically handle the greater paperwork burden of frequent assessment for formative purposes, classroom teachers would need more assistance. In an effort to establish ways of helping teachers implement formative practices more effectively, Black and William (2009) designed a framework for formative assessment. Their purpose was to offer a rationale and unify formative practices, discuss theories of pedagogy in connection with formative interactions, and recommend ways to improve formative practices within the classroom. The framework by Black and William suggested five key strategies:

1. Clarifying and sharing learning intentions and criteria for success;

2. Engineering effective classroom discussions and other learning tasks that elicit evidence of student understanding;
3. Providing feedback that moves learners forward;
4. Activating student as instructional resources for one another; and
5. Activating students as the owners of their own learning. (p. 8)

In a meta-analysis of studies exploring the effects of formative assessment, Fuchs and Fuchs (1986) reviewed the literature to determine first, the effectiveness of individualized instruction, and second, the usefulness of formative assessment with consideration of the time required for administration versus the benefits. The review included 21 studies with a total of 3,835 subjects in pre-school through high school. Participants diagnosed with disabilities were included in 83% of the investigations. Of these participants, 98% had mild to moderate disabilities, and 2% had severe disabilities. Results indicated the use of formative assessment significantly increased achievement with an effect size of .70. Findings indicated systematic formative assessment was effective regardless of age, treatment duration, frequency, or handicapped status. Effect sizes were higher when teachers employed systematic, explicit rules when evaluating the data (i.e., implementing a change in program if progress slope was not as steep as goal line) in comparison to teacher judgment. Effect sizes were also higher when data were graphed and presented to the student in comparison to data recorded by the teacher with no feedback to students. Additionally, when behavior modification was added, typical achievement outcomes were boosted. Researchers suggested the use of systematic formative assessment was worthwhile, despite additional time required for teachers.

Fuchs, Deno, and Mirkin (1984) examined the effects of repeated measurement and continuous evaluation on student achievement using the *Data-based Program Modification* (DBPM; Mirkin et al., 1981). Additionally researchers examined measures of pedagogy and student knowledge of learning. Analysis of covariance (ANCOVA) revealed that students with teachers who employed ongoing measurement had superior reading progress on passage reading as well as decoding and comprehension measures. In addition to better student achievement, results also suggested teacher pedagogy and students' own knowledge about their learning improved as a result of systematic measurement.

The relationship between assessment and student performance has traditionally focused on the use of standardized tests. In a synthesis of research, Black and William (1998) reviewed studies on the impact of improved classroom assessment on student success on summative assessments. Findings suggested that enhancement in teachers' classroom assessment practices was directly associated with differences in standardized test scores. Specifically, enhanced assessment raised student performance on standardized tests in England by 0.4 to 0.7 of a standard deviation, which is comparable to 15 percentile points on U.S. standardized tests. Overall findings indicated that the development of formative assessment raised standards and led to large learning gains. Black and William advocated for formative assessment to be included as an essential feature of classroom work.

Later, Black and William (2006) stated formative assessment may be "particularly effective, in part because the quality of interactive feedback is a critical feature in determining the quality of a learning activity, and is therefore a central feature of

pedagogy" (p.100). Similarly, Shepard (2000) asserted that in order to genuinely increase learning, it was necessary to use in-depth and ongoing assessments. However, in a review of how the concepts of "formative" and "summative" assessment have developed over time and the implications for student learning, Kennedy et al. (2008) stated that the "valorizing of formative assessment over other forms of assessment could be problematic if it is assumed that the promotion of formative assessment somehow solves the problems of summative assessment" (p. 198).

Actually, it is believed that summative and formative assessment could lead into each other as one continuous process since the gap between expected standards, goals, and criteria not met on summative assessment can be reached through instruction informed by formative assessment (Shepard, 2006; Taras, 2007). As stated by Shepard, "summative assessments can be thought of as important milestones on the same learning continua that undergrid formative assessment" (p. 638). When used effectively, formative assessments allow teachers to check student understanding and support learning prior to the external, large-scale assessments that are required for high-stakes accountability.

In a review of research from the Center for the Improvement of Early Reading Achievement (CIERA), Paris and Hoffman (2004) examined promising reading assessments available for teachers and confirmed the use of formative assessment to be a valuable tool for promoting instruction appropriate for individual students' needs. However, they also verified that even though measurement of motivation, self-concept, and critical thinking are difficult, large-scale standardized tests were not to be ignored. According to findings of the review, all assessments should "contribute to theory building that ultimately informs effective teaching and learning" (p. 215). Through ongoing

assessment of students' knowledge, teachers can improve students' opportunities to learn and improve instruction to promote learning (Junker & Matsumura, 2006).

Curriculum-Based Measurement in Reading

Curriculum-based measurement (CBM; Deno, 1985) is an alternative approach to assessing students. This approach combines the advantages of commercial, standardized tests and teachers' informal observations. CBM has been characterized as a general outcome measure (Fuchs & Deno, 1991) due to the information it can provide for evaluation of the overall effectiveness of an instructional program in reading, written expression, math computation, math application, and spelling (Tindal, 1989). Reading CBM probes are brief, 1- minute measures sampled from grade-level curriculum materials, which are sensitive to student growth over time. The administration and scoring of CBM probes is standardized. All CBM probes are designed to assess fluency rather than just accuracy.

Historical Background of Curriculum-Based Measurement

CBM has its origins in a federally funded project from the late 1970s. Deno and colleagues at the University of Minnesota developed the data-based program modification (DBPM; Deno & Mirkin, 1977) model to help teachers improve performance. The researchers sought an "alternative" to the widely used, commercially developed, standardized, and norm referenced measures of reading achievement. They focused their efforts on the separation between measurement and instruction (Deno, 2003). Their objective was to develop valid and reliable assessment methods that would enable educators to use student achievement data to make instructional decisions. This approach became curriculum-based measurement (CBM; Deno, 1985), which is now

widely used for data-based decision making about the effectiveness of instructional programs, development of instructional goals, and improvement in student achievement (Fuchs & Deno, 1991).

Characteristics of Curriculum-Based Measurement

During the research and development of curriculum-based measurement (CBM), characteristics and criteria were specified for the measures in order to establish a simple method for monitoring student progress and achievement in the curriculum. In a report, Jenkins, Deno, and Mirkin (1979) outlined important characteristics for a data system to be used in providing appropriate educational programs, making eligibility decisions, making program planning decisions, and adjusting programs for effective decision-making to improve pupil progress. In their report, they outlined desirable characteristics for the data system. From the initial planning, the system was designed to be (a) relevant, (b) sensitive, (c) flexible, (d) repeatedly administrable, and (e) easily administrable.

> Deno (1985) stated the measures would have to be:
>
> (1) Reliable and valid if the results of their use were to be accepted as evidence regarding student achievement and the basis for making instructional decisions.
>
> (2) Simple and efficient if teachers were going to use them, or teach others to use them, to frequently monitor student achievement.
>
> (3) Easily understood so that the results could be clearly and correctly communicated to parents, teachers, and students.
>
> (4) Inexpensive since multiple forms were to be required for repeated measurement. (p. 221)

The ultimate outcome of the early development stages of the research program was to create a formative evaluation system to help teachers and promote effective teaching for students with academic disabilities (Deno, 2003).

Reliability and Validity of Curriculum-Based Measurement (CBM)

The reliability and validity of using CBM as an indicator of students' academic skill level were evident throughout the literature. Marston (1989) summarized studies examining the criterion-related validity and reliability of CBM in a review of literature on reading CBM. Correlations between CBM-ORF measures and global skills on criterion tests of reading ranged from .63 to .90 and the subtests of global measures ranged from .53 to .91. Correlations between CBM oral fluency measures and criterion-referenced mastery tests from basal reading series ranged from .57 to .86. Correlations between fluency CBM oral fluency measures and word lists were .76. Test-retest reliability estimates ranged from .82 to .97; parallel form estimates ranged from .84 to .96; and interrater agreement coefficients were .99. Findings indicated CBM reading measures are reliable and valid. Using CBM, performance can be compared to a standard of mastery; therefore, measurement can occur throughout the year as a valid and reliable indicator of growth toward grade level performance.

In the establishment of reliability and validity for CBM, Deno et al. (1982) conducted the initial three concurrent CBM validity studies to determine the relationship between performance on formative reading measures and performance on standardized reading achievement measures. Correlational analyses were examined for student performance on five formative measures and three standardized measures. In the first of the three concurrent studies, Deno et al. used student performance data from 18 students

in regular class and 15 resource students with learning disabilities in first through fifth grades from a suburban elementary school in Minnesota to determine what measures would generate valid, continuous evaluation of reading progress. Words in Isolation, Words in Context, Oral Reading, Cloze Comprehension, and Word Meaning were the formative measures administered. The standardized tests included the reading comprehension subtest of the *Stanford Diagnostic Reading Test* (SDRT; Karlsen, Madden, & Gardner, 1975) and the word identification and word comprehension subtests of the *Woodcock Reading Mastery Test* (WRMT; Woodcock, 1973). Results indicated correlations ranged from .60 and .91 between formative measures and criterion measures. Oral reading rate had trends at a higher level than other measures. For the resource group, correlations on cloze and word meaning measures were somewhat lower ranging from .48 to .67.

The purpose of the second concurrent study was to determine if the grade level of materials selected or duration of test changed correlations. Participants included 27 students in regular class and 18 resource students with learning disabilities in first through sixth grade from two public schools in Minneapolis. Formative measures used in this study were additional forms developed, which were identical to measures used in the first study with the exception of the cloze comprehension passages. In the cloze measure, fewer words were omitted in an attempt to increase the amount of correct responses; therefore, every 10^{th} word was omitted instead of every fifth word. Results indicated strong correlations between the third and sixth grade materials as well as between the 30 second and 1-minute tests on the three word recognition measures.

The purpose of the third concurrent study was to replicate and integrate findings from the first two studies using a larger subject pool. Participants included 43 students in regular class and 23 resource students with learning disabilities in first through sixth grade from three inner-city schools in Minneapolis. Alternate forms of all materials used in the previous study were used as well as the reading comprehension subtest from the *Peabody Individual Achievement Test* (PIAT; Dunn & Markwardt, 1970). The PIAT and SDRT were individually administered to each student. Results indicated the resource group performed relatively low on all measures. The performance of regular class students was three to five times higher on all measures and nine times higher on cloze measures.

Overall findings from the three concurrent studies by Deno et al. (1982) indicated using 1 minute samples of reading performance on oral reading, isolated word, and cloze comprehension measures related to student performance on standardized reading tests. For word reading, results indicated the difficulty of the words did not determine the validity of the data. Additionally, reading proficiency was best measured by collecting data of correct performance rather than error performance, but a combination of correct and incorrect performance could economically and easily be obtained at the same time. In this series of criterion validity studies, all curriculum-based measures except for the word meaning task were highly correlated with student performance on the standardized reading achievement tests.

In another study, Fuchs et al. (1988) examined four informal measures of reading comprehension for criterion, construct, and concurrent validity. Question answering tests, recall measures, oral passage reading tests, and cloze techniques were used to determine

relations with the reading comprehension and word study skills subtests of the Stanford Achievement Test with 70 middle and high school boys. Results indicated moderate to high correlations between the informal reading measures and standardized measures. The oral passage reading test correlated more strongly with the comprehension subtest ($r = .91$) and word skills subtest ($r = .80$) of a standardized achievement test. Reading aloud was the most feasible and useful method for indexing reading improvement for students, including reading comprehension. However, acceptable alternatives included written recall and written cloze.

A study conducted by Fuchs et al. (1982) was designed to investigate the reliability and validity of curriculum-based informal reading inventories. Specifically, using correlational and congruency analyses, researchers explored the reliability and validity of (a) using 95% accuracy standard to determine instructional level, (b) arbitrarily selecting a passage, and (c) employing one-level floors and ceilings. Participants included 91 students in grades 1-6 who were administered the word identification and passage comprehension subtests of the Woodcock Reading Mastery Tests (Woodcock, 1973) and passages from Ginn 720 (1976) and Scott-Foresman Unlimited (1976). Teachers' placement of instructional level was also reported for analysis. Results indicated high correlations between achievement tests and teacher placements with a standard of 95% accuracy of word recognition, which supports this standard for determining instructional level. However, results indicated one-level ceilings and floors were inadequate, and the practice of selecting articles arbitrarily was insufficient. Researchers suggested highly structured procedures for creation and administration of curriculum-based informal reading inventories (IRIs).

Later, Fuchs and Fuchs (1992) summarized a research program examining alternative CBM reading measures that incorporated tasks which required comprehension. In their research, four reading measures (i.e., question answering tests, recall procedures, cloze techniques, and maze methods) were examined. The reading Maze task was identified to be a useful measure for monitoring student growth. Criterion validity was strong for Maze. Also, teacher satisfaction of Maze was high, since teachers reported that the Maze measure reflected decoding, comprehension, and fluency skills for students.

To establish criterion-related validity and provide cross-validation across measures and reading curricula, Bain and Garlock (1992) examined first, second, and third grade students' performance on CBM reading passages subtest developed from Macmillan Series and the Comprehensive Test of Basic Skills (CTBS). Moderate to high correlations were found between the CBM measures and total reading scores on CTBS for all three grade levels. Findings supported the use of the CBM reading measurement technique and provided evidence for cross-validation for CBM reading passage measures.

Research on CBM emphasized the technical adequacy of alternative reading measures. Tindal and Marston (1996) examined the technical adequacy of reading CBM in two concurrent studies. In the first study, participants included a total of 772 students, representing 20 students from each grade level of 13 elementary schools. Researchers examined the validity of seven alternative reading measures including the following: letter identification, dolch word list, ORF using Holt passages, ORF using literature based passages, reading comprehension Maze, reading comprehension idea units, and reading expression. Outcome measures included teacher judgment and the California

Achievement Test. Using multiple regression analyses, ORF was the leading contributor at every grade level to the prediction of overall reading skills with the variance accounted for slightly less in fifth and sixth grades. Maze reading comprehension demonstrated strong correlations with criterion measures at third, fourth, and fifth grades. Findings supported the use of formative reading measures in classroom assessment practices. In the second concurrent study, participants included 1 student with a learning disability and 1 teacher. Data were collected in order to focus on instructional decision-making based on measurement information. Results suggested the measurement of student performance and progress led to effective instructional programs and fluency. Also, prosody actually improved, which validated the relevance of measurement data.

Recently, Wayman, Wallace, Wiley, Ticha, and Espin (2007) synthesized the literature on CBM research since Marston's (1989) review. The technical adequacy, effects of text materials, and growth for CBM reading measures was examined and described. In their review, reading aloud, Maze, and word identification measures were included. Results of the technical adequacy section indicated the CBM read aloud measure was a better indicator of reading comprehension than the other measures. The read aloud measure demonstrated strong relationships with overall reading proficiency. Correlations between read aloud measures and criterion measures were moderate to strong. The strongest correlations were in the early elementary grades, with diminishing relationships in the intermediate grades. However, Maze results stayed constant across grade levels with moderate to strong reliability and criterion-related validity.

CBM measures in reading have been empirically validated, and scientific evidence has suggested CBM is a valid and reliable method for assessment of current

student performance (Deno et al., 1982), rate of progress (Jenkins et al., 1979) and growth (Fuchs, & Fuchs, 1997; Shin, Deno, & Espin, 2000). Research clearly supports (a) ongoing measurement systems are important; (b) CBM measures can be used as a formative measure of student achievement; and (c) student outcomes can be enhanced by frequent measurement on curriculum tasks and instructional decision-making based on data (Fuchs et al., 1984).

Curriculum-Based Measures of Oral Reading Fluency

The history of informal reading assessment has been outlined back for nearly a century (Powell, 1971). Research clearly demonstrates measures of oral reading fluency are highly related to overall reading (Tindal & Marston, 1996) and CBM reading probes are valid indicators of reading competence (Good & Jefferson, 1998). Throughout the literature, there is evidence of significant correlations between measures of fluency, especially ORF, and standardized measures of overall reading achievement (Crawford et al., 2001; Fuchs et al., 2001; Good et al., 2001). There is general agreement in the literature on assessment of fluency that it is necessary for a comprehensive assessment of fluency to include measures of oral reading accuracy, rate, and reading comprehension (Pikulski & Chard, 2005).

Oral reading fluency (ORF) is the most widely used curriculum-based measure of reading competence (Good et al., 2001). Measures of oral reading fluency are "administratively feasible" and "economically affordable" (Tindal & Marston, 1996). Scientific evidence of the reliability and validity of the measure is even documented with researchers other than those typically associated with studying the measure (Deno, 1985; Fuchs et al., 1984). In fact, Anderson, Wilkinson, and Mason (1991) examined small

group reading lessons and found that "among the measures of reading ability, it was group fluency that was most strongly related to outcome measures" (p. 439). Results indicated individual comprehension, individual fluency, group fluency, story emphasis, and teaching emphasis significantly influenced recall of important story elements.

Deno (1989) promoted the development and use of standardized, locally-normed curriculum-based assessment for decision making. Hasbrouck and Tindal (1992) developed large-scale norms for ORF in order to address concerns with local norms. With large-scale norms for ORF, students' ORF scores can be compared to norms from a large group of students at the same grade level who took the same test. Standardized CBM procedures were used to conduct 1-minute timed oral reading samples from at least two grade level passages for 7,000 to 9,000 students in second through fifth grades in five states. The curriculum-based norms established by Hasbrouck and Tindal "serve as benchmarks to rank student performance" (p. 42). Through extensive study of ORF, Hasbrouck and Tindal (2006) recently published updated norms. Norms for each grade level can be found in Table 1. Updated norms for each grade level can be found in Table 2.

Table 1

Curriculum-Based Norms in Oral Reading Fluency for Grades 2-5 (50th Percentile)

Grade	Fall WCPM	Winter WCPM	Spring WCPM
2	53	78	94
3	79	93	114
4	99	112	118
5	105	118	128

*WCPM = words correct per minute

Table 2

Oral Reading Fluency Norms for Grades 1-5 (50th Percentile)

Grade	Fall WCPM	Winter WCPM	Spring WCPM	Avg. Imp./ wk
1	N/A	23	53	1.9
2	51	72	89	1.2
3	71	92	107	1.1
4	94	112	123	0.9
5	110	127	139	0.9

*WCPM = words correct per minute

Curriculum-Based Measurement from a Theoretical Perspective

Despite the number of studies conducted to examine oral reading fluency as a reliable and valid measure of reading skills (Deno et al., 1982; Fuchs et al., 1988; Tindal, Fuchs, Fuchs, Shinn, Deno, & Germann, 1985), few studies have explored CBM reading fluency from a theoretical perspective. In an effort to evaluate a model of underlying processes of reading and reading comprehension development, Lomax (1983) examined the causal relationships among phonological word recognition, word recognition, reading rate, and reading comprehension. Researchers used structural equation modeling (SEM) procedures to examine relationships suggested by previous research. Participants included 101 students with learning disabilities in 11 self-contained classrooms in an urban school district. Results indicated phonological skills had a direct causal influence on word recognition. Additionally, word recognition had a causal influence on reading comprehension. The causal model was replicated within the study and results indicated the model remained the same, which provided support for the model of reading comprehension based on component processes.

Shinn et al. (1992) investigated the contribution of CBM-ORF to theoretical process models of reading. Participants included 238 students in third (n=114) and fifth grade (n=124) from 13 elementary schools in a public school in the Northwest. The relationships among eight reading measures were examined using confirmatory factor analysis to test the contribution of oral reading in a model. Results indicated the three-factor model explained the obtained relationships for third and fifth grades. However, for third grade, the best explanation was a single factor model of reading identified as Reading Competence. In this model, two CBM reading measures correlated the highest

(r = .88 and .90), but all measures significantly contributed to the model, as SDRT inferential and literal comprehension also had strong correlations (r = .71 and .72). For fifth grade, a two-factor model with decoding and comprehension best fit the common conception of reading, with fluency as part of decoding. Decoding and comprehension were highly correlated (r = .83), but could also be differentiated as constructs. High correlations were also found for CBM measures (r = .74 and .76), SDRT measures (r = .73 and .76), and cloze procedures (r = .86). Study results strongly supported ORF as an index of reading proficiency and validated ORF as a measure of general reading achievement and comprehension. Also, support was demonstrated for the theoretical models of various authors who have proposed the pivotal role of fluency in the reading process (e.g., LaBerge & Samuels, 1974).

Curriculum-Based Measures Used to Measure General Outcomes

Dynamic Indicators of Basic Early Literacy Skills (DIBELS). DIBELS are standardized, individually administered, general outcome fluency measures developed by researchers at the University of Oregon (DIBELS; Kaminski & Good, 1998). The subtests were designed to evaluate the development of early literacy development and are available for download free of charge at https://dibels.uoregon.edu for grades K-6. Seven subtests make up the DIBELS curriculum-based assessment: (a) Initial Sound Fluency (ISF), (b) Letter Naming Fluency (LNF), (c) Phoneme Segmentation Fluency (PSF), (d) Nonsense Word Fluency (NWF), (e) Oral Reading Fluency (ORF), (f) Retell Fluency (RTF), and (g) Word Use Fluency (WUF).

For Universal Screening (e.g. all students in school or grade level), the DIBELS system allows 3 or 4 benchmark assessment periods during each school year. While being

timed for 1 minute, students are asked to complete tasks including the following: (a) identifying the correct picture based on initial letter sound (ISF), (b) naming upper and lower-case letters (LNF), (c) segmenting words into individual sounds (PSF), (d) reading CVC nonsense words (NWF), (e) reading connected text at appropriate grade level (ORF), (f) retelling passage (RTF), and (g) using words in sentences (WUF). Additionally, multiple forms of ISF, PSF, NWF, and ORF are available as progress monitoring measures to allow for more frequent assessment of students whose scores are below benchmark level.

Scores for each of the measures are reported with a level of risk. Also, for each student, scores are combined and individually weighted for an overall level of risk. Schools can use scores to identify students in need of supplementary instruction. ORF is emphasized from the winter of first grade through sixth grade. The benchmark goal for ORF is 40 correct words per minute (wcpm) in first grade, 90 wcpm in second grade, 110 wcpm in third grade, 118 wcpm in fourth grade, 124 wcpm in fifth grade, and 125 wcpm in sixth grade. There are not established benchmark goals for RTF; however, in order to demonstrate adequate comprehension, it is recommended that students meet ORF goal and retell at least 25% of passage.

Despite criticism from some researchers (i.e., Goodman, 2006; Manning, Kamii, & Kato, 2006; Flurkey, 2006), Dynamic Indicators of Basic Early Literacy Skills (DIBELS) has proven to be useful as an indicator of student performance on measures of overall achievement (Burke & Hagan-Burke, 2007; Riedel, 2007; and Schatschneider, Wagner, & Crawford, 2008). DIBELS has been useful for predicting student achievement level on statewide, mandated, high-stakes assessments (Baker et al., 2008; Catts, et al.,

2009; Chard et al., 2008; Good et al., 2001; Roehrig et al., 2008; Schilling et al., 2007; Shapiro et al., 2008; Wood, 2006; and Wood, 2009). An assessment committee from the Institute for the Development of Educational Achievement was formed to conduct an analysis of reading instruments. The committee found DIBELS to be an appropriate reading assessment instrument tool for local education agencies to use in screening and progress monitoring for one or more essential reading components at one or more grade levels (Kame'enui et al., 2002).

AIMSweb Maze Curriculum-Based Measurement. With some research to suggest limitations of ORF as an indicator of overall reading skills as students advance in grade levels (Shinn et al., 1992), the Maze task is another form of CBM that has been shown to be reliable and valid for measuring student reading skills (Fuchs & Fuchs, 1992; Shin et al., 2000). Using data obtained from 43 second grade students, Shin et al. examined the technical adequacy of Maze for assessing student growth. The students were assessed with 10 different forms of Maze passages collected monthly. Hierarchical linear modeling was used to determine sensitivity of the Maze task. Validity was examined by looking at the relationship between growth rates on the Maze task and student performance on the California Achievement Test (CAT) using HLM. Results indicated the Maze task can be used to assess reading growth and findings supported the use of the Maze task as a reliable, sensitive, and valid measure.

The AIMSweb Maze Curriculum-Based Measures (Edformation, 2009) are standard reading comprehension assessment passages based on the cloze task. Maze-CBM is a measure of reading comprehension and can be used as a supplemental measure of reading skills. For each AIMSweb Maze passage, the first sentence is left intact and

every seventh word is replaced with three words in parenthesis. One word is a near distracter that does not make sense in the passage, even though it is the correct part of speech. Another word is a far distracter that is not of the same part of speech, and it does not make sense in the passage. Finally, one word is the exact word from the original passage. Following standardized directions read aloud by the administrator, the students are expected to read the passage and circle the correct word choice from the words in parenthesis. The students complete the task by reading silently for 3 minutes. The score is the number of correctly circled word choices.

Curriculum-Based Measurement and Severe Deficits in Reading

Few studies have been conducted to examine measures to assess students who are severely deficit in reading. One single-subject study was located that examined the most sensitive and efficient CBM strategy for measurement of student progress and instructional decision making for this population. Faykus and McCurdy (1998) investigated effective assessment practices for students with severe deficits in reading. Participants included 6 students with mental retardation and emotional/behavioral disorders in self-contained classrooms at a residential school in Philadelphia. Two curriculum-based reading measures (i.e., ORF and a computer program with a modified cloze measure called Maze) were used to measure student progress in reading. Using an A-B-C design, slope and lines of best fit were calculated during graphic feedback and instructional intervention conditions. Data were plotted on graphs and analyzed to compare sensitivity of the two measures. Results of the investigation indicated that ORF is a more sensitive measure than Maze for index of reading progress in low performing readers, despite the fact that teachers preferred the Maze measure. Implications for

practice from the study indicated using oral reading rate would result in more efficient decision making for this population of students.

Curriculum-Based Measurement in Relation to Student Performance on Standardized Measures of Overall Reading Achievement

Researchers have explored the use of Curriculum-Based Measurement (CBM) and its relation to student performance on measures of overall reading achievement. Studies in the following section examined the use of CBM and its relation to student performance on standardized tests of overall reading achievement.

Review of Published Studies Using CBM-ORF to Predict Performance on Measures of Overall Reading Achievement

Eleven studies were located in peer reviewed journals between 1993 and 2008 that investigated the use of various reading CBM measures as a formative assessment of student achievement and its relation to student performance on overall reading achievement. All studies reviewed were located in peer-reviewed journals and are described in this section in terms of purpose, participants and setting, measures, data analysis, and results. Studies reviewed in this section are summarized in Table 3.

Description of Studies. The purpose of a study by Jenkins and Jewell (1993) was to investigate the relationship between performance on informal reading measures (reading aloud and Maze) and performance on standardized reading assessments. Nolet and McLaughlin (1997) examined growth of reading and written expression skills over the school year and performance on a performance task similar to the statewide performance assessment program. The focus of the study was to address questions about the potential problem of important instruction lost in classrooms with schools facing

statewide accountability testing. The purpose of a study by Kranzler et al. (1998) was to determine whether general cognitive ability, processing speed and efficiency, and oral reading fluency have a significant role in predicting student performance on measures of reading comprehension. Kranzler, Miller, and Jordan (1999) examined racial/ethnic and gender bias on CBM as an estimate of performance on a measure of reading comprehension.

Similarly, Hintze et al. (2002) replicated and extended research by Kranzler et al. (1998) by examining predictability of CBM-ORF on reading comprehension. In another study, Roberts, Good, and Corcoran (2005), investigated the efficiency and effectiveness of a curriculum-based measure of oral reading fluency. Retell fluency was also examined in order to maximize effective instruction for students whose reading fluency rates were higher than typical performance on comprehension tasks. The focus of the study was an examination of the relationship between ORF with retell and a measure of overall reading competence. The purpose of a study by Yovanoff et al. (2005) was to determine the importance of fluency and vocabulary in relation to performance on measures of comprehension.

Riedel (2007) investigated the relationship between DIBELS measures and reading achievement at the end of first and second grade and determined optimal cut scores for performance. Burke and Hagan-Burke (2007) examined convergent validity of first grade DIBELS measures and Test of Word Reading Efficiency. Schatschneider, et al. (2008) compared the predictive validity of measures of achievement and growth in achievement, as well as the validity of using a combination of achievement status and growth for predicting future reading achievement. Most recently, Klauda and Guthrie

(2008) explored the relationships of word, syntactic, and passage level fluency with reading comprehension.

Participants. All studies reviewed in this section included participants at the elementary level, but two studies also included older children. Participants in the study conducted by Jenkins and Jewell (1993) included 335 students in second grade through sixth grade in two elementary schools in the Pacific Northwest. The participants in a study conducted by Nolet and McLaughlin (1997) included 58 students in fifth-grade in an urban elementary school in Maryland. In the study conducted by Kranzler et al. (1998), participants included 57 fourth grade students in an elementary school in North Central Florida. Another study, Kranzler et al. (1999) included 326 students in second grade through fifth grade at an elementary school in North Central Florida.

Also at the elementary level, the participants in Hintze et al. (2002) included 136 students in second grade through fifth grade in an urban school in the Northeastern United States. Roberts et al. (2005) collected first grade data from six schools in an urban school district in the Southeastern United States. Of the 86 students included, 100% received free or reduced lunch and 90% were African American. Yovanoff et al. (2005) included a total of 6,012 students in fourth grade through eighth grade in a school district in the Pacific Northwest. Riedel (2007) included 1,518 first grade students in Memphis City School district. Demographic information of students revealed 92% of students participating were African American and 85% received free or reduced lunch. Participants in the study conducted by Burke and Hagan-Burke (2007) used 213 first grade students in a K-2 primary school in Georgia. Schatschneider et al. (2008) included 23,438 first grade students in Reading First schools in Florida. Participants in Klauda and

Guthrie (2008) were 278 fifth grade students in 13 classrooms from three different schools in a mid-Atlantic state.

