2008

The reliability and validity of screening measures in reading

James Albert Van Hook
Louisiana State University and Agricultural and Mechanical College, jimvh3@aol.com

THE RELIABILITY AND VALIDITY OF SCREENING MEASURES IN READING

A Dissertation

Submitted to the Graduate Faculty of the
Louisiana State University and
Agricultural and Mechanical College
in partial fulfillment of the
Requirements for the degree of
Doctor of Philosophy

in

The Department of Psychology

By
James A. Van Hook, III
B.A. Centenary College, Shreveport, LA., 1993
M.S. Northwestern State University, Natchitoches, LA 1996
May, 2008

DEDICATION

This dissertation was completed in memory of my Grandfather, James A. Van Hook, and is dedicated to my two children, James and Olivia.

ACKNOWLEDGMENTS

The completion of this work has been possible only through the support and encouragement of many people, who are now due my thanks.

A warm thanks to Joe Witt for giving me the opportunity for doctoral study and for his guidance as a researcher. I also thank my wife Catherine for her ever-present love and support.

I wish to also extend thanks to my family and friends, particularly my parents, in-laws and my brother. Finally, sincere thanks to all of those that have provided me friendship and professional supervision during this long journey to this last level in the study of psychology, especially: Tom Staats, Web Sentell, Ray Adomaitis, Catherine Hansen, Miyo Chun, Jim Connell and the late Sandy Saulter.

I wish to acknowledge Amanda Vanderheyden for her insightful comments and helpful suggestions for this work. Finally, thanks to Chisato Komatsu, Scott Johnson, and Georgene Johnson, all of whom were fundamental players in the data collection for this project.

TABLE OF CONTENTS

ABSTRACT

National educational groups have recommended the use of universal screening to assist in the early identification of reading problems. One of the most widely used measures used for the universal screening of reading is oral reading fluency (ORF) (Fewster & Macmillan, 2002). However, ORF is somewhat time consuming to administer and has been reported to lack "face validity" with teachers (Fuchs, Fuchs & Maxwell, 1988). The purpose of this study was to investigate maze and other group-administered reading assessments because of their potential as a time efficient assessment that is as psychometrically valid as ORF. In this study, maze and a variation of maze known as sentence maze, both group-administered measures of basic reading performance and comprehension, were studied. A third assessment, picture word fluency, which measures a combination of site word reading and simple vocabulary, was also evaluated. The study consisted of two experiments. In the first experiment, these assessments were evaluated based on their psychometric adequacy, as well as their utility and accuracy for decision-making in the context of the requirements for universal screening. The purpose of the second experiment was to examine the generality of the results to another state with different criterion measures. A total of 789 regular education first, third and fifth grade students in two states participated in the two experiments. Students were administered CBM assessments and a criterion achievement measure. Two groups of validity analyses were reported: (a) those pertaining to concurrent/predictive validity, and (b) those pertaining to classification accuracy. These analyses revealed validity estimates for the two maze assessments similar to those shown in previous research studies. Similarly, the validity analyses for picture word fluency were also promising. Most germane to the evaluation of the screening measures was the classification accuracy analyses. Although the results were somewhat variable by grade, the results indicated that there was a moderate to high degree of concordance between those students identified as at risk by the group-administered CBM measures and the criterion measures used in this study, including ORF and the state accountability tests. The limitations of the study are discussed with suggestions for future research.

CHAPTER 1. INTRODUCTION

The number of children experiencing problems learning to read has reached staggering proportions in the United States and has led to an increase in those identified as having reading disabilities. (Rithchey and Speece, 2004). In 2000, 3.9 million children, or 8% of those enrolled in public elementary and secondary schools, were classified as disabled under the Individuals with Disabilities Education Act (IDEA). Of those classified, the largest group consisted of 2.8 million children classified as learning disabled (LD), followed by lesser numbers of children classified as mentally retarded and emotionally disturbed (U.S Department of Education 2005). In 2002, the United States Department of Education reported that the number of students receiving services for LD diagnoses increased by almost 300% from 1976-77 to 2001-2002 (U.S Department of Education 2002). MacMillan, Gresham, Siperstein & Bocian (1996) suggested that the increase in LD diagnoses has resulted in a disability "epidemic" in our pubic schools. This marked increase in the number of students identified as LD in the school system continues to be a major concern, because the general educational outcomes of LD students are poor, especially in the areas of school dropout rate and subsequent employment challenges (U.S. Department of Education, 1998).

Among those designated as LD, reading is the most common area of difficulty (Lyon, 1985; Siegel, 1989; Vellutino, Scanlon, & Lyon, 2000). Importantly, however, a consensus report by the National Reading Panel (2000) indicated that many potential reading problems may be prevented with early identification and intervention. Schools that have incorporated early identification and intervention measures have shown a substantial reduction in the number of students that might otherwise be found eligible for LD services (Torgesen, 2000). Because of the success of such measures in preventing and eliminating LD determinations, screening in reading has become relatively commonplace in U.S. schools.

Fuchs & Fuchs (1998) indicated that most schools use the curriculum-based measurement of oral reading fluency (ORF) in their screening procedures. ORF consists of the individual assessment of each child using three one-minute assessments. Although ORF is a relatively efficient assessment measure, it

can require up to five minutes per student to administer, including transition time. Many screening measures, such as the Dynamic Indicators of Basic Early Literacy Skills (DIBELS), the most common form of screening, are administered by classroom teachers. Because of their significant administration time, these screening measures sometimes take valuable instruction time away from the students, thereby lessening the valuable role of teacher instruction in preventing reading problems. If the administration time for screening could be reduced without affecting the accuracy of the screening process, the overall prevalence of screening could be increased, especially in low capacity schools, thereby maximizing prevention efforts while minimizing the overhead associated with screening administration.

The purpose of this study was to examine the utility and efficacy of alternative screening methods, which require less administration time. The literature review that follows will begin with a discussion of the problem that screening is designed to ameliorate: the over-identification of students as LD. The first section of the literature view will address the possible reasons for the surge in students who have been identified as LD and emerging efforts to assist schools in the prevention of erroneous LD determinations. The second section of the literature review will discuss federal recommendations for screening for the identification and early intervention of reading problems. The third section of the literature review will discuss the status of screening in identifying reading deficiencies, the different methods for identifying students with reading problems, and the continued development of screening measures. Following the literature review, the rationale of this study—the continued investigation of alternative techniques as more efficient means of screening in reading—will be presented. The final section will introduce the alternative techniques under study for this purpose.

1.1 Reasons Cited for the Rise in LD

Experts frequently discuss reasons for the notable rise in students identified as LD. Many allude to problems in eligibility determination, edumetric classification, bias, and instruction as major factors. (Carlberg & Kavale, 1980; Donovan & Cross, 2002). Other specific reasons cited for the increase include inconsistencies in decision-making and referral practices, and problems with early classroom instruction

that leave children ill equipped to handle educational challenges. (Gresham, 2002; Lyon, Fetcher, Shaywitz, Torgenson, Wood, Schulte, & Olson 2002; MacMillan, Gresham, Bocian & Siperstien, 1997; MacMillan, Siperstien & Gresham 1996). What follows is a discussion of the proposed reasons for the dramatic increase in the students identified as LD and the recognized shortcomings of the policies and procedures that have provided researchers an impetus for change.

1.1.1 Inconsistent Local Decision-Making Practices

The referral and classification procedures used in public schools to diagnose learning disabilities are often inconsistent because they are highly reliant on teacher referral. Specifically, general education teachers typically apply local or classroom norms to judge whether a child's academic performance is below normal. This inclination is referred in the literature as "teachers as imperfect test" (Gerber & Semmel, 1984; Gresham, MacMillan, & Bocian, 1997; Gresham, Reschly, & Carey, 1987). Bocian et al. (1999) suggested that when, in the teacher's judgment, there a gap between a student's academic performances and the class' academic performance that cannot be closed using extant classroom resources, referral for psychoeducational evaluation is probable. After the psychoeducational referral is made, testing and subsequent eligibility for special education services is likely (Algozzine, Yesseldyke, and Christenson, 1983). Thus, the preliminary diagnostic measure—the teacher's own subjective judgment—often is the key factor in determining the likelihood or propriety of special education services.

Gresham and colleagues (Gresham, VanDerHeyden & Witt, 2005) suggested that the increase in students qualifying as LD is the product of inconsistent decision-making strategies that place undue emphasis on the degree (or perceived amount) of disability rather than differences in kind (or perceived type of) disability. In doing so, disability evaluators may inadvertently place greater emphasis on the extent of the weakness of the academic performance instead of the reasons why the academic performance is so weak. The tendency to focus on degree of the disability rather than the kind of disability is not surprising, given that many traditional measures used in the schools for identifying leaning challenges, such as discrepancy indicators, do not measure "exclusionary" variables such as

environmental, cultural, and economic factors that may also account for apparent low achievement (Lyon, Fetcher, Shaywitz, Torgenson, Wood, Schulte, & Olson 2002).

This over-reliance on discrepancy indicators constitutes what Shaywitz and associates refer to as the "wait-to-fail" model. This occurs in practice when school officials ignore the effects of these exclusionary variables on a student's performance until the student gets far enough behind to justify intervention. At that point, however, the student's failure is likely. The "wait-to-fail" model has profound consequences for younger students of reading because years may pass while school officials wait before a discrepancy is realized. In the meantime, the student has missed valuable opportunities for remediation, which could have bettered the student's reading performance had they been initiated promptly (Torgeson, 2000). When such difficulties persist unabated, the outcome is generally poor. Mounting evidence suggests that if a student fails to develop reading fluency by the third grade, he will likely continue to struggle with reading into adulthood (Shaywitz, Fletcher, Holohan, Schneider, Marchione, Stuebing, Francis and Shaywitz, 1999).

School practices in classifying students as LD using objective appraisal methods are also inconsistent. This may be because school officials, while permitted to consider all relevant factors in referring a student for an LD determination, may be overly influenced by teacher referral. The influence of teacher referral for LD classification frequently means that testing is inevitable (Algozzine, Yesseldyke, and Christenson, 1983). It is at this point, according to Gresham et al. (2005), that the LD label becomes misapplied because school officials' decisions are based on *degree* (significantly low achievement) rather than *kind* of low achievement (low achievers versus disabled learners). With regard to the latter, Gresham indicated that the research that contrasts these two groups has not been able to make a reliable distinction, which will result in low achievers being falsely identified as disabled learners.