Measures. Jenkins and Jewell (1993) used two informal measures (Maze passages and oral reading measures) as well as teacher judgment to examine their relationship with standardized reading achievement tests. Students were given three Maze passages with 2.3 readability and three narrative passages to read aloud with a mean readability of 1.7. The standardized reading achievement tests included Metropolitan Achievement Tests (MAT) at three different levels (primary for second grade, elementary for third and fourth grades, and intermediate for fifth and sixth grades) and Gates-MacGinitie Reading Tests at three different levels (Level B for second grade, Level C for third grade, and Level D for fourth through sixth grades). Nolet and McLaughlin (1997) administered CBM-ORF probes with retell measures in fall, winter, and spring as well as CBM written expression measures in winter and spring. ORF probes were developed using narrative stories from curriculum materials and individually administered using standardized directions and scoring. A performance task, which was a publicly released alternate form of the Maryland School Performance Assessment Program (MSAP) was also administered to students. In Kranzler et al. (1998), students were administered six curriculum-based measures of reading fluency from the Ginn Basal Readers. Within 3 weeks of CBM administration, students were given psychometric and chronometric tests. The psychometric test was the Kaufman Brief Intelligence Test (K-BIT; Kaufman & Kaufman, 1990) and the chronometric tests consisted of four short tests of cognitive processing speed and efficiency called elementary cognitive tasks (ECTs) (e.g., simple reaction time, choice reaction time, odd-man-out paradigm, and inspection time). The

measure used for reading comprehension was Kaufman Test of Educational Achievement (KTEA; Kaufman & Kaufman, 1985). Kranzler et al. (1999) used six CBM measures of reading fluency from Ginn Basal Readers and the California Achievement Test (CAT).

In another study, Hintze et al. (2002) used three CBM measures of reading fluency from Silver, Burdett, & Ginn Reading Series. The measure of reading comprehension was Woodcock Johnson Psychoeducational Battery-Revised (WJ-R; Woodcock & Johnson, 1989). Roberts et al. (2005) developed two CBM-ORF passages and tracking procedures for counting the number of words correct during the retell of the story. In this study, the letter-word identification, word attack, and passage comprehension subtests of the Woodcock Diagnostic Reading Battery (WDRB; Woodcock, Mather, & Schrank, 2004) were used as the measure of overall student reading achievement. Yovanoff et al. (2005) used 250-word, grade-level appropriate CBM-ORF passages, and vocabulary measures developed with 70 items using one correct response, one far-response, and one near-response for answer choices. The measure of reading comprehension used in the study was a different form of the same passages developed for oral fluency, followed by 15 selected response questions. In a study conducted by Riedel (2007) DIBELS measures (LNF, PSF, NWF, ORF, and RF) were used to examine their relationship with comprehension on the Group Reading Assessment and Diagnostic Evaluation (GRA + DE) and the TerraNova Reading Subtest. The GRA + DE is a standardized, group administered, multiple-choice test of reading ability administered in the spring of first grade. The Terra Nova is also a standardized, group administered, multiple-choice test of achievement which is used as a measure of second grade comprehension and includes timed subtests. Burke and Hagan-Burke

(2007) administered DIBELS measures (PSF, NWF, DORF, RF, WUF) and the Sight Word Reading Efficiency (SWE) and Phoneme Decoding Efficiency (PDE) subtests of the TOWRE. The PDE subtest measures phonological decoding ability and student ability to decode non-words. The SWE subtest measures the student's ability to read sight words and sight word reading fluency. Schatschneider et al. (2008) obtained data from Florida's Progress Monitoring and Reporting Network (PMRN) for DIBELS ORF measures. Researchers also used the Stanford Achievement Test, which is a multiple-choice measure of reading comprehension administered in a group format.

In a study conducted by Klauda and Guthrie (2008), three measures of fluency and single measures of reading comprehension were used to show relationships. The tests of comprehension used were the Gates-MacGinitie Reading Test (GMRT: MacGinitie, MacGinitie, Maria, & Dreyer, 2000), an Inference Assessment, and Background Knowledge Assessment. The three measures of fluency included Woodcock-Johnson III Reading Fluency Test (WJ-III) to measure fluency at the syntactic level, Passage Oral Reading Assessment (PORA) to measure fluency at the passage level, and Word Recognition Assessment (WRA) to measure fluency at the word level.

Data Analysis. In the study conducted by Jenkins and Jewell (1993), cross-grade correlations, grade level correlations, and teacher judgment correlations were computed to determine relations with standardized measures of reading proficiency. Nolet and McLaughlin (1997) used repeated measures ANOVA for three related samples to determine if there were significant differences in each administration time. Paired t-tests were used to determine significant increases in correct word sequences. Kranzler et al. (1998) used two simultaneous multiple regression analyses to explain relationships

among the variables. The first analysis was the regression of Reading Comprehension on Matrices, Reading Fluency, and Mental Speed. The second analysis was the regression of Reading Comprehension on Matrices, Reading Fluency, and Reaction Time Parameters on all ECTs. In the study conducted by Kranzler et al. (1999), simultaneous multiple regression analyses were used to examine group differences on CBM to estimate performance on reading comprehension.

Hintze et al. (2002) used a series of hierarchical multiple regression analyses to determine the significance of age, ORF, SES, gender, and ethnicity on the prediction of reading comprehension. Later, Roberts et al. (2005) used correlation analyses to examine relationships between ORF, retell, and overall reading competence. In order to investigate the differential importance of fluency and vocabulary for measurement of reading comprehension as a function of grade level, Yovanoff et al. (2005) used structural equation modeling (SEM).

In a study conducted by Riedel (2007), Receiver Operating Characteristic (ROC) Analysis was used to examine the relationship between DIBELS subtests and reading comprehension. ROC was also used to determine cut scores. Logistic Regression was used in the study to determine whether the predictive power of DIBELS improved as subtests were added to the regression equation. ANOVA, chi-square, and logistic regression analyses were used to examine students for which DIBELS was a poor predictor of comprehension. Burke and Hagan-Burke (2007) examined the concurrent validity with correlation analysis and used regression analyses to examine the amount of variance from the subtests of TOWRE that was explained by DIBELS measures. Two exploratory principal axis factor analyses were also conducted to determine the relation

of the DIBELS measures to the construct they are meant to represent. Schatschneider et al. (2008) examined individual growth curves to estimate individual differences and growth. Using hierarchical multiple regression analyses, researchers reversed the order of the entry of slope and EOY ORF to predict performance on SAT 10 at the end of first grade, SAT 10 at the end of second grade, and ORF at the beginning of second grade.

Hierarchical regression analyses were used in the study by Klauda and Guthrie (2008) in order to (a) determine the extent to which word, syntactic, and passage fluency predict reading comprehension, (b) analyze the extent to which cognitive variables (inference and background knowledge) mediated the association between fluency and reading comprehension, (c) examine the relationship of each type of fluency and reading comprehension when controlling for other types of fluency, and (d) determine the extent to which fluency predicted change in comprehension over time as well as the associations of change in fluency with comprehension over time.

Results. Jenkins and Jewell (1993) found strong, statistically significant correlations between oral reading, Maze, and the achievement tests. A negative trend in scores with significant differences between grade levels was found across grade levels for the relationship between oral reading and the achievement tests; however, there was not a negative trend found across grade levels for Maze. In the study conducted by Nolet and McLaughlin (1997), researchers found significant gains in ORF rates from fall to winter, but no progress from winter to spring. Results suggested that gains in writing from practicing performance tasks much like the ones included on the MSAT were achieved at the cost of reading performance. Researchers questioned the relative benefits of implementation of statewide performance programs due to the undesirable consequences.

Kranzler et al. (1998) found processing speed and efficiency or general cognitive ability could not be used to explain the relationship between reading fluency and comprehension. The contribution of reading fluency in the prediction of reading comprehension was significant, but ORF did not have significant correlations with any of the Reaction Time Parameters of the ECTs used in the study. Results from a study conducted by Kranzler et al. (1999) indicated evidence of bias, as intercept bias was found for racial/ethnic groups in fourth and fifth grade and intercept and slope bias was found for gender in fifth grade. However, no bias was found for second and third grades. Findings suggested CBM reading is a biased test. Therefore, performance on CBM may overestimate or underestimate reading comprehension of students depending on race/ethnicity and gender.

Contradictory to findings of Kranzler et al. (1999), results of a study by Hintze et al. (2002) indicated CBM-ORF scores significantly predicted performance on reading comprehension and SES did not contribute to the prediction of scores. Findings suggested there was no differential predictive bias of reading comprehension skills across racial/ethnic groups when age was taken into account for performance on CBM. Findings from Roberts et al. (2005) indicated support for the efficiency of using retell fluency with fluency measures. Results of the study by Roberts et al. suggested ORF explained 57% of variance in prediction of scores on Broad Reading and a small amount of additional variance (58%) was explained with retell fluency. Yovanoff et al. (2005) indicated the importance of fluency and vocabulary in explaining comprehension. Based on findings, Yovanoff et al. suggested fluency was more important in fourth grade than in fifth grade and beyond. However, results suggested vocabulary was not constant across grade levels.

In another study, Riedel (2007) found DIBELS ORF was the single best predictor of satisfactory performance at the end of first grade and a good predictor of comprehension at the end of second grade. Researchers recommended a cut score of 38 at the end of first grade and suggested a strong relationship between DIBELS ORF and comprehension. Results also indicated the importance of vocabulary in comprehension since the group with poor comprehension scored 20 points lower on the vocabulary subtest. Burke and Hagan-Burke (2007) found DIBELS ORF to be the best predictor of both subtests on TOWRE when examined individually. Strong associations were found between RF and both subtests, which indicated the importance of a measure of comprehension. Schatschneider et al. (2008) found growth on ORF did not add to the prediction of future reading skills over and above prediction based on one final ORF assessment at the end of the year. In fact, ORF slope (growth) made little or no contribution to prediction of outcomes for first and second grade students because information about growth is already indicated on the final assessment. Results from Klauda and Guthrie (2008) indicated each type of fluency (word, syntactic, and passage) was significantly related to reading comprehension and associations were partially mediated by cognitive variables (background knowledge and inference). Reading fluency and comprehension had a bidirectional relationship at the syntactic level.

Table 3

Studies Published in Peer Reviewed Journals demonstrating the relationship between CBM and measures of overall reading achievement

Author(s)	Participants / Setting	Predictor Variable(s)	Outcome Measure(s)	Data Analysis
Jenkins & Jewell (1993)	335 students in second - sixth grade in Pacific Northwest	ORF Maze Passages, Teacher Judgment	MAT / Gates-MacGinitie Reading Tests	Cross-Grade Correlations Grade- Level Correlations
Purpose: To examine the relationship between informal measures of reading and measures of reading proficiency with heterogeneous and grade homogeneous samples. Results: Significant correlations were found between both informal reading measures and standardized achievement tests particularly at the earlier grade levels. Oral reading and performance on achievement tests were less strongly correlated at higher grade levels. Performance on measures was highly correlated to teacher judgment of proficiency.				

Table 3 (continued)

Author(s)	Participants / Setting	Predictor Variable(s)	Outcome Measure(s)	Data Analysis
Nolet & McLaughlin (1997)	58 students in fifth grade classrooms in urban Maryland	CBM-ORF, CBM Reading Retell, and CBM Written Exp.	Performance Assessment Task	Repeated measures ANOVA
Purpose: To examine growth of reading and written expression skills over the school year and performance on a task similar to the statewide performance assessment program. Results: Growth patterns in ORF indicated growth from fall to winter, but an overall decline in ORF scores from fall to spring. Gains in written expression were significant and researchers suggested both CBM and performance assessment were important.				
Kranzler, Brownell, & Miller (1998)	57 students in fourth grade in North Central Florida	CBM-ORF, K-BIT, ECTs	KTEA	Simultaneous Multiple Regression Analyses
Purpose: To examine the role of general cognitive ability, processing speed and efficiency, and ORF in the prediction of reading comprehension. Results: CBM-ORF was found to significantly predict reading comprehension. However, none of the ECT parameters were significantly correlated to reading fluency.				

Table 3 (continued)

Author(s)	Participants / Setting	Predictor Variable(s)	Outcome Measure(s)	Data Analysis
Kranzler, Miller, & Jordan (1999)	326 students in second - fifth grade in North Central Florida	CBM-ORF	CAT	Simultaneous Multiple Regression Analyses
Purpose: To examine group differences on CBM as an estimate of reading comprehension and examine racial/ethnic and gender bias on reading CBM. Results: Results indicated evidence of bias. Findings suggested performance on CBM may overestimate or underestimate reading comprehension of students depending on race/ethnicity and gender at particular grade levels.				
Hintze, Callahan, Matthews, & Tobin (2002)	136 students in second - fifth grade in Northeastern US	CBM-ORF	WJ-R	Hierarchical Multiple Regression Analyses
Purpose: To examine the differential prediction of reading comprehension based on performance on CBM-ORF for African American and Caucasian students. Results: Findings indicated age and CBM-ORF were the only significant predictors for reading comprehension scores. Neither SES nor ethnicity significantly added to the prediction of reading comprehension for African American or Caucasian students.				

Table 3 (continued)

Author(s)	Participants / Setting	Predictor Variable(s)	Outcome Measure(s)	Data Analysis
Roberts, Good, & Corcoran (2005)	86 students in first grade from six schools in urban southeast	CBM-ORF Reading Passages and Retell Protocols	WDRB	Correlation
Purpose: To determine the efficiency and effectiveness of using retell fluency with oral fluency measures, and to examine their relationship with overall reading competence. Results: Findings indicated ORF explained 57% of variance in prediction of scores on Broad Reading with retell fluency contributing a very small amount of additional explained variance (58%). Results suggested some support for inclusion of retell fluency to ORF measures as an efficient, useful tool.				
Yovanoff, Duesbery, Alonzo, & Tindal (2005)	6,012 students in fifth - eighth grades in the Northwest	CBM-ORF/ Vocabulary Measures/ Comprehension	District Reading Comprehension Test	Structural Equation Modeling
Purpose: To determine the importance of measurement of vocabulary and ORF in prediction of overall reading comprehension. Results: Findings suggested ORF was a significant predictor of reading comprehension, but effects diminished as grade level increased. Vocabulary knowledge was a significant predictor despite grade level increases.				

Table 3 (continued)

Author(s)	Participants / Setting	Predictor Variable(s)	Outcome Measure(s)	Data Analysis
Riedel (2007)	1,518 students in first grade in Memphis City Schools	DIBELS (LNF) DIBELS (PSF) DIBELS (NWF) DIBELS (ORF) DIBELS (RF)	(GRA + DE)/ TerraNova	ROC Analysis/ Logistic Regression
Purpose: To examine the relationship between DIBELS measures and overall reading achievement and determine optimal cut scores. Results: ORF was the single best predictor of comprehension at the end of first grade. MOY and EOY first grade results predicted second grade. Vocabulary was important.				
Burke & Hagan-Burke (2007)	213 students in first grade in one public K-2 primary school in Georgia	DIBELS (PSF) DIBELS (NWF) DIBELS (ORF) DIBELS (RF) DIBELS (WUF)	TOWRE PDE SWE	Regression Analyses/ Factor Analysis
Purpose: To examine the technical adequacy of early literacy measures as predictors of Phoneme Decoding and Sight Word Reading ability. Results: Moderate to strong correlations found between DIBELS ORF and PDE and SWE. DIBELS ORF was the best predictor of both subtests. DIBELS ORF had the highest factor loading of all measures. Strong associations between RF and both subtests.				

Table 3 (continued)

Author(s)	Participants / Setting	Predictor Variable(s)	Outcome Measure(s)	Data Analysis
Schatschneider, Wagner, & Crawford (2008)	23,438 students in first grade from Reading First in Florida	DIBELS ORF	SAT-10	Growth Curves / Multiple Regression
Purpose: To compare predictive validity of measures of achievement status and growth in achievement. Results: At the end of year, ORF made contribution to prediction, but slope (growth) made little or no contribution. Results suggested growth does not give additional information above and beyond one assessment point at the end of the year.				
Klauda & Guthrie (2008)	278 students in fifth grade from 13 classrooms in 3 schools in mid-Atlantic	IA BKA WRA PORA WJ-III	GMRT	Hierarchical Regression Analyses
Purpose: To investigate the extent to which word, syntactic, and passage fluency correlated with reading comprehension. Results: Results indicated a bidirectional relationship between comprehension and fluency. Fluency related significantly to performance on comprehension measures at each level, controlling for background knowledge, and inference skills.				

Predicting Performance on High-Stakes, Statewide Assessments Using Reading Curriculum-Based Measurement (R-CBM)

With research to indicate reading curriculum-based measurement (R-CBM) has high correlations with other standardized measures of overall reading achievement (Marston, 1989), researchers have focused efforts on student performance on CBM to predict performance on high-stakes assessments. Given that evidence has suggested ORF is an excellent indicator of overall reading competence (Fuchs et al., 1988; Shinn et al., 1992), further studies have focused on CBM-ORF. Despite the fact that CBM-ORF is a brief screening tool, there are significant implications for its predictive utility for future high-stakes assessments. This has been explored in the literature recently and research clearly suggests CBM can be used when attempting to determine whether a student will be successful on a future high-stakes assessment. Studies in the following section examined the use of CBM-ORF to predict performance on high-stakes statewide assessments of reading used for accountability.

Review of Published Studies Using CBM-ORF to Predict Performance on High-Stakes Statewide Assessments

Eight studies were located in peer reviewed journals between 2001 and 2008 that investigated the use of CBM-ORF measures to predict performance on statewide, high-stakes assessments used for accountability. A technical report and a paper presented at an annual conference, which are frequently referenced in the literature were also included in the review. The eight studies located in peer-reviewed journals are described in terms of purpose, participants and setting, measures used, data analysis, and results in the

following section and reviewed in Table 4. The additional two studies are included in the next section and summarized in Table 5.

Description of Studies. The purpose of each of the studies summarized in Table 4 was to investigate the relationship between CBM-ORF measures and performance on statewide assessments. The purpose of a study by Stage and Jacobsen (2001) was to determine whether CBM-ORF performance informed educators about performance on the Washington Assessment of Student Learning (WASL) reading assessment. In another study, Crawford et al. (2001) predicted performance on statewide assessment in reading and math using CBM. In their study, reading rates across two successive years were measured to determine the utility of oral reading rate in providing useful information and making predictions of student performance. Hixson and McGlinchey (2004) investigated the contribution of ORF, socioeconomic status, and race in predicting student performance on a high-stakes statewide test and a reading comprehension measure. In the same year, McGlinchey and Hixson (2004) studied the relationship between CBM and performance on high-stakes assessment. In their study, CBM was used to predict performance on Michigan Educational Assessment Program's (MEAP) as a replication of Stage and Jacobsen, including more students in a different state.

The purpose of two studies, Hintze and Silberglitt (2005) and Silberglitt and Hintze (2005) was to determine the relationship between high-stakes testing and R-CBM. Hintze and Silberglitt replicated and extended research on the relationship between high-stakes testing and R-CBM and compared statistical approaches to setting standards and determining cut scores. In the same year, Silberglitt and Hintze examined the extent to which R-CBM predicted performance on state mandated high-stakes tests and examined

the accuracy and appropriateness of various approaches in setting standards and determining cut scores.

In the next study reviewed, Shapiro et al. (2006) examined the relationship between CBM and standardized assessments including state mandated tests and norm-referenced standardized tests in reading and math. Finally, Keller-Margulis, Shapiro, and Hintze (2008) investigated the long-term relation between CBM and statewide achievement tests in order to identify students as early as first grade who were at risk for not passing statewide testing so that educators had sufficient time to change the trajectory of performance. In their study, they examined the relationship between benchmark data and the rate of growth for CBM in reading, math application, and math computation with student outcomes on statewide achievement tests.

Participants and Setting. Elementary school age students in various states were the participants in each of the studies summarized in Table 4. Stage and Jacobsen (2001) included 173 fourth grade students from an elementary school in Washington. Crawford et al. (2001) gathered data from 51 third grade students in blended classrooms in a rural school district in Oregon. All students participated in 2 years of the study. In Michigan, Hixson and McGlinchey (2004) included 376 fourth grade students from an urban school district. Also in Michigan, McGlinchey and Hixson (2004) used a total of 1,362 fourth grade students in one elementary school across eight years. In Minnesota, Hintze and Silberglitt (2005) included 1,766 students from seven elementary schools. Also in Minnesota, Silberglitt and Hintze (2005) used a total of 2,191 students from rural and outer-ring suburban elementary schools. Shapiro et al. (2006) included participants in two school districts in Pennsylvania. In District 1, a stratified random sample of third, fourth,

and fifth grade students was drawn from six elementary schools with a total of 617 students for reading and 475 students for math. In District 2, a sample of 431 students for reading and math was drawn across all schools. Keller-Margulis et al. (2008) included 1,461 elementary students in the starting reading normative sample and 1,477 elementary students in the math sample from a local norming project in an urban-suburban district in Pennsylvania.

Measures. Curriculum-based measurement oral reading fluency (CBM-ORF) measures and various statewide high-stakes assessments were used in each of the studies summarized in Table 4. Stage and Jacobsen (2001) used CBM-ORF measures developed from the Silver Burdette & Ginn Reading Series (Pearson et al., 1989). The three fluency passages were given at 3 benchmark times during the school year (fall, winter, and spring). The outcome measure used in the study was the Washington Assessment of Student Learning (WASL), which is an untimed test with multiple-choice, short answer, and extended response items to determine whether students are meeting state standards. Crawford et al. (2001) used three CBM passages from the Houghton Mifflin Basal Reading Series (1989) in January for each year of the study. During the second year of the study, a criterion-referenced statewide test of proficiency was used. Students were administered the reading and math sections of the Oregon Assessment of Knowledge and Skills (OAKS).

Two studies, Hixson and McGlinchey (2004) and McGlinchey and Hixson (2004) used CBM passages from the Macmillan Connections Reading Program (Arnold & Smith, 1987) and the Michigan Educational Assessment Program (MAEP) high-stakes, multiple-choice test designed to assess student progress toward meeting essential goals

and objectives in Michigan. In the first mentioned study conducted by Hixson and McGlinchey, researchers also used the Vocabulary and Reading Comprehension subtest of the Metropolitan Achievement Tests (MAT) for a Total Reading Score in the analysis.

In another two studies, Hintze and Silberglitt (2005) and Silberglitt and Hintze (2005) used standardized R-CBM benchmark passages developed by AIMSweb (Edformation, 2002) as the predictive measure and the Minnesota Comprehensive Assessment (MCA) as the criterion measure. The third grade reading MCA is an untimed, criterion-referenced test of reading proficiency administered over the course of 2 days in multiple-choice and constructed response format.

Shapiro et al. (2006) used standardized R-CBM benchmark reading probes developed by AIMSweb (Edformation, 2005) and math probes from Monitoring Basic Skills Progress (MBSP), including Math Computation (Fuchs, Hamlett, & Fuchs, 1998) and Math Concepts and Applications (Fuchs, Hamlett, & Fuchs, 1999) problems for each grade level. For outcome measures, researchers used The Pennsylvania System of School Assessment (PSSA), Stanford Achievement Test-9 (SAT-9), Metropolitan Achievement Test-8 (MAT-8), and Stanford Diagnostic Reading Test (SDRT). The PSSA is the statewide high-stakes achievement test in Pennsylvania to test student performance on state standards in multiple-choice format. Similarly, Keller-Margulis et al. (2008) used CBM-ORF probes developed by AIMSweb (Edformation, 2002), math computation probes from Monitoring Basic Skills Progress Math Computation (Fuchs et al., 1998), and math application probes from Monitoring Basic Skills Progress Math Concepts and Applications (Fuchs et al., 1999) for predictive measures. The outcome measures used in the study were the Pennsylvania System of School Assessment (PSSA)

and TerraNova Achievement Test, which is a standardized test of reading and mathematics achievement.

Data Analysis. Stage and Jacobsen (2001) used growth curves analysis using hierarchical linear modeling to examine individual student slopes in ORF across the school year. Researchers used multiple regression analyses to determine if ORF performance at different benchmark times or across the year better predicted WASL reading performance. Crawford et al. (2001) used descriptive statistics, correlations between oral readings and statewide testing in reading and math, and chi-square analyses to determine which level of oral reading rates were most predictive.

Hixson and McGlinchey (2004) used simultaneous multiple regression of racial group, lunch status, and CBM-ORF to predict performance on MEAP and MAT scores. This type of analysis allowed researchers to hold other variables constant while testing the significance of each variable. Stepwise regression procedure was used to determine the contribution of each variable in the prediction of performance. McGlinchey and Hixson (2004) determined accuracy of cut scores using diagnostic efficiency statistics including sensitivity, specificity, positive predictive power, negative predictive power, and overall correct classification. In two studies, Hintze and Silberglitt (2005) and Silberglitt and Hintze (2005) used descriptive statistics, logistic regression, ROC Curves, and discriminant analysis to determine the relationship between R-CBM and high-stakes testing and compare the three common approaches of establishing cut scores. In the later mentioned study, Silberglitt and Hintze also used equipercentile methods to determine the accuracy and appropriateness of the procedures.

Shapiro et al. (2006) used hierarchical regression analysis and ROC curves analysis to determine the contribution of CBM scores to outcomes on PSSA and norm-referenced, standardized achievement tests. Finally, Keller-Margulis et al. (2008) used correlation analyses to determine the relation between growth rate and performance on the statewide test after 1 year and after 2 years. Additionally, Receiver Operator Characteristic (ROC) Curves were used to identify specific cut scores for CBM probes in reading, math computation, and math application, as well as for determining cut scores for the rate of growth.

Results. Findings from Stage and Jacobsen (2001) indicated students below benchmark cut scores can be identified as at risk for failure on WASL reading assessment. Diagnostic efficiency was calculated to be 34% above chance for overall accuracy of ORF cut scores in prediction of performance on WASL. In their study, ORF scores in fall predicted scores on WASL more accurately than growth in ORF across the year. Results of the study by Crawford et al. (2001) indicated strong correlations between oral reading in second grade and reading rates in third grade along with moderate correlations between scores on criterion-referenced reading and math tests. Results of nonparametric analyses indicated students who read at least 119 wcpm passed the statewide reading test in third grade; furthermore, students who read at least 72 wcpm in second grade passed the statewide test in third grade. Hixson and McGlinchey (2004) found ORF, lunch status, and race each made a significant contribution to the prediction of performance on both reading comprehension measures, with CBM reading score as the strongest predictor of MEAP and MAT performance. Using stepwise regression analysis, no bias was found. ORF accounted for most of the variation with very little addition of

predictive power with lunch and race. Results suggested the addition of a comprehension measure may provide unbiased prediction, especially for older elementary level students. McGlinchey and Hixson (2004) found a moderately strong relationship between oral reading rates and performance on MEAP. Diagnostic efficiency statistics indicated 72% of students who reached 100 wcpm made a satisfactory score on MEAP.

Results from a study conducted by Hintze and Silberglitt (2005) indicated the predictive validity of R-CBM to MCA was significant at each benchmark time, but measures given with closer proximity in time yielded stronger results. Results of the study indicated all three statistical procedures identified cut scores that yielded diagnostic accuracy and efficiency which indicates R-CBM can be used as a predictor of MCA performance. In their comparison of the statistical procedures, ROC curves allowed diagnostic accuracy and efficiency while still providing flexible means for determination of cut scores across many different assessment decisions (i.e., screening, classification, entitlement). Researchers suggested when using scores only for classification or prediction purposes, an alternative method would be to use logistic regression with R-CBM as the predictor and high-stakes test as the criterion. Similarly, Silberglitt and Hintze (2005) found R-CBM to be a strong tool for predicting performance on MCA with strong correlations and greater than 80% chance of making accurate predications back to spring of first grade. Of the four methods used to generate cut scores, the strongest were logistic regression and ROC curves analysis, with ROC curve analysis providing the most flexibility. Shapiro et al. (2006) found strong relationships between CBM reading measures and PSSA as well as norm-referenced standardized tests. The strongest contributors to outcomes on the high-stakes assessment were the measures obtained in

winter and spring. CBM math computation was also a moderate predictor of performance on state assessments. Finally, results from Keller-Margulis et al. (2008) indicated data for reading and math benchmarks had moderate significant correlations with student outcomes on both statewide achievement tests. Also, the rate of reading growth in first grade had moderate, significant relationship to performance on statewide assessments in third grade. Findings suggested CBM can be used to identify students at risk and provided diagnostic accuracy for prediction of performance on statewide testing 1 year later and 2 years later. However, researchers suggested cut scores may need to be reexamined in relation to district norms.

Table 4

Studies Published in Peer-Reviewed Journals Using CBM-ORF to Predict Student Performance on High-Stakes Statewide Assessments

Author(s)	Participants / Setting	Predictor Variables	Outcome Measure(s)	Data Analysis
Stage & Jacobsen (2001)	173 students in fourth grade in an elementary school in Washington	CBM-ORF	WASL	ANOVA Growth Curves /HLM Regression
Purpose: To determine if CBM-ORF rates inform performance on the WASL for fourth grade students. Results: Findings indicate students who fall below the cut score are at risk for failure on WASL assessment. Slope of ORF also statistically significant, but level of ORF in fall, winter, or spring predicted performance better than growth.				
Crawford, Tindal, & Stieber (2001)	51 students in third grade in rural Oregon	(CBM-ORF)	OAKS	Correlation / Chi Square Analyses
Purpose: To analyze the relationship between ORF and scores on statewide achievement tests in reading and math. Results: The results supported using data from CBM to predict performance on testing. Students reading 119 wcpm in third grade and 72 wcpm in 2nd grade were proficient.				