1.1.2 Inadequate State and Federal Eligibility Practices

State and federal regulations for special education identification complicate the determination of a legitimate LD label because these regulations do not adequately address factors that may often better

explain low achievement, such as bias, ineffective attempts at remediation and/or poor instruction (Gresham, VanDerHeyden, and Witt, 2005). For example, with the reauthorization of IDEA in November 2004, the definition of learning disabilities became codified as a condition noticeable by either the presence of "a severe discrepancy between achievement and intellectual ability in oral expression, listening comprehension, reading recognition . . . [or a] response to a scientific research based intervention" (United States Department of Education, 2002). The definition of LD in the IDEA, because it is stated in the disjunctive, is ambiguous at best.

Further, although the identification methods of LD students under state and federal regulations appear to be objective, these regulations often introduce a subjective component as well (Macmillan and Siperstein, 2002). For example, state and federal identification methods are often described with objective "measurement bound" characteristics, including specific cut off scores, required discrepancies, and other objective psychometric profiles. However, the evaluation teams making the determinations are permitted to consider other relevant information, such as local/classroom norms, national norms, and sociocultural factors, all of which involve elements of subjectivity. Consideration of these other, more subjective factors vitiates reliable decision-making (Bocian et al., 1999; MacMillan &Siperstein, 2002).

1.1.3 Inadequate Edumetric Approaches

Many school districts employ the use of norm-based achievement tests at the conclusion of the school year to hold schools accountable for educating students. These norm-based tests, such as the Iowa Test of Basic Skills (IOWA) and the Louisiana Educational Assessment Program (LEAP 21) are frequently used as achievement criterion measures. A student's performance on these measures may be used to help determine whether the student's academic achievement is deficient. However, while these measures may provide some picture of the student's current level of academic functioning, the use of these tests for instructional planning, LD determination, and/or remediation of students' skill deficits is problematic. This is because these tests represent a "one-shot" opportunity for a student to show

competence, Any inference that a student is LD from this "one-shot" opportunity would be a vague inference at best.

Others studying the increase in children classified as LD have voiced concerns regarding the unreliability of standardized assessment instruments to discriminate learning disabled students from low achieving students (Algozzine & Yesseldyke, 1983; Yesseldyke, Algozzine, Shinn & McGue, 1982). The primary problem with standardized tests is that a "low" score may mean that the student either cannot learn or has not learned. Further, the tests are incapable of discriminating between students who have not learned because of poor instruction and those who have failed to learn despite excellent instruction. Most often, diagnostic professionals will infer that a low score means the student cannot learn without considering the possibility that observed low achievement is an instructional problem.

Gresham and Witt (1997) criticized standard assessment batteries for their lack of treatment utility and social validity because these standard assessment approaches are based on a psychometrically-exclusionary approach—relying on standardized scores for diagnostic purposes only. Standardized approaches, according to Gresham and Witt, have failed to adequately differentiate between impaired learners whose difficulty is due to cognitive deficits and those whose difficulties are due to inadequate instruction and/or preliterary experiences. The standard assessment approaches also fail to provide information that is beneficial in remediating deficits and training necessary skills. The use of inadequate edumetric approaches raises multiple ethical and potentially legal concerns regarding the practice of identifying and placing students in programs that do not improve student outcomes (e.g., Marshall v Georgia, 1984). For example, in the case of Marshall v. Georgia (1984), a case brought in federal district court in Georgia, the plaintiffs sued the state, claiming that minorities were overrepresented in special education. The court however, rejected the racial challenge to the state's classification procedures because the state's classification methods were accurate and aided the state in providing necessary educational services.

1.1.4 Instruction

Recently, attention has shifted away from eligibility practices and toward the variables that may impact whether a student is referred and tested in the first place (Maheady, Algozine, & Yesseldyke, 1984). Yesseldyke & Thurlow (1984), citing lack of appropriate instruction in the regular education setting as a common denominator, asserted that many students are inappropriately identified as LD. In relation to reading, Clay (1987) stated that the overwhelming majority of reading problems are due to inadequate or inappropriate preliterary experiences and/or reading instruction. Clay further suggested that these limited experiences are often mistaken for basic cognitive deficits, such as specific learning disabilities, during subsequent LD testing. Vaughn (2003) described these students as "instructional casualties." James Patton, Professor of Leadership and Special Education at the College of William & Mary, more colorfully referred to these children, especially minority children, as "ABT," an acronym for "ain't been taught."

If these experts are correct, then the first question in an evaluation of a student for an LD determination is whether the classroom instruction has been adequate and appropriate. Such a determination is in keeping with the "Law of Parsimony" (Whaley & Surrat, 1968), which provides that all simple, logical explanations for a problem—in this case, the lack of opportunity to read—must be ruled out before more complex explanations—in this case, a student's cognitive deficits—may be accepted. The violation of this rule by failing to take into account the level of classroom instruction inevitably results in inappropriate, biased, and inaccurate teacher referrals (Shinn, Tindal, & Spira, 1987).

1.2 Recommendations for the Problem of Overidentification

Whatever the causes for the rise in LD determinations, there is no doubt that the increase has affected the overall education of the nation's children because of the poor academic skills generally acquired in LD programs. In response to the dramatic rise in LD determinations and the poor outcomes of LD students, Congress in 2001 passed the "No Child Left Behind" Act (PL 107-110, 2001), which mandates that our educational systems be responsible for educating all students, including those in both

general and special education. With the requirements of No Child Left Behind have come federal initiatives to solve the problems associated with the increase in LD determinations

The National Research Council (2002) and the President's Commission on Excellence in Special Education (2002) studied the problem of over-identification extensively and provided recommendations. Many of their proposals focus on catching small problems early and preventing them from becoming more serious. Of these recommendations, a proposal of universal screening to assist in the early identification of reading problems is currently the center of concentration. This section will discuss universal screening and some of the techniques that are useful for this purpose.

1.2.1 Universal Screening

The proper classification of children has long been the focus of federal attention. In 2002, the President's Commission on Excellence in Special Education found that locally driven, universal (or classwide) screening of young children, in which prerequisite skills are equitably assessed classwide, is associated with better outcomes and results for all children. This finding was based on research indicating that effective and reliable screening of young children can identify those most at risk for later achievement and behavioral problems (Coyne et al 2001; Gresham 2001; Langenberge 1999), including those most likely to be referred and placed in special education programs (National Reading Council, 2002). Testimony provided to the Commission gave compelling evidence of how early intervention can prevent disabilities once they are identified. The same testimony also showed how early intervention can ameliorate the impact of disabilities in those who develop them. Most impressive to the Commission were the results of large-scale clinical trials indicating that early intervention of reading skills combined with positive behavior programs resulted in improved academic achievement and a reduction in behavioral difficulties in high-risk, predominantly minority children (Kellman et al., 1994). The Commission suggested the use of school-wide universal screening procedures for students in late kindergarten and early first grade to identify those who are at risk for reading problems. It further suggested that school officials continually

monitor these students on indicators that predict later reading difficulties (National Research Council, 2002; President's Commission on Excellence in Special Education, 2002).

In light of the benefits of early identification, various nationally recognized groups (National Research Counsel, 2002; President's Commission on Excellence in Special Education, 2002) have called for the use of procedures that can be used for the screening of all students schoolwide. The data generated by universal screening of all students would assist primarily in detecting individual problems but also in identifying problems associated with a class's core curriculum. In this regard, universal screening has expanded to the use of measurement of prerequisite skills classwide and schoolwide for purposes of comparing performances across classes and grades.

1.2.2 Screening for Reading Problems

Various practices have emerged in response to congressional mandates expanding the responsibilities of the educational system, scholarly recognition of the weaknesses inherent in traditional assessment approaches, and the Commission's call for universal techniques. However, attempts to develop an array of techniques to handle an increasing number of assessment challenges have remained imprecise (Ritchey & Speece, 2004). One notable factor in the development of universal techniques, according to Parker & Hasbrouk (1992), is the usability of assessments in terms of the time to create the measure and the ease of administration of the measure. Another notable factor is the degree to which teachers view the information provided by the technique as meaningful (Allinder & Oats, 1997; Fuchs & Fuchs, 1992).

The following sections will outline the current status of procedures that have been used for screening in reading, the implementation of curriculum-based assessment (CBA) to directly assess student achievement through existing course material, and the emergence of curriculum-based measurement (CBM), and more specifically, oral reading fluency (ORF) measurement, out of the CBA paradigm. The final section will discuss the strengths and limitations of oral reading fluency and suggest alternatives to oral reading fluency as a measure culminating with particular emphasis on maze.

CHAPTER 2. EXISTING STATUS FOR SCREENING IN READING

Ritchey and Speece (2004) recognized that, despite the advances in instructional interventions, assessment practices used to determine which students actually need intervention remain elusive. Researchers continue to work to identify assessments methods that are valid, efficient, and cost effective and that meet the Commission's goal of identifying all students who are at risk for reading failure. To accomplish this goal, researchers administer screening measures that they believe can determine risk status for reading problems. The utility of these approaches is evaluated in terms of how well the screening device is able to identify a "true" reading problem in terms of the student's outcome status (Ritchey & Speece, 2004). In doing so, researchers have traditionally evaluated techniques in terms of their ability to have high sensitivity and specificity together with low false positive and false negative error rates.

Although this has been the mission in the development of approaches to identify reading problems, perfect classification without error is unrealistic (O'Connor & Jenkins, 1999). Based on the premise that it is better to develop a technique that never misses, or only misses some of the time, researchers have strived to find some degree of balance among degrees of error. However, both the false positive error rate and the false negative error rate come with costs that create a dilemma for the researcher. The administration of an extensive battery of tests may improve accuracy—i.e., the tests never fail to identify a student with a reading problem—but these tests require a great amount of time and money that could otherwise be applied toward instruction and learning. A more abbreviated testing process may be more efficient in terms of time and money, but it may fail to correctly identify those who have legitimate deficiencies. Ritchey and Speece (2004) suggested that this latter option might be the more insidious because these errors represent children who have problems but who were not identified during the screening process. In fact, both error types must be taken into account when developing screening approaches, because the over-identification of students with reading problems stresses limited school

resources, while the under-identification of such students excludes students who need intervention (Fletcher & Satz, 1984; Torgesen & Burgess, 1998).