Table 4 (continued)

Author(s)	Participants / Setting	Predictor Variable(s)	Outcome Measure(s)	Data Analysis
Hixson & McGlinchey (2004)	376 students in fourth grade in urban Michigan	CBM-ORF	MAEP MAT	Simultaneous Multiple Regression Stepwise Regression
Purpose: To investigate the relationship between ORF rate, socioeconomic status (SES), and race in prediction of performance on state reading assessment. Results: Findings indicated ORF, lunch status, and race made significant contributions in the prediction of performance on both measures of reading comprehension.				
McGlinchey & Hixson (2004)	1,362 students in fourth grade in Michigan	CBM-ORF	MEAP	Diagnostic Efficiency Statistics
Purpose: To investigate predictive validity of CBM reading probes in relation to MEAP performance and replicate study by Stage & Jacobsen (2001) in a different state. Results: Study results indicated a moderately strong relationship between oral reading rates and performance on MEAP and extended findings of Stage & Jacobsen with higher correlations. The percent agreement between wcpm and cut scores on MEAP performance was 74% overall correct classification.				

Table 4 (continued)

Author(s)	Participants / Setting	Predictor Variable(s)	Outcome Measure(s)	Data Analysis
Hintze & Silberglitt (2005)	1,766 students in seven elementary schools in Minnesota	R-CBM-ORF	MCA	Logistic Regression/ ROC Curves/ Discriminant Analysis
Purpose: To extend research on R-CBM and its relationship with high-stakes testing and compare statistical approaches to determine cut scores. Results: Findings indicate R-CBM strongly associated with MCA performance predicts performance on high-stakes tests from first grade. Consistent results across three statistical approaches. ROC curves yielded higher sensitivity, specificity, PPP, and NPP.				
Silberglitt & Hintze (2005)	2,191 students in first, second, and third grade in Minnesota	R-CBM-ORF	MCA	ROC Curves Analysis
Purpose: To examine usefulness of R-CBM for prediction of state-mandated tests and compare methods of setting standards and examine different approaches for cut scores. Results: Findings indicated moderate to high predictive and concurrent validity with high degree of diagnostic accuracy. R-CBM predicted with greater than 80% accuracy students likely to pass MCA. Stronger relationships found with closer administrations.				

Table 4 (continued)

Author(s)	Participants / Setting	Predictor Variable(s)	Outcome Measure(s)	Data Analysis
Shapiro, Keller, Lutz, Santoro, & Hintze (2006)	1048 students in third, fourth, and fifth grade in Pennsylvania	R-CBM-ORF, Math Concepts Math Computations	PSSA SAT-9 MAT-8 SDRT	Hierarchical Regression ROC Curves Analysis
Purpose: To examine the relationship between statewide standardized achievement tests and reading and math CBM. Results: Moderate to strong relationships found between CBM and high-stakes assessment and norm-referenced standardized tests. Scores in winter were the most powerful predictor with 125 and 126 wcpm having the highest sensitivity and specificity.				
Keller-Margulis, Shapiro, & Hintze (2008)	1,461 students in reading sample and 1,477 students in math	CBM-ORF, Math Computation & Application	PSSA TerraNova	ROC Curves Analysis
Purpose: To examine the relationship between benchmark data and growth on CBM in reading, math application, and math computation with outcomes on statewide tests. Results: Reading and math had moderate significant correlations with student outcomes on both statewide achievement tests and first grade reading growth rate had a moderate, significant relationship to performance on third grade statewide assessments. CBM provide diagnostic accuracy for prediction, but cut scores may need to be reexamined.				

Review of Technical Reports and Presentations Using CBM to Predict Performance on High-Stakes Statewide Assessments

Description of Studies. The purpose of both of the additional studies located in technical reports and papers presented at presentations was to determine the utility of measures of oral reading fluency to predict performance on statewide assessments. Sibley et al. (2001) examined the utility of established benchmarks for predicting student performance on high-stakes achievement testing in Illinois. Results were reported in a paper presented at the Annual Meeting of the National Association of School Psychologists in Washington, D.C. Buck & Torgesen (2003) conducted a study to determine whether performance on ORF measures were predictive of student achievement on Florida Comprehensive Assessment Test Sunshine State Standards (FCAT-SSS). Results were reported in a Florida Center for Reading Research Technical Report.

Participants and Setting. Each of the studies included students at the elementary school level. Sibley et al. (2001) included 112 fifth-grade students in two elementary schools in Illinois. The schools were located in a suburban school district. Buck and Torgesen (2003) included 1,102 third grade students from one school district in Florida.

Measures. Both studies used CBM-ORF measures and a statewide achievement test. Sibley et al. (2001) used CBM-ORF and Illinois Standards Achievement Test (ISAT) in reading, which is a group administered, multiple-choice test given in third, fifth, and eighth grade in Illinois. Additionally, researchers used the Level Test for Reading, which is a multiple-choice standardized test of achievement from local district goals administered to all students beginning in third grade. Buck and Torgesen (2003)

used CBM-ORF measures and the Reading Comprehension section of the Florida Comprehensive Assessment Test-Sunshine State Standards (FCAT-SSS), which is a norm-referenced test of student achievement.

Data Analysis. Sibley et al. (2001) reported correlation coefficients between reading fluency probes and high-stakes achievement tests. Linkages between established benchmarks for each grade level (second, third, and fourth) and performance on ISAT and Level Reading Test were reported. In the study conducted by Buck and Torgesen (2003), correlation coefficients were reported and sensitivity and specificity were calculated for predicting reading FCAT-SSS scores from ORF scores for White students, African-American students, Hispanic students, and students who do and do not receive free/reduced lunch.

Results. Strong correlations were found between CBM-ORF and ISAT in the study conducted by Sibley et al. (2001). Researchers found the established benchmarks for ORF accurately predicted performance on high-stakes state and local achievement measures with very strong links between CBM-ORF and ISAT as well as Level Reading Test. Buck and Torgesen (2003) found moderate to strong, significant correlations between ORF and FCAT-SSS. Results indicated ORF could be used to predict performance on FCAT. There were not significant interactions between racial background and free/red lunch status. ORF predicted scores on FCAT-SSS equally well for students of different races and SES groups.

Table 5

Technical Reports and Papers Presented at Conferences on Using CBM to Predict Performance on High-Stakes Statewide Assessments

Author(s)	Participants / Setting	Predictor Variables	Outcome Measure(s)	Data Analysis
Sibley, Biwer, & Hesch (2001)	112 students in fifth grade in Illinois	CBM-ORF	ISAT	Correlation
Purpose: To examine the utility of established benchmarks to student performance on state and local assessment instruments. Results: Strong correlations were found between CBM-ORF and ISAT with very strong links between CBM-ORF and state achievement and Level Reading Test. Established benchmarks for ORF had high utility for prediction of performance on high-stakes.				
Buck & Torgesen (2003)	1,102 students in third grade in Florida	ORF Measures	FCAT-SSS	Correlation Multiway Frequency
Purpose: To determine whether ORF performance is predictive of achievement on FCAT-SSS. Results: Moderate to strong, significant correlations were found between ORF and FCAT-SSS. ORF predicted scores on FCAT-SSS equally well for students of different races and SES groups. Results suggested ORF predicted performance on FCAT.				

Predicting Performance on High-Stakes, Statewide Assessments Using the Addition of Maze Measures with Oral Reading Fluency Measures

The Maze measure has been identified as an efficient measure of students' reading progress, and Maze has demonstrated sensitivity to growth (Shin et al., 2000). Recently, studies have been conducted in order to examine whether the addition of Maze measures increases the predictive power of oral reading fluency measures on high-stakes statewide assessments. This information can be useful to educators since it is a group administered assessment and uses limited instructional time to administer.

Review of Studies with Addition of Maze to CBM-ORF to Predict Performance on High-Stakes Statewide Assessments

Three studies published in peer reviewed journals between 2004 and 2006 were located in which researchers used Maze measures to predict performance on statewide assessments. The three studies are described in terms of purpose, participants and setting, measures, data analysis, and results. Information from each is reviewed and analyzed in relation to the present study and summarized in Table 6.

Description of Studies. The first study reviewed was conducted by Ardoin et al. (2004) to examine the predictive validity of CBM versus a group administered achievement test, the contribution of administering Maze, and the use of one versus three probes. Wiley and Deno (2005) conducted a study to determine whether the addition of the Maze procedure added to the predictive power of oral reading fluency measures for English Language Learners on the Minnesota Comprehensive Assessment (MCA) high-stakes state assessment. The purpose of Silberglitt et al. (2006) was to determine whether

the strength of the relationship between R-CBM, Maze, and state accountability tests changed as a function of grade.

Participants and Setting. All of the studies reviewed included elementary level students in third and fifth grade, and one study also included students in seventh and eighth grade. Participants in the study conducted by Ardoin et al. (2004) included 77 third grade students in one elementary school in the Southeast. Wiley and Deno (2005) included 36 students in third grade and 33 students in fifth grade in one urban elementary school in Minnesota. Also in Minnesota, Silberglitt et al. (2006) used a total of 5,472 students in third, fifth, seventh, and eighth grades from five rural and suburban districts.

Measures. In the study conducted by Ardoin et al. (2004), reading curriculum-based measurement (R-CBM) and Maze were used as predictors of Woodcock-Johnson-III (WJ-III; Woodcock, McGrew, & Mather, 2001) and Iowa Test of Basic Skills (ITBS). In a study conducted by Wiley and Deno (2005), researchers used Standard Reading Passages (Children's Educational Services, 1987) and Maze measures from the Basic Academic Skill Samples (BASS: Deno, Espin, Maruyama, & Cohen, 1989). The Maze measures had the first sentence left intact and every seventh word replaced with a choice of three words. The score was the number of correct word choices made in 1 minute. Silberglitt et al. (2006) used R-CBM grade level passages from Silver Burdett and Ginn Reading Series (Pearson et. al., 1989) and ORF probes by AIMSweb (Edformation, 2002). The Standard Reading Assessment Passages (Howe & Shinn, 2002) and Maze measures used by Silberglitt et al. (2006) were formatted similar to those used by Wiley and Deno. Both studies used the Minnesota Comprehensive Assessment (MCA). Also,

Silberglitt et al. used the Basic Standards Test – Reading (BST-R) for eighth grade students.

Data Analysis. In the study conducted by Ardoin et al. (2004), researchers used T-tests to examine significant differences between dependent correlations, Z-tests to determine if significant differences existed between predictors, and hierarchical multiple regression to examine the validity of using Maze in addition to R-CBM. Wiley and Deno (2005) determined whether Maze added to ORF in the prediction of scores on MCA using multiple regression analyses. Silberglitt et al. (2006) used correlation and Fisher Transformation to determine the amount of variance between scores.

Results. Ardoin et al. (2004) found high correlations between the predictors (CBM and Maze) and reading achievement and comprehension, but CBM was a better predictor of reading achievement and comprehension than Maze. Findings also suggested administration of only one R-CBM probe was effective for identifying risk. Additionally, results suggested CBM was a more accurate predictor of overall reading achievement than WJ-III, but ITBS-RC was a better predictor of reading comprehension. Wiley and Deno (2005) found moderate to strong correlations between ORF and MCA and Maze and MCA for third and fifth grade students and provided evidence that oral reading and Maze measures were predictive of student performance on MCA. Maze was a better predictor of performance than oral reading for fifth grade non-EL students and slightly better for third grade non-EL students, accounting for significant variance in scores beyond oral reading for non-EL students. For EL students, Maze did not account for additional variance. Silberglitt et al. (2006) found strong correlations between CBM and Maze in all grade levels indicating R-CBM scores were significantly related to state

accountability scores. In later grades, R-CBM continued to account for a substantial amount of variance, but the value of R-CBM diminished as the grade level increased.

Table 6

Review of Studies Using Addition of Maze to CBM-ORF to Predict Performance on High-Stakes Statewide Assessments

Author(s)	Participants / Setting	Predictor Variables	Outcome Measure(s)	Data Analysis
Ardoin et al. (2004)	77 student in third grade in the Southeast	R-CBM Maze	WJ-III ITBS	Hierarchical Multiple Regression/ Simultaneous Regression
Purpose: To examine the contribution of Maze in addition to CBM, the predictive validity of CBM, and the use of three versus one reading probe. Results: Results indicated CBM and Maze had high correlations with reading achievement and comprehension, but CBM was a better predictor of overall reading achievement than maze. Findings also suggested administering only one R-CBM probe is an effective way to identify students at risk for reading difficulty.				

Table 6 (continued)

Author(s)	Participants / Setting	Predictor Variable(s)	Outcome Measure(s)	Data Analysis
Wiley & Deno (2005)	36 third grade 33 fifth graders in Minnesota	ORF Maze	MCA	Multiple Regression Analyses
Purpose: To determine if the Maze procedure adds to the predictive power of General Outcome Measures of Oral Reading on MCA high-stakes state assessment for ELL. Results: Findings indicated oral reading and Maze measures predict performance on the MCA in reading with moderate to moderately strong correlations. For EL students, oral reading was a better predictor of performance and the Maze task did not add to the prediction; however, for non-EL students in fifth grade, Maze was a better predictor than oral reading.				
Silberglitt, Burns, Madyun, & Lail (2006)	5,472 third, fifth, seventh, and eighth grade students in Minn.	R-CBM Maze	MCA-R BST-R	Correlation A Fisher Transformation
Purpose: To analyze the relationship between R-CBM, Maze, and state accountability tests and to determine if strength of relationship changes as a function of grade. Results: Statistically significant correlations were found between R-CBM and Maze in all grade levels indicating R-CBM scores were significantly related to state accountability scores. Value of R-CBM diminished as grade level increased.				

Predicting Performance on High-Stakes Statewide Assessments Using Dynamic Indicators of Basic Early Literacy Skills (DIBELS) Measures

With the passage of current legislation, schools face the reality of high-stakes assessment. In response to increased accountability, many schools use Dynamic Indicators of Basic Early Literacy Skills (DIBELS) for universal screening. DIBELS are standardized, individually administered tests of accuracy designed to identify children in need of additional support. DIBELS has benchmark and progress monitoring probes available for educators to monitor progress toward instructional goals (Good & Kaminski, 2002). When used as a screening instrument, DIBELS are highly useful for data-based decision making. Recently, studies have been conducted in order to examine the utility of the DIBELS ORF measure in predicting student achievement on high-stakes assessments. However, DIBELS were not created to predict outcomes on the types of assessments that were designed to measure progress toward state curriculum standards for accountability purposes. For educators, this is important because it can provide meaningful information early enough to make data-based decisions to improve student outcomes. Goals can be established and instruction can be altered based on progress toward optimal cut scores. This section includes a review of studies published in peer reviewed journals and technical reports of studies in which researchers used DIBELS ORF scores to predict performance on statewide assessments. Studies are described in terms of purpose, participants and setting, predictor and outcome measures, data analysis, and results. Nine studies were located in peer reviewed journals published between 2001 and 2009, and four additional studies were located in technical reports between 2002 and 2005. Studies published in peer reviewed journals are summarized in Table 7, and

technical reports are summarized in Table 8. Information from each study is reviewed and analyzed in relation to the present study in the following two sections.

Review of Published Studies Using Dynamic Indicators of Basic Early Literacy Skills (DIBELS) to Predict Performance on High-Stakes Statewide Assessments

Description of Studies. The purpose of each of the studies reviewed in the following section was to determine the relationship between scores on DIBELS and statewide, high-stakes assessments. The purpose of a study conducted by Good et al. (2001) was to explore the utility of DIBELS fluency-based indicators to predict reading outcomes, inform educational decisions, and change outcomes for students. Later, Wood (2006) examined the relationship between DIBELS ORF and performance on a statewide reading test. Schilling et al. (2007) examined the predictive validity of DIBELS fluency based measures on year-end reading assessment and the predictive utility of using established DIBELS benchmarks to identify students. Roehrig et al. (2008) evaluated the validity of DIBELS ORF in predicting performance on measures of reading comprehension and the utility of established ORF cutoffs for predicting high-stakes outcomes. Baker et al. (2008) investigated the relationship between ORF and high-stakes reading tests and examined whether slope of performance added to prediction of performance above and beyond initial performance. Additionally, researchers in this study investigated how well ORF stood up in prediction models for predicting performance on high-stakes tests the second year. Shapiro et al. (2008) examined the diagnostic accuracy of DIBELS ORF and 4Sight Benchmark Assessment and utility for identification of students at risk for reading difficulty. Researchers in this study also determined the degree to which the additional measure of comprehension enhanced

prediction of high-stakes assessment performance over DIBELS ORF alone. Chard et al. (2008) conducted a study within schools implementing a school-wide prevention model to examine reading development for students. The focus of the study was which variables in first grade students predict later reading achievement on high-stakes assessment in third grade. Wood (2009) analyzed the relationship between cognitive and reading measures using path modeling in order for multiple direct and indirect effects between predictors and outcome variables to be tested simultaneously. The outcome measures used in the path models were word identification, ORF, and reading comprehension. The predictor measures were vocabulary knowledge, orthographic speed, pseudo word reading, and rapid naming digits. In order to determine the accuracy of universal screening tools used within an RTI framework, Catts et al. (2009) examined the impact of floor effects on the predictive validity of DIBELS, which is a highly used screening instrument. Most recently, Goffreda, et al. (2009) conducted a study to investigate predictive validity of scores on DIBELS.

Participants and Setting. Each of the studies reviewed in this section involved data gathered from students at the elementary school level in various states. Good et al. (2001) included 4 cohorts of students from kindergarten through third grade in six elementary schools in Oregon. In their study, they included a total of 3,478 individual scores on Phoneme Segmentation Fluency (PSF), Nonsense Word Fluency (NWF), and Oral Reading Fluency (ORF) between 1998 and 2000. In a study conducted in Colorado, Wood (2006) included 281 participants in a public elementary school in third (n=82), fourth (n=101), and fifth (n=98) grades. Schilling et al. (2007) gathered data from first grade (n= 2,588), second grade (n = 2,437), and third grade (n=2,527) students attending

44 schools in nine districts in Michigan that made up the first Reading First cohort during the 2003-2004 school year. Roehrig et al. (2008) included 35,207 students in third grade in Florida Reading First Schools. Participants were split into two samples, with 17,409 students in the first calibration and 17,798 in the cross-validation sample. However, participants without FCAT score were removed, which reduced the number of participants in the calibration (n=16,539) and cross-validation (n=16,908) samples. Baker et al. (2008) also conducted a study in Oregon. Four cohorts of students in kindergarten through third grade from 34 Oregon Reading First schools participated in this study, with approximately 2,400 students in each cohort. Shapiro et al. (2008) collected data from a total of 1,000 students in six elementary schools in Pennsylvania. Students were in third grade (n=401), fourth grade (n=394), and fifth grade (n=205) across three districts. Chard et al. (2008) included longitudinal data from 668 students in first grade in Oregon and Texas. Wood (2009) included 74 students who were followed longitudinally from third grade through fourth grade in an elementary school in Colorado. Catts et al. (2009) obtained data from the Progress Monitoring and Reporting Network (PMRN) in Florida for 18,667 students enrolled in Florida Reading First schools. Data were gathered for students who began kindergarten in the 2003-2004 school year. In the most recent study, Goffreda et al. (2009) included longitudinal data from a total of 67 first grade students from a rural school district in Pennsylvania.

Measures. All of the studies reviewed in this section used Dynamic Indicators of Basic Early Literacy Skills (DIBELS) fluency measures, and a statewide, high-stakes assessment. DIBELS are standardized, individually administered tests of accuracy designed to identify children in need of additional support and monitor progress toward

instructional goals (Good & Kaminski, 2002). In the study conducted by Good et al. (2001), Phoneme Segmentation Fluency (PSF), Nonsense Word Fluency (NWF), and Oral Reading Fluency (ORF) DIBELS measures were used. Additionally, researchers used the Test of Reading Fluency (Children's Educational Services, 1987) to assess ORF for third grade. The Test of Reading Fluency is a standardized set of passages for individual administration using standardized administration procedures. The high-stakes outcome measure used in this study was Oregon Statewide Assessment (OSA), which is a standardized, multiple-choice measure of comprehensive reading achievement. OSA is used in Oregon used to assess individual achievement levels and compare performance with Oregon performance standards. Wood (2006) used DIBELS ORF and Colorado Student Assessment Program (CSAP) Reading Test. The CSAP is a measure of reading comprehension designed to assess whether students attain state standards at each level. The test, which includes multiple-choice and constructed response questions, is administered to all students in third, fourth, and fifth grades in Colorado.

Schilling et al. (2007) gathered data from DIBELS measures appropriate for each grade level. For first grade students, the measures included Letter Naming Fluency (LNF) for the fall administration, Phoneme Segmentation Fluency (PSF) for each administration, Nonsense Word Fluency (NWF) for each administration, Oral Reading Fluency (ORF) for the spring administration, and Word Use Fluency (WUF) for each administration. For second grade students, the measures included NWF for the fall administration, ORF for each administration, and WUF for each administration. For third grade students, the measures included ORF and WUF for each administration. For each grade level, the outcome measure was Iowa Test of Basic Skills (ITBS), which includes

vocabulary, word analysis, listening, language, and reading comprehension subtests. Roehrig et al. (2008) used DIBELS ORF to predict performance on the Florida Comprehensive Assessment Test (FCAT-SSS) and Stanford Achievement Test (SAT-10) reading comprehension measures. The Florida Comprehensive Assessment Test (FCAT-SSS) is a group administered, criterion-referenced test consisting of reading passages followed by multiple-choice items in the areas of main idea, words and phrases in context, comparison/cause and effect, and reference/research. The SAT-10 is an untimed, multiple-choice, group-administered test of overall reading proficiency with word study skills, word reading, sentence reading, and reading comprehension subtests. In that same year, Baker et al. (2008) also used DIBELS ORF measures to predict performance on Stanford Achievement Test-Tenth Edition (SAT-10) for first and second grade students, but used Oregon Statewide Reading Assessment (OSRA) for third grade students. The Oregon Statewide Reading Assessment is an untimed, multiple-choice test of reading achievement administered to all students in third grade in Oregon.

Using an additional measure of comprehension, Shapiro et al. (2008) used DIBELS ORF and 4Sight Benchmark Assessment (4Sight) to predict outcomes on the Pennsylvania System of School Assessment (PSSA). With a format similar to a statewide assessment, 4Sight is a group administered, multiple-choice test of reading comprehension designed to be predictive of outcomes on the statewide assessment. The PSSA, including multiple-choice and open ended tasks, is the statewide measure of accountability designed to assess whether students are meeting state standards in reading. Chard et al. (2008) used DIBELS Letter Naming Fluency (LNF), Phonemic Segmentation Fluency (PSF), Nonsense Word Fluency (NWF), and Oral Reading

Fluency (ORF) measures (Good & Kaminski, 2002) as well as Growth Modeling Oral Reading Fluency Passages (GMORF; Fuchs, 2003) to predict performance on the Reading Vocabulary and Reading Comprehension subtests of the Stanford Achievement Test-10 (SAT-10), as well as the Word Identification , Word Attack, and Passage Comprehension subtests of the Woodcock Reading Mastery Tests-Revised (WRMT-R; Woodcock, 1987). Chard et al. also used a Social Skills Rating System (SSRS; Gresham & Elliot, 1990) to document the teachers' perceived academic competence. Wood (2009) used DIBELS ORF to measure oral reading fluency, the Word Identification subtest of the Woodcock-Johnson III Test of Achievement to measure word identification, and the Word Attack subtest from the Woodcock-Johnson III Test of Achievement (WJ-III; Woodcock, et al., 2001) to measure pseudo word reading. To measure orthographic speed, a form of the Orthographic Coding Test was used. Rapid Naming was administered from the Comprehensive Test of Phonological Processing (Wagner, Torgesen, & Rashotte, 1999) and vocabulary knowledge was measured from the Vocabulary subtest of the Wechsler Intelligence Scale for Children –IV (WISC; Wechsler, 2003). The statewide reading test, Colorado Student Assessment Program (CSAP), was used to measure reading comprehension. Catts et al. (2009) used DIBELS measures (ISF, LNF, PSF, NWF) to predict outcomes on the DIBELS ORF measure, while the Reading Comprehension subtest of the SAT-10 served as the outcome measure for DIBELS ORF. Finally, Goffreda et al. (2009) used DIBELS first grade benchmark measures (LNF, PSF, NWF, ORF) to predict scores on TerraNova California Achievement Test, Second Edition (TerraNova; CTB/McGraw-Hill, 2005) and Pennsylvania System of School Assessment (PSSA; Pennsylvania Department of

Education, 2005). The TerraNova consists of measures in reading/language arts, mathematics, science, and social studies, but only scores for second grade reading/language arts were examined. PSSA consists of measures in reading, mathematics, and writing, but only scores in reading for third grade students were examined.

Data Analysis. Good et al. (2001) used a series of longitudinal studies linking the four cohorts and examined correlation coefficients and percentage of variance to determine the strength of relations among foundational reading measures and the statewide reading assessment in third grade. Wood (2006) used hierarchical linear modeling to analyze the relationships. Information at level 1 included the individual student's ORF scores, level 2 included students nested in classrooms, and level 3 included classrooms nested within grade levels. Both 2- level and 3-level models were used for analyses. Schilling et al. (2007) determined the extent to which DIBELS predicted scores using hierarchical regression analyses. Additionally, Receiver Operating Characteristic (ROC) functions were analyzed to assess the optimal decision rule for identifying student level of risk. Roehrig et al. (2008) also generated ROC curves with a calibration and cross validation sample to examine sensitivity and specificity of cut score values. Optimal cut scores were determined and tested in a 2 x 2 contingency table. Baker et al. (2008) used growth curve analyses to test the intercept and slopes of ORF trajectories prediction on SAT-10 performance. The initial growth model was compared to a set of models that predicted performance on comprehensive reading tests and fit within a structural equation modeling framework. Shapiro et al. (2008) generated ROC curves to determine accuracy and probability of correct classification of risk.

Additionally, researchers used logistic regression to determine whether performance on PSSA was enhanced by scores on 4Sight and ORF versus ORF alone. Chard et al. (2008) used growth modeling and path analysis to determine significant predictors of reading comprehension and vocabulary achievement as well as growth in oral reading fluency. Wood (2009) analyzed relationships between cognitive and reading measures using path modeling. Catts et al. (2009) used quantile regression, which is similar to ordinary least squares analysis, to show the change in correlation between the predictor and outcome variables at various administration points. Logistic regression analysis was used to examine the predictability of DIBELS measures. Similarly, Goffreda et al. (2009) used logistic regression to determine the predictive validity of risk categories identified by first-grade DIBELS indicators and third grade PSSA proficiency as well as second grade TerraNova proficiency. Receiver Operating Characteristic (ROC) Curves were also used to determine cutoff scores. Inspections of the area under the curve (AUC) were used to determine sensitivity and specificity levels for each measure.

Results. Good et al. (2001) found high correlations between earlier and later skills with variance explained ranging from 12% to 67%. Results supported fluency as an important foundation for reading competence. Students who read 110 words were likely to meet or exceed expectations on the state assessment, and students who read only 70wcpm were not likely to meet expectations. Similarly, Wood (2006) found strong relationships with ORF and performance on statewide reading proficiency assessments. Results of the study indicated ORF predicted performance equally well for CSAP in third, fourth, and fifth grade. ORF was a unique predictor of CSAP performance above previous year performance. The study provided the first evidence that classroom level

variables can influence how well ORF predicts performance on statewide testing. Schilling et al. (2007) found performance on DIBELS ORF was significantly related to performance on ITBS across all administrations. Results indicated the importance of foundational skills for prediction of performance on ITBS decreased as grade level increased. In addition, the association of fluency with comprehension decreased as grade level standards of vocabulary knowledge and text inferences increased. Researchers in the study also found the overall discrimination of ORF was better when a combination of the some risk and at risk rules were used. This rule improved identification of students below the 50th percentile on ITBS at the end of the year.

Strong correlations between ORF and statewide achievement testing were also found in the study conducted by Roehrig et al. (2008). ORF was found to be the most significant predictor of risk on FCAT-SSS and SAT-10, with the third administration having the strongest correlations. In their study, the analyses showed no evidence of predictive bias across demographic groups. Additionally, race, SES, and language were not significant contributors to performance. Additionally, researchers suggested more students could be identified using recalibrated scores. Baker et al. (2008) also found strong correlations between ORF and SAT-10 high-stakes test in second grade as well as between ORF and OSRA in third grade. Results indicated ORF provided a stronger index of overall reading proficiency in second grade than in third grade, and ORF slope added to the accuracy of predicting performance.

Consistent with other studies using DIBELS to predict performance on statewide assessments, Shapiro et al. (2008) found significant correlational relationships between ORF, 4Sight, and PSSA. The combination of using ORF and 4Sight improved the

accuracy of prediction of student performance. In order to maximize sensitivity and specificity for fluency measures, researchers suggested established DIBELS benchmark cut points may need to be adjusted. Results of the study by Chard et al. (2008) indicated growth from first grade to third grade on ORF is curvilinear with deceleration in growth as grade level increased. Also, results of the model suggested ORF slope and spring of first grade passage comprehension as the strongest predictors of performance on comprehension and vocabulary on SAT-10 at the end of third grade. Significant predictors of spring of first ORF which jointly accounted for 75% of ORF initial status variance included fall of first LNF, spring of first AP, spring of first academic competence rating, and AP by competence interaction. Significant predictors of ORF slope which accounted for 11% of slope variance included spring of first AP and AP by competence interaction.

Wood (2009) found the relationship between ORF and CSAP was not significant with third grade students when vocabulary knowledge and word identification measures were included in the model. However, with fourth grade students, the path from ORF was significant, indicating ORF, word identification, and vocabulary knowledge were significant predictors of CSAP. In the study by Catts et al. (2009), strong floor effects were found for the measures with the initial administration, but the floor effects lessened across subsequent administrations. ORF was a good predictor of reading outcomes on the SAT-10, but optimal rates of predictability were not reached until second grade. Finally, Goffreda et al. (2009) found performance on ORF to be the only statistically significant predictor of PSSA proficiency. Also, using classification accuracy values for each DIBELS indicator and DIBELS recommended cutoff scores, ORF was the only measure

with adequate sensitivity and specificity (77% sensitivity, 88% specificity). Optimal cut scores were determined using ROC Curves and ORF was still the only indicator to demonstrate adequate sensitivity and specificity (88% sensitivity, 88% specificity). Using ROC Curves, the optimal ORF cutoff score for PSSA proficiency was 23, compared to DIBELS-recommended benchmark cutoff score of 20.

Table 7

Studies Published in Peer Reviewed Journals Using Dynamic Indicators of Basic Early Literacy Skills (DIBELS) to Predict Student Performance on Statewide Assessments

Author(s)	Participants / Setting	Predictor Variables	Outcome Measure(s)	Data Analysis
Good, Simmons, & Kame'enui (2001)	3,478 students in kindergarten – third grade in Oregon	DIBELS (OnRF) DIBELS (PSF) DIBELS (NWF) DIBELS (ORF) DIBELS (WUF)	CBM-ORF OSA	Correlation
Purpose: To investigate the utility of DIBELS benchmark goals for decision-making and to determine the strength of the relationship between CBM-ORF and high-stakes reading. Results: Findings support utility of DIBELS benchmark goals. Students who attained earlier goals were likely to meet subsequent goals. Students reading 110 wcpm were likely to meet or exceed expectations. Students not reading 70 wcpm were not likely to meet expectations.				

Table 7 (continued)

Author(s)	Participants / Setting	Predictor Variable(s)	Outcome Measure(s)	Data Analysis
Wood (2006)	281 students in third, fourth, and fifth grade in Colorado	DIBELS (ORF) Previous year CSAP scores	CSAP	Hierarchical Linear Modeling
Purpose: To investigate classroom and grade level variation and evaluate ORF assessment as a valid index of performance on statewide reading proficiency tests. Results: Found strong relationship between ORF and statewide assessment across grade levels with ORF significant and unique predictor above previous year performance.				
Schilling, Carlisle, Scott, & Zeng (2007)	2,588 first, 2,437 second and 2,527 third grade in Michigan	DIBELS (ORF)	ITBS	Hierarchical regression ROC Curves
Purpose: To examine effectiveness of DIBELS measures as predictors of reading achievement on statewide assessment and determine predictive validity of established DIBELS benchmarks for identifying below grade level performance at the end of the year. Results: Results indicated DIBELS performance at each administration and across all three administrations significantly related to performance on ITBS. Overall discrimination of ORF based ROCs indicated combination of at risk and some risk is best prediction.				

Table 7 (continued)

Author(s)	Participants / Setting	Predictor Variable(s)	Outcome Measure(s)	Data Analysis
Roehrig, Petscher, Nettles, Hudson, & Torgesen (2008)	35,207 students in third grade in Florida Reading First Schools	DIBELS (ORF)	SAT-10 FCAT-SSS	ROC Curve Logistic Regression Analysis
Purpose: To determine predictive and concurrent validity of ORF, to investigate DIBELS cut scores and adjust, and to evaluate ORF for predictive bias. Results: Results indicated strong correlations of ORF with SAT-10 and FCAT-SSS, with third administration strongest. Recalibrated scores identified more students at risk. No evidence of predictive bias. Race, SES, and language not significant contributors to risk.				
Baker et al. (2008)	2,400 students in kindergarten through third grade in Oregon	DIBELS (ORF)	SAT-10 ORSA	Growth Curve Analysis SEM
Purpose: To examine relationship between ORF and high-stakes reading tests, examine slope of ORF, and to test the predictive performance of ORF on high-stakes reading tests. Results: Study results indicated strong association between ORF and high-stakes tests. ORF slope added to the accuracy of prediction, accounting for over 95% of variance.				

Table 7 (continued)

Author(s)	Participants / Setting	Predictor Variable(s)	Outcome Measure(s)	Data Analysis
Shapiro, Solari, & Petscher (2008)	1,000 students in third, fourth, and fifth grade in Pennsylvania.	DIBELS (ORF)	4Sight Benchmark Assessment/ PSSA	ROC Curves Analysis / Logistic Regression
Purpose: To examine predictive value of ORF at fall and winter administration on PSSA performance in late winter and investigate the addition of 4Sight Benchmark Assessment. Results: Results indicated strong correlations between ORF, 4Sight, and PSSA for third and fourth grade. Prediction of benchmark level students fairly accurate, but those below less likely to be predicted accurately. Addition of 4Sight improved accuracy of prediction.				
Chard et al. (2008)	688 students in first, second, and third grade in Oregon and Texas	DIBELS (LNF) DIBELS (PSF) DIBELS (NWF) DIBELS (ORF) GMORF	WRMT-R SAT-10 SSRS	Path Analysis
Purpose: To examine reading development, determine predictive variables for performance on standardized, high-stakes reading comprehension measures and vocabulary achievement. Results: Findings indicated growth from first to third grade on ORF is curvilinear with deceleration in growth as grade level increases. ORF slope and spring of first grade passage comprehension had the strongest effects on comprehension and vocabulary on SAT-10.				

Table 7 (continued)

Author(s)	Participants / Setting	Predictor Variable(s)	Outcome Measure(s)	Data Analysis
Wood (2009)	74 students in fourth grade in Colorado	WJ-III Test of Phonological Processing WISC-IV	WJ-III DIBELS (ORF) CSAP	Path Modeling
Purpose: To analyze the relationship between cognitive and reading measures. Results: The effects of orthographic speed, pseudo word reading, and rapid naming on comprehension were mediated through ORF and word identification. Path from ORF to comprehension not significant in third grade, but significant in fourth grade.				
Catts, Petscher, Schatschneider, Bridges, & Mendoza (2009)	18,667 students who began kindergarten in 2003-2004 in Florida	DIBELS (ISF) DIBELS (LNF) DIBELS (PSF) DIBELS (NWF) DIBELS (ORF)	DIBELS (ORF) SAT-10	Quantile Regression Logistic Regression Analyses
Purpose: To examine distribution of scores and the impact that floor effects have on predictive validity of a common screening instrument (DIBELS). Results: Floor effects of DIBELS measures were found in initial administrations and lessened across administrations. ORF level by second grade was found to be a good predictor of outcomes on the SAT-10 for third grade students.				

Table 7 (continued)

Author(s)	Participants / Setting	Predictor Variable(s)	Outcome Measure(s)	Data Analysis
Goffreda, Diperna, & Pedersen (2009)	67 students in first grade	DIBELS (LNF) DIBELS (PSF) DIBELS (NWF) DIBELS (ORF)	TerraNova CAT PSSA	Logistic Regression ROC Curves Analysis
Purpose: To examine the predictive validity of students' risk categories established by DIBELS in first grade and reading proficiency on district tests in second grade and state standardized assessments in third grade. Results: DIBELS ORF was the only measure to yield adequate levels of sensitivity and specificity. ORF yielded high levels of sensitivity and specificity with DIBELS-recommended (77%, 88%, respectively) and optimal cutoff scores (88%, 88%, respectively). DIBELS ORF found to be effective tool to use when predicting later reading proficiency. Optimal and DIBELS-recommended cutoff scores not significantly different.				

Review of Technical Reports Using Dynamic Indicators of Basic Early Literacy Skills (DIBELS) Measures to Predict Performance on High-Stakes, Statewide Assessments

Description of Studies. All of the studies reviewed in this section were conducted to determine the relationship between DIBELS and high-stakes, statewide assessments in various states. The purpose of a study reported in a technical report by Shaw and Shaw (2002) was to determine the utility of DIBELS ORF for prediction of placement level on the Colorado State Assessment Program (CSAP) test of reading comprehension. Barger

(2003) reported a similar study in a technical report conducted to determine the connection between ORF benchmark scores and student performance on the North Carolina End-of-Grade Reading Comprehension Test. In a research brief by the Assessment and Evaluation Department of Tempe School District No. 3, Wilson (2005) reported a study conducted to determine the usefulness of reaching benchmark level on ORF to influence the likelihood of students meeting standards on Arizona Instrument to Measure Standards (AIMS). In the same year, VanderMeer, et al. (2005) reported results of a study in an Ohio Technical Report. The study was conducted to examine the end of third grade as well as the beginning and end of fourth grade benchmark DIBELS goals, CBM-ORF goals, in relation to student performance on the Fourth Grade Ohio Proficiency Test (OPT) in Reading.

Participants and Setting. All of the technical reports involved students at the elementary school level. Shaw and Shaw (2002) obtained scores for 52 students in a third grade elementary school in Colorado. Barger (2003) included thirty-eight third grade students in an elementary school in Buncombe County in North Carolina. The analysis conducted by Wilson (2005) also included third grade students (n=241) from an elementary school in Arizona. Finally, a total of 364 students who were in third grade and tracked to fourth grade the following year from a suburban elementary school in southwest Ohio were included in the study by VanderMeer et al. (2005).

Measures. In one study, Shaw and Shaw (2002) assessed students using DIBELS ORF for fall, winter, and spring benchmarks. The Colorado State Assessment Program (CSAP) was used as the outcome measure. Barger (2003) used DIBELS ORF spring benchmark measures and the North Carolina End-of-Grade (NC EOG) Reading

Comprehension Assessment. The NC EOG is a test of reading comprehension with passages followed by multiple-choice questions. Similarly, Wilson (2005) used DIBELS ORF spring benchmark assessment. The outcome measure used by Wilson was the Arizona Instrument to Measure Standards (AIMS), which is a multiple-choice test of reading proficiency with an emphasis on comprehension. In a technical report, VanderMeer et al. (2005) described a study using three measures. DIBELS ORF fall and spring benchmarks were used along with CBM-ORF and Ohio Fourth Grade Reading Proficiency Test (OPT). The OPT is a multiple-choice, short answer, and extended response reading test to determine whether students have met fourth grade level literacy proficiency.

Data Analysis. All of the technical reports reviewed reported correlation coefficients between DIBELS ORF and a statewide assessment. Shaw and Shaw (2002) reported correlation coefficients for DIBELS ORF and CSAP by benchmark assessment time (fall, winter, spring) and displayed median DIBELS scores for performance levels of CSAP. Barger (2003) reported correlations between DIBELS ORF and NC EOG. The number of students at each EOG level was reported by ORF score. Wilson (2005) reported correlation coefficients between AIMS and DIBELS ORF. Using scaled scores and DIBELS ORF, percentages of students in each category (low risk, at risk, some risk) were displayed to reflect who met or did not meet proficiency. Additionally, the cross-classifications and correlation between AIMS and ORF were reported for demographic subgroups. In the study by VanderMeer et al. (2005), correlation coefficients among DIBELS ORF and OPT scores were reported and percentages were shown for students in each risk category who met or did not meet proficiency.

Results. The technical report of DIBELS ORF and CSAP by Shaw and Shaw (2002) had strong correlations ranging between .80 and .93 for DIBELS spring and fall scores and CSAP. Authors reported that 90% of students who reached the benchmark goal of 110 scored proficient or advanced on CSAP. Also, 43% of students who scored below 110 on DIBELS ORF scored below proficiency on CSAP. Results indicated using DIBELS ORF scores to predict performance on CSAP correctly classified 74% of student scores. Results from Barger (2003) indicate ORF could be accurate predictor of proficiency on NC EOG. Students who read at least 100 wcpm scored proficient, with 92% of students who read at least 110 wcpm achieving Level IV scores. The dividing line for making accurate predictions was 100 wcpm with a target goal of 110. For students who read below 69 wcpm, prediction of performance was more difficult. Wilson (2005) found moderately strong, positive correlations between DIBELS ORF and AIMS. ORF accurately identified students likely to meet proficiency and those who were unlikely to reach proficiency. Results of student performance in demographic subgroups did not vary from overall results which showed that students in low risk category are likely to score above proficiency and students in the at risk category are likely to score below proficiency regardless of subgroup. Finally, VanderMeer et al. (2005) also found significant correlations between ORF and OPT reading test. Overall, high percentages of students in third grade and fourth grade who scored the benchmark score on DIBELS ORF and CBM-ORF scored proficient on ORT. Researchers concluded that benchmark goals at each level were sufficient for accurate prediction of performance level on OPT.

Table 8

Technical Reports of Studies Using Dynamic Indicators of Basic Early Literacy Skills (DIBELS) to Predict Student Performance on High-Stakes Statewide Assessments

Author(s)	Participants / Setting	Predictor Variables	Outcome Measure(s)	Data Analysis
Shaw & Shaw (2002)	52 students in third grade in Colorado	DIBELS -ORF	CSAP	Correlation
Purpose: To examine the utility of DIBELS ORF as a predictor of placement level in third grade reading Colorado State Assessment Program (CSAP). Results: DIBELS spring and fall scores have strong correlation with CSAP, with correlations ranging between .80 and .93. Most students (90%) who reached the benchmark goal of 110 scored proficient or advanced on CSAP.				
Barger (2003)	38 students in third grade in North Carolina	DIBELS -ORF	NC EOG	Correlation
Purpose: To determine the connection between ORF and achievement on North Carolina End-of-Grade Reading Comprehension Test. Results: Study shows ORF could be accurate predictor of proficiency on NC EOG. Students who read at least 100 wcpm scored proficient. The dividing line for making prediction was 100 wcpm and correlation below this level was less clear.				

Table 8 (continued)

Author(s)	Participants / Setting	Predictor Variable(s)	Outcome Measure(s)	Data Analysis
Wilson (2005)	241 students in third grade in Arizona Reading First Schools	DIBELS ORF	AIMS	Correlation
Purpose: To determine if ORF scores influence the likelihood of meeting standards on AIMS Reading test. Results: Correlation between ORF and AIMS was positive and moderately strong. ORF identified students likely to meet proficiency and those who were unlikely to reach proficiency with good accuracy. Student performance in demographic subgroups was similar to overall results.				
VanderMeer, Lentz, & Stollar (2005)	364 students in third grade in Ohio	DIBELS ORF CBM-ORF	OPT	Correlation
Purpose: To examine DIBELS benchmark goals in comparison to expectations on OPT and examine relationship of ORF with OPT. Results: Significant correlations were found between ORF and OPT for reading. Overall, high percentages of students in third grade and fourth grade who scored benchmark score on DIBELS ORF and CBM-ORF scored proficient on ORT.				

Summary

Research on CBM-ORF has a long history of being a valid and reliable measure of student achievement in reading. Moderate to strong relationships have been found between CBM-ORF and scores on overall reading achievement tests. In addition, research has demonstrated that the Maze measure is a reliable, sensitive, and valid procedure (Shin et al., 2000). In fact, scores on Maze measures have been linked to student performance on high-stakes tests. Recent research examining the relationship between DIBELS ORF and states' reading tests has shown that scores on DIBELS ORF can be used to predict performance on statewide reading assessments. Further investigation to determine the relationship between Maze, DIBELS, and outcomes on high-stakes assessment was warranted.

DIBELS, like other general outcome measures, was not designed to predict performance on statewide assessments. However, it has been used extensively in elementary schools to determine which level of support is needed. These measures are used to make decisions about appropriate supplemental reading instruction and to determine whether students respond to instruction. Therefore, recent research has provided critical information about the relationship between ORF and high-stakes assessment. In each of the studies reviewed, ORF was associated with performance on high-stakes testing, with moderate to strong correlations. All studies supported fluency as an important foundation for reading competence and predictor of performance on statewide assessments.

Throughout the literature, research has consistently demonstrated the link between fluency and comprehension. Theory supports the assumption that efficient low-level

word recognition frees up capacity for higher level comprehension processing of text (LaBerge & Samuels, 1974) and research has suggested that gains in fluency have been shown to generalize to gains in reading comprehension (Fuchs et al., 2001). Literature reviewed demonstrated a strong association between fluency and comprehension. All studies linked ORF rates to performance on measures of overall reading achievement. Additionally, strong correlations were found between ORF and statewide, high-stakes reading assessments. Research on such formative measures that impact student learning and promote improvement of student outcomes on summative assessments is vital. Using formative measures as predictors of reading achievement on summative measures, such as the End-of-Grade test in reading comprehension, is critical to allow educators to make data-based instructional decisions for students with and without disabilities.

The next logical step in this area of research is to examine the degree that a measure of oral reading fluency that is used by many systems (DIBELS ORF) and a measure of reading comprehension (AIMSweb Maze) predict performance on a statewide reading assessment. Both DIBELS ORF and AIMSweb Maze are short, accurate indicators of overall reading competence. There is a need to investigate the relationship between scores on DIBELS ORF and AIMSweb Maze and outcomes on a large-scale assessment of reading comprehension. Additionally, there is a need to determine whether grade level differences exist in the relationship, to examine the accuracy of established DIBELS benchmark goals, and to determine optimal cut scores to predict proficiency on the statewide assessment for third, fourth, and fifth grade students.

CHAPTER 3: METHOD

The purpose of the current study was to investigate the relationship between outcomes from a state's large-scale reading comprehension assessment and scores on the Dynamic Indicators of Basic Indicators of Basic Early Literacy Oral Reading Fluency Measure (DIBELS ORF) and AIMSweb Maze curriculum-based measurement (Maze-CBM). Specifically, a nonexperimental research design using correlational methodology was used to examine the degree that DIBELS ORF and AIMSweb Maze-CBM predict standard scores on the comprehension measures of the North Carolina End-of-Grade (EOG) Reading Comprehension assessment. Other objectives that guided the study were to examine differences in the magnitude of the relationship as a function of grade, to determine the accuracy of established DIBELS benchmark cutoff scores, and to establish optimal cut scores to predict proficiency on the statewide assessment for third, fourth, and fifth grades. The following four research questions were investigated.

Research Questions

1. Using DIBELS Oral Reading Fluency (ORF) and AIMSweb Maze-CBM universal screening scores, which measure or combination of measures are the best predictors of standard scale scores on a state developed reading accountability measure for third, fourth, and fifth grade students?
2. Is there a difference in the magnitude of the relationship between EOG and ORF and Maze among third, fourth, and fifth grade?

3. How accurate are published DIBELS ORF risk level cutoff scores for ORF and AIMSweb Maze scores for identifying third, fourth, and fifth grade students who will or will not be proficient as measured by the statewide grade level NC EOG Reading Comprehension test?
4. What are the optimal DIBELS ORF and AIMSweb Maze-CBM cut scores to use when attempting to predict satisfactory reading comprehension by the end of third, fourth, and fifth grade level as measured by EOG performance?

Description of Participants, Setting, and Measures

Participants

Participants included in this study were 336 students enrolled in third, fourth, and fifth grade in one public elementary school during the 2008-2009 school year. The school had a total enrollment of 645 students in kindergarten through fifth grade. There were 117 students in third grade, 115 students in fourth grade, and 122 students in fifth grade. Overall, the school population was comprised of 78.5% White, 11.5% African American, 5.8% Hispanic, 1.9% Asian, 1.7% Multi, and less than 1% American Indian. The school had a total of 14% of the population identified in the Exceptional Children's Program (6% AIG, 8% EC). Students who received free/reduced lunch made up 34% of the school population.

All data were de-identified by school administration prior to the researcher receiving the information. Therefore, the characteristics of the specific sample included in the study could not be determined. Additionally, scores for students who participated in NCEXTEND 2 Reading Comprehension assessment were not included in data obtained from the school. Therefore, the total number of students participating in the

study included 110 third grade students, 111 fourth grade students, and 115 fifth grade students.

The universal screening used for all students was Dynamic Indicators of Basic Early Literacy Skills Oral Reading Fluency (DIBELS ORF) and AIMSweb Maze Curriculum-based Measures (Maze-CBM). Participants were eligible for participation if they meet the following selection criteria: (a) enrolled in grades 3-5, (b) obtained a DIBELS ORF score from spring benchmark assessment, (c) obtained an AIMSweb Maze-CBM score from spring benchmark assessment, and (d) obtained a standard score from North Carolina End-of-Grade Reading Comprehension statewide assessment. Students with disabilities and students coded as English Language Learners were included in the analyses as long as they were not tested using NCEXTEND 2 for reading.

Setting

All assessment took place in one public, elementary school in a suburban school district in the southeastern United States. The school was selected to participate based on a sufficient number of students in each grade level using DIBELS and AIMSweb Maze for benchmark assessments, permission from the principal, and a vested interest of the researcher as the special education teacher at the school. The school was the only school in the school district using DIBELS for universal screening in kindergarten through fifth grade. All other elementary schools in the district used AIMSweb Curriculum-Based Measures in reading for first grade only. During the 2008-2009 school year, participants from 16 third, fourth, and fifth grade classrooms were individually assessed by a trained benchmark team using the DIBELS EOY benchmark. All DIBELS measures were administered one-on-one in the media center of the school in an enclosed area, free of

distractions. The student could not see or hear any other students during assessment. All Maze measures were group administered by a trained general education teacher. Participants were assessed as a group by their individual classroom teacher in their general education classroom using Benchmark #3 AIMSweb Maze-CBM. All EOG Reading Comprehension Assessment took place in the general education classroom with a trained administrator and proctor present in at all times. Students who required testing accommodations were tested in the setting specified on their Individualized Education Plan (IEP) or 504 Plan with a certified Teacher of Exceptional Children and a proctor.

Measures

Dynamic Indicators of Basic Early Literacy Skills (DIBELS) Oral Reading Fluency. Dynamic Indicators of Basic Early Literacy Skills (DIBELS; Kaminski & Good, 1998). Oral Reading Fluency (ORF) measures are 1-minute timed, fluency measures that take into account accuracy and speed of reading grade-level connected text. The measures assess progress on aspects of reading development important for students at each grade level in fall, winter, and spring. DIBELS ORF is standardized and individually administered, requiring students to read three passages aloud for 1 minute each. The score is the number of words read correctly per minute. Words read incorrectly or omitted, and hesitations of more than 3 seconds are counted as errors, but words self-corrected within 3 seconds are counted as accurate.

The measures were designed to be (a) sensitive to change in student performance, (b) easy to administer, (c) time efficient, (d) cost effective, (e) capable of frequent administration, and (f) representative of important skill areas (Kaminski & Good, 1996). Reliability and validity data on DIBELS ORF indicate alternate form reliability for one

probe is .90 and criterion-related validity is .70 - .80 (Good & Kaminski, 2002; Rouse & Fantuzzo, 2006).

Administration of DIBELS ORF occurred during the second week of May for End-of-year (EOY) Spring Assessment. Each student was individually assessed using three grade-level specific passages and the median score across the three passages was recorded as the overall score. The school had a "Benchmark Team" that consisted of teachers and assistants who had received at least 5 hours of training on the administration and scoring of the ORF measures. Members of the team were checked for reliability and validity prior to each benchmark assessment and had to score within two words correct on reliability in order to achieve acceptable validity during checkouts. Any member of the team who did not have acceptable reliability and validity was used as a runner instead of an assessor for that benchmark assessment. Data were collected from the DIBELS spring benchmark reading passages for third, fourth, and fifth grade with the median score across the three passages recorded as the variable to reflect reading fluency.

AIMSweb Maze Curriculum-Based Measurement. AIMSweb Maze Curriculum-based measures (Edformation, 2009) are fluency based assessments developed by Pearson as part of the AIMSweb system. The measures are curriculum independent in order to assess student skills regardless of differences in curriculum (Edformation). Maze-CBM reading is a standardized, group-administered, multiple-choice cloze task to measure comprehension skills. The student reads one standard 150-400 word, grade-level reading passage silently. The first sentence is left intact, while the following sentences have each seventh word replaced with three words inside parenthesis for the student to choose the correct word. One of the choices is correct and the other two choices are

distracters, one near and one far. The near distracter does not make sense in the sentence, but is the same part of speech as the word that makes sense (e.g., noun, verb, adjective) and the far distracter is a randomly selected word that does not make sense. Students are given 3 minutes to complete the task. The number of words correctly circled is recorded as the score.

The concurrent validity of Maze and oral reading rate indicates the Maze measure has within-grade correlations ranging from .63 to .76 with other standardized measures of reading (Jenkins & Jewell, 1993). Correlations with all grades combined range from .80 to .85. This implies that Maze measures similar constructs as Gates MacGinitie Reading Tests, Metropolitan Achievement Tests, and Oral Reading Measures within each grade and across grade levels.

Administration of AIMSweb Maze occurred during the second week of May. The EOY spring benchmark for Maze was given by the general education classroom teachers, who each received at least 2 hours of training on the group administration and scoring of AIMSweb measures. Each student was assessed using three grade-level specific passages. The median score across the three passages was recorded as the overall score. Data were collected from the AIMSweb Maze spring benchmark passages for third, fourth, and fifth grade. The overall score was recorded as the variable to reflect reading comprehension performance.

North Carolina End-of-Grade Reading Comprehension Edition 3 Test. The North Carolina End-of-Grade (NC EOG) Test is administered annually to students in each grade level during the last 3 weeks of school. NC EOGs are designed to measure student performance on grade-level goals and objectives and to assess whether students are

attaining state standards in the 2004 North Carolina English Language Arts Standard Course of Study (NC Department of Instruction, Division of Accountability Services, Testing Section, 2009). Each student reads eight reading selections and answers six to nine questions following each selection for a total of 58 questions. There are four literary selections (two fiction, one nonfiction, one poem), three informational selections (two content and one consumer), and one embedded experimental selection for students to complete in order to assess their ability to read for (a) literary experience, (b) gaining information, and (c) performing a task.

The third edition of the End-of-Grade Reading Comprehension test administration provides an estimated time schedule of 158 minutes to complete the assessment, but allows students to take as long as they need to complete the test (up to twice the estimated time required or 4 hours). Student raw scale scores are reported within achievement level ranges of Level I through Level IV. Students must achieve at least a Level III to demonstrate grade-level reading comprehension skills as required in the North Carolina Standard Course of Study. In order to achieve a Level III, third grade students must have at least 66 to 68% correct across forms. Students in fourth and fifth grade must have at least 62 to 64% correct across forms.

Technical information about the NC EOG Reading Comprehension test indicates that the test is highly reliable as a whole. All of the forms have high reliability coefficient alpha indices averaged across forms by grade, with results indicating third grade = 0.925, fourth grade = 0.912, and fifth grade = 0.900. Reliability indexes across forms for males and females and various ethnic groups indicate a high degree of reliability across gender and ethnicity, with averages ranging from .873 to .927. The standard error of

measurement for a given score ranges from 3 to 6 points for third and fourth grade and 3 to 5 points for fifth grade, which indicates high accuracy of an obtained score.

Additionally, North Carolina Department of Public Instruction (NCDPI, 2009) found moderate to strong criterion-related validity correlations in test scores and predicted scores for third, fourth, and fifth grades, with results indicating coefficients of 0.66, 0.63, and 0.61, respectively. Moderate to strong validity correlations in test scores and predicted achievement by raw score were also reported for third, fourth, and fifth grade, with coefficients of 0.69, 0.68, and 0.67, respectively. The NC EOG Reading Comprehension Test has a moderate to strong correlation between scale scores and external variables, with correlation coefficients ranging from 0.50 to 0.69.

Data Analysis

Data Analysis

This section describes the procedures that were used in analyzing the data in order to address the research questions. This study used a nonexperimental design with data collected and entered into a SPSS, which stands for *Statistical Package for the Social Sciences* (Landau & Everitt, 2004). SPSS is a widely used package for analyzing, manipulating, and presenting data. NCSS Statistical Analysis and Graphics Software (Hintze, 2007) was used for Receiver Operating Characteristic (ROC) Curve Analysis.

Procedure

Research Question One: Using DIBELS Oral Reading Fluency (ORF) and AIMSweb Maze-CBM measures, which measure or combination of measures are the best predictors of standard scale scores on a state developed reading accountability measure for third, fourth, and fifth grades?

In order to answer the first research question, data obtained from the school were entered into SPSS. All data were de-identified by school administration in order to protect the identity of the students. The database was examined by a second viewer for accuracy of data entry. A simultaneous multiple regression analysis was used to examine which variable or combination of variables best predicted standard scale scores on the NC EOG Reading Comprehension test for each grade level. Each grade level was analyzed and reported separately. Prior to analysis, data for each grade level were screened for missing variables, outliers, normality, and assumptions. Descriptive statistics and correlations were examined for strength of associations between the predictor variables (DIBELS ORF and Maze-CBM) and the outcome measure (NC EOG). Results of the multiple regression for each grade level were reported, including the unstandardized regression coefficients (*B*) and intercept, the standardized regression coefficients (β), and semipartial correlations (sr_i). The variance accounted for (R^2) and adjusted R^2 values were also reported to determine the variability in EOG standard scores predicted by ORF and Maze measures.

Research Question Two: Is there a difference in the magnitude of the relationship between EOG and ORF and Maze among third, fourth, and fifth grade?

In order to answer the second research question, the strength of the relationship between EOG, ORF, and Maze was examined with a Fisher transformation. Fisher's z' transformation converted Pearson's r to the normally distributed variable z. Once the Fisher transform was computed, the transformed data was analyzed in terms of its deviation from the mean. Correlation coefficients between grade levels were compared using a Fisher Transformation to determine if grade differences existed in the relationship

between EOG, Maze, and ORF. Coefficients for ORF and Maze to the EOG for each grade level were compared, with an alpha level of .05 necessary to demonstrate a significant finding.