Research on the early identification of students who may have problems reading has examined the efficacy of using single measures versus using multiple skill batteries as predictors for reading success. Single predictors for younger children (kindergarteners) may include, for example, skills that are closely associated with conventional reading, such as letter-names and letter-sounds. Single predictors for older children (first grade and up) may include, for example, letter-sounds and word reading (Ritchey & Speece, 2004; Scarborough, 1998). Multiple skill batteries, on the other hand, are usually comprised of a range of skills that are associated with later reading achievement, for example, visual discrimination, visual motor copying, and directional orientation. Ritchey & Speece (2004), after comparing both single and multiple approaches for identifying reading problems, reported that the multiple skill batteries are the most common.

Rithchey and Speece (2004), however, suggested that although multiple skill batteries are more sensitive than single measures, they are often time consuming to administer and/or may require specialized equipment or statistical expertise for proper use. The need for specialized equipment and/or statistical expertise may limit the implementation of multiple skill batteries in the schools. Ritchey & Speece (2004) suggested that new techniques for early identification of reading problems should consider both the accuracy of the techniques in classifying students, as well as the convenience and accessibility of the instrumentation. In relation to classification accuracy, Ritchey & Speece (2004) proposed that developers of techniques used in the early identification of reading problems should insure that their methodologies are capable of examining both growth rates in the development of reading skills and the level of performance of the students. One approach that has consistently worked well to satisfy these suggestions has been curriculum-based measure (CBM) of reading, particularly curriculum-based measurement of oral reading fluency (ORF).

2.1 Curriculum-Based Measurement

Universal (classwide) screening typically incorporates school use of curriculum-based measurement (CBM), which is administered to all students (Donovan & Cross, 2002). CBM has typically been used as a means of formative evaluation for purposes of determining eligibility for special education. CBM enhances the integration of services between general education and special education by providing a common measurement process that can be applied to assess normal progress through a curriculum and to substantiate that the child's performance is discrepant from his or her peers who have had the same educational opportunities (Ardoin, Witt, Suldo, 2004). As federal legislation has encouraged the integration of special and regular education for insurance of problem identification and remediation, CBM functions as a "bridge" between disciplines. Schools using CBM may screen for students who are at risk for academic failure (Deno, 2003). CBM provides a measurement system that (1) is efficient for teachers; (2) produces accurate information that records academic performance and progress in a meaningful way; (3) provides feedback about the impact of instruction and intervention on student progress, and (4) produces information that could improve instructional planning (Deno, 1985; Deno, Fuchs, Martson, & Shinn, 2001).

2.2 Oral Reading Fluency as a Curriculum-Based Measure

CBM screening in reading (CBM-R) traditionally measures a student's oral reading fluency (ORF). This method arose from the Precision Teaching Movement, which originated at the University of Kansas and the University of Washington; ORF later gained popularity as a curriculum-based measurement (CBM) method at the University of Minnesota. CBM-ORF has historically stood alone as being the leader for the direct assessment of reading. School officials gathering ORF data benefit from a useful appraisal method for the assessment of a student's general reading ability, including their comprehension skills (Fuchs, Fuchs, and Maxwell, 1988).

The research surrounding ORF is extensive and validates the use of ORF for assessment of basic reading skill development and reading comprehension. Fuchs, Fuchs & Maxwell (1988) showed that

ORF possesses strong validity as a measure of reading. The authors examined the correlations of non-commercial measures of reading (e.g., simple question answering, recall procedures, cloze techniques and oral passage reading) with commercially available norm-referenced reading measures, such as Word Study Skills and Reading Comprehension subtests of the Stanford Achievement Test. Their results showed that ORF surpassed all of the measures sampled in terms of criterion-related validity. Furthermore, the average number of words read correctly per minute correlated with each of the Stanford Achievement Tests subtest at an average of .89. The authors showed that ORF correlations surpassed the correlations of the other non-commercial comprehension measures and the Stanford Reading Comprehension subtest.

In a follow up study, Jenkins & Jewell (1993) examined the relationship among non-commercial measures (including ORF), numerous commercially available norm-referenced reading measures, and teacher judgment. The authors described the correlations between ORF and the commercial measures as "quite impressive." More so, they found that teacher judgment correlated strongest with ORF than any of the other measures. Jenkins & Jewell concluded that:

> These results are indeed striking given the *level of importance* of commercially available norm-referenced tests of achievement to the special education eligibility decision-making process. (Emphasis added)

Today, ORF continues to be a respected measure of general reading performance for classwide screening and intervention progress monitoring (Fewster & Macmillan, 2002; Hintz, Conte, Shapiro & Basile, 1997; Shinn, Good, Knutson, Tilly, & Collins, 1992). Although more time efficient than many other commercial and non-commercial options, ORF must nonetheless be administered individually. Thus, one disadvantage of using ORF as a classwide assessment is the considerable time required to collect data on the students individually (Wesson, Fuchs, Tindal, Mirkin, & Deno, 1986). Another disadvantage of using ORF as a screening and intervention measure appears to be its lack of face validity as an index of reading comprehension, especially with teachers (Fuchs, Fuchs & Maxwell, 1988). This is

because oral fluency is not widely accepted as a proxy for text understanding, despite a strong correlation between the two (Fuchs, Fuchs, & Maxwell, 1988; Yesseldyke, 1979).

2.3 Prospective Alternatives for Screening in Reading

In light of the perceived disadvantages of ORF, researchers have pursued alternative reading measures (Fuchs & Fuchs, 1992). To gain teacher acceptance for routine use, curriculum-based reading assessments should be easy to produce, administer, and score (Ritchey & Speece, 2004). Secondly, curriculum-based reading assessments must also measure what they claim to measure (Fuchs, Fuchs & Maxwell, 1988). Fuchs and Fuchs (1992), after reviewing previous research, identified four reading measures—recall procedures, cloze techniques, and maze procedures—that appeared potentially useful in compensating for the limitations of ORF. Each of these methods will be briefly reviewed below.

- Recall Procedures. Recall procedures are constructed by taking a sample of suitable reading material from the students' curriculum. Students are typically required to read the selection and then recreate it using their own words (Fuchs et al., 1988). Although recalls are undemanding in their preparation and administration, scoring can be time consuming, difficult, and technically unsound (Fuchs & Fuchs, 1992). Many of the scoring methods are so burdensome that they are infrequently used, thus making it difficult to determine their reliability. (Parker, Tindal, & Hasbrouk, 1989). Recall correlations reported by Fuchs et al. (1998) with Reading Comprehension on the Stanford Achievement Test have ranged from .59 to .79 (Fuchs & Fuchs, 1992).

- Cloze Techniques. Cloze was first introduced in the 1950's as a reliable and valid measure of text comprehension in a teaching context (Bormuth, 1968; Jongsma, 1971). When constructing cloze, the first sentence of the passage is left intact but thereafter every nth word is omitted and replaced with a blank (Fuchs & Fuchs, 1992). Students are required to fill in the blanks with meaningful words. Unlike recall, cloze may be administered either individually or in groups. Cloze probes, however, may be difficult to create, because cloze procedure requires reproduction of the passage with deletions.

According to Parker and Hasbrouk (1992), cloze has a moderate relationship with other reading tests but has disadvantages that limit its usefulness in reading research. These disadvantages include dependency on writing skills and time consumption. Cloze has also been characterized as too frustrating for low achievers (Parker, Tindal, & Hasbrouk, 1989). In 2004, DuBay reported that cloze showed adequate validity but that it was suitable for use with intermediated to advanced readers only. Reliability estimates as reported by Spear-Swerling (2004), were in the high .90's, with validity coefficients with listening and comprehension tests in the .70's. As for the use of close in measuring reading progress, Fuchs et al, (1992) opined that cloze lacks the needed technical integrity to make it a feasible and accurate measure of student growth.

- Maze Procedures. Maze, a variation of cloze, is a classroom-based measure developed in the early 1970's. It was originally used with students who were culturally disadvantaged, who were learning English as a second language, or who were demonstrating signs of reading disabilities (Parker & Hasbrouck, 1992). With maze, the first sentence of a passage remains intact. Thereafter, the student reads sentences in which every nth word has been selectively omitted. The student is asked to select, in a multiple-choice format, the one word which best completes the sentence. A student's maze score comprises the total correct words marked. Maze has been shown to have sound reliability with internal consistency estimates of .85-.87 (Bruning, 1985; Cranney, 1972; Gutherie, 1973) and test-retest estimates of .79-.91 (Kingston and Weaver, 1970).

Maze as an approach for reading assessment has been criticized for being frustrating, excessively difficult and of limited usefulness (Pikulski and Pikulski, 1977). Jenkins and Jewell (1993), however, reported validity coefficients between .65 and .76 with standardized reading test. Further, Fuchs and Fuchs (1992), based on investigation of the above-mentioned screening procedures, found maze to be the best alternative to oral reading fluency for classwide screening. Their findings showed strong criterion validity for maze and technical features similar to those of oral reading fluency. They also showed that maze was more efficient in terms of administration and scoring than the other alternatives.

CHAPTER 3. RATIONALE OF THE CURRENT STUDY

CBM-ORF, because of the factors discussed above, offers several advantages over other forms of screening in reading. However, despite the advantages surrounding the use of ORF as an index of reading proficiency, it has its drawbacks. First, the administration of ORF is time-consuming. A class of 30 students will require at least 90 minutes to administer DIBELS, excluding set up, transition, and scoring time. Wesson, King, & Deno (1984) noted that most educators recognize the benefits of using direct measurement to enhance the development of student's reading skills, but that the time and skill requirements limit their use of it. Second, ORF suffers from a lack of face validity as a measure of reading comprehension (Fuchs et al, 1988; Jenkins & Jewell, 1993; Shinn, Good, Knutson, Tilly, & Collins, 1992). Wiley and Deno (2005) discussed the problem of "word calling," in which a student reads fluently but without understanding the text. Because of the phenomenon of "word calling," that may arise during the administration of ORF, teachers do not believe that the data from ORF probes provide users with an accurate indicator of reading comprehension. Teachers view maze as an improvement over ORF in this regard because maze requires students to understand the content of the passage they are reading (Fuchs, Fuchs, Hamlett, & Furguson, 1992).

The rationale of this study is to further investigate maze and other related measures to gain better understanding of the utility of maze as a universal screening measure. What follows below is a discussion of research pertaining to maze and two additional single measures—sentence reading and picture word fluency, which could potentially serve as screening instruments in reading. A review of each measure follows with additional examination of the development and technical characteristics that underlie the utility of both approaches.

3.1 Paragraph Maze

As set forth earlier, prior research has shown maze to be an efficient and reliable measure of reading proficiency with sufficient reliability and validity (Parker & Hasbrouck, 1992). Maze has been shown to have sound reliability with internal consistency estimates in the .85-.87 range (Bruning, 1985; Cranney,

1972; Gutherie, 1973) and test-retest estimates between .79-.91 (Kingston and Weaver, 1970). Alternate forms reliability has presented a challenge because of differences in maze construction; i.e., differences in format. Alternate forms reliability coefficients, however, have ranged from .62-.92 to .86-.91 (Bradley, Ackerson, & Ames, 1978) and .70-.93 to .54-.73 (Parker, Hasbrouck, & Tindal, 1989). Shin, Deno, and Espin (2000) showed alternate forms coefficients in the .80's.

Fuchs, Fuchs, Hamlett, & Ferguson (1992), in a cross-grade concurrent validity study (under timed conditions) of maze compared to a standardized reading test found validity coefficients of .82 for number correct (marked correct) falling to .43 for percentage correct (of all choices). Jenkins and Jewell (1993), in a within grade concurrent validity study, compared maze to standardized reading tests and showed coefficients between .65 and .76.

Parker and Hasbrouck (1992) described maze as an easy and straightforward task for most students to complete but found that student achievement on maze could produce a skewed distribution of scores, with poor discrimination among the higher achieving 30% to 40% of the students. The authors suggested that further refinement of maze would make this measure more suitable for use as a general classroom tool. Fuchs, Fuchs, Hamlett & Ferguson (1992), provided a solution to this dilemma, which was to set a time limit for maze completion. Use of a time limit reduced the skewness of the distribution and raised the validity coefficients from .43 to .82.

3.2 Sentence Maze

Developed by Witt (2005), sentence maze is a variation of paragraph maze. Sentence reading is a measure of basic reading performance and comprehension. Two preliminary studies support the use of sentence reading as a curriculum-based reading measure.

In the first study of sentence reading, results indicated concurrent validity coefficients with ORF at .638 ($p<.01$). In addition, preliminary findings also showed sufficient relation of sentence reading to standardized tests with coefficients ranging from .341 to .594 for grades 3-5. In another study, concurrent

validity coefficients with ORF were .74 to .83 for grades 3 and 5. Validity coefficients with standardized reading tests ranged from .46 to .55.

3.3 Picture Word Fluency

Picture word fluency is a variation of curriculum-based measurement of word identification fluency (CBM-WIF). Developed by Witt (2005) as a measure of site word reading, picture word fluency shows potential as an alternative screening measure of reading due to its similarities to CBM-WIF and its success in two pilot field trials. What follows below is a description of CBM-WIF, a review of studies supporting CBM-WIF as a reading measure, and preliminary findings from field trials supporting use of picture word fluency as a curriculum-based reading measure.

Deno, Mirkin, & Chiang (1982) investigated word identification fluency as one of the CBM measures for reading at the University of Minnesota Institute for Research on Learning Disabilities. With CBM-WIF, students have one minute to read isolated words, which are presented in lists. These words are randomly selected from high frequency word lists. The test is scored based on the number of words read correctly. CBM-WIF tests for automatic word recognition skills, one of the hallmarks of competent reading behavior. Deno et al. (1982) examined the concurrent validity of word identification fluency with 66 children in Grades 1 to 6. The criterion measures were the reading comprehension subtest of the Peabody Individual Achievement Test and the phonetic analysis and inferential and literal reading comprehension subtests of the Stanford Diagnostic Reading Test. They found that when students read third grade word lists, correlations with these four respective measures were .76, .68, .71, and .75. When they read sixth grade words lists, respective correlations were .78, .71, .68 and .74.

In a related study, Fuchs, Fuchs, & Campton (2004), studying first graders with severe reading difficulties, contrasted the concurrent and predictive validity for two alternative CBM early reading measures, CBM-WIF and nonsense word fluency. To explore concurrent validity, the authors ran correlations between the two CBM measures and the Word Attack Subtest and the Word Identification Subtest of the Woodcock Reading Mastery Test-Revised (Woodcock, 1987). Their findings showed that

the CBM-WIF was statistically significantly higher than nonsense word fluency (.77 vs. .58) but comparable with the Woodcock Word Attack Subtest (.59 for CBM-WIF and .50 for nonsense word fluency). Later, Fuchs, Fuchs & Hamlett (1989) included a greater variety of criterion measures for reading fluency and comprehension with inclusion of the Comprehensive Reading Assessment Battery (CRAB). The authors found that across Woodcock Word Identification, CRAB fluency and CRAB comprehension, correlation coefficients ranged from .73 to .93 for word identification fluency and from .51 to .80 for nonsense word fluency.

Fuchs, Fuchs, & Campton (2004) showed that CBM-WIF outperformed CBM nonsense word fluency for predicting performance on the CRAB fluency and comprehension criterion measures. Consequently, for identifying first graders at risk for poor end-of-year reading outcomes in October, the CBM-WIF measure showed superior predictive strength compared to the CBM nonsense word fluency task. Specifically, predictive coefficients with CBM-WIF with CRAB fluency and comprehension was .80 to .66, respectively whereas CBM nonsense word fluency was .64 to .50. The authors concluded that this study demonstrates the superiority of CBM-WIF over CBM nonsense word fluency

Picture word fluency, like CBM-WIF, is a measure of sight word recognition. It differs from CBM-WIF in that is uses pictures to prompt word recognition and require basic vocabulary skills. Preliminary data from two field trials support picture word fluency as an alternative to CBM-WIF as a reading measure. In one study comparing performance on picture word fluency to performance on ORF in grades 1-5 analysis indicated concurrent validity coefficients ranging from .54 to .77 ($p<.01$) with the Iowa Test of Basic Skills. In addition, picture word fluency had promising coefficients with standardized tests for grades three through five ($p<.01$). In the second study, concurrent validity coefficients were .73 to .76 for grades three and five with subsequent correlation to the Mississippi Curriculum Test, a high stakes measure of reading proficiency, at .56 to .46, respectively for the two grades.

CHAPTER 4. PURPOSE

The foregoing review has provided a context and support that universal screening is an important tool in a process of improving achievement and in reducing special education referrals. At present, ORF is the most common means of screening in reading. ORF involves a one-to-one interaction with an adult, and because of this, there is a need to examine alternative methods for screening which may accomplish the same purpose more efficiently. Group-administered CBM assessment in the areas of writing and math allow for a time- and cost-efficient measure of a student's proficiency in these areas. Although ORF has historically been the standard for identifying students with reading deficits, administration is time-consuming (Wesson, King, & Deno, 1984). Additionally, ORF may lack face validity with teachers as a reading comprehension measure. A potentially more efficient alternative for identifying students with reading problems is maze. Maze possesses similar technical characteristics as ORF, but because administration is class-wide, maze may significantly reduce the time required to administer the assessment measure. The present study will evaluate both paragraph maze and sentence maze as alternative universal screening measures. A second group-administered reading assessment, picture word fluency, a derivative of CBM-WIF, which has received recent research attention, will also be evaluated because of its initial promise as an accurate yet efficient tool for screening.

One purpose of this study is to evaluate the technical adequacy of group-administered CBM screening assessments to assist in the identification of students who may need intervention services. A second purpose of this study is to explore the predictive utility of group-administered CBM probes for their ability to estimate performance of students on state tests. Assessments (independent variables) will include paragraph maze, sentence maze, and picture word fluency as alternative curriculum-based measures. Oral reading fluency will also be included in this study and will be used as both an independent variable and a criterion variable across various analyses. Criterion variables will include the Group Reading Assessment and Diagnostic Evaluation (GRADE), Integrated Louisiana Educational Assessment Program (*i*LEAP), and the Mississippi Curriculum Test (MCT).

CHAPTER 5. EXPERIMENT 1

Experiment 1 was designed to accomplish several goals. First, this study explored the psychometric characteristics of various measures by examining the concurrent and predictive validity of oral reading fluency, paragraph maze, sentence maze, and word identification fluency. The author then compared these variables of study to established criterion measures. Reliability estimates examined the internal consistency of test items, whereas concurrent, predictive, and classification validity coefficients examined the relationship between CBM data and the established criterion measurements. Second, this study examined the "diagnostic" accuracy of the various assessments to determine the degree of agreement across the various measures with respect to traditional metrics such as sensitivity and specificity. The author evaluated the classification accuracy of these measures to appraise the extent of correct classification of students into two groups – those "at risk" and those "not at risk" for reading problems. As a result of this analysis, appropriate cut scores were determined for screening children who may be "at risk" for reading problems. ORF served as a secondary criterion variable.

5.1 Method

5.1.1 Participants and Setting

A total of 482 students participated in the first experiment and were enrolled in first, third and fifth grade regular education classrooms in elementary schools in the Southeastern United States. Of the participants, 73% to 92% received free/reduced lunch, 42% were African American, 3% were Hispanic, 42% were Caucasian and 13% were Asian/Pacific Islander.

5.1.2 Measures

The focus of this study was on the reliability and validity of the following group-administered curriculum-based screening measures (CBM): paragraph maze, sentence maze, and picture word fluency. The foregoing CBM measures were evaluated along with the most commonly used method for screening in the elementary school, oral reading fluency (ORF). The study investigated the concurrent, predictive and classification validity of paragraph maze, sentence maze, picture word fluency, and ORF, using either

the Group Reading Assessment and Diagnostic Evaluation (GRADE) or the Mississippi Curriculum Test (MCT). Classification accuracy estimates were determined for paragraph maze, sentence maze, and picture word fluency using ORF as the criterion measure.

- Oral Reading Fluency (ORF). Oral reading fluency (ORF) is a well-researched, individually administered measure of reading performance and achievement for school aged students. ORF requires students to read a passage aloud for one minute while simultaneously being scored by an examiner. Words omitted, substituted, and hesitations of more than three seconds are scored as errors. Words self-corrected within three seconds are scored as accurate. The number of words read per minute, minus the number of errors, equals the student's score on ORF.

Numerous studies and research reviews support the validity and reliability of ORF (Elliot & Fuchs, 1997; Fewster, Macmillian, & Peter, 2002; Hintz & Conte, 1997; Shinn, 1989; Shinn & Good III, 1992). These studies also show a strong relationship between ORF and established achievement tests (Fuchs, Fuchs, Hosp, & Jenkins, 2001; Good, Simmons, & Kame'enui, 2001; McGlinchey & Hixon, 2004; Stage, Jacobsen, & Michael, 2001), and teacher's judgment of reading competence (Jenkins & Jewell, 1993).