Research Question Three: How accurate are published DIBELS benchmark risk level cutoff scores for ORF and AIMSweb Maze Aggregate Norm 50th percentile scores for identifying third, fourth, and fifth graders who will or will not be proficient as measured by the North Carolina End-of-Grade (EOG) Reading Comprehension test?

Research question three was investigated using Receiver Operating Characteristic (ROC) Analysis (Swets, Dawes, & Monahan, 2000). ROC curve analysis was used to examine the relationship between DIBELS ORF, Maze, and comprehension. This analysis was chosen because ROC curve analysis has been shown to demonstrate more flexibility in estimation of diagnostic accuracy and predictive power (Silberglitt & Hintze, 2005) than discriminant analysis and has been successful in identifying cut scores resulting in higher sensitivity, specificity, positive and negative predictive power (Silberglitt et al., 2006). Diagnostic accuracy of DIBELS ORF and AIMSweb Maze measures were tested by generating a receiver operating characteristic (ROC) curve.

Since ROC curves were originally used in electronic signal-detection theory and have recently become widely used in the psychology and medical field (Swets et al., 2000), the terminology used typically relates to the presence of a disease (positive) or absence of a disease (negative). In the medical field, the accuracy of diagnostic tests used to predict breast cancer and prostate cancer have been assessed using ROC curves as well as various other diagnostic tests. In education, with prediction of a dichotomous outcome, such as satisfactory or poor performance on a measure of reading comprehension, there is

a possibility of four possible outcomes. The term *positive* indicates there is a problem with comprehension. The term *negative* indicates there is no problem with comprehension when ROC curves are used with DIBELS ORF to predict outcomes on standardized reading comprehension measures. The first possible outcome with DIBELS is a *true positive* if there is indication of a problem (low ORF score) and there truly is a problem with comprehension (low EOG score). The second possible outcome is a *false negative* if there is no indication of a problem (high ORF score) and there is a problem with comprehension (low EOG score). The third possible outcome is a *true negative* if there is no indication of a problem (high ORF score) and there is no problem with comprehension (high EOG score). Finally, the fourth possible outcome is a *false positive* if there is indication of a problem (low ORF score) and there is no comprehension problem (high EOG score). The term *sensitivity* refers to the proportion of positives correctly identified as positives. The term *specificity* refers to the proportion of negatives correctly identified as negatives. Optimal prediction would result in 100% sensitivity (i.e., predict all students from the not proficient EOG group as at risk for proficiency) and 100% specificity (i.e., predict all students from the proficient EOG as not at risk for proficiency).

In order to answer research question three, conditional probability indices were calculated using NCSS Statistical Software (Hintze, 2007). The cutoff for Level III performance on the EOG for each grade level was used as the cutoff for binary outcome of not proficient (0) and proficient (1). Positive predictive values (PPV) were calculated to show the probability that a student who is identified as being at risk is truly at risk. Negative predictive power (NPP) was calculated to show the chances that a student who

is identified not at risk is truly not at risk. Accuracy rate (AR) was calculated to show the percentage of students who were correctly classified. Misclassification rate was calculated to show the percentage of students who were incorrectly classified. The misclassification rate (MR) was computed as the proportion of all misclassified students (the sum of false positives and false negatives) out of all students (Gonen, 2007). The area under the curve (AUC) was generated as part of the ROC analysis in order to provide the probability of the independent variable correctly classifying a pair of individuals when one student is at risk and the other is not.

No studies were located that established cutoff scores for risk levels for Maze measures; however, AIMSweb provides norms for student performance at the 25^{th}, 50^{th}, 75^{th}, and 90^{th} percentiles (Edformation, 2009) that were used as cutoff scores. Cut points by the publishers of DIBELS were established by using the percentage of students who achieved subsequent literacy goals (Good et al., 2002). The publishers of DIBELS used ROC curves and target percentages of students in the risk categories (low risk, some risk, at risk) to determine cutoff scores for benchmark, strategic, or intensive support. For example, the prediction of students who would achieve Nonsense Word Fluency (NWF) goals based on their performance on the preceding critical skill of Phoneme Segmentation Fluency (PSF) were used in the determination of cutoff scores for risk levels.

For the current study, logistic regression was also used to check correct classification rates. Logistic regression allows prediction of a discrete outcome (e.g., poor comprehension / no poor comprehension) and shows the probability of an outcome for each case. For the analyses, the dependent variable was dichotomously coded to indicate

proficient or not proficient comprehension for each grade level. Results were summarized and the significant predictor(s) and classification accuracy were reported.

Research Question Four: What are the optimal DIBELS ORF and AIMSweb Maze-CBM cut scores to use when attempting to predict satisfactory reading comprehension by the end of third, fourth, and fifth grade level as measured by EOG performance?

ROC curve analysis was used to examine the relationship between DIBELS ORF, Maze, and comprehension for each grade level and to determine optimal cut scores. ROC curves were used to examine the proportion of students correctly classified as at risk on both ORF and NC EOG (i.e., sensitivity or true positives) and the proportion of students correctly classified as ***not*** at risk on both measures (i.e., specificity or true negatives). Sensitivity and 1-specificity pairs were plotted on the ROC Curve using NCSS (Hintze, 2007). Optimal cut scores for the samples were determined by examination of sensitivity and specificity values for each cutoff value. The optimal values are typically represented at the shoulder of the ROC curve (Swets, 2000). The scores were tested in a 2 X 2 contingency table to determine DIBELS ORF and AIMSweb Maze scores that would identify the greatest proportion of students as true positives and true negatives (i.e., at risk and not at risk).

CHAPTER 4: RESULTS

The purpose of this study was to investigate the relationship between outcomes from a state's large-scale reading comprehension assessment and scores on DIBELS ORF and Maze-CBM, to determine if there are grade level differences in the magnitude of the relationship, to examine the accuracy of benchmark cut scores, and determine optimal cutoff scores to predict proficiency on the statewide assessment for third, fourth, and fifth grades. This chapter describes the results of the data analyses used to examine each of the four research questions. The collected data were entered into SPSS and NCSS statistical programs to examine the research questions. The statistical procedures used in the study are described in this chapter. First, procedures used to screen data are described. This section is followed by a description of statistical analyses used to address each research question.

Data Screening Procedures

Prior to conducting the major analysis, all data were entered into the SPSS database and examined by a second viewer for accuracy of data entry. Reliability of data entry was 100%. When data entry was complete and validated, data were copied from the SPSS database to the NCSS database. Students who participated in NC EXTEND 2 testing were removed from the data set during the de-identification process by school administration; therefore, data for these students were not included in the final database. There were a total of 336 participant cases included in the study, with 110 participants in

the third grade data set, 111 in the fourth grade data set, and 115 in the fifth grade data set. All cases had complete data with no missing variables; therefore, all cases were included in the analyses.

Research Questions

Research Question One: Using DIBELS Oral Reading Fluency (ORF) and AIMSweb Maze-CBM universal screening scores, which measure or combination of measures are the best predictors of standard scale scores on a state developed reading accountability measure for third, fourth, and fifth grade students?

Third Grade. Using the statistical program, SPSS, a standard multiple regression was conducted between scores on third grade EOG as the outcome variable and DIBELS ORF and AIMSweb Maze as the predictor variables. SPSS EXPLORE was used for evaluation of assumptions and analysis was performed using SPSS REGRESSION. Prior to analysis, data were screened for missing data, outliers, and assumptions. Of the total cases (N = 110), there were no missing variables. Two univariate outliers were detected, but all cases were retained for analysis. With the use of $p < .001$ criterion for Mahalanobis distance, no multivariate outliers were found among the cases.

Descriptive statistics including mean, median, standard deviation, skewness, and kurtosis for the variables are reported in Table 9. An examination of the skewness values and visual inspection of boxplots and frequency distributions suggested that the distributions of all variables were approximately normally distributed. Examination of bivariate scatterplots indicated that there were linear relationships between all the variables. A preliminary regression was used to create a residual plot. The shape of the scatterplot did not indicate a violation of any of the assumptions of regression. Normality

and homoscedasticity can be assumed because the scatterplot indicated an approximately normal distribution of residuals. The correlation coefficients among the variables are reported in Table 10. As shown in Table 10, correlation coefficients for ORF and Maze met or exceeded .50 and were significant ($p < .01$). Multicollinearity was not a significant problem with this dataset, as the matrix of correlations between the variables did not indicate any correlations above .90. Collinearity diagnostics indicated all dimensions with a condition index under 15, which verified no possible multicollinearity problems with the data. However, dimension 3 (eigenvalue = .017) could indicate an ill-conditioned crossproduct matrix, with an eigenvalue close to zero. Dimension 1 (eigenvalue = 2.910) and dimension 2 (eigenvalue = 0.073) did not indicate collinearity problems.

Results of the multiple regression indicated both of the independent variables (or predictor variables) contributed significantly to the prediction of EOG. The unstandardized regression coefficients (*B*) and intercept, the standardized regression coefficients (β), and semipartial correlations (sr_i) are reported in Table 11. The variance accounted for (R^2) equaled .349 (adjusted $R^2 = .337$), which was significantly different from zero ($F = 28.72$, $p < .01$). The adjusted R^2 value of .349 indicated that more than a third of the variability in EOG standard scores is predicted by ORF and Maze measures. ORF had the largest positive standardized beta ($\beta = .363$) and semipartial correlation coefficient ($sr_i = .259$), but Maze had a similar, statistically significant positive standardized beta ($\beta = .276$) and semipartial coefficient ($sr_i = .197$).

Fourth Grade. A standard multiple regression was conducted between scores on fourth grade EOG, DIBELS ORF, and AIMSweb Maze. Prior to conducting the analysis, SPSS EXPLORE was used for evaluation of assumptions. Analysis was performed using

SPSS REGRESSION. Data were screened for missing data and outliers. In the total cases (N = 111), there were no missing variables. Three univariate outliers were detected, but all cases were retained for analysis. With the use of $p < .001$ criterion for Mahalanobis distance, one multivariate outlier was found among the cases, but retained for analysis.

Descriptive statistics including mean, median, standard deviation, skewness, and kurtosis for the variables are reported in Table 9. An examination of the skewness values and visual inspection of boxplots and frequency distributions suggested that the assumption of univariate normality of Maze may be slightly questionable since the skewness and kurtosis of this measure were greater than 1.0. However, distributions of ORF and EOG were approximately normally distributed with skewness and kurtosis only slightly greater than or less than zero. Despite the increased chance of Type I error, a preliminary regression was used to create a residual plot. Evaluation of the scatterplot indicated an approximately normal distribution of residuals. Examination of bivariate scatterplots indicated that there were linear relationships between all the variables. Correlation coefficients for the variables are reported in Table 10. As shown in Table 10, correlation coefficients for ORF and Maze met or exceeded .50 and were significant ($p < .01$). Assumption of multicollinearity was satisfactory, as all variables indicated correlations below .90. Collinearity diagnostics indicated no dimensions with a condition index over 15. However, dimension 3 (eigenvalue = .023) was a condition with an eigenvalue close to zero, which could indicate an ill-conditioned crossproduct matrix. Dimension 2 (eigenvalue = 2.890) and dimension 2 (eigenvalue 0.087) did not signify any collinearity problems.

The unstandardized regression coefficients (*B*) and intercept, the standardized regression coefficients (β), and semipartial correlations (sr_i) are reported in Table 11. The variance accounted for (R^2) equaled .481 (adjusted R^2 = .472), which was significantly different from zero (F = 50.14, p < .01). The adjusted R^2 value of .481 indicated that nearly half of the variability in EOG standard scores was predicted by ORF and Maze measures. Both of the independent variables (or predictor variables) contributed significantly to the prediction of EOG. ORF had the largest positive standardized beta (β = .441) and semipartial correlation coefficient (sr_i = .317), but Maze had a similar, statistically significant positive standardized beta (β = .310) and semipartial coefficient (sr_i = .223).

Fifth Grade. A standard multiple regression was conducted between scores on fifth grade EOG, DIBELS ORF, and AIMSweb Maze. Prior to conducting the analysis, SPSS EXPLORE was used to screen data for missing data, outliers, and evaluation of assumptions. Analysis was performed using SPSS REGRESSION. There were no missing variables in any of the total cases (N = 115). One univariate outlier was detected, but all cases were retained for analysis. With the use of p < .001 criterion for Mahalanobis distance, no multivariate outliers were found among the cases.

Descriptive statistics including mean, median, standard deviation, skewness, and kurtosis for the variables are reported in Table 9. An examination of the skewness values and visual inspection of boxplots and frequency distributions suggested that the assumptions of univariate normality and linearity were satisfactory. Distributions of ORF and EOG were approximately normally distributed with skewness and kurtosis only slightly greater than or less than zero. A preliminary regression was used to create a

residual plot. Evaluation of the scatterplot indicated approximately normal distribution of residuals. Examination of bivariate scatterplots indicated there were linear relationships between all the variables. Correlation coefficients are reported in Table 10. As shown in Table 10, correlation coefficients for Maze and ORF met or exceeded .50 and were significant ($p < .01$). Assumption of multicollinearity was satisfactory, as all variables indicated correlations below .90. Collinearity diagnostics indicated no condition index over 15. However, dimension 3 (eigenvalue = .016) was a condition with an eigenvalue close to zero, which could indicate an ill-conditioned crossproduct matrix. Dimension 1 (eigenvalue = 2.932) and dimension 2 (eigenvalue = 0.052) did not indicate any collinearity problems.

The unstandardized regression coefficients (*B*) and intercept, the standardized regression coefficients (β), and semipartial correlations (sr_i) are reported in Table 11. The variance accounted for (R^2) equaled .570 (adjusted $R^2 = .563$), which was significantly different from zero ($F = 74.31$, $p < .01$). The adjusted R^2 value of .570 indicated that over half of the variability in EOG standard scores is predicted by ORF and Maze measures. However, only one of the two independent variables (predictor variables), ORF, contributed significantly to the prediction of EOG. ORF had a statistically significant standardized regression coefficient ($\beta = .708$) and semipartial correlation ($sr_i = .545$). Maze did not have a statistically significant standardized beta and semipartial correlation coefficient was close to zero.

Table 9

Means, Standard Deviations, Skewness, Kurtosis, and Variance

Grade	Measure	Mean	SD	Skewness	Kurtosis	R^2
3rd						
	EOG	340.04	11.14	- .39	- .21	.349
	ORF	111.51	25.93	- .09	- .15	
	Maze	15.17	06.14	.54	- .39	
4th						
	EOG	344.84	09.68	- .22	- .63	.481
	ORF	126.91	34.70	.39	- .29	
	Maze	15.33	06.85	1.12	2.18	
5th						
	EOG	351.38	08.70	- .20	- .20	.570
	ORF	136.30	28.24	- .23	- .04	
	Maze	22.00	7.37	.33	- .55	

Table 10

Intercorrelations Between Measures; EOG, ORF, and Maze

Grade	Measure	ORF	Maze	EOG
3rd (N = 110)				
	ORF		.702**	.557**
	Maze			.531**
4th (N = 111)				
	ORF		.696**	.657**
	Maze			.617**
5th (N = 115)				
	ORF		.639**	.753**
	Maze			.523**

*All correlation coefficients were statistically significant at the 0.01 level (2-tailed)

Table 11

Simultaneous Multiple Regression Analysis for Measures Predicting EOG Scores

Grade	Variable	*B*	SE	β	sr_i	*t*-value	*p*-value
3rd							
	ORF	.156	.047	.363	.259	3.316	.001
	Maze	.502	.199	.276	.197	2.522	.013
4th							
	ORF	.123	.027	.441	.317	4.575	.000
	Maze	.438	.136	.310	.223	3.216	.002
5^{th}							
	ORF	.218	.025	.708	.545	8.793	.000
	Maze	.083	.095	.070	.054	.872	.385

Research Question Two: Is there a difference in the magnitude of the relationship between EOG and ORF and Maze among third, fourth, and fifth grade?

A Fisher Transformation was used in order to answer question two. The strength of the relationship between EOG, ORF, and Maze was examined to determine if grade differences existed in the relationship between EOG, Maze, and ORF. The correlation coefficients between grades were compared using A Fisher Transformation. Coefficients for ORF and Maze to the EOG for each grade level were compared, with an alpha level of .05 necessary to demonstrate a significant finding. As shown in Table 12, the only significant difference in magnitude of relationship was found between ORF and EOG in third and fifth grade. The correlation among fifth graders (r = .753) was significantly larger than the correlation among third graders (r = .557). Results indicated there were no significant differences between coefficients for Maze in any grade.

Table12

Results of Fisher's z Transformation Comparing Coefficients between ORF, Maze, and North Carolina End-of-Grade Reading Assessment for Grade Levels

	Measure Relationship	3rd Grade	4th Grade	5th Grade
3rd Grade				
	ORF and EOG	-	- 1.17	- 2.60*
	Maze and EOG	-	- 0.94	0.08
4th Grade	ORF and EOG		-	- 1.43
	Maze and EOG		-	1.04

* $p < .01$.

Research Question Three: How accurate are published DIBELS benchmark risk level cutoff scores for ORF and AIMSweb Maze Aggregate Norm 50th percentile scores for identifying third, fourth, and fifth grade students who will or will not be proficient as measured by the statewide grade level NC EOG Reading Comprehension test?

Using NCSS Statistical Software (Hintze, 2007), Receiver Operating Characteristic (ROC) Curves were constructed for graphical representation. For each grade level, the score Level III performance on the EOG was used as the cutoff for binary outcome of not proficient (0) and proficient (1). Sensitivity, specificity, and predictive values were calculated using a 2 x 2 contingency table. In the medical field, a test would be considered positive if it showed a disease is present and negative if it does not. In the current study, poor comprehension is considered "disease" (positive) and no poor comprehension is considered "no disease" (negative). Therefore, a low ORF or Maze score predicted students at risk for poor comprehension (not proficient EOG score). A high ORF or Maze score predicted students who were not at risk for poor comprehension (proficient EOG score).

Diagnostic efficiency of ORF and Maze was tested by examining sensitivity (i.e., the proportion of students correctly classified as at risk using DIBELS or Maze and EOG) and specificity (i.e., the proportion of students correctly classified as not at risk using DIBELS or Maze and EOG) of cut score values. Cases were considered t*rue positives* if poor comprehension was predicted (low ORF/Maze), and poor comprehension was actually observed (not proficient EOG score). Cases were considered *true negatives* if poor comprehension was not predicted (high ORF/Maze), and poor comprehension was not observed (proficient EOG score). Cases were considered *false positives* if poor

comprehension was predicted (low ORF/Maze), but poor comprehension was not observed (proficient EOG score). Finally, cases were considered *false negatives* if poor comprehension was not predicted (high ORF/Maze), but poor comprehension was observed (not proficient EOG score). The accuracy rate (AR) was computed as the proportion of all correctly classified students (the sum of all true positives and true negatives) out of all students. The misclassification rate (MR) was computed as the proportion of all misclassified students (the sum of false positives and false negatives) out of all students (Gonen, 2007).

ROC Curves visually represent the statistical accuracy for all possible cutoff scores on a measure (Swets, 1988). The ROC curve in this study represents the probability that a random pair of students will be correctly ranked as to their proficiency level on the EOG using DIBELS ORF and AIMSweb Maze scores. According to Swets, values for AUC range from .50 (no discrimination) to 1.0 (perfect discrimination). Results were interpreted using Simon (1999) suggested interpretation of AUC values: 0.97 – 1.00 (excellent); 0.92 – 0.97 (very good); 0.75 – 0.92 (good); 0.500 – 0.75 (fair). Sensitivity and specificity values for DIBELS ORF were calculated and reported using the published DIBELS benchmark level cutoff scores recommended by Good and Kaminski (2002) for each grade level. Sensitivity and specificity values for Maze were calculated and reported. The 50^{th} percentile score from AIMSweb Maze Aggregate Norm (Edformation, 2009) was used as a cutoff score for Maze.

Conditional probability indices were also calculated using NCSS ROC CURVES. Positive predictive value (PPV) was calculated to show the probability that a student who is identified as being at risk is truly at risk. Negative predictive value (NPV) was

calculated to show the chances that a student who is identified not at risk is truly not at risk. The area under the curve (AUC) was generated as part of the ROC analysis in order to provide the probability of the independent variable correctly classifying a pair of individuals when one student is at risk and the other is not.

Third Grade. Reading EOG scores were dichotomized so that scores 338 and above were considered proficient, and scores below were considered not proficient. There were 38 students who did not meet the minimal acquisition of skills as measured by the standard on the NC EOG in Reading Comprehension Assessment, whereas 72 students met EOG reading proficiency. Predicted group memberships were compared based on performance on DIBELS ORF and AIMSweb Maze measures.

The ROC Curve Plot is shown in Figure 1. Inspection of the area under the curve suggested sensitivity and specificity values for DIBELS risk level cutoff scores for ORF (AUC = .809, p = .00) and AIMSweb Maze (AUC = .788, p = .00) were statistically significant. The 95% confidence interval for ORF was .718 - .899 and the confidence interval for Maze was .695 - .881. Sensitivity and specificity values for the recommended cutoff scores are shown in Table 13. As shown in the table, sensitivity levels for both ORF (.816) and Maze (.868) measures were adequate. However, both Maze (.486) and ORF (.653) demonstrated fair specificity levels (e.g., 50% - 75%).

Sensitivity, specificity, and predictive values were calculated using a 2 x 2 contingency table. As seen in Table 14, for ORF, there were 31 students for which poor comprehension was predicted (low ORF), and poor comprehension was observed (not proficient EOG). These students represent the true positives (sensitivity). There were 50 students for which poor comprehension was not predicted (high ORF) or observed

(proficient EOG). These students represent the true negatives (specificity). There were 7 students for which poor comprehension was not predicted (high ORF), but was observed (not proficient EOG). These students represent the false negatives. There were 22 students for which poor comprehension was predicted (low ORF), but not observed (proficient EOG). These students represent the false positives. For Maze, there were 28 true positives, 35 true negatives, 10 false negatives, and 37 false positives. Overall, for third grade ORF, the accuracy rate (AR) was 74%, misclassification rate (MR) was 26%, positive predictive value (PPV) was 55%, and negative predictive value (NPV) was 87%. For Maze, AR was 57%, MR was 42%, PPV was 47%, and NPV was 88%.

A direct logistic regression analysis was performed on EOG as outcome (coded 0= not proficient and 1=proficient) and two predictors: ORF and Maze. Analysis was performed using SPSS. There were a total of 72 students who performed at the proficient level and 38 students who were not proficient on the Third Grade End-of-Grade Assessment of Reading Comprehension.

A test of the full model with both predictors against a constant-only model was statistically reliable $\chi 2$ (2, N=110) = 34.54, $p<.001$, indicating that ORF and Maze reliably distinguished between students who were proficient and not proficient on the EOG. The variance in EOG accounted for is moderate, with Cox and Snell R^2 equal to .269 and Nagelkerke R^2 equal to .372. Predicted success was adequate with correct identification of 89% of the students who were proficient and correct identification of 61% of the students who were not proficient. Predicted success had an overall success rate of 79%.

According to the Wald criteria, ORF and Maze reliably predicted proficiency on the EOG in reading for third grade students. Table 15 shows the regression coefficients, Wald statistics, statistical significances, and odds ratios for each of the predictors. The odds ratio indicated that for every word read correctly per minute on DIBELS ORF, students were 1.039 times more likely to be proficient on the EOG. Therefore, for every one word increase in oral reading fluency, there was a 3.9% increase in odds of proficiency on the EOG. For every word identified correctly per minute on AIMSweb Maze, there was a 1.125 greater chance that a student would be proficient on the EOG. In other words, for every 1 word increase in fluency of comprehension, there was a 12.5% increase in odds of proficiency on the EOG.

Fourth Grade. Reading EOG scores were dichotomized so that scores 343 and above were considered proficient, and scores below were considered not proficient. There were 43 students who did not meet the minimal acquisition of skill standard on the NC EOG in Reading Comprehension Assessment in fourth grade, whereas 68 students met EOG reading proficiency. Predicted group memberships were compared based on performance on DIBELS ORF and AIMSweb Maze measures.

The ROC Curve Plot is shown in Figure 2. Inspection of the area under the curve suggested statistically significant sensitivity and specificity values for DIBELS ORF (AUC = .879, p = .00) and AIMSweb Maze (AUC = .839, p = .00). The 95% confidence interval for ORF was .764 - .914, and the confidence interval for Maze was .818 - .941. Sensitivity and specificity values for the recommended cutoff scores are shown in Table 13. As shown in Table 13, only ORF demonstrated adequate sensitivity and specificity using the recommended cutoff score of 118. Both ORF and Maze demonstrated adequate

sensitivity using the recommended risk cutoff score, with a very good Maze sensitivity level (.953) and a good ORF sensitivity (.860) and specificity (.765) level. However, specificity levels for Maze were less than adequate, as Maze demonstrated poor specificity levels (.324).

Sensitivity, specificity, and predictive values were calculated using a 2 x 2 contingency table. Results for ORF and Maze are presented in Table 14. For ORF, there were 37 students for which poor comprehension was predicted (low ORF) and observed (not proficient EOG). These students represent the true positives (sensitivity). There were 52 students for which poor comprehension was not predicted (high ORF) or observed (proficient EOG). These students represent the true negatives (specificity). There were 6 students for which poor comprehension was not predicted (high ORF), but was observed (not proficient EOG). These students represent the false negatives. There were 16 students for which poor comprehension was predicted (low ORF), but not observed (proficient EOG). These students represent the false positives. For Maze, there were 40 true positives, 26 true negatives, 3 false negatives, and 42 false positives. Overall, for fourth grade ORF, the accuracy rate (AR) was 80%, misclassification rate (MR) was 20%, positive predictive value (PPV) was 70%, and negative predictive value (NPV) was 90%. For Maze, AR was 59%, MR was 41%, PPV was 47%, and NPV was 92%.

A direct logistic regression analysis was performed on EOG as outcome (coded 0= not proficient and 1= proficient) and two predictors: ORF and Maze. Analysis was performed using SPSS. Results indicated there were a total of 68 students who were proficient and 43 students who were not proficient on the Fourth Grade End-of-Grade Assessment of Reading Comprehension.

A test of the full model with both predictors against a constant-only model was statistically reliable $\chi 2$ (2, N=111) = 61.38, p<.001, indicating that ORF and Maze reliably distinguished between students who were proficient and not proficient on the EOG. The variance in EOG accounted for is moderate, with Cox and Snell R^2 equal to .425 and Nagelkerke R^2 equal to .576. Predicted success was adequate with 85% of the proficient students and 74% of the students who were not proficient identified correctly and an overall success rate of 81 %.

The regression coefficients, Wald statistics, statistical significances, and odds ratios for each of the predictors are presented in Table 15. According to the Wald criteria, ORF and Maze reliably predicted proficiency on the EOG in reading for students in fourth grade. The odds ratio indicated that when holding all other variables constant, for every word read correctly per minute on DIBELS ORF, students were 1.055 times more likely to be proficient on the EOG. In other words, for every one word increase in reading fluency, there was 5.5% increase in odds of proficiency on EOG. For every word identified correctly per minute on AIMSweb Maze, there was a 1.169 greater chance that a student would be proficient on the EOG. Therefore, for every one word increase in words identified correctly, there was a 16.9% increase in odds of proficiency on EOG.

Fifth Grade. Reading EOG scores were dichotomized so that scores 349 and above were considered proficient, and scores below were considered not proficient. There were 44 students who did not meet the minimal acquisition of skill standard on the NC EOG in Reading Comprehension Assessment, whereas 71 students met EOG reading proficiency. Predicted group memberships were compared based on performance on DIBELS ORF and AIMSweb Maze measures.

The ROC Curves Plot is shown in Figure 3. Inspection of the area under the curve suggested statistically significant sensitivity and specificity values for DIBELS ORF (AUC = .900, p = .00) and AIMSweb Maze (AUC = .814, p = .00). The 95% confidence interval for ORF was .838 - .963. The 95% confidence interval for Maze was .736 - .892. Sensitivity and specificity values for the recommended cutoff scores are shown in Table 13. As demonstrated in the table, only ORF demonstrated adequate sensitivity and specificity using the recommended cutoff scores. Both measures demonstrated adequate sensitivity using the recommended risk cutoff scores, with a very good sensitivity level for Maze (.955) and good sensitivity level for ORF (.795). The specificity level for ORF (.929) was very good. However, the specificity level for Maze was less than adequate (.521), falling within the fair level range (e.g., 50% - 75%).

Sensitivity, specificity, and predictive values were calculated using a 2 x 2 contingency table. Results for ORF and Maze are presented in Table 14. For ORF, there were 35 students for which poor comprehension was predicted (low ORF), and poor comprehension was observed (not proficient EOG). These students represent the true positives (sensitivity). There were 66 students for which poor comprehension was not predicted (high ORF) or observed (proficient EOG). These students represent the true negatives (specificity). There were 9 students for which poor comprehension was not predicted (high ORF), but was observed (not proficient EOG). These students represent the false negatives. There were 5 students for which poor comprehension was predicted (low ORF), but not observed (proficient EOG). These students represent the false positives. For Maze, there were 42 true positives, 40 true negatives, 2 false negatives, and 31 false positives. Overall, for fifth grade ORF, the accuracy rate (AR) was 88%,

misclassification rate (MR) was 12%, positive predictive value (PPV) was 88%, and negative predictive value (NPV) was 88%. For Maze, AR was 71%, MR was 29%, PPV was 55%, and NPV was 95%.

A direct logistic regression analysis was performed on EOG as outcome (coded 0= not proficient and 1= proficient) and two predictors: ORF and Maze. Analysis was performed using SPSS. Results indicated in fifth grade, there were a total of 71 students who were proficient and 44 students who were not proficient. A test of the full model with both predictors against a constant-only model was statistically reliable X^2 (2, N=115) = 66.57, p<.001, indicating that ORF and Maze reliably distinguished between students who were proficient and not proficient on the EOG. The variance in EOG accounted for is moderately strong, with Cox and Snell R^2 equal to .439 and Nagelkerke R^2 equal to .597. Predicted success was adequate with 86% of the proficient students and 77% of the students who were not proficient identified correctly and an overall success rate of 83 %.

Table 15 shows the regression coefficients, Wald statistics, statistical significances, and odds ratios for each of the predictors. According to the Wald criteria, only one of the predictors, ORF, reliably predicted proficiency on the EOG in reading for fifth grade students. The odds ratio for ORF indicated that when holding all other variables constant, students were 1.082 times more likely to be proficient on the EOG for every word read correctly per minute on DIBELS ORF. Therefore, each word increase in oral reading fluency equaled an 8.2% increase in odds of proficiency on the EOG.

Table 13

Decision-Making Accuracy for Recommended DIBELS ORF Cutoff Scores for Each Grade Level When Predicting NC EOG Reading Comprehension Proficiency

Grade	Measure	Cutoff	TPF	TNF	PPV	NPV	AR	MR	AUC
3rd									
	ORF	110	.816	.653	.553	.870	.736	.263	.808
	Maze	16	.868	.486	.471	.875	.572	.427	.788
4th									
	ORF	118	.860	.765	.700	.900	.802	.198	.879
	Maze	19	.953	.324	.471	.917	.595	.405	.839
5th									
	ORF	124	.795	.929	.875	.880	.878	.121	.900
	Maze	25	.955	.521	.553	.949	.713	.287	.814

Note: Good and Kaminski (2002) recommended cutoff scores
AIMSweb Maze 50th percentile Norm Scores (Edformation, 2009)

TPF = True Positive Fraction (sensitivity); TNF = True Negative Fraction (specificity); PPV = Positive Predictive Value; NPV= Negative Predictive Value; AR = Accuracy Rate; MR = Misclassification Rate; AUC = Area Under the Curve (ROC)

Table 14

Reporting Accuracy for Prediction of Proficiency on NC EOG Using ORF and Maze

		Observed	
Predicted	Positive	Negative	Total
Positive	True Positive (TP)	False Positive (FP)	TP + FP
Negative	False Negative (FN)	True Negative (TN)	FN + TN
Total	TP + FN	FP + TN	TP + FP + FN + TN

3rd Grade ORF (Cutoff = 110)	Observed		
Predicted	Poor Comp. (low EOG)	No Poor Comp. (high EOG)	Total
Poor Comp.	31	22	53
No Poor Comp.	7	50	57
Total	38	72	110

3rd Grade Maze (Cutoff = 16)	Observed		
Predicted	Poor Comp. (low EOG)	No Poor Comp (high EOG)	Total
Poor Comp.	28	37	65
No Poor Comp.	10	35	45
Total	38	72	110

Table 14 (continued)

4th Grade ORF (Cutoff = 118)	Observed		
Predicted	Poor Comp. (low EOG)	No Poor Comp (high EOG)	Total
Poor Comp.	37	16	53
No Poor Comp.	6	52	58
Total	43	68	111

4th *Grade Maze (Cutoff = 19)*	Observed		
Predicted	Poor Comp. (low EOG)	No Poor Comp (high EOG)	Total
Poor Comp.	40	42	82
No Poor Comp.	3	26	29
Total	43	68	111

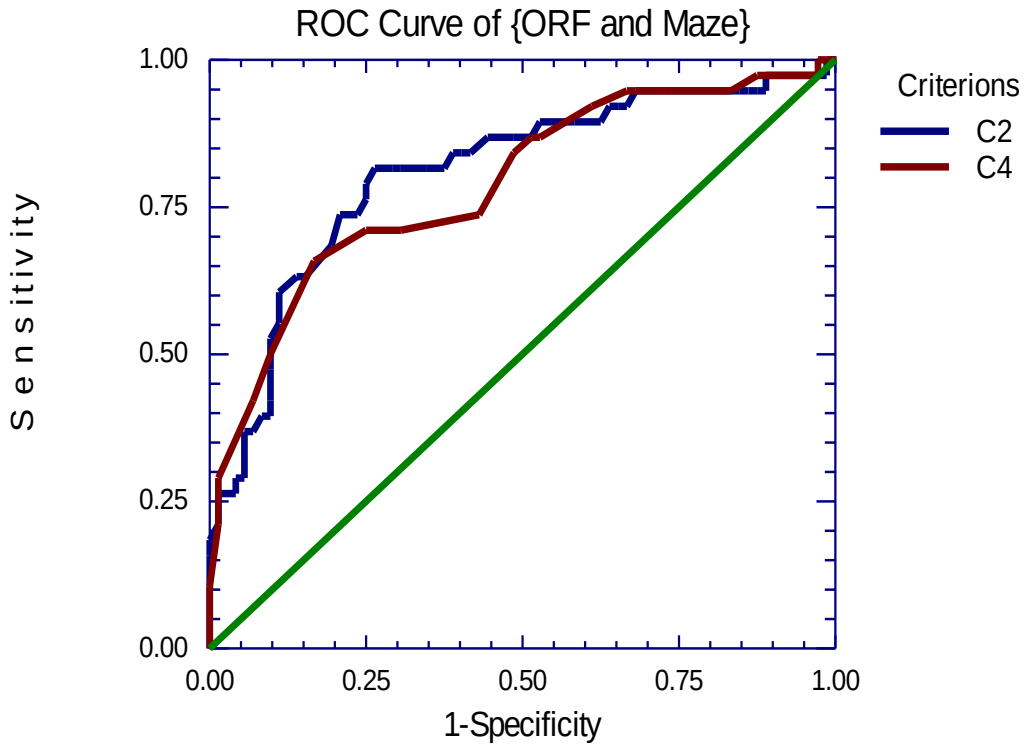
5th Grade ORF (Cutoff = 124)	Observed		
Predicted	Poor Comp. (low EOG)	No Poor Comp (high EOG)	Total
Poor Comp.	35	5	40
No Poor Comp.	9	66	75
Total	44	71	115

5th *Grade Maze (Cutoff = 25)*	Observed		
Predicted	Poor Comp. (low EOG)	No Poor Comp (high EOG)	Total
Poor Comp.	42	31	73
No Poor Comp.	2	40	42
Total	44	71	115

Table 15

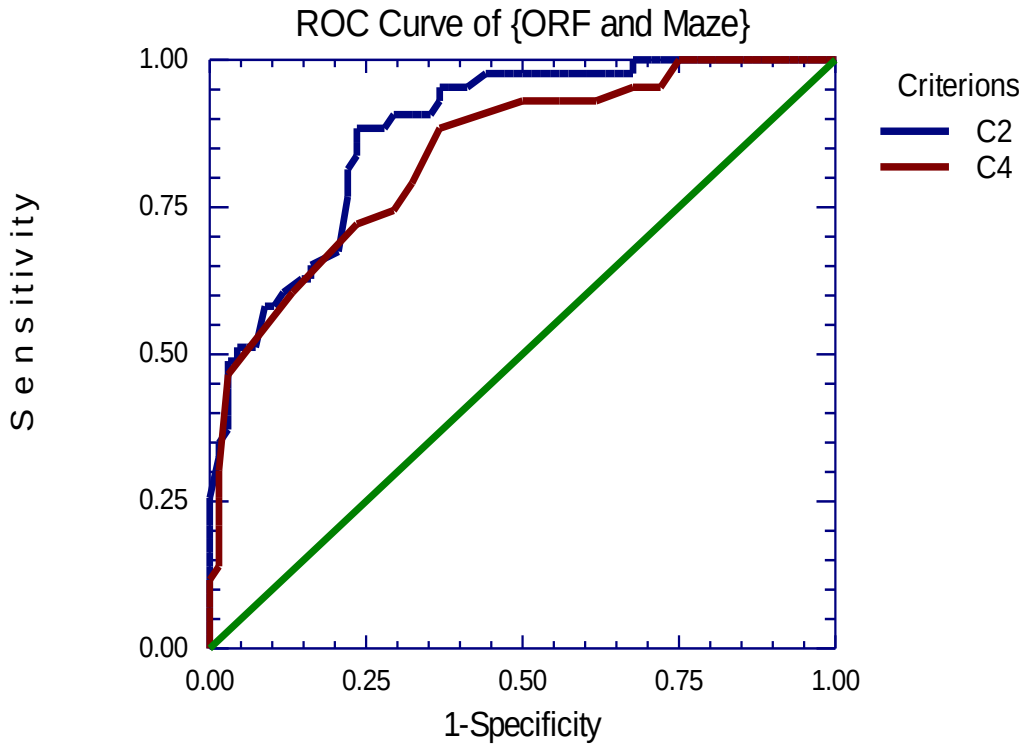
Logistic Regression Coefficients, Standard Errors, Wald Statistics, Statistical Significance, Odds Ratio, 95% Confidence Interval for Correct Classification for Each Grade Level

Grade	Predictor	β	S.E.	Wald	df	Sig	Odds Ratio	95% C.I.	
3rd									
	ORF	.039	.014	7.530	1	.006	1.039	1.011	1.069
	Maze	.118	.059	3.923	1	.048	1.125	1.001	1.264
	Constant	-5.139	1.333	14.867	1	.000	.006		
4th									
	ORF	.054	.015	12.686	1	.000	1.055	1.025	1.087
	Maze	.156	.074	4.501	1	.034	1.169	1.012	1.350
	Constant	-8.093	1.674	23.369	1	.000	.000		
5th									
	ORF	.074	.018	17.004	1	.000	1.077	.979	1.197
	Maze	.079	.051	2.386	1	.122	1.082	1.040	1.115
	Constant	-10.962	2.156	25.845	1	.000	.000		



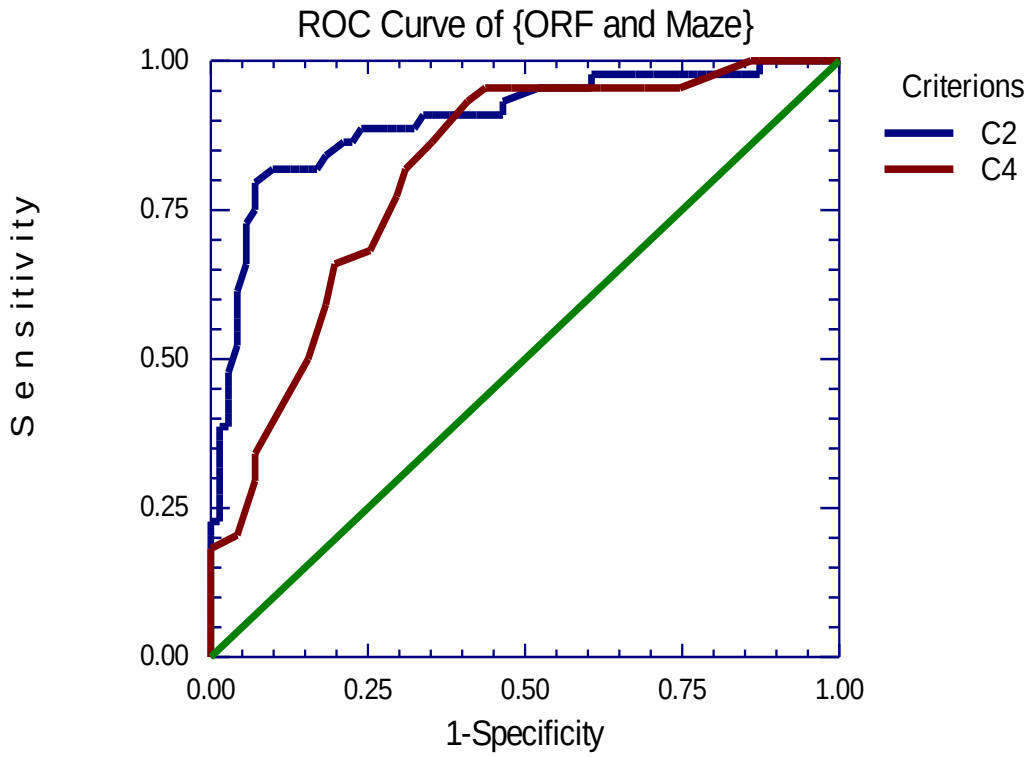
Note: Criterion C2 = ORF; Criterion C4 = Maze

Figure 1: ROC Curve Plot for Proficiency on 3rd Grade NCEOG in Reading using DIBELS ORF and Maze



Note: C2 Criterion = ORF; C4 Criterion = Maze

Figure 2. ROC Curve Plot for Proficiency on 4th Grade NCEOG in Reading using DIBELS ORF and Maze



Note: Criterion 2 = ORF; Criterion 4 = Maze

Figure 3. ROC Curve Plot for Proficiency on 5th Grade NCEOG in Reading using DIBELS ORF and Maze

Research Question Four: What are the optimal DIBELS ORF and AIMSweb Maze-CBM cut scores to use when attempting to predict satisfactory reading comprehension by the end of third, fourth, and fifth grade level as measured by EOG performance?

Receiver Operating Curves were used for determination of the optimal cut scores for DIBELS ORF and Maze Measures to use in predicting proficiency on the NC EOG in Reading Comprehension. Using ROC CURVES, optimal cutoff scores were identified for each measure and tested in a 2 x 2 contingency table. The optimal scores yielded maximum levels of sensitivity (students with low ORF and below proficient on EOG) and specificity (students with high ORF and above proficient on EOG). In other words, optimal scores were identified by considering two things: (a) sensitivity, which is the percentage of students who performed below proficient levels on EOG and were correctly identified as at risk by DIBELS (i.e., presence of a problem), and (b) specificity, which is the percentage of students who performed at or above proficient levels on EOG and were correctly identified as low risk by DIBELS (i.e., absence of a problem). Generally, the optimal cut score is near the shoulder of the ROC curve (Swets et al., 2000). According to Swets, the rule of thumb for identification of optimal cutoff scores is "a large benefit associated with finding true cases generally argues for a lenient threshold . . . a high cost for false alarms generally calls for a strict threshold" (p. 84).

In the field of education, the index of interest is sensitivity; the ability of the measure (ORF and Maze) to detect children who are later identified as exhibiting poor comprehension as measured by the EOG (true positives). From an educational perspective, these are the students who need to be provided with intervention. However, an increase in sensitivity means a decrease in specificity. The challenge of identifying

optimal cutoff levels is to "set cut-scores that maximize each characteristic to its fullest potential" (Hintze, Ryan, & Stoner, 2003, p. 548). According to Swets et al. (2000), it is questionable to allow an unreasonable number of false positives in order to identify true positives. Therefore, the value was located to balance sensitivity and specificity.

Third Grade. Receiver Operating Curves were used to determine the optimal cutoff score for third grade that would result in a balance between sensitivity and specificity. The tradeoff for sensitivity and specificity for DIBELS and Maze are reported in Table 16. The optimal score to balance sensitivity and specificity at adequate levels was located for ORF and Maze and tested in a 2 x 2 contingency table, which is presented in Table 17. The optimal cutoff score for ORF in third grade to balance sensitivity and specificity levels was 107. This yielded an adequate (e.g., above 75%) sensitivity level (.816), but slightly lower than adequate specificity level (.736). For Maze, a cutoff value of 15 maximized sensitivity and specificity levels. The cutoff score of 15 yielded an adequate sensitivity level (.842), but less than adequate specificity level (.514). Optimal cutoff scores for DIBELS ORF and AIMSweb Maze are presented in Table 18 for third grade.

Fourth Grade. Receiver Operating Curves were used to determine the optimal cutoff score for fourth grade. The tradeoff for sensitivity and specificity for DIBELS and Maze are reported in Table 16. The optimal score to balance sensitivity and specificity at adequate levels was located for ORF and Maze and tested in a 2 x 2 contingency table, which is presented in Table 17. For ORF, a cutoff score of 120 resulted in adequate sensitivity (.884) and specificity (.765) levels. For Maze, the optimal cutoff score of 15 that balanced sensitivity and specificity in fourth grade yielded a less than adequate level

of specificity (.632), but adequate sensitivity (.884). It was not possible to determine a cutoff score with adequate levels of both sensitivity and specificity because both sensitivity and specificity were not at adequate levels anywhere along the continuum of scores. Lower cutoff scores yielded lower than adequate levels of sensitivity. Results for optimal cut scores for fourth grade DIBELS ORF and Maze are presented in Table 18.

Fifth Grade. Receiver Operating Curves were used to determine the optimal cutoff score in fifth grade. The tradeoff for sensitivity and specificity for DIBELS and Maze are reported in Table 16. The optimal score to balance sensitivity and specificity at adequate levels was located for ORF and Maze and tested in a 2 x 2 contingency table, which is presented in Table 17. For ORF, using 132 as the cutoff score resulted in adequate sensitivity (.841) and specificity (.817) levels. However, the optimal cutoff score for Maze in fifth grade does not have adequate levels of both sensitivity and specificity. Although adequate levels of sensitivity were observed using cut-scores in the range of 21 - 27 on the Maze task, less than adequate levels of specificity were noted across the continuum of cut scores. Therefore, the optimal cutoff score yielded adequate sensitivity, but less than adequate specificity. The optimal cutoff score of 21 had adequate sensitivity (.818), but demonstrated a lower than adequate specificity level (.690). Results are presented in Table 18 for optimal DIBELS and AIMSweb Maze cutoff scores for fifth grade.

Table 16

Tradeoff of Sensitivity and Specificity for Possible Cutoff Values

3rd Grade ORF			3rd Grade Maze		
Cutoff	Sensitivity	Specificity	Cutoff	Sensitivity	Specificity
102	.684 (68%)	.805 (80%)	10	.500 (50%)	.903 (90%)
103	.737 (74%)	.791 (79%)	11	.658 (66%)	.833 (83%)
104	.737 (74%)	.763 (76%)	12	.711 (71%)	.750 (75%)
105	.763 (76%)	.750 (75%)	13	.711 (71%)	.694 (69%)
106	.789 (79%)	.750 (75%)	14	.737 (74%)	.569 (57%)
107	.816 (82%)	.736 (74%)	15	.842 (84%)	.514 (51%)
108	.816 (82%)	.708 (71%)	16	.868 (87%)	.486 (49%)
109	.816 (82%)	.694 (69%)	17	.868 (87%)	.472 (47%)
110	.816 (82%)	.653 (65%)	18	.921 (92%)	.389 (39%)
111	816 (82%)	.625 (63%)	19	.947 (95%)	.333 (33%)
112	.842 (84%)	.611 (61%)	20	.947 (95%)	.319 (32%)
113	.842 (84%)	.597 (60%)	21	.947 (95%)	.264 (26%)
114	.842 (84%)	.583 (58%)	22	.947 (95%)	.208 (21%)
115	.868 (87%)	.556 (56%)			
116	.868 (87%)	.514 (51%)			
117	.868 (87%)	.514 (51%)			
118	.868 (87%)	.486 (49%)			
119	.895 (90%)	.472 (47%)			
120	.895 (90%)	.431 (43%)			

Table 16 (continued)

4th Grade ORF			4th Grade Maze		
Cutoff	Sensitivity	Specificity	Cutoff	Sensitivity	Specificity
106	.581 (58%)	.897 (90%)	10	.465 (47%)	.971 (97%)
107	.605 (61%)	.882 (88%)	11	.605 (61%)	.868 (87%)
108	.628 (63%)	.853 (85%)	12	.721 (72%)	.765 (77%)
109	.628 (63%)	.838 (84%)	13	.744 (74%)	.706 (71%)
110	.651 (65%)	.838 (84%)	14	.791 (79%)	.676 (68%)
111	.651 (65%)	.838 (84%)	15	.884 (88%)	.632 (63%)
112	.674 (67%)	.794 (79%)	16	.930 (93%)	.500 (50%)
113	.674 (67%)	.794 (79%)	17	.930 (93%)	.441 (44%)
114	.767 (77%)	.779 (78%)	18	.930 (93%)	.383 (38%)
115	.814 (81%)	.779 (78%)	19	.953 (95%)	.324 (32%)
116	.837 (84%)	.765 (77%)	20	.953 (95%)	.279 (28%)
117	.860 (86%)	.765 (77%)	21	1.00 (100%)	.250 (25%)
118	.860 (86%)	.765 (77%)			
119	.884 (88%)	.765 (77%)			
120	.884 (88%)	.765 (77%)			
121	.884 (88%)	.721 (72%)			
122	.884 (88%)	.721 (72%)			
123	.907 (91%)	.706 (71%)			
124	.907 (91%)	.676 (68%)			
125	.907 (91%)	.662 (66%)			

Table 16 (continued)

5th Grade ORF			5th Grade Maze		
Cutoff	Sensitivity	Specificity	Cutoff	Sensitivity	Specificity
118	.568 (57%)	.958 (96%)	16	.500 (50%)	.845 (85%)
119	.614 (61%)	.958 (96%)	17	591 (59%)	.817 (82%)
120	.659 (66%)	.944 (94%)	18	.659 (66%)	.803 (80%)
121	.659 (66%)	.944 (94%)	19	.682 (68%)	.746 (75%)
122	.727 (73%)	.944 (94%)	20	.773 (77%)	.704 (70%)
123	.750 (75%)	.930 (93%)	21	.818 (82%)	.690 (69%)
124	.795 (80%)	.930 (93%)	22	.864 (86%)	.648 (65%)
125	.795 (80%)	.930 (93%)	23	.932 (93%)	.592 (59%)
126	.818 (82%)	.901 (90%)	24	.955 (96%)	.563 (56%)
127	.818 (82%)	.873 (87%)	25	.955 (96%)	.521 (52%)
128	.818 (82%)	.873 (87%)	26	.955 (96%)	.451 (45%)
129	.818 (82%)	.859 (86%)	27	.955 (96%)	.380 (38%)
130	.818 (82%)	.845 (85%)			
131	.818 (82%)	.831 (83%)			
132	.841 (84%)	.817 (82%)			
134	.864 (86%)	.789 (79%)			
135	.864 (86%)	.775 (78%)			
136	.886 (89%)	.761 (76%)			
137	.886 (89%)	.718 (72%)			
138	.886 (89%)	.704 (70%)			

Table 17

Optimal Cutoff Scores to Balance Sensitivity and Specificity for Each Grade Level Tested in 2 x 2 Contingency Table

3rd Grade ORF (Cutoff=107)		Observed	
Predicted	Poor Comp. (low EOG)	No Poor Comp. (high EOG)	Total
Poor Comp.	31	17	48
No Poor Comp.	8	54	62
Total	39	71	110

3rd Grade Maze (Cutoff= 15)		Observed	
Predicted	Poor Comp. (low EOG)	No Poor Comp. (high EOG)	Total
Poor Comp.	28	33	61
No Poor Comp.	10	39	49
Total	38	72	110

4th Grade ORF (Cutoff = 120)		Observed	
Predicted	Poor Comp. (low EOG)	No Poor Comp. (high EOG)	Total
Poor Comp.	38	16	54
No Poor Comp.	5	52	57
Total	43	68	111

Table 17 (continued)

4th Grade Maze (Cutoff = 15)	Observed		
Predicted	Poor Comp. (low EOG)	No Poor Comp. (high EOG)	Total
Poor Comp.	34	24	58
No Poor Comp.	9	44	53
Total	43	68	111

5th Grade ORF (Cutoff = 132)	Observed		
Predicted	Poor Comp. (low EOG)	No Poor Comp. (high EOG)	Total
Poor Comp.	36	12	48
No Poor Comp.	8	59	67
Total	44	71	115

5th Grade Maze (Cutoff = 21)	Observed		
Predicted	Poor Comp. (low EOG)	No Poor Comp. (high EOG)	Total
Poor Comp.	34	21	55
No Poor Comp.	10	50	60
Total	44	71	115

Table 18

Optimal DIBELS ORF and AIMSweb Maze Cutoff Scores to Maximize Sensitivity and Specificity for Predicting NC EOG Reading Comprehension Proficiency for Each Grade

Grade	Measure	Cutoff	TPF	TNF	PPV	NPV	AR	MR
3rd	ORF	107	. 816	.736	.795	.761	.772	.227
	Maze	15	.842	.514	.737	.542	.609	.391
4th	ORF	120	.884	.765	.884	.765	.811	.189
	Maze	15	.884	.632	.791	.647	.703	.297
5th	ORF	132	.841	.817	.818	.831	.826	.174
	Maze	21	.818	.690	.773	.831	.730	.270

TPF = True Positive Fraction (sensitivity); TNF = True Negative Fraction (specificity); PPV = Positive Predictive Value; NPV= Negative Predictive Value; AR = Accuracy Rate; MR = Misclassification Rate

Summary of Results

The relationship between outcomes on NC EOG and scores on DIBELS ORF and AIMSweb Maze was examined to determine which measure or combination of measures best predicted scores on EOG. Results of the current study indicated a moderate, significant relationship between scores on DIBELS ORF, AIMSweb Maze Measures, and NC EOG in each grade level with correlations ranging from .523 to .753. Results of the multiple regression suggested DIBELS ORF and AIMSweb Maze were significant predictors of NC EOG. In third grade, both variables accounted for more than a third of the variability ($R = .349$) in EOG scores. For students in third grade, ORF had the largest contribution, but Maze also made a similar, statistically significant contribution to the prediction of scores on the EOG. In fourth grade, ORF and Maze contributed significantly to the prediction of EOG scores, accounting for almost half of the variability ($R^2 = .481$) in scores. For students in fourth grade, scores on ORF made the largest contribution to prediction, but Maze made a similar, statistically significant contribution. In fifth grade, ORF and Maze together accounted for more than half of the variability of the variability ($R^2 = .570$) in EOG scores. However, only ORF made a significant contribution to the prediction of EOG scores. When ORF was in the equation, Maze did not make a significant contribution to the prediction of scores on EOG for students in fifth grade.

A Fisher Transformation was used to examine the strength of the relationship between EOG, ORF and Maze for each grade level to determine if grade level differences existed. The only significant difference in magnitude of relationship was found between ORF and EOG in third and fifth grade. The correlation for ORF among fifth grade

students (r = .753) was significantly larger than the correlation among third grade students (r = .557), but no significant differences were found between coefficients in any grade for Maze.

Receiver Operating Characteristic (ROC) Curves were constructed to determine the accuracy of published DIBELS ORF benchmark risk level cutoff scores for proficiency on the NC EOG Reading Comprehension assessment. ROC Curves were also used to determine the accuracy of AIMSweb Maze Aggregate Norm Scores (50^{th} percentile). For third grade, inspection of the area under the curve suggested sensitivity and specificity values for DIBELS ORF (AUC = .809, p = .00) and AIMSweb Maze (AUC = .788, p = .00) were statistically significant. Sensitivity levels for ORF and Maze were adequate, and specificity levels were fair. Results of logistic regression suggested ORF and Maze scores reliably distinguished between students who were proficient and not proficient on the EOG, with an overall success rate of 79%. Both ORF and Maze reliably predicted proficiency for third grade students. For fourth grade, inspection of the area under the curve suggested sensitivity and specificity values for DIBELS ORF (AUC = .879, p = .00) and AIMSweb Maze (AUC = .839, p = .00) were statistically significant. ORF demonstrated adequate sensitivity and specificity levels, but Maze demonstrated poor specificity levels with very good sensitivity. Results of logistic regression suggested ORF and Maze scores reliably distinguished between students who were proficient and not proficient on the EOG, with an overall success rate of 81%. Both measures reliably predicted proficiency on EOG. For fifth grade, inspection on the area under the curve suggested sensitivity and specificity values for DIBELS ORF (AUC = .900, p = .00) and AIMSweb Maze (AUC = .814, p = .00) were statistically significant. ORF had adequate

sensitivity and very good specificity levels, and Maze had very good sensitivity with fair specificity levels. Results of logistic regression suggested ORF and Maze reliably distinguished between students who were proficient and not proficient on the EOG, with an overall success rate of 83%. However, only one of the predictors, ORF, reliably predicted proficiency for fifth grade.

Optimal cutoff scores for DIBELS ORF and AIMSweb Maze were determined using ROC Curves for each grade level. The optimal score for determining proficiency on the NC EOG was determined by balancing sensitivity and specificity at adequate levels. For third grade, the optimal cutoff scores for ORF (107) and Maze (15) were similar to published cutoff levels for ORF (110) and Maze (16), with a slightly lower ORF and Maze cutoff score. For fourth grade, optimal cutoff scores for ORF (120) was slightly higher than published cutoff levels for ORF (118) and Maze (15) was slightly lower than cutoff levels for Maze (19). For fifth grade, the optimal cutoff score for ORF (132) was higher than the published cutoff level for ORF (124), but the optimal cutoff score for Maze (21) was lower than the published cutoff level (25). A comparison of optimal cutoff scores for determining who will or will not be proficient on NC EOG and recommended DIBELS and AIMSweb cut scores is presented in Table 19.

Table 19

Publisher Recommended Cutoff Scores Versus Optimal Cutoff Scores to Balance Sensitivity and Specificity When Predicting NC EOG Reading Proficiency Using DIBELS ORF and AIMSweb Maze

	Recommended Score			Optimal Score		
Predictor	Cutoff	Sensitivity	Specificity	Cutoff	Sensitivity	Specificity
3rd ORF	110	.816	.653	107	.816	.736
3rd Maze	16	.868	.486	15	.842	.514
4th ORF	118	.860	.765	120	.884	.765
4th Maze	19	.953	.324	15	.884	.632
5th ORF	124	.795	.930	132	.841	.817
5th Maze	25	.955	.521	21	.818	.690

Note: Good & Kaminski (2002) recommended cutoff scores for DIBELS ORF; AIMSweb Maze 50th percentile Norm Scores recommended cutoff scores for Maze (Edformation, 2009)

CHAPTER 5: DISCUSSION

The purpose of the current study was to investigate the relationship between outcomes from a state's large-scale reading comprehension assessment and scores on DIBELS ORF and Maze-CBM. A nonexperimental research design using correlational methodology was used to determine the degree that DIBELS ORF and AIMSweb Maze-CBM can be used to predict the comprehension measures on the statewide assessment. Next, this study examined differences in the magnitude of the relationship as a function of grade. An additional purpose of the study was to examine the accuracy of published DIBELS ORF risk level cutoff scores and AIMSweb 50th percentile norm scores for prediction of proficiency. Finally, the study was conducted to determine optimal cut scores for prediction of proficiency.