Fuchs, Fuchs, and Maxwell (1988) concluded that ORF is a psychometrically useful method of assessing reading ability, including comprehension. Test retest and interrater reliability estimates have been at 95% to 100% (Hintze & Conte, 1997), and correlations with both oral reading measures and commercially available norm-referenced tests of reading were between .80 to .89 (Fuchs et al., 1988; Hintze & Conte, 1997).

- Paragraph Maze. Paragraph maze is a paper-and-pencil measure of basic reading performance and comprehension. Administration is individual or group. The paragraph maze requires a student to read related passages in which words have been selectively omitted. Administration of paragraph maze asks the student to select, in a multiple-choice format, the one word that best completes the sentence. Students

have three minutes to complete the paragraph maze. A student's paragraph maze score comprises the total correct words marked.

> Tomorrow is my best friend's birthday. She is going to be turning _____ (ten, red, sat) years old. My other friends and I _____ (pen, have, disk) decided to have a surprise birthday _____ (party, phone, mouse) for her. Tonight we are all _____ (orange, super, going) to meet at my house and ____ (set, car, page) up for the party.

Figure 1—Example of paragraph maze

Paragraph maze has shown to possess sound reliability with internal consistency estimates in the .85-.87 range (Bruning, 1985; Cranney, 1972; Gutherie, 1973) and test-retest estimates between .79-.91 (Kingston and Weaver, 1970). Alternate forms reliability coefficients have ranged from .62-.92 to .86-.91 (Bradley, Ackerson, & Ames, 1978) and .70-.93 to .54-.73 (Parker, Hasbrouck, & Tindal, 1989). Shin, Deno, and Espin (2000) showed alternate forms coefficients in the .80's. A validity study by Fuchs, Fuchs, Hamlett, & Ferguson (1992) of maze compared to a standardized reading test showed coefficients of .82 for number correct (marked correct) falling to .43 for percentage correct (of all choices). Jenkins and Jewell (1993), showed better concurrent validity study using percent correct, comparing maze to a standardized reading tests with coefficients between .65 and .76.

- Sentence Maze. Sentence maze is a variation of paragraph maze. Sentence maze is a paper-and-pencil measure of basic reading performance and comprehension, which is administered to students individually or in groups. Sentence maze requires a student to read unconnected sentences wherein the last word is omitted. For sentence maze, the student is asked to select, in a multiple-choice format, the one word that best completes the sentence. Students have three minutes to complete the sentence maze. A student's sentence maze score comprises the total correct words marked.

1. My school just started a football team and I plan to try-out for the ____ (went, team, same).
2. My name is Todd and today is my tenth _____ (birthday, together, same).
3. When class started, Miss Jones told us to take out our pencils and _____ (together, person, notebooks).
4. I like playing football, but practice makes me so ____ (car, tired, friend).
5. My school had a contest last week to see which class could bring the most can goods for the food _____ (drive, part, seem).

Figure 2—Example of sentence maze

Preliminary studies with sentence maze (Van Hook & Witt, unpublished) have indicated concurrent validity coefficients with ORF at .64 ($p<.01$) for grades 1-5. A rundown of these coefficients by grade shows .76 ($p<.01$) for first, .71 ($p<.01$) for second, .59 ($p<.01$) for third, .67 ($p<.01$) for fourth, and .49 ($p<.01$) for fifth. In addition, preliminary findings also show that sentence maze is adequately correlated with the Iowa Test of Basic Skills for third and fifth grade students at .47 ($p<.01$ and .34 ($p<.05$), respectively. Preliminary findings also show that sentence maze correlates adequately with the Louisiana Assessment Program for fourth grade students at .59 ($p<.01$).

- Picture Word Fluency. Picture word fluency is a measure, which requires both sight word reading skills and vocabulary. This procedure was developed using words from frequently used word lists in children's basal reading series including the Fry and Dale-Chall word lists. The student is asked to select, in a multiple-choice format, the word that best matches the picture. Administration is individual or in groups. Students have three minutes to complete picture word fluency. A student's picture word fluency score comprises the total correct words marked.

a. whispering

b. stapler

c. pineapple

d. twinkle

Figure 3—Example of picture word fluency

Preliminary studies by Van Hook and Witt (unpublished) comparing performance on picture word fluency to performance to ORF in grades 1-5 have indicated concurrent promising validity coefficients. A rundown of these coefficients by grade showed .77 ($p<.01$) for first, .63 ($p<.01$) for second, .45 ($p<.01$) for third, .45 ($p<.01$) for fourth, and .47 ($p<.01$) for fifth. In addition, picture word fluency data was shown to correlate adequately with Iowa Test of Basic Skills with third grade and fifth grade students at .37 ($p<.01$), and .51 ($p<.01$), respectively. Picture word fluency was shown to correlate adequately with Louisiana Educational Assessment Program for fourth grade students at .64 ($p<.01$).

- Group Reading Assessment and Diagnostic Evaluation (GRADE). GRADE is a group-administered, standardized, and norm-referenced achievement test published by *AGS Publishing* (Williams, 2001). GRADE is geared for assessing core reading skills, instructional planning, and progress monitoring of reading skills from kindergarten through adulthood. GRADE subtests focus on the critical early reading skills identified by the National Reading Panel. GRADE scores have demonstrated internal reliability range from .95 to .99, alternate forms reliability between .81 to .94, and test retest reliability ranging from .77 to .98. Concurrent validity coefficients between GRADE and other recognized group-administered achievement measures is sound. For example, coefficients between GRADE and the *Iowa Test of Basic Skills* Total Reading scores ranging from .69 to .83. Likewise, coefficients on GRADE and the *California Achievement Test* Total Reading scores were .82 to .87. Concurrent validity coefficients between GRADE and individually administered achievement tests are also adequate. For example, coefficients with the *Peabody Individual Achievement Test–Revised* General

Information, Reading Recognition, Reading Comprehension, and Total Reading scores with GRADE Vocabulary, Comprehension Composite, and Total Test ranged from .47 (for General Information) to .80 (for Total Reading) with a median of .74 (Williams, 2001).

- Mississippi Curriculum Test. The Mississippi Grade Level Testing Program consists of the Mississippi Curriculum Test (MCT) in Reading, Language Arts Mathematics, and the Mississippi Writing Assessment for grades 2-8. There are three forms of the MCT, which are used in each subject area in grades 2 through 8. The Mississippi Department of Education has reported Cronbach's alpha reliability coefficients across forms and grade levels ranging between 0.88 and 0.90 for Reading, 0.87 and 0.91 for Language, and 0.85 to 0.90 for Mathematics in 2001 (MSDOE, 2002). MCT performance data allows for a comparison of gains across grade levels and the tracking of individual growth patterns by providing vertically equated scaled scores using the same metric for student performance across grade levels (Tomkowicz & Schaeffer, 2002). Content validity of the MCT was addressed by statewide teacher committees, who formed a consensus about what specific skills and objectives are to be taught to particular subjects and to particular grade levels (MSDOE, 2002). For this study, MCT subtests in Reading were used to measure student achievement in grades three and five. Cronbach's alpha coefficients have been reported to range from .88 to .90 for third grade and from .89 to .90 for fifth grade. Unadjusted criterion cut points for student performance on the MCT in the third grade are 429 for "minimal/basic," 452 for "basic/proficient," and 519 for "proficient/advanced." Unadjusted criterion cut points student performance on the MCT in the fifth grade are for these criterions are 465 for "minimal/basic," 483 for "basic/proficient," and 551 for "proficient/advanced.

5.1.3 Procedure

- The curriculum-based screening probes and the criterion measures were administered within one month of each other. The author administered the GRADE and school personnel administered the MCT. The author, together with trained school personnel, administered paragraph maze, sentence maze, and picture word fluency. First grade students were administered ORF, sentence maze, picture word fluency,

and the GRADE. The administration order of sentence maze and picture word fluency was counterbalanced to control for order effects. Students in third and fifth grade were administered oral reading fluency, paragraph maze, sentence maze, picture word fluency, and the MCT. Administration order of paragraph maze, sentence maze and picture word fluency was counterbalanced to control for order effects.

5.1.4 Assessment Administration

The author administered the GRADE and school personnel administered the MCT. The author, together with trained school personnel, administered paragraph maze, sentence maze, and picture word fluency. First grade students were administered ORF, sentence maze, picture word fluency, and the GRADE. The administration order of sentence maze and picture word fluency was counterbalanced to control for order effects. Students in third and fifth grade were administered oral reading fluency, paragraph maze, sentence maze, picture word fluency, and the MCT. Administration order of paragraph maze, sentence maze and picture word fluency was counterbalanced to control for order effects.

- Oral Reading Fluency (ORF). Assessment teams administered ORF following the procedures described by Shinn (1989). In doing so, assessors asked students to read individually for one minute from grade appropriate reading probes.

- Paragraph Maze. The author, using the procedures described in Appendix A, administered one grade-level paragraph maze probe to third and fifth grade students. The procedure calls for probes to be distributed to students face down. Directions are read to students who then have three minutes to work. At the end of three minutes, students are asked to stop and hold their papers in the air. The probes are collected and scored using a scoring key.

- Sentence Maze. Using procedures identical to that of paragraph maze (see appendix B), the author administered one sentence maze probe to first, third and fifth grade students.

- Picture Word Fluency. Using procedures similar to sentence maze (see appendix C), the author administered one grade-level picture word fluency probe to first, third and fifth grade students.

- Group Reading Assessment and Diagnostic Evaluation (GRADE). Within one month of the administration of curriculum-based measures, the author administered GRADE to first grade students, according to the procedures described in the administration manual. Administration of GRADE occurred in close time proximity to district accountability testing and administration of CBM measures.
- Mississippi Curriculum Test. School personnel administered the Mississippi Curriculum Test according to the procedures prescribed by the state of Mississippi.