The following sections of the chapter provide a discussion of the findings of the study organized around the four major research questions. Additionally, this chapter addresses limitations of the study, implications for practice at various levels, and recommendations for future research. The investigation was guided by the following research questions:

1. Using DIBELS Oral Reading Fluency (ORF) and AIMSweb Maze-CBM universal screening scores, which measure or combination of measures are the best predictors of standard scale scores on a state developed reading accountability measure for third, fourth, and fifth grade students?

2. Is there a difference in the magnitude of the relationship between EOG and ORF and Maze among third, fourth, and fifth grade?
3. How accurate are published DIBELS ORF risk level cutoff scores for ORF and AIMSweb Maze scores for identifying third, fourth, and fifth grade students who will or will not be proficient as measured by the statewide grade level NC EOG Reading Comprehension test?
4. What are the optimal DIBELS ORF and AIMSweb Maze-CBM cut scores to use when attempting to predict reading comprehension by the end of third, fourth, and fifth grade level as measured by EOG performance?

The current study demonstrated a significant relationship between scores on DIBELS ORF, AIMSweb Maze-CBM, and outcomes on the NC EOG Reading Comprehension assessment. First, results indicated that both ORF and Maze measures significantly predicted proficiency on the NC EOG in Reading Comprehension. ORF measures accounted for the greatest amount of variance and significantly predicted reading proficiency in third, fourth, and fifth grades. Maze measures significantly predicted reading proficiency in third and fourth grades, but did not account for a significant amount of variance in scores for fifth grade when considered with ORF. Next, there were significant grade level differences in the magnitude of the relationship between the measures. Fifth grade had significantly higher correlation coefficients with EOG than third grade. Additionally, the results of the study supported the recommended DIBELS risk level cutoff scores and Maze 50th percentile Norm scores as significant predictors of outcomes on EOG in third, fourth, and fifth grades. Finally, optimal cutoff scores were determined by considering the tradeoff between sensitivity and specificity.

Once sensitivity and specificity were maximized, the optimal cutoff scores yielded slightly different cutoff values than the recommended cutoff scores.

Relationship of Measures

Question 1: Using DIBELS Oral Reading Fluency (ORF) and AIMSweb Maze-CBM universal screening scores, which measure or combination of measures are the best predictors of standard scale scores on a state developed reading accountability measure for third, fourth, and fifth grade students?

The results of this study indicate that performance on DIBELS and Maze measures was significantly related to performance on EOG. In each grade level ORF and Maze measures were associated with performance on EOG. Clearly, these findings highlight the potential for using curriculum-based measures to predict performance on high-stakes accountability testing and reinforce the ability to identify students who need additional support. A moderate relationship was found between performance on measures of oral reading fluency and statewide assessments with statistically significant correlation coefficients between ORF and EOG scores ranging from .531 to .753. The relationship between Maze measures and EOG was also moderate with statistically significant correlation coefficients ranging from .523 to .617.

Findings from the current study are consistent with previous research highlighting the strong association between ORF and statewide assessments (Baker et al., 2008; Barger, 2003; Catts et al., 2009; Chard et al., 2008; Good et al., 2001; McGlinchey & Hixson, 2004; Roehrig et al., 2008; Schilling et al., 2007; Shapiro et al., 2008; Shaw & Shaw, 2002; VanderMeer et al., 2005; Wilson, 2005; Wood, 2006; Wood, 2009). In each grade level, ORF made the largest contribution to the prediction of EOG standard scores.

Scores on ORF accounted for 36%, 44%, and 70% of the variance in standard scores in third, fourth, and fifth grades, respectively. Therefore, results suggest it is feasible to use ORF measures to detect students who may require more targeted instructional support to meet grade level proficiency requirements. Based on results, ORF measures can be used to monitor progress toward grade-level expectations as measured by the statewide assessment.

Results also support previous research on the association between Maze and statewide assessments (Ardoin et al., 2004; Silberglitt et al., 2006; Wiley & Deno, 2005). Current findings provide evidence for the use of Maze measures for third and fourth grades. However, results suggest questionable use of Maze measures for fifth grade to predict proficiency because when ORF is included in the equation, Maze does account for a significant amount of variance on EOG for fifth grade students. These results seem to be compatible with those of Ardoin et al., who found measures of oral reading fluency as a valid predictor in third grade, but indicated CBM was a better predictor than Maze measures.

Since reading fluency has been identified as an important component in reading (NRP, 2000), the fact that reading fluency rates are associated with performance on high-stakes assessment is encouraging. Other studies have also emphasized the importance of ORF (Good et al., 2001) and Maze (Ardoin et al., 2004; Silberglitt et al., 2006; Wiley & Deno, 2005). In previous studies, DIBELS ORF was found to have moderate to strong predictability of statewide reading comprehension tests (Barger, 2003; Crawford et al., 2001; Shaw & Shaw, 2002; Schilling et al., 2007).

Despite the lack of significance for Maze measures in fifth grade, results confirm ORF and Maze together accounted for 35%, 48%, and 57% of the variance in outcomes on NC EOG for third, fourth, and fifth grades, respectively. As such, it is feasible for teachers to use data from both ORF and Maze to help determine which students are at risk for proficiency on the EOG since both measures require relatively small amounts of time (1-minute for ORF and 3 minutes for Maze). Using formative data, educators can increase instructional support in response to the magnitude of student need in order to help students to attain proficiency on state-mandated testing.

It is valuable to know that both measures account for a significant amount of the variance in EOG scores. However, it is even more advantageous to know that ORF alone accounted for most of the variance in each grade level. Therefore, it seems reasonable to suggest DIBELS ORF alone could be used in schools with limited resources, time, and personnel. This is especially true for fifth grade since Maze measures did not account for a significant amount of variance in EOG scores.

Magnitude of the Relationship as a Function of Grade

Question 2: Is there a difference in the magnitude of the relationship between EOG and ORF and Maze among third, fourth, and fifth grade?

The question of a difference in the magnitude of the relationship between the fluency measures and standardized, statewide assessments is important to address because ORF is consistently used in schools across the nation an indicator of reading ability for other purposes, such as Response to Intervention (RTI). Previous research has suggested significant grade level differences in the magnitude of the relationship between fluency and scores on state accountability tests. Many researchers report a decrease in the

magnitude of the correlation as students gain experience reading connected text. As such, fluency was less closely associated with comprehension as students gained experience (Baker et al., 2008; Fuchs et al., 2001; Jenkins & Jewell, 1993; Schilling et al., 2007; Silberglitt et al., 2006; Yovanoff et al., 2005).

In the current study, ORF was more closely associated with performance on EOG as grade level increased, with significant differences between coefficients in third and fifth grades. The mean ORF scores from the current data increased approximately 15 words from third to fourth grade and increased another 10 words from fourth to fifth grade, with correlations increasing from .557 to .657 to .753 in third, fourth, and fifth grades, respectively. Therefore, results suggest the relationship between fluency and comprehension remain strong in later elementary grades with the overall value of the predictor increasing significantly. Results of this study were consistent with studies that found a consistent or increasing relationship between ORF and comprehension, despite grade level increases (Sibley et al., 2001; Wood, 2006).

The increasing relationship between ORF and state test scores found in the current study demonstrates that students continue to develop fluency skills through fifth grade. In fact, the mean increased from 111 words correct per minute to 136 words correct per minute between third and fifth grades. Similar to findings of Wood (2006) who found fluency rates to increase 16 words per minute each year, findings from the current study suggest a consistent relationship across later elementary grades between ORF and comprehension. However, it should be noted that findings from Fuchs et al (2001) suggest fluency rates slow down during these years. More research is needed to resolve

these differences, but current school level initiatives to improve fluency and increased teacher knowledge of fluency may have contributed to differences in results.

Diagnostic Efficiency of DIBELS and Maze

Question 3: How accurate are published DIBELS ORF risk level cutoff scores for ORF and AIMSweb Maze scores for identifying third, fourth, and fifth grade students who will or will not be proficient as measured by the statewide grade level NC EOG Reading Comprehension test?

With consistent findings from research to suggest students who do not learn to read by second grade are likely to continue to struggle (Juel, 1988), it is imperative to use predictive measures in order to monitor progress and change student outcomes, particularly for those students with consistent underachievement in reading. Information about students who are (or are not) truly at risk is necessary and beneficial for educators to target needs and provide the level of support necessary. Diagnostic efficiency statistics are useful for predicting whether students are likely to pass or fail high-stakes assessments.

ROC Curves Analysis can be used to determine the true positive rate (sensitivity) and the false positive rate (1-specificity) for different cut off points. A test with perfect discrimination has a ROC plot that passes through the upper left corner. This would indicate 100% sensitivity and 100% specificity, which is the highest overall accuracy (Zweig & Campbell, 1993). In this study, sensitivity represented the true positive rate, which was the percentage of students who performed below proficient levels on EOG and were correctly identified as at risk by DIBELS / Maze (i.e., the chance the diagnostic test will show the presence of a problem). Specificity represented the true negative rate,

which was the percentage of students who performed at or above proficient levels on EOG and were correctly identified as low risk by DIBELS / Maze (i.e., the chance the diagnostic test will show the absence of a problem).

Additionally, positive predictive value (PPV), negative predictive value (NPV), accuracy rate (AR), and misclassification rate (MR) are other characteristics that express the usefulness of a diagnostic test. PPV is the probability of poor comprehension when comprehension problems were predicted. This represents the chance that a student with a positive result (low ORF/Maze) actually has a problem with comprehension (below Level III EOG score). NPV is the probability of no comprehension problems when no comprehension problems were predicted. This represents the chance that a student with a negative diagnostic test (high ORF/Maze) actually doesn't have a problem with comprehension (below Level III EOG score). Accuracy rate is the proportion of all correctly identified students. This is represented by the sum of the true positives (students below Level III who were identified at risk) and true negatives (students not below Level III who were identified not at risk), out of the total number of students. Finally, the misclassification rate is the proportion of all misclassified students. This is represented by the sum of the false negatives (students below Level III who were not identified at risk) and false positives (students not below Level III who were identified at risk), out of the total number of students.

For diagnostic accuracy in the current study, the following questions (Gohen, 2007) are helpful for interpretation: (a) sensitivity - If the student has comprehension problems (low EOG scores), what is the chance that ORF/Maze diagnostic test will show that the student has problems? (b) specificity – If the student doesn't have comprehension

problems (proficient EOG scores), what is the chance that ORF/Maze diagnostic test will show that the student does not have problems? (3) PPV – The student has low ORF/Maze scores (positive), what is the chance that the student actually has comprehension problems (low EOG scores)? (4) NPV- The student has high ORF/Maze scores (negative), what is the chance that the student actually doesn't have comprehension problems (proficient EOG scores)? (5) AR – How many students were correctly classified? (6) MR – How many students were misclassified?

Using current data, diagnostic accuracy of DIBELS ORF and AIMSweb Maze evaluate how well the DIBELS recommended risk level cutoff scores and the Maze 50th percentile norm scores differentiate between students who will or will not exhibit comprehension problems as measured by high-stakes, statewide assessments. For recommended cutoff scores for each measure in each grade level the inspection of the area under the curve (AUC) suggested that the cut scores resulted in levels of sensitivity and specificity above .75, which is considered adequate (Swets, 1988). For ORF, results of ROC curves indicated AUC index of .809, .879, and .900 for third, fourth, and fifth grades, respectively. For Maze, results of ROC curves indicated AUC index of .788, .839, and .814 for third, fourth, and fifth grades, respectively. Therefore, findings support the DIBELS recommended risk level cutoff scores from Good and Kaminski (2002) and AIMSweb Maze 50th percentile norm cutoff scores (Edformation, 2009) as accurate in prediction of students who will or will not be proficient on the NC EOG in Reading Comprehension for third, fourth, and fifth grades. However, using the recommended cutoff scores, sensitivity levels were adequate in each grade level for ORF and Maze, but specificity levels for Maze were less than adequate in each grade level. Therefore,

sensitivity and specificity levels were maximized in order to determine optimal cutoff scores for prediction of proficiency, which yielded slightly different cutoff values.

Many researchers have validated the recommended risk level DIBELS ORF cutoff scores. Findings from a study conducted by Good, Simmons, and Kame'enui (2001) provided strong support for the utility of the DIBELS benchmark goals. Sibley et al. (2001) found the recommended risk level cutoff scores for DIBELS ORF accurately predicted performance on high-stakes state and local achievement measures. Wood (2006) found the established ORF cut scores were accurate in determining whether students would pass or fail. Additionally, Goffreda et al. (2009) found ORF yielded high levels of sensitivity and specificity when using recommended cutoff scores. In fact, alternative optimal cutoff scores found in Goffreda et al. were not significantly different from those recommended by Good and Kaminski (2002).

However, Hintze et al. (2003) noted standard DIBELS cutoff scores yielded low levels of specificity, which is consistent with current findings. Shapiro et al. (2008) indicated there was a need to maximize sensitivity and specificity for fluency measures, which is also consistent with the results of the current study. Additionally, Roehrig et al. (2008), suggested adjusting the at risk category to <45 and low risk category to >76 in order to improve efficiency with higher values in sensitivity, specificity, and overall correct classification. Shaw and Shaw (2002) recommended lowering the cutoff score for third grade to 90 in order to yield greater sensitivity and specificity levels.

Even though DIBELS ORF and AIMSweb Maze were never intended to be predictive of statewide assessments, it is important to note that schools use data from these indicators to make instructional decisions and to determine the impact of

instructional practices. Based on ROC Analysis, the accuracy rate (AR) for ORF ranged from 74% - 88% and the misclassification rate (MR) ranged from 12% - 26%. For Maze, the accuracy rate (AR) ranged from 57% - 71% and the misclassification rate (MR) ranged from 28% - 42%. Clearly, the current data suggests ORF is a more accurate predictor of proficiency for third, fourth, and fifth grade students. Nonetheless, inspection of the ROC Plot indicates both recommended DIBELS cutoff scores (Good & Kaminski, 2002) and the 50th percentile AIMSweb Norm Scores (Edformation, 2009) have adequate AUC (e.g., .greater than .75). Therefore, recommended cutoff scores for both measures provide accurate prediction of overall reading proficiency as measured by the high-stakes accountability tests. However, through inspection of various cutoff scores, sensitivity and specificity levels can be maximized using alternate cutoff scores for the sample included in the current study.

Determination of Optimal Cut Scores to Predict Proficiency

Question 4: What are the optimal DIBELS ORF and AIMSweb Maze-CBM cut scores to use when attempting to predict satisfactory reading comprehension by the end of third, fourth, and fifth grade level as measured by EOG performance?

Educators are faced with the importance of meeting grade level standards as measured by state-mandated testing under NCLB. By identifying optimal cutoff scores to predict outcomes on high-stakes assessments, students at risk for problems with comprehension can work toward target goals. Identification of optimal cutoff scores provides educators with a target goal. In essence, having a target goal that reliably predicts outcomes on high-stakes testing empowers educators to use data to inform instruction and change student outcomes prior to the summative assessment. With the

growing importance of formative data, the current study adds to the literature on diagnostic accuracy of two measures widely used by schools: DIBELS ORF and AIMSweb Maze.

With ROC Analysis, the choice of the cut-score depends on the purpose of the decision and "the definition of the assessment situation are subjective to the researcher" (Hintze & Silberglittt, 2005, p. 376). Varied cutoff scores may be necessary for different types of classification decisions (Goffreda et al., 2009; Hintze et al., 2003). Using the current data, cut scores were adjusted to maximize sensitivity and specificity (Swets, 2000). For each grade level and measure, cut scores were adjusted so that they were as balanced as possible, with a preference for over identification of students at risk rather than under identification.

Results of ROC Analysis suggest DIBELS ORF and AIMSweb Maze probes are reliable measures that can be used to predict outcomes on statewide assessments. Based on inspection of ROC Curve Plot, the outcomes of this study suggest that the cutoff scores for both measures had adequate sensitivity levels, but not adequate specificity levels for each grade level when predicting outcomes on statewide assessments. Therefore, using the current data, optimal cutoff scores were slightly different than DIBELS ORF recommended scores (Good & Kaminski, 2002) and AIMSweb Maze 50th percentile Norm scores (Edformation, 2009) for prediction of high-stakes assessments. Findings support the need to look carefully at recommended DIBELS ORF benchmarks (Shapiro et al., 2008), especially in fifth grade since current data indicate a much higher optimal cut score is necessary in order to maximize levels of sensitivity and specificity.

In order to determine the optimal cut scores for predicting outcomes, the question of whether to set a lower or higher threshold depends on the need to identify more students potentially at risk who may truly not be at risk (false positives) versus the risk of missing students who may need remediation but were not identified at risk (false negatives). In schools with adequate personnel and resources, providing extra support to students identified at risk who may not really need extra support is not a significant problem in comparison to not identifying students who may truly need that level of support. However, when schools lack sufficient resources and personnel, it may be more important to set a strict threshold in order to limit the number of false positives (Swets et al., 2000). The tradeoff between sensitivity and specificity should be based on the particular needs of the school or district because large numbers of false positives (students identified at risk who were not truly at risk) would be costly. On the other hand, large numbers of true positives (students identified at risk who were truly at risk) could provide educators with information to make data-based instructional decisions to change student outcomes on state mandated measures of reading comprehension. Another main concern would be limiting the number of false negatives (students not identified at risk who were truly at risk) because these students would not be identified and would not receive any supplemental instruction, but they truly need support.

According to Swets et al. (2000), setting a higher cut score on the predictive measure decreases the probability of predicting failure and increases the probability of predicting success on the criterion measure. A more lenient threshold allows the likelihood of missing students to be decreased and the likelihood of identifying students who need additional interventions to be increased. Findings from this study were

consistent with findings of Goffreda et al. (2009), who found optimal cut scores to maximize sensitivity and specificity resulted in slightly different cut off values for DIBELS and Maze in each grade level. Findings were also consistent with findings of Roehrig et al. (2008), who suggested more students could be identified using recalibrated scores according to ROC curve results.

Limitations

Although results were consistent with results of previous research, the findings of the present study have some limitations that should be considered. First, this study represents results from only one elementary school in North Carolina. Therefore, the ability to generalize results to different school districts and other states may be limited due to the small sample. The school was in a small, suburban district with no other schools in the district using DIBELS ORF or Maze formative measures of reading. Replication of the study including a greater number of participants from other schools in other districts could provide generalizability of findings.

Second, data gathered included only the spring benchmark for DIBELS ORF and AIMSweb Maze from 1 year. Since the spring benchmark for DIBELS ORF and AIMSweb Maze occurred only a few weeks prior to EOG testing, results are more concurrent in nature, as there was not time in between to change outcomes of EOG based on results of ORF or Maze. This is a potential limitation of the current study, despite the fact that the third administration typically has the strongest correlations with accountability tests (Roehrig et al, 2008; Shaw & Shaw, 2002). In future studies, this potential limitation could be overcome by using fall, winter, and spring benchmark data for prediction of proficiency on EOG.

Finally, all studies of diagnostic accuracy incorporate a gold standard of 100% accuracy (Swets, 1996). However, every measure of student progress has shortcomings, which is a possible limitation of the current study. For example, one more question right or wrong on the EOG can increase or decrease the standard score on the EOG above or below the cutoff score for proficiency. In turn, the student would be in a different category of proficiency for analysis (i.e., proficient, Level III or not proficient, Level II). In fact, students who score within 1 SE on the NCEOG are considered to "pass" the NCEOG for purposes of accountability. However, when scores were dichotomized for the current study, the Level III cutoff score for each grade level was used to determine proficient (1) or not proficient (0). For students who were on the borderline of proficiency, this is a possible limitation. One way to overcome this error would be cross-validating the findings of one study with findings from another in order to provide more confidence in generalization of findings (Sideridis, Morgan, Botsas, Padeliadu, & Fuchs, 2006).

Suggestions for Future Research

Future replications of this study across different schools in other districts are needed in order to provide generalizability of the findings. Alternate, optimal cut scores were determined based on balancing identification of students at risk for not being proficient on high-stakes assessments (sensitivity) with identification of students not at risk for being proficient on high-stakes assessment (specificity). Another consideration in optimal cutoff scores was balancing the cost of identifying students who may really not need extra support (false positives) with the risk of not identifying students who may need the support (false negatives). Since this may require allocation of resources to

students who were inaccurately identified, replication studies or longitudinal follow-up studies would be beneficial in order to determine if alternate cut scores can be generalized to meet the needs of schools, regardless of resources available. If a limited amount of resources are available, the cut-score may need to be more conservative. As recommended by Swets et al. (2000), the optimal cut-score should be chosen with an understanding of the risks involved with incorrect classification.

Future research to determine how well CBM predicts proficiency on statewide systems of accountability in reading and math is another area that warrants further empirical investigation. Research is warranted to identify students at risk for proficiency in both academic areas prior to summative assessments, such as the EOG in Reading and Math. Findings from previous research have suggested data from reading and math benchmarks had significant correlations with student outcomes on statewide achievement tests in reading and math (Crawford et al., 2001; Keller-Margulis et al., 2008; Shapiro et al., 2006).

Another area of research that deserves further empirical investigation is the use of student progress toward target goals (growth) as an additional predictor of student outcomes on high-stakes assessments. As an extension of work by Baker et al. (2008), research is warranted to determine whether growth in ORF can predict performance on statewide assessments. As noted by Stanovich (1986), students who fall behind in reading have a more difficult time bridging the gap over time, but when difficulties are recognized, skills can be remediated to change their learning trajectory. Longitudinal data following a cohort of students from early grades may depict the impact of growth in the development of fluency.

Finally, research is warranted with other student populations, including students in other grade levels. Research with younger and older students could help determine the value of curriculum-based measures for predicting high-stakes accountability measures in early elementary, middle, and high school. Additional research to investigate relations between ORF, Maze, and high-stakes tests in other grade levels with a focus on a subset of students, such as ELL, special education, or students at risk for reading failure, or academically gifted is an area that warrants further empirical investigation.

Implications for Practice

Despite limitations, the results of this study offer practical implications for administrators and practitioners at various levels. This study has a number of implications for administration at the district and school level. Additionally current findings offer practical implications for practitioners in the classroom, such as general education and special education teachers. Wayman, Midgley, and Stringfield (2006) suggest data initiatives are possible "when they are built with proper supports at all levels and help educators in this learning endeavor" (p.2). Implications for administration at the district and school level are presented in the following section. In addition, implications for practitioners, such as general education and special education classroom teachers are presented.

Implications for District Level Administration

No Child Left Behind Act of 2001 (NCLB, 2001) mandated statewide systems of assessment in place to gauge Adequate Yearly Progress (AYP). Therefore, the relationship of DIBELS ORF and AIMSweb Maze to standardized, high-stakes assessments in reading is relevant to school districts across the nation. Additionally,

initiatives such as Response to Intervention (RTI) and Professional Learning Communities (PLC) promote school-wide data collection and data-based (data-driven) decision making. Initiatives with a focus on data have caused a paradigm shift. The focus has changed from making sure students are provided with instruction to making sure all students learn (DuFour, 2004). Solutions for ensuring that all students learn require problem solving at the systems level and the individual student level (Tilly, 2008). In essence, due to federal and state initiatives, there is a growing need for the use of formative assessment to inform instruction and best meet the needs of all students.

Specifically, findings of the current study demonstrate significant relationships and add to the literature by demonstrating brief measures of oral reading fluency and comprehension can be used in the prediction of performance on high-stakes assessments in elementary grades. With increased importance of statewide assessments under NCLB (2001), it is important to have reliable data to target students in need of additional support in reading prior to high-stakes assessment. Findings of the current study support that DIBELS ORF and AIMSweb Maze measures assess skills that are necessary for students to demonstrate proficiency on statewide, standardized assessment. Both measures offer a quick and efficient way to monitor student progress toward grade level expectations.

Furthermore, results of this study provide confirmation of the importance of formative assessment. Findings are consistent with previous research indicating results of formative data are useful for determining the likelihood of proficiency on high-stakes, summative assessments based on a particular level of oral reading fluency (Baker et al., 2008; Buck & Torgesen, 2003; Catts et al., 2009; Chard et al., 2008; Crawford et al., 2001; Goffreda et al., 2009; Good et al., 2001; McGlinchey & Hixson, 2004; Schilling et

al., 2007; Shapiro et al., 2006; Shapiro et al., 2008; Roehrig et al., 2008; Stage & Jacobsen, 2001; Wood, 2006; Wood, 2009).

With information to support ORF and Maze measures as valid predictors of statewide assessment outcomes, it seems reasonable to suggest that district level administration focus staff development opportunities around target goals directed at the implications of data. Additionally, to change outcomes on statewide assessments based on data, allocation of resources by school district administrators may need to be prioritized to sustain gathering of data, interpretation of data, and delivery of instruction based on data. For example, districts should carefully consider financial priorities to provide schools with resources necessary to provide (a) training for teachers, (b) measurement of student skills (universal screening instrument), (c) research-based interventions to target needs (trained teacher and materials), (d) measurement of progress (formative, progress monitoring measures), (e) interpretation of data (data manager), and (f) fidelity of implementation (literacy coach/data manager). Essentially, by carefully considering the relationship between DIBELS ORF, Maze, and high-stakes assessment, administrators can use many of the same principles used within an RTI framework with a focus on data-based decision making to change student outcomes on statewide assessments.

Implications for School Level Administration

The current study provides administrators at the school level information about using curriculum-based measures as predictors of EOG scores. Specifically, the study is useful to administrators because it provides evidence of the diagnostic accuracy of DIBELS ORF and Maze for identifying those students who may or may not be at risk for

proficiency on state-mandated high-stakes assessment in reading. Results of the current study suggest that DIBELS ORF and AIMSweb Maze measures are useful for differentiating between students who are likely to be proficient and those who are not likely to be proficient on EOG testing. When used together, these measures offer 79%, 81%, and 83% correct classification in third, fourth, and fifth grades, respectively.

In light of federal and state mandates, such as NCLB (2001) and IDEA (2004), such information is relevant and meaningful to administrators for decisions for AYP. When the measures are administered with high fidelity, administrators can use data to inform educational decisions such as instructional programs, allocation of resources, staff distribution, staff development, and scheduling. Similar to the implications of the study for the district, the impact of recent initiatives at the school level compels administrators to run a data-driven school.

It is important for administrators to realize scores on ORF and Maze measures accounted for 35%, 48%, and 57% of the variance in NC EOG reading comprehension scores for third, fourth and fifth grades, respectively. Moderate correlations between DIBELS and NCEOG as well as Maze and NCEOG provide evidence that both DIBELS ORF and Maze evaluate similar skills and abilities as the high-stakes, statewide assessment of reading comprehension.

Findings of this study support fluency as an important goal and confirm the use of ORF for the purpose of making decisions about who is on track for proficient performance (or not proficient performance) on the NC EOG in third, fourth, and fifth grades. DIBELS ORF accounted for most of the variability in EOG standards scores for each grade level (36%, 44%, 70% for third, fourth, and fifth, respectively). Therefore, use

of DIBELS ORF to measure progress toward grade level expectations should be a top priority for administrators. These findings are consistent with other studies emphasizing the importance of oral reading fluency (Baker et al., 2008; Buck & Torgesen, 2003; Good et al., 2001; McGlinchey & Hixson; Riedel, 2007; Roehrig et al., 2008; Schilling et al., 2007; Shapiro et al., 2006; Shapiro et al., 2008; Stage & Jacobsen, 2001; Wood, 2006; Wood, 2009)

Administrators may also want to consider that results of the study support the additional use of Maze measures in third and fourth grades. In the current study, Maze measures together with ORF predicted proficiency for third, fourth, and fifth grade levels. However, in fifth grade, Maze scores alone did not significantly increase the odds of predicting reading proficiency as measured by the EOG. Therefore, the usefulness of Maze measures in fifth grade is questionable. Overall, for third and fourth grades, findings were consistent with previous studies suggesting that an additional measure of comprehension provided accuracy of prediction (Ardoin et al., 2004; Shaprio et al., 2008; Silberglitt et al., 2006; Wiley & Deno, 2005).

Implications for General Education and Special Education Teachers

Results of the current study have practical significance at the classroom level. The National Reading Panel (NRP, 2000) found that the components of (a) phonemic awareness, (b) alphabetic understanding, (c) vocabulary, (d) comprehension, and (e) accuracy and fluency are all necessary components of effective reading instruction. The importance of developing reading fluency has been highlighted as an important goal by the NRP and research has consistently shown the association between reading fluency and overall reading proficiency and comprehension (Burke & Hagan-Burke, 2007; Hintze

et al., 2002; Jenkins & Jewell, 1993; Riedel, 2007; Roberts et al., 2005). Findings of the current study validate the importance of reading fluency.

Recent research and federal initiatives have pressed teachers to use formative assessment to make data-based instructional decisions in order to meet adequate yearly progress (AYP) goals (NCLB, 2002; IDEA, 2004). Therefore, at the classroom level, general education and special education teachers in third and fourth grades may find both instruments useful for assessing skill development, since DIBELS ORF and AIMSweb Maze provided 79%, 81%, and 83% correct classification for third, fourth, and fifth grades, respectively.