5.1.5 Data Collection and Scoring

- Mississippi Curriculum Test. The Mississippi Curriculum Test was scored by the test manufacturers.
- Oral Reading Fluency (ORF). Assessment teams scored ORF following the procedures described by Shinn (1989). ORF assessment involved an assessor listening to the student read and marking errors. Reading errors included mispronounced words, skipped words, word transpositions, word substitutions, words told to the student after a three-second hesitation, and skipping a whole role of words (not counted in total words read). Words not counted as errors included words read correctly, insertions, repetitions, self-corrections, and words read incorrectly due to dialectical, articulation or foreign accent issues. The number of words read per minute, minus the number of errors, equaled the student's score on ORF.
- Paragraph Maze. The author scored paragraph maze passages by tallying the number of correct choices a student made during a three-minute timed assessment. A student's paragraph maze score equaled the total correct words circled.
- Sentence Maze. The author scored sentence maze by tallying the number of correct choices a student made during a three-minute timed assessment. A student's sentence maze score equaled the total correct words circled.
- Picture Word Fluency. A student's picture word fluency score equaled the total correct words marked during a three-minute timed assessment.

- Group Reading Assessment and Diagnostic Evaluation (GRADE). The author scored GRADE protocols for first grade students, according to the procedures described in the administration manual.

5.2 Results

The purpose of Experiment 1 was to examine the extent to which group-administered CBM screening assessments provided reliable and valid measurement of reading. Descriptive statistics are provided in the first section for the CBM assessments and criterion variables in the grades studied. In the second section, organized by grade, are data pertaining to reliability and validity of the group screening measures. Specifically, split-half coefficients were used for each grade to examine internal consistency reliability. In the following section, correlation and regression analyses are presented for each grade to examine the extent to which the CBM assessments are associated with and are predictors of student performance on criterion assessments (i.e., GRADE and MCT). In the final section, data are presented on the classification accuracy of the group-administered CBM assessments.

5.2.1 Descriptive Statistics

Descriptive statistics for all variables in Experiment 1 are presented by grade in Table 1. According to Kline (1988), skewness values greater than an absolute value of three and kurtosis values greater than an absolute value of eight indicate normality problems. Absolute values of skewness and kurtosis levels in this study were acceptable (skewness less than 3 and kurtosis less than 8) for all variables, thus satisfying the assumption of normality (Kline 1988). Additionally, following the power analysis guidelines recommended in Cohen and Cohen (1983), sample size was sufficient for multiple correlation and regression analysis (Faul, Erdfelder, Lang, & Buchner, 2007).

Table 1—Descriptive Statistics of Target Variables

First Grade

Variable	Mean (SD)	Skewness	Kurtosis
GRADE	64.03 (20.140)	0.14	1.05

(table 1 continued)

ORF	41.23 (31.93)	0.36	1.45
Sentence Maze	11.73 (8.42)	0.82	0.12
Picture Word Fluency	25 (10.15)	0.21	0.51
Third Grade			
MCT	507.89 (48.71)	0.01	1.477
ORF	124.36 (34.77)	0.144	0.353
Paragraph Maze	20.06 (7.071)	0.757	1.615
Sentence Maze	29.57 (8.020)	0.123	0.579
Picture Word Fluency	46.85 (10.48)	0.712	0.498
Fifth Grade			
MCT	538.96 (50.60)	0.457	4.02
ORF	142.26 (38.56)	0.351	0.173
Paragraph Maze	24.85 (9.192)	0.178	0.169
Sentence Maze	25.62 (8.37)	0.462	1.005
Picture Word Fluency	45.93 (11.82)	0.776	0.74

5.2.2 First Grade Reliability and Validity Analyses

- Reliability. Of the 73 first grade students who participated in the study, 30 were randomly selected for inclusion in the reliability sample. Internal consistency reliability proceeded with split-half reliability analysis (Crocker & Algina 1986). For this, each response opportunity on the group-administered CBM assessments was dichotomously coded into a statistical database using "1" to indicate a correct response and "0" to indicate an incorrect response. The researcher then created two forms for each group-administered CBM measure by assigning all odd-numbered items to form 1 and all even-

numbered items to form 2. Correlation between forms was .97 for sentence maze and .99 for picture word fluency. Guttman Split-Half coefficients were respectively .98 and .99.

- Concurrent Validity. Bivariate correlation coefficients, presented in Table 2, were computed between ORF, sentence maze, picture word fluency, and GRADE to show association among these variables. Pearson correlations among all CBM measures were statistically significant ($p \leq .01$). The researcher then employed the Steiger's Z-test for correlated-correlations (Meng, 1992) to examine whether one or more of the correlations between individual CBM assessments and GRADE were significantly stronger than others. Difference tests suggested that correlations between ORF and GRADE were significantly greater than the correlation between sentence maze ($Z = 4.86$, $p < .01$) and GRADE or picture word fluency ($Z = 3.62$, $p<.01$) and GRADE. The correlation between sentence maze and GRADE was significantly less than the correlation between picture word fluency and GRADE ($Z = 1.8$, p, <.05).

Table 2—First Grade Bivariate Correlations Between Independent Variables

		Variable		
Variable	n	2	3	4
1. ORF	73	0.76**	0.69**	0.73**
2. Sentence Maze	73		0.68**	0.37**
3. Picture Word Fluency	73			0.53**
4. GRADE	73			

**$p < .01$.

- Predictive Validity. Individual regression coefficients were computed using the GRADE as the criterion measure. Regression equations showed that the percentage of variance accounted for by all CBM measures was statistically significant. ORF performance explained 53% of the variance (F (1, 71) = 81.07, $p < .01$) followed by picture word fluency, which explained 28% of the variance in GRADE

scores F (1, 71) = 27.06, $p < .01$) and sentence maze, which explained 14% of the variance in student's GRADE scores F (1, 71) = 11.3, $p < .01$).

The relative contribution of the various CBM(s) in explaining variance in GRADE scores was examined. Simultaneous multiple regressions are presented in Table 3. In Part I, a simultaneous multiple regression utilizing ORF, sentence maze and picture word fluency was computed with the goal of evaluating the additional contribution of each group-administered CBM assessments over ORF alone. In Part II, a simultaneous multiple regression with sentence maze and picture word fluency investigated which CBM assessment explained the most variance. Part I of Table 3 shows simultaneous multiple regressions comparing the relative contributions of ORF, sentence maze and picture word fluency in explaining GRADE performance. This analysis showed that all three variables together accounted for 57% of the variance but ORF uniquely accounted for 29% of the variance in student GRADE performance over and beyond sentence maze and picture word fluency. Sentence maze uniquely predicted an additional 3% of the variance but picture word fluency had little influence in explaining GRADE performance and yielded an insignificant beta weight and squared semipartial correlation coefficients. Part II of Table 3 shows a simultaneous multiple regression analysis directly comparing the relative influence of sentence maze and picture word fluency together in the prediction of GRADE performance. This analysis indicated that together both predictors accounted for 28% of the variance in student GRADE performances. In this analysis, only picture word fluency yielded statistically significant beta weights and squared semipartial correlation coefficients thus uniquely accounting 14% of the variance.

Hierarchical multiple regression coefficients, presented in Table 4, were computed to evaluate the added predictive value of the group-administered CBM compared to ORF with GRADE as the criterion measure. First, findings were examined whereby group-administered CBM scores (picture word fluency and sentence maze) alone were used to predict GRADE performance. ORF was then entered to see if it

Table 3—First Grade Simultaneous Multiple Regression Analysis Comparing the Utility of CBM Assessments in Predicting GRADE Performance

Part I	Standard β	sr^2
ORF	0.88	0.29**
Sentence maze	-0.258	0.03*
Picture word fluency	0.027	0
		R^2 = .57
		Adjusted R^2 =.55
Part II		
Sentence maze	0.03	0
Picture word fluency	0.51	0.14**
		R^2 = .28
		Adjusted R^2 =.26

*p < .05
**p<.01.

accounted for any additional variance after the group-administered CBM measures. Second, findings were examined whereby ORF scores alone were used to predict performance on GRADE. After that, the group-administered CBM assessments were entered to see if they explained any additional variance. In the first analysis, picture word fluency was entered into the equation first, followed by sentence maze. Picture word fluency accounted for 28% of the variance. The addition of sentence maze in Step 2 was nonsignificant. In Step 3, ORF added significantly to prediction with an R^2 change value of 0.29. In the second analysis, ORF alone entered at Step 1 and accounted for 53% of the variance in the students'

GRADE performances. In Step 2, the inclusion of picture word fluency did not contribute significantly to understanding the variance in GRADE. In Step 3, however, the inclusion of sentence maze did contribute significantly to understanding the variance in GRADE.

Table 4—First Grade Hierarchical Multiple Regression Analyses Comparing the Utility of Group-administered CBM Versus ORF in Predicting GRADE Performance

Predictor		*R*	*R* Square	Adjusted *R* Square	R^2 Change	*F* Change
	First Analysis					
Step 1						
	Picture Word Fluency	0.53	0.28	0.27	0.28	27.1**
Step 2						
	Sentence Maze	0.53	0.28	0.26	0	0.03
Step 3						
	ORF	0.75	0.57	0.55	0.29	46.13**
	Second Analysis					
Step 1						
	ORF	0.73	0.53	0.53	0.53	81.07**
Step 2						
	Picture Word Fluency	0.73	0.54	0.52	0.002	0.304
(table 4 continued)						
Step 3						
	Sentence Maze	0.75	0.57	0.55	0.03	5.0*

*$p<.05$
**$p<.01$

- Classification Accuracy. Individual logistic regression equations were computed to determine the usefulness of the CBM measures assessments for categorical prediction. This analysis used student GRADE performance to indicate a positive or a negative problem. A student was categorized as having a problem on the GRADE when his/her performance yielded a standard score at or below 85, which coincides with performance below the 16th percentile, indicating a significant risk status. When a student's performance yielded a standard score above 85, or above the 16th percentile, the student was categorized as negative for a problem on GRADE. Of the 73 students who entered into the logistic regression analysis, five student performances were indicated to be at risk on the GRADE. Table 4 presents classification accuracy statistics determined according to the "pass/fail" rates on the GRADE as predicted by the student's scores on the CBM assessments.

Categorical prediction was examined by three sets of analyses. First, findings were examined in terms of variance accounted for by each individual predictor. Second, findings were examined in terms of goodness-of-fit between actual student performance on the GRADE and student performance as predicted by each CBM assessment. Third, findings were examined in terms of classification accuracy with attention devoted to each CBM assessment's sensitivity, specificity, positive predictive power and negative predictive power. In the first set of logistic analyses only ORF accounted for a significant portion of the variance (92%) in student GRADE performance. Picture word fluency accounted for 10% of the variance in GRADE performance, whereas sentence maze accounted for only 0.1%. The second set of analyses utilized the Hosmer and Lameshow chi-square test showing the goodness-of-fit between actual and predicted student performance. Goodness-of-fit indices, which indicated there was little difference between the number of students who were actually scored below 85 on the GRADE and the number students who were predicted to have a problem on the GRADE. Analyses indicated Hosmer and Lameshow chi-square for ORF (χ^2 (8, N = 73) = 0.00, p > 1.00) followed by Hosmer and Lameshow chi square (χ^2 (8, N = 73) = 7.71, p= .46 for picture word fluency and Hosmer and Lameshow chi-square (χ^2 (8, N = 73) = 9.25, p = .32) for sentence maze

The third set of analyses investigated the classification accuracy of ORF, sentence maze and picture word fluency in evaluating the extent of correct categorical classification of students into two groups—those at risk and those not at risk on the GRADE. Sensitivity of each measure was calculated as the proportion of true positives and specificity was calculated as the proportion of true negatives. Positive predictive power was calculated to determine the utility of finding a CBM-positive finding as an inclusionary criterion; that is, indicating that the same student was categorized as having "a problem" on the GRADE. Positive predictive power was calculated as the probability that a problem GRADE performance indicated a validated problem given a CBM-positive finding. Negative predictive power was calculated to determine the utility of finding of a CBM-negative finding as an exclusionary criterion; i.e., indicating that the same student was not categorized not a problem on the GRADE. Negative predictive power was calculated as the probability that students categorized "not a problem" on the GRADE indicated a non-validated problem given a CBM-negative problem. ORF sensitivity was 1.0 and specificity was .99. Positive predictive power for ORF was .83, where as negative predictive power was 1.0, Sensitivity for sentence maze and picture word fluency was .0 yet respective specificity were 1.0 with negative predictive power estimates at 1.0.

- Screening Accuracy. A primary question in this study was the degree to which group-administered screening measures would identify the same students as ORF. A two-step process for generating cut scores for screening accuracy was modeled after the procedures described by Siberglitt and Hintze (2005), which used a combination of Logistic Regression and ROC curve analysis. ORF was utilized as the criterion measure for this analysis using a cut score of 40 words correct per minute as proposed by Hasbrouk and Tindal, 1992. In Step 1, logistic regression was used for category prediction ("at risk" or "not at risk") to identify overall classification rate and the probability of various errors. Step 2 used ROC analysis, which yielded hit rate juxtaposed to the likelihood of false positives to determine the final cut scores. In Step 1, logistic regression showed sensitivity for picture word fluency and sentence maze at 0.85. Specificity was 0.85 for picture word fluency and 0.79 for sentence maze. Table 5

shows the coordinates of the curve for Step 2, whereby the ROC analysis was used to generate the final set of decision rules for sentence maze and picture word fluency with paramount importance geared for sensitivity; i.e., identifying too many students as at risk, rather than not identifying truly at-risk students. This analysis showed picture word fluency sensitivity and specificity at 0.90 and 0.82, respectively. Sensitivity and specificity for sentence was 0.90 and 0.67. Table 6 shows the resultant cut scores for each measure accompanied by additional information about their predictive accuracy. Diagnostic statistics are presented in Table 7.

Table 5—First Grade ROC Analysis Coordinates of the Curve

Picture Word Fluency

Positive if Less Than or Equal To	Sensitivity	1 - Specificity
2	0	0
4.5	0.025	0
7.5	0.05	0
10.5	0.1	0
12.5	0.175	0
13.5	0.225	0
14.5	0.25	0
15.5	0.3	0.03030303
16.5	0.4	0.03030303
17.5	0.475	0.03030303
18.5	0.55	0.03030303
19.5	0.575	0.03030303
20.5	0.65	0.03030303
21.5	0.75	0.06060606

(table 5 continued)

22.5	0.775	0.06060606
23.5	0.85	0.09090909
25	0.85	0.12121212
26.5	0.85	0.15151515
27.5	0.9	0.18181818
28.5	0.95	0.27272727
29.5	0.95	0.33333333
30.5	0.95	0.42424242
32	0.95	0.45454545
33.5	0.95	0.48484848
34.5	0.95	0.57575758
35.5	0.975	0.63636364
36.5	0.975	0.6969697
37.5	0.975	0.75757576
38.5	0.975	0.81818182
40	1	0.87878788
41.5	1	0.90909091
45	1	0.93939394
49	1	0.96969697
51	1	1

Coordinates of the Curve

Sentence Maze

Positive if Less Than or Equal To	Sensitivity	1 – Specificity
0	0	0

table 5 continued

1.5	0.05	0
2.5	0.15	0
3.5	0.3	0
4.5	0.45	0.06060606
5.5	0.525	0.06060606
6.5	0.6	0.09090909
7.5	0.625	0.12121212
8.5	0.75	0.15151515
9.5	0.775	0.18181818
10.5	0.8	0.18181818
11.5	0.825	0.21212121
12.5	0.85	0.21212121
13.5	0.875	0.27272727
14.5	0.9	0.33333333
15.5	0.9	0.39393939
16.5	0.925	0.42424242
17.5	0.95	0.42424242
18.5	0.975	0.54545455
19.5	0.975	0.66666667
20.5	1	0.6969697
22	1	0.75757576
23.5	1	0.78787879
24.5	1	0.81818182
25.5	1	0.84848485
26.5	1	0.87878788
28	1	0.90909091
32	1	0.93939394

(table 5 continued)

35.5	1	0.96969697
37	1	1

Table 6—First Grade CBM Cut Scores and Predictive Accuracy

CBM Measure	Cut Score	Percent at or above cut score passing the criterion measure	Percent below cut score on criterion measure failing the criterion measure
Picture Word Fluency	24	83%	89%
Sentence Maze	11	79%	83%

Table 7—First Grade Diagnostic Statistics

	Diagnostic Statistics for Cut Scores							
CBM Measure	Cut Score	Positive Predictive Power	Negative Predictive Power	Sensitivity	Specificity	Kappa	Standard Error of Kappa	Phi Coefficient
Picture Word Fluency	24	0.89	0.83	0.83	0.88	0.7	0.08	0.7
Sentence Maze	11	0.83	0.79	0.83	0.78	0.61	0.09	0.7

5.2.3 Third Grade Reliability and Validity Analysis

- Reliability. Of the 236 third grade students who participated in the study, 30 were randomly selected for inclusion in a sample to evaluate reliability. Internal consistency reliability was conducted using split-half reliability analysis (Crocker & Algina 1986). For this, each response opportunity on the group-administered CBM assessments was dichotomously coded into a statistical database using "1" to indicate a correct response and "0" to indicate an incorrect response. The researcher then created two forms for the screener by assigning all odd-numbered items to form 1 and all even-numbered items to form 2. Correlation between forms was .95 for paragraph maze, .97 for sentence maze and .99 for picture

word fluency. Respective Guttman Split-Half coefficients were .98 for paragraph maze and .99 for sentence maze and picture word fluency.

- Concurrent Validity. Bivariate correlation coefficients, presented in Table 8, were computed between ORF, paragraph maze, sentence maze, picture word fluency, and MCT to show associations among these variables. Pearson correlations among all CBM measures achieved statistical significance ($p \leq .01$). The author subsequently employed Steiger's Z-test for correlated-correlations (Meng, 1992) to examine if certain correlations between individual CBM assessments and MCT were stronger than others. ORF and paragraph maze showed the same correlation with MCT. Difference tests suggested that correlations between ORF and MCT were significantly greater than sentence maze and MCT ($Z = 2.39$, $p<.01$), yet not significantly greater than the correlation between picture word fluency and MCT ($Z = 1.03$, $p>.05$). Difference test showed that correlations between paragraph maze and MCT were also significantly greater than the correlation between sentence maze and MCT ($Z = 1.84$, $p<.05$), yet failed to achieve statistical significance for the correlation between picture word fluency and MCT ($Z = .987$, $p>.05$). However, correlation between sentence maze and MCT and picture word fluency and MCT was similar ($Z=.79$, $p>.05$).

Table 8—Third Grade Bivariate Correlations Between Target Variables

	Variable				
Variable	N	2	3	4	5
1. ORF	236	0.8	0.85	0.72	0.59
2. Paragraph Maze	236		0.739	0.694	0.59
3. Sentence Maze	236			0.75	0.52
4. Picture Word Fluency	236				0.55
5. MCT	236				

Note. All coefficients are significant ($p < .01$)

- Predictive Validity. Individual regression analyses were computed using the MCT as the criterion measure. Regression equations showed that all CBM(s) accounted for a significant portion of the variance in MCT performance. ORF and paragraph maze performance each explained 35% of the variance in student's MCT performance (F (1, 235) = 124.90, $p < .01$; F (1, 235) = 123.20, $p < .01$). Sentence maze accounted for 27% of the variance in student's MCT performance (F (1, 235) = 88.45, $p < .01$), and picture word fluency accounted for 30% (F (1, 235) = 99.62, $p < .01$).

The relative contributions of the various CBM measures were then examined as predictors of MCT performance. Simultaneous multiple regressions, presented in Table 6, show the relative influence of the various CBM measures for predicting MCT performance. In Part I, a simultaneous multiple regression utilizing ORF, paragraph maze, sentence maze, and picture word fluency was computed with the goal of evaluating the additional contribution of each group-administered CBM assessment over ORF alone. In Part II, a simultaneous multiple regression with paragraph maze, sentence maze, and picture word fluency investigated which assessment explained the most variance. Part I of Table 9 shows that all four variables together accounted for 41% of the variance F (4, 231) = 39.05, $p < .01$). Paragraph maze uniquely accounted for 3% of the variance in a student's MCT performance. ORF and picture word fluency each accounted for approximately an additional 2%. With all four variables used simultaneously, sentence maze had little additional influence in predicting MCT performance, yielding insignificant beta weights and squared semipartial correlations. Part II of Table 9 shows simultaneous multiple regression analysis directly comparing the relative influence of paragraph maze, sentence maze, and picture word fluency together in predicting MCT performance. This analysis indicated that the three predictors when used together accounted for 39% of the variance in the students' MCT performance F (3, 232) = 48.64, $p < .01$). In this analysis, both paragraph maze and picture word fluency exclusively accounted for significant variance, while sentence maze failed to achieve statistical significance.