This research demonstrates statistically significant AUC values in third, fourth, and fifth grades for DIBELS ORF and Maze. Sensitivity levels (students at risk who were identified at risk) were adequate for both measures using recommended cut scores. However, with such low specificity levels (students at proficient level who were identified at proficient level), there are a significant number of students identified as at risk who were truly not at risk (false positives). For classroom teachers, this results in a significant number of students who require a substantial amount of support in order to be successful on EOG testing. For this reason, teachers may need to use alternate, optimal cutoff scores when making decisions about proficiency on high-stakes assessments.

The optimal cutoff values have significant meaning for general and special education teachers. Since the choice of cutoff scores depends on the purpose of the decision (Hintze & Silberglitt, 2005), a threshold was set to maximize the number of at risk students who were identified at risk (sensitivity) and the number of students not at risk who were identified as such (specificity). This threshold maximized sensitivity and

specificity levels, which resulted in more students identified correctly. At the general education classroom level, providing instructional support to students who actually do not need that level of support is time consuming and unnecessary. In fact, for special education teachers, the implications of over identification could be detrimental to other students who truly require more individualized, intensive instruction.

However, one important consideration is that when ORF and Maze measures are both included as part of universal screening in schools, any student identified below a specified level would require frequent progress monitoring. With consistent monitoring using alternate probes, misidentification of students may be minimized. In turn, the large number of false positives and false negatives may not be a significant concern. Based on current data, the alternate, optimal scores identified more students correctly. Therefore, these cutoff scores may provide valuable information to teachers who make instructional decisions based on data.

Summary

Overall, findings of the current study have theoretical and practical implications. The theoretical foundation for fluency in reading can be traced to LaBerge and Samuels (1974) Automatic Information Processing Model. According to the automaticity theory, good reading comprehension depends on developing skills toward an automatic level in order to develop higher skills such as comprehension (Samuels, 1994). Fluent oral reading is a significant factor in overall reading ability (Strecker et al., 1998) and results of the current study indicate a relationship between fluency and scores on a statewide assessment of comprehension.

Even though curriculum-based measures were not designed for the purpose of prediction of high-stakes assessment, these brief, 1-minute measures are useful when attempting to determine whether a student will be successful on high-stakes assessments Specifically, findings of the current study were consistent with previous research suggesting DIBELS ORF scores can be used to predict performance on high-stakes statewide assessments of reading used for accountability purposes. Furthermore, findings suggest student outcomes on Maze-CBM measures, together with student outcomes on ORF measures can predict student outcomes on high-stakes, statewide assessments.

Results of the study clearly support the use of DIBELS ORF in third, fourth, and fifth grades as an effective screening tool to use for prediction of reading proficiency. Additionally, AIMSweb Maze was useful in third and fourth grades for prediction of reading proficiency, as measured by high-stakes assessment. Depending on resources available, educators may choose to administer DIBELS ORF alone or DIBELS ORF with AIMSweb Maze to gauge student progress toward meeting grade level standards as measured by End-of-Grade assessments. However, for fifth grade, the use of Maze measures is questionable.

In determining the diagnostic accuracy of recommended (Good & Kaminski, 2002) cutoff scores, the recommended DIBELS benchmark level cutoff scores and AIMSweb Maze 50th percentile scores were found to be adequate in predicting student outcomes on NC EOG. Both ORF and Maze recommended cutoff scores were accurate for prediction, but Maze yielded less than adequate specificity levels in each grade level. Optimal cutoff scores were determined to maximize sensitivity and specificity levels. The alternate, optimal cutoff scores were only slightly different than recommended cutoff

scores in each grade level, with fifth grade requiring a higher cutoff score to maximize sensitivity and specificity.

Results of this study should be of interest to educators at various levels. Administrators can use information about the importance of formative measures as predictors of performance on high-stakes assessment to inform educational decisions. General education and special education teachers can use information to change learning trajectories and improve student outcomes by providing support necessary prior to the end of the learning cycle (Kennedy, et al., 2008). The significance of the use of fluency measures, specifically DIBELS ORF, is important for administrators, general educators, and special educators.

REFERENCES

Adams, M. J. (1990). *Beginning to read: Thinking and learning about print.* Cambridge, MA: MIT Press.

Anderson, R. C., Hiebert, E. H., Scott, J. A., & Wilkinson, I. A. G. (1985). *Becoming a nation of readers: The report of the commission on reading*. Washington, DC: The National Institute of Education.

Anderson, R. C., Wilkinson, I. A. G., & Mason, J. M. (1991). A microanalysis of the small-group, guided reading lesson: Effects of an emphasis on global story meaning. *Reading Research Quarterly, 26,* 417- 441.

Ardoin, S. P., Witt, J. C., Suldo, S. 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.

Armbruster, B. B., Lehr, F., & Osborn, J. (2001). *Put reading first: The research building blocks for teaching children to read.* Washington, DC: Center for the Improvement of Early Reading Achievement.

Arnold, V. A., & Smith, C. B. (1987). *Macmillan connections reading program.* New York: Macmillan.

Bain, S. K., & Garlock, J. W. (1992). Cross-validation of criterion-related validity for CBM reading passages. *Diagnostique, 17,* 202-208.

Baker, S. K., Smolkowski, K., Katz, R., Fien, H., Seeley, J. R., Kame'enui, E. J., et al. (2008). Reading fluency as a predictor of reading proficiency in low-performing, high-poverty schools. *School Psychology Review, 37,* 18-37.

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

Biggs, J. (1998). Assessment and classroom learning: A role for summative assessment? *Assessment in Education: Principles, Policy & Practice, 5,* 103-110.

Black, P., & William, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy, & Practice, 5,* 7-74.

Black, P., & William, D. (2006). Developing a theory of formative assessment. In J. Gardner (Ed.), *Assessment and learning* (p. 81-100). London: Sage.

Black, P., & William, D. (2009). Developing a theory of formative assessment. *Educational Assessment, Evaluation and Accountability, 21,* 5-31.

Bloom, B. S., Hastings, J. T., & Madaus, G. F. (1971). *Handbook on the formative and summative evaluation of student learning.* New York: McGraw-Hill.

Buck, J., & Torgesen, J. (2003). *The relationship between performance on a measure of oral reading fluency and performance on the Florida Comprehensive Assessment Test* (Technical Report No. 1). Tallahassee: Florida State University, Florida Center for Reading Research.

Burke, M. D., & Hagan-Burke, S. (2007). Concurrent criterion-related validity of early literacy indicators for middle of first grade. *Assessment for Effective Intervention, 32,* 66-77.

Caffrey, E., Fuchs, D., & Fuchs, L. S. (2008). The predictive validity of dynamic assessment: A review. *Journal of Special Education, 41,* 254-270.

Catts, H. W., Petscher, Y., Schatschneider, C., Bridges, M. S., & Mendoza, K. (2009). Floor effects associated with universal screening and their impact on the early identification of reading disabilities. *Journal of Learning Disabilities, 42,* 163-176.

Chard, D. J., Stoolmiller, M., Harn, B. A., Wanzek, J., Vaughn, S., Linan-Thompson, S., et al. (2008). Predicting reading success in a multilevel schoolwide reading model. *Journal of Learning Disabilities, 41,* 174-188.

Children's Educational Services. (1987). *Test of reading fluency.* Minneapolis, MN: Author.

Coleman, M. R., Buysse, V., & Neitzel, J. (2006). *Recognition and response: An early intervening system for young children at-risk for learning disabilities.* Full report. Chapel Hill: University of North Carolina, Frank Porter Graham Child Development Institute.

Cowen, J. E. (2003). *A balanced approach to beginning reading instruction: A synthesis of six major U.S. research studies.* Newark, DE: International Reading Association.

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

CTB/McGraw-Hill. (2002). TerraNova, the second edition. Monterey, CA: Author

Denton, C. A., Fletcher, J. M., Anthony, J. L., & Francis, D. J. (2006). An evaluation of intensive intervention for students with persistent reading difficulties. *Journal of Learning Disabilities, 39,* 447-466.

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

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

Deno, S. L. (2003). Curriculum-based measures: Development and perspectives. *Assessment for effective intervention, 28,* 3-12.

Deno, S. L., Espin, C., Maruyama, G., & Cohen C. (1989). *Basic Academic Skills Samples (BASS).* (U.S. Department of Education Grant #G00873025588) Minneapolis, MN: University of Minnesota.

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

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

Dorn, S. (2007). *Accountability Frankenstein: Understanding and taming the monster.* Charlotte, NC: Information Age.

DuFour, R. (2004). Schools as learning communities. What is a "Professional learning community?" *Educational Leadership, 61,* 6-11.

Dunn, L. M., & Markwardt, F. C. (1970). *Peabody individual achievement test.* Circle Pines, MN: American Guidance Service.

Edformation. (2002). *AIMSweb standard benchmark reading assessment passages.* Eden Prarie, MN: Author.

Edformation. (2005). *AIMSweb progress monitoring and improvement system.* Available from http://www.aimsweb.com

Edformation. (2009). *AIMSweb Maze Curriculum-based Measures.* Available from http://www.aimweb.com

Faykus, S. P., & McCurdy, B. L. (1998). Evaluating the sensitivity of the maze as an index of reading proficiency for students who are severely deficient in reading. *Education and Treatment of Children, 21,* 1-21.

Flurkey, A. (2006). What's "normal" about real reading? In K. S. Goodman (Ed.), *The truth about DIBELS: What it is, what it does* (pp. 40-49). Portsmouth, NH: Heinemann.

Francis, D. J., Shaywitz, S. E., Stuebing, K. K., Shaywitz, B. A., & Fletcher, J. M. (1996). Developmental lag versus deficit models of reading disabilities: A longitudinal individual growth curves analysis. *Journal of Educational Psychology, 88,* 3-17.

Fuchs, L. S. (2003). *Growth modeling oral reading fluency passages: Reusable stimulus materials.* Unpublished, Nashville, TN.

Fuchs, L. S., & Deno, S. L. (1991). Paradigmatic distinctions between instructionally relevant measurement models. *Exceptional Children, 57,* 488-499.

Fuchs, L. S., Deno, S. L., & Mirkin, P. K. (1984). The effects of frequent curriculum-based measurement and evaluation on pedagogy, student achievement, and student awareness of learning. *American Educational Research Journal, 21,* 449-460.

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

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

Fuchs, L. S., & Fuchs, D. (1997). Use of curriculum-based measurement in identifying students with disabilities. *Focus on Exceptional Children, 30,* 1-14.

Fuchs, D., Fuchs, L. S., Compton, D. L., Bouton, B., Caffrey, E., & Hill, L. (2007). Dynamic assessment as responsiveness to intervention: A scripted protocol to identify young at-risk readers. *Teaching Exceptional Children, 39,* 58-63.

Fuchs, L. S., Fuchs, D., & Deno, S. L. (1982). Reliability and validity of curriculum-based informal reading inventories. *Reading Research Quarterly, 18,* 6-26.

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.

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

Fuchs, L. S., Hamlett, C. L., & Fuchs, D. (1998). *Monitoring basic skills progress: Basic math computation* (2nd ed.). Austin, TX: Pro-Ed.

Fuchs, L. S., Hamlett, B., & Fuchs, D. (1999). *Monitoring basic skills progress: Basic math concepts and applications.* Austin, TX: Pro-Ed.

Fuchs, L., Tindal, G., & Deno, S. (1981). *Effects of varying item domain and sample duration on technical characteristics of daily measures in reading.* (Res. Rep. No. 48). Minneapolis: University of Minnesota, Institute for Research on Learning Disabilities.

Ginn and Company. (1976). *Reading 720.* Lexington, MA: Author.

Goffreda, C. T., Diperna, J. C., & Pedersen, J. A. (2009). Preventive screening for early readers: Predictive validity of the dynamic indicators of basic early literacy skills (DIBELS). *Psychology in the Schools, 46,* 539 – 552.

Gonen, M. (2007). *Analyzing Receiver Operating Characteristic Curves with SAS.* Cary, NC: SAS Institute, Inc.

Good, R. H., & Jefferson, G. (1998). Contemporary perspectives on curriculum-based measurement validity. In M. R. Shinn (Ed.), *Advanced applications of curriculum-based measurement* (pp. 61-88). New York: Guilford Press.

Good, R. H., & Kaminski, R. A. (Eds.). (2002). *Dynamic Indicators of Basic Early Literacy Skills* (sixth ed.). Eugene, OR: Institute for the Development of Educational Achievement.

Good, R. H., Gruba, J., & Kaminski, R.A. (2002). Best practices in using Dynamic Indicators of Basic Early Literacy Skills in an outcomes-driven model. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology IV* (pp. 699-720). Bethesda, MD: National Association of School Psychologists.

Good, R. H., Simmons, D., & Kame'enui, E. (2001). The importance and decision-making utility of a continuum of fluency-based indicators of foundational reading skills for third grade high-stakes outcomes. *Scientific Studies of Reading, 5,* 257-288.

Goodman, K. S. (2006). A critical review of DIBELS. In K. S. Goodman (Ed.), *The truth about DIBELS: What it is, what it does* (pp. 1-39). Portsmouth, NH: Heinemann.

Gorlewski, J. (2008). Research for the classroom. *English Journal, 98,* 94-97.

Gresham, F. M., & Elliot, S. N. (1990). *Social skill rating scale.* Circle Pines, MN; American Guidance Services.

Harcourt. (2003). *Stanford Achievement Test – 10.* New York: Author.

Hasbrouck, J. E., & Tindal, G. (1992). Curriculum-based oral reading fluency norms for student in grades 2 through 5. *Teaching Exceptional Children, 24,* 41-44.

Hasbrouck, J. E., & Tindal, G. (2006). Oral reading fluency norms: A valuable assessment tool for reading teachers. *The Reading Teacher, 59,* 636-644.

Hintze, J. (2007). *NCSS 2007*. Kaysville, UT; NCSS, LLC. Retrieved on June 29, 2009 from www.ncss.com

Hintze, J. M., Callahan, J. E., Matthews, W. J., Williams, S. A., & 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., 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., Ryan, A. L., & Stoner, G. (2003). Concurrent validity and diagnostic accuracy of the dynamic indicators of basic early literacy skills and the comprehensive test of phonological processing. *School Psychology Review, 32,* 541- 556.

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

Hixon, M. D., & McGlinchey, M. T. (2004). The relationship between race, income, and oral reading fluency and performance on two reading comprehension measures. *Journal of Psychoeducational Assessment, 22,* 351-373.

Houghton Mifflin Basal Reading Series. (1989). *Journeys (grade 3). Discoveries (grade 2).* Boston: Author.

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* [Technical Manual]. Eden Prarie, MN: Edformation.

Individuals with Disabilities Education Improvement Act (IDEIA) of 2004, PL 108-446, 20 U.S.C. §§ 1400 et seq.

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

Jenkins, J. R., Deno, S. L., & Mirkin, P K. (1979). *Measuring pupil progress toward the least restrictive environment* (Monograph No. 10). Minneapolis: University of Minnesota, Institute for Research on Learning Disabilities. (ERIC Document Reproduction Service No. ED 185 767)

Juel, C. (1988). Learning to read and write: A longitudinal study of 54 children from first through fourth grade. *Journal of Educational Psychology, 80,* 437-447.

Junker, B. & Matsumura, L. C. (2006). *Beyond summative evaluation: The instructional quality assessment as a professional development tool* (CSE Technical Report 691). Los Angeles: University of California.

Kame'enui, E. J., Francis, D. J., Fuchs, L., Good, R. H., O'Connor, R. E., Simmons, D.C., et al. (2002). *Analysis of reading assessment instruments for K-3.* Washington, DC: National Institute for Literacy. Available at http://idea.uoregon.edu/assessment/index.html

Kaminski, R. A., & Good, R. H. (1996). Toward a technology for assessing basic early literacy skills. *School Psychology Review, 25,* 215-227.

Kaminski, R. A., & Good, R. H. (1998). Assessing early literacy skills in a problem solving model: Dynamic indicators of basic early literacy skills. In M. R. Shinn (Ed.), *Advanced applications of Curriculum-Based Measurement (pp.113-* 142). New York: Guilford.

Karlsen, B., Madden, R., & Gardner, E. F. (1975). *Stanford diagnostic reading test (Green Level Form B).* New York: Harcourt Brace Jovanovich.

Kaufman, A. S., & Kaufman, N. L. (1985). *Kaufman test of educational achievement*. Circle Pines, MN: American Guidance Service.

Kaufman, A. S., & Kaufman, N. L. (1990). *Kaufman brief intelligence test.* Circle Pines, MN: American Guidance Service.

Keller-Margulis, M. A., Shapiro, E. S., & Hintze, J. M. (2008). Long-term diagnostic accuracy of curriculum-based measures in reading and mathematics. *School Psychology Review, 37,* 374-390.

Kennedy, K. J., Chan, J. K., Fok, P. K., & Yu, W. M. (2008). Forms of assessment and their potential for enhancing learning: Conceptual and cultural issues. *Educational Research Policy and Practice, 7,* 197-207.

Klauda, S. L., & Guthrie, J. T. (2008). Relationships of three components of reading fluency to reading comprehension. *Journal of Educational Psychology, 100,* 310-321.

Kranzler, J. H., Brownell, M. T., & Miller, D. (1998). The construct validity of curriculum-based measurement of reading: An empirical test of a plausible rival hypothesis. *Journal of School Psychology, 36,* 399-415.

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

Kuhn, M. R., & Stahl, S. A. (2003). Fluency: A review of developmental and remedial practices. *Journal of Educational Psychology, 95,* 3-21.

LaBerge, D., & Samuels, S. (1974). Toward a theory of automatic information processing in reading. *Cognitive Psychology, 6,* 292-323.

Landau, S., & Everitt, B. (2004). *A handbook of statistical analyses using SPSS.* Boca Raton, FL: Chapman & Hall.

Lane, H. B., Hudson, R. F., Leite, W. I., Kosanovich, M. L., Strout, M. T., Fenty, N. S., et al. (2009). Teacher knowledge about reading fluency and indicators of students' fluency growth in reading first schools. *Reading & Writing Quarterly, 25,* 57-86.

Lomax, R. (1983). Applying structural modeling to some component processes of reading comprehension development. *Journal of Experimental Education, 52,* 33-40.

MacGinitie, W. H., MacGinitie, R. K, Maria, K., & Dreyer, L. G. (2000). *Gates-MacGinitie reading tests: Manual for scoring and interpretation.* Itasca, IL: Riverside.

Manning, M., Kamii, C., & Kato, T. (2006). DIBELS: Not justifiable. In K. S. Goodman (Ed.), *The truth about DIBELS: What it is, what it does* (pp. 71-78). Portsmouth, NH: Heinemann.

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

Marston, D., & Magnusson, D. (1985). Implementing curriculum-based measurement in special and regular education settings. *Exceptional Children, 52,* 266-276.

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

Mehrens, W. A., & Clarizio, H. F. (1993). Curriculum-based measurement: Conceptual and psychometric considerations. *Psychology in the Schools, 30,* 241-254.

Mirkin, P. K., Deno, S. L., Fuchs, L. S., Wesson, C., Tindal, G., Marston, D., et al. (1981). *Procedures to develop and monitor progress on IEP goals.* Minneapolis: University of Minnesota.

National Center for Education Statistics (2007). *The Condition of Education*. Washington: DC, National Academy Press.

National Center on Response to Intervention (2009). *Progress monitoring tools chart.* Retrieved August 9, 2009, from www.rti4success.org

National Institute of Child Health and Human Development (2000). *Report of the National Reading Panel. Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction: Reports of the subgroups* (NIH Publication No. 00-4754). Washington, DC: U. S. Government Printing Office.

National Joint Committee on Learning Disabilities. (2005). *Responsiveness to intervention and learning disabilities.* Retrieved May 13, 2009, from http://www.ldonline.com

Nichols, S. L., & Berliner, D. C. (2007). *Collateral damage: How high-stakes testing corrupts America's schools.* Cambridge: Harvard Education, 2007.

No Child Left Behind Act of 2001, PL 107-110, 115 Stat. 1425, 20 U.S.C. §§ 6301 et seq.

Nolet, V., & McLaughlin, M. (1997). Using CBM to explore a consequential basis for the validity of a statewide performance assessment. *Diagnostique, 22,* 146-163.

North Carolina Department of Instruction, Division of Accountability Services, Testing Section. (2009). Retrieved May 14, 2009 from www.ncpublicschools.org/accountability/testing

Paris, S. G., & Hoffman, J. V. (2004). Reading assessments in kindergarten through third grade: Findings from the center for the improvement of early reading achievement. *The Elementary School Journal, 105,* 199-217.

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

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

Pennsylvania Department of Education. (2005). Assessment. Retrieved June, 24, 2009, from http://www.pde.state.pa.us

Pikulski, J. J., & Chard, D. J. (2005). Fluency: Bridge between decoding and reading comprehension. *The Reading Teacher, 58,* 510-519.

Pinnell, G. S., Pikulski, J. J., Wixson, K. K., Campbell, J. R., Gough, P. B., & Beatty, A. S. (1995). *Listening to children read aloud: Data from NAEP's integrated reading performance record (IRPR) at grade 4* (NAEP Report No. 23-FR-04). Washington, DC: National Center for Education Statistics.

Powell, W. R. (1971). The validity of the I.R.I. reading levels. *Elementary English, 48,* 637-642.

Pressley, M. (2006). *Reading instruction that works: The case for balanced teaching.* (3rd ed.). New York: Guilford Press.

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

Roberts, G., Good, R., & Corcoran, S. (2005). Story retell: A fluency-based indicator of reading comprehension. *School Psychology Quarterly, 20,* 304-317.

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

Rouse, H. L., & Fantuzzo, J. W. (2006). Validity of the dynamic indicators for basic early literacy skills as an indicator of early literacy for urban kindergarten children. *School Psychology Review, 35,* 341-355.

Samuels, S. J. (1976). Automatic decoding and reading comprehension. *Language Arts, 53,* 323-325.

Samuels, S. J. (1979). The method of repeated readings. *Reading Teacher, 32,* 403-408.

Samuels, S. J. (1987). Information processing abilities and reading. *Journal of Learning Disabilities, 20,* 18-22.

Samuels, S. J. (1994). *Toward a theory of automatic information processing in reading, revisited.* In R. B. Ruddell, & N. J. Unrau (Eds.). Theoretical Models and Processes of Reading (pp. 1127-1148). Newark, DE: International Reading Association.

Samuels, S. J. (1997). The method of repeated readings. *Reading Teacher, 50,* 376-381.

Schatschneider, C., Wagner, R. K., & Crawford, E. C. (2008). The importance of measuring growth in response to intervention models: Testing a core assumption. *Learning and Individual Differences, 18,* 308-315.

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

Scott-Foresman Systems, Revised. (1976). *Unlimited Series.* Glenview, IL: Scott Foresman & Co.

Shapiro, E. S., Keller, M. A., Lutz, G., Santoro, L. E., & Hintze, J. M. (2006). Curriculum-based measures and performance on state assessment and standardized tests: Reading and math performance in Pennsylvania. *Journal of Psychoeducational Assessment, 24,* 19-35.

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

Shaw, R., & Shaw, D. (2002). *DIBELS oral reading fluency-based indicators of third grade reading skills for Colorado State Assessment Program (CSAP)* (Technical Report). Eugene: University of Oregon.

Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational Researcher, 29,* 4-14.

Shepard, L. A. (2006). *Classroom Assessment*. In R. L. Brennan (Ed.). Educational Measurement (pp. 627-646). Portsmouth, NH: Greenwood Publishing Group, Inc.

Shinn, M. R. (1989). *Identifying and defining academic problems: CBM screening and eligibility procedures.* In M.R. Shinn (Ed.). Curriculum-based measurement: Assessing special children (pp. 90-129) New York: Guilford Press.

Shinn, M. R., & Bamonto, S. (1998). Advanced applications of curriculum-based measurement: "Big ideas" and avoiding confusion. In M. R. Shinn, (Ed.), *Advanced applications of curriculum-based measurement* (p. 1-31). New York: Guilford Press.

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

Sibley, D. Biwer, D., & Hesch, A. (2001, April). *Establishing curriculum-based measurement oral reading fluency performance standards to predict success on local and state tests of reading achievement.* Paper presented at the Annual Meeting of the National Association of School Psychologists. Washington, DC.

Sideridis, G. D., Morgan, P. L., Botsas, G., Padeliadu, S., & Fuchs, D. (2006). Predicting LD on the basis of motivation, metacognition, and psychopathology: An ROC analysis. *Journal of Learning Disabilities, 39,* 215 – 229.

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

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.

Simon, S. (1999). ROC curve. Children's Mercy Hospitals and Clinics. Retrieved December 20, 2009, from http://www.childrens-mercy.org/stats/ask/roc.asp

Snow, C. E., Burns, M., & Griffin, P. (1998). *Preventing reading difficulties in young children.* Washington, DC: National Academy of Sciences.

Speece, D. (2007). *How progress monitoring assists decision making in a response to instruction framework.* National Center on Student Progress Monitoring. Downloaded from www.rti4success.org on June 3, 2009.

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

Stanovich, K. E. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Reading Research Quarterly, 21,* 360-407.

Stiggins, R. (2002). Assessment crisis: The absence of assessment for learning. *Phi Delta Kappan, 83,* 758-765.

Strecker, S. K., Roser, N. L., & Martinez, M. G. (1998). Toward understanding oral reading fluency. *National Reading Conference Yearbook, 47,* 295-310.

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

Swets, J. A. (1996). *Signal detection theory and ROC analysis in psychology and diagnostics collected papers.* Mahwah, NJ: Lawrence Erlbaum Associates.

Swets, J. A., Dawes, R. M., & Monahan, J. (2000). Better decisions through science. *Scientific American, 283*, 82-87.

Taras, M. (2007). Assessment for learning: Understanding theory to improve practice. *Journal of Further and Higher Education, 31*, 363-371.

Tilly, D. (2008). The evolution of school psychology to science based practice. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology V* (pp. 17-36). Bethesda, MD: National Association of School Psychologists.

Tindal, G., Fuchs, L. S., Fuchs, D., Shinn, M. R., Deno, S. L., & Germann, G. (1985). Empirical validation of criterion-referenced tests. *Journal of Educational Research, 78,* 203-209.

Tindal, G. (1989). Evaluating the effectiveness of educational programs at the systems level using curriculum-based measurement. In M. R. Shinn (Ed.), *Curriculum-based measurement: Assessing special children* (pp. 202-238). New York: Guilford Press.

Tindal, G., & Marston, D. (1996). Technical adequacy of alternative reading measures as performance assessments. *Exceptionality, 6,* 201-230.

Tracey, D. H., & Morrow, L. M. (2006). *Lenses on reading: An introduction to theories and models.* New York: Guilford Press

U.S. Department of Education. (1983). *A nation at risk: The imperative for school reform*. Washington, DC: Author, Commission on Excellence in Education.

U. S. Department of Education (2002, October). *Strategies for making adequate yearly progress using curriculum-based measurement.* Paper presented at the Student Achievement and School Accountability Conference. Retrieved on May 14, 2009, from http://www.ed.gov/admins/lead/account/aypstr/edlite-slide010.html

Vander Meer, C. D., Lentz, F. E., & Stollar, S. (2005). *The relationship between oral reading fluency and Ohio proficiency testing in reading* (Technical Report). Eugene: University of Oregon.

Wagner, R. K., Torgesen, J. K., & Rashotte, C. A. (1999). *Comprehensive test of phonological processing.* Austin, TX: ProEd.

Wayman, J.C., Midgley, S., & Stringfield, S. (2006). Leadership for data-based decision making: Collaborative educator teams. Paper presented at the 2006 Annual Meeting of the American Educational Research Association. Retrieved on January 20, 2010, from http://edadmin.edb.utexas.edu/datause/papers/Wayman-Midgley-Stringfield-AERA2006.pdf

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

Wechsler, D. (2003). *Wechsler intelligence scale for children* (fourth ed.). San Antonio, TX: The Psychological Corporation.

Whalley, K., & Hansen, J. (2006). The role of prosodic sensitivity in children's reading development. *Journal of Research in Reading, 29,* 288-303.

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.

William, D., & Black, P. (1996). Meanings and consequences: A basis for distinguishing formative and summative functions of assessment. *British Educational Research Journal, 22,* 537-548.

Wilson, J. (2005). *The relation of dynamic indicators of basic early literacy skills (DIBELS) oral reading fluency to performance on Arizona instrument to measure standards.* Tempe, AZ: Tempe School District No. 3.

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

Wood, D. E. (2009). Modeling the relationships between cognitive and reading measures in third and fourth grade children. *Journal of Psychoeducational Assessment, 27,* 96-112.

Woodcock, R. W. (1973). *Woodcock reading mastery tests.* Circle Pines, MN: American Guidance Service.

Woodock, R. W. (1987). *Woodcock reading mastery test – revised.* Circle Pines, MN: American Guidance Service.

Woodcock, R. W., & Johnson, M. B. (1989). *Woodcock-Johnson Psychoeducational Battery-Revised*. Allen, TX: DLM Teaching Resources.

Woodcock, R. W., Mather, N. & Schrank, F. A. (2004). *Woodcock-Johnson III Diagnostic Reading Battery.* Rolling Meadows, IL: Riverside Publishing Company.

Woodcock, R. W., McGrew, K. S., & Manther, N. (2001). *Woodcock-Johnson III Test of Achievement*. Itasca, IL; Riverside Publishing.

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

Zimmerman, B. J., & Dibenedetto, M. K. (2008). Mastery learning and assessment: Implications for students and teachers in an era of high-stakes testing. *Psychology in the Schools, 45,* 206-216.

Zweig, M. H., & Campbell G. (1993). Receiver-operating characteristic (ROC) plots: A fundamental evaluation tool in clinical medicine. *Clinical Chemistry, 39,* 561-577.