Table 9—Third Grade Simultaneous Multiple Regression Analyses Comparing the Utility of CBM Assessments in Predicting MCT Performance

I. Predictors	Standard β	sr^2
ORF	0.282	0.02**
Paragraph Maze	0.28	0.03**
Sentence Maze	-0.076	-0.001
Picture Word Fluency	0.205	0.02*
		R^2 = .408
		Adjusted R^2 =.393
II. Predictors		
Paragraph Maze	0.37	0.06**
Sentence Maze	0.08	0.002
Picture Word Fluency	0.23	0.02**
		R^2 = .386
		Adjusted R^2 =.378

*p < .05.
**p < .01.

Hierarchical multiple regression coefficients, presented in Table 10, were computed to evaluate the added predictability of the group-administered CBM measures compared to ORF with MCT as the criterion measure. First, findings were examined whereby group-administered CBM measures alone were used to predict MCT performance. ORF was then entered to see if it accounted for any additional variance after the group-administered CBM measures. Second, findings were examined where by ORF was used solely to predict performance on the MCT. After that, the group-administered CBM measures were entered to see if they explained and additional variance. In the first analysis, paragraph maze entered first followed by picture word fluency, sentence maze and ORF. Paragraph maze accounted for

35% of the variance. The addition of picture word fluency at Step 2 was significant, with both predictors combined accounting for 38% of the variance in MCT performance. The addition of sentence maze in Step 3 was insignificant; however, the added influence of ORF in Step 4 was significant, with all four variables now accounting for nearly half of the variance in student MCT performance. In the second set of analysis, ORF was entered at Step 1 and accounted for 35% of the variance in student MCT performance. Paragraph maze was entered in Step 2, resulting in a significant increase, with both variables combined accounting for 39% of the variance. The inclusion of picture word fluency resulted in additional increase; however, the addition of sentence maze was insignificant.

Table 10—Third Grade Hierarchical Multiple Regression Analyses Comparing the Utility of Group-administered CBM Versus ORF in Predicting MCT Performance

Predictor		*R*	*R* Square	Adjusted *R* Square	R^2 Change	*F* Change
Step 1						
	Paragraph Maze	0.6	0.35	0.35	0.35	124.9**
Step 2						
	Picture Word Fluency	0.62	0.38	0.38	0.04	13.67**
(table 10 continued)						
Step 3						
	Sentence Maze	0.62	0.39	0.38	0.002	0.753
Step 4						
	ORF	0.64	0.4	0.39	0.02	6.70**
Step 1						
	ORF	0.59	0.35	0.34	0.35	123.12**
Step 2						

(table 10 continued)

Paragraph Maze	0.62	0.39	0.38	0.04	15.99**
Step 3					
Picture Word Fluency	0.63	0.4	0.39	0.02	5.9*
Step 4					
Sentence Maze	0.64	0.4	0.39	0.001	0.54

***p<.01*

- Classification Accuracy. Individual logistic regression equations were computed to determine the usefulness of the CBM assessments for categorical prediction. This analysis used a student's MCT performance to indicate a positive or negative problem. A student was categorized as having a problem on the MCT when his/her performance was indicated to be below "minimal/basic." When a student's performance on the MCT indicated to be "minimal/basic" or above, the student was categorized as negative for a problem on the MCT. Of the 229 students who entered into the logistic regression analyses seven student performances were classified as below "minimal/basic" on the MCT. Table 11 presents classification accuracy statistics determined according to pass/fail rates on the MCT as predicted by the student's scores on the CBM assessments.

Categorical prediction was examined by three sets of analyses. First, findings were examined in terms of variance accounted for by each individual predictor. Second, findings were examined in terms of goodness-of-fit between a student's actual performance on the MCT and their predicted performance per each CBM assessment. Third, findings were examined in terms of classification accuracy with attention devoted to each CBM assessments sensitivity, specificity, positive predictive power, and negative predictive power. In the first set of logistic analyses, all CBM assessments accounted for a significant portion of the variance in student MCT performances. Picture word fluency was the strongest predictor, accounting for 73% of the variance in student MCT performance. Paragraph maze accounted for 69% of

the variance while ORF accounted for 66% of the variance, followed by sentence maze, which accounted for 65% of the variance. The second set of analyses utilized the Hosmer and Lameshow chi-square test showing the goodness-of-fit between the actual and predicted student performance. Analysis indicated Hosmer and Lameshow χ^2 (8, N = 236) = .495, p > 1.00 for picture word fluency, followed by Hosmer and Lameshow χ^2 (8, N = 236) = .472, p > .1.00 for paragraph maze, Hosmer and Lameshow χ^2 (8, N = 236) = .405, p > 1.00 for ORF, and Hosmer and Lameshow χ^2 (7, N = 236) = .387, p > .1.00 for sentence maze. The third set of analyses investigated the classification accuracy of ORF, paragraph maze, sentence maze, and picture word fluency to evaluate the extent of correct categorical classification of students into two groups—those "at risk" and those "not at risk" on the MCT. ORF sensitivity was .43 and specificity was .99. Positive and negative predictive power for ORF were .60 and .98, respectively. Paragraph maze sensitivity was .70 and specificity was 1.0. Paragraph maze positive and negative predictive power was .88 and .99 respectively. For sentence maze, sensitivity and specificity were .43 and 1.00, respectively, whereas positive and negative predictive power were .75 and .98. Picture word fluency sensitivity was .71 and specificity was 1.0. Positive and negative predictive power for picture word fluency were .83 and .99.

Table 11—Third Grade Logistic Regression for Classification Accuracy of CBM Assessments

ORF	0.6	0.98	0.43	99	0.49	0.12	0.49
Paragraph Maze	0.88	0.99	0.7	1	0.77	0.11	0.77
Sentence Maze	0.75	0.98	0.75	1	0.54	0.18	0.56
Picture Word Fluency	0.71	1	0.83	0.99	0.77	0.13	0.77

- Screening Accuracy. A two-step process for generating cut scores for screening accuracy was modeled after procedures described by Siberglitt and Hintze (2005), using a combination of Logistic Regression and ROC curve analysis. ORF was utilized as the criterion measure for this analysis, using a cut score of 100 words correct per minute as proposed by Hasbrouk and Tindal (1992). In Step 1, Logistic Regression was used for category prediction ("at risk" or "not at risk") to identify overall

classification rate and the probability of various errors. Step 2 used ROC analysis, which yielded hit rate juxtaposed to the likelihood of false positives to determine the final cut scores.

In Step 1, the logistic regression showed sensitivity for paragraph maze, sentence maze and picture word fluency at 0.63 and 0.62, and 0.45 respectively. Specificity was 0.90 for paragraph maze and 0.93 for sentence maze. Picture word fluency specificity was 0.92. Table 12 shows the coordinates of the curve for Step 2, whereby the ROC analysis was used to generate the final set of decision rules for picture maze, sentence maze, and picture word fluency, with paramount importance geared for sensitivity. Summary diagnostic statistics are presented in Table 14.

Table 12—Third Grade ROC Analysis Coordinates of the Curve

Paragraph Maze

Positive if Less Than or Equal To	Sensitivity	1 - Specificity
4	0	0
5.5	0.05	0
6.5	0.083333333	0
7.5	0.1	0
8.5	0.116666667	0
9.5	0.183333333	0
10.5	0.233333333	0
11.5	0.366666667	0.010869565
12.5	0.45	0.043478261
13.5	0.566666667	0.081521739
14.5	0.633333333	0.10326087

(table 12 continued)

15.5	0.666666667	0.119565217
16.5	0.7	0.14673913
17.5	0.75	0.190217391
18.5	0.766666667	0.25
19.5	0.866666667	0.282608696
20.5	0.9	0.358695652
21.5	0.95	0.505434783
22.5	1	0.695652174
23.5	1	0.72826087
24.5	1	0.766304348
25.5	1	0.798913043
26.5	1	0.815217391
27.5	1	0.858695652
29	1	0.869565217
30.5	1	0.885869565
31.5	1	0.902173913
32.5	1	0.923913043
33.5	1	0.934782609
34.5	1	0.945652174
35.5	1	0.951086957
36.5	1	0.961956522
37.5	1	0.967391304
38.5	1	0.972826087
40	1	0.97826087
41.5	1	0.983695652
42.5	1	0.989130435

(table 12 continued)

46	1	0.994565217
50	1	1

Coordinates of the Curve

Sentence Maze

Positive if Less Than or Equal To	Sensitivity	1 - Specificity
2	0	0
5	0.016666667	0
8.5	0.033333333	0
11.5	0.066666667	0
13.5	0.083333333	0
15	0.116666667	0
16.5	0.183333333	0
17.5	0.216666667	0
18.5	0.316666667	0.005434783
19.5	0.35	0.005434783
20.5	0.383333333	0.010869565
21.5	0.45	0.016304348
22.5	0.533333333	0.027173913
23.5	0.616666667	0.065217391
24.5	0.683333333	0.076086957
25.5	0.766666667	0.152173913
26.5	0.85	0.184782609
27.5	0.916666667	0.255434783
28.5	0.966666667	0.320652174

(table 12 continued)

29.5	0.966666667	0.358695652
30.5	0.983333333	0.429347826
31.5	1	0.527173913
32.5	1	0.586956522
33.5	1	0.630434783
34.5	1	0.684782609
35.5	1	0.72826087
36.5	1	0.766304348
37.5	1	0.798913043
38.5	1	0.826086957
39.5	1	0.842391304
40.5	1	0.880434783
41.5	1	0.902173913
42.5	1	0.913043478
43.5	1	0.929347826
44.5	1	0.934782609
46	1	0.956521739
47.5	1	0.961956522
49	1	0.972826087
51	1	1

Coordinates of the Curve

Sentence Picture Word Fluency

Positive if Less Than or Equal To	Sensitivity	1 - Specificity
10	0	0
13.5	0.033333333	0

(table 12 continued)

17	0.05	0
18.5	0.066666667	0
20	0.083333333	0
21.5	0.1	0
23	0.116666667	0
25.5	0.133333333	0
27.5	0.166666667	0
28.5	0.216666667	0.005434783
29.5	0.233333333	0.005434783
30.5	0.25	0.005434783
31.5	0.283333333	0.005434783
32.5	0.3	0.005434783
33.5	0.333333333	0.02173913
34.5	0.4	0.027173913
35.5	0.45	0.038043478
36.5	0.45	0.076086957
37.5	0.516666667	0.076086957
38.5	0.55	0.086956522
39.5	0.55	0.097826087
40.5	0.616666667	0.119565217
41.5	0.7	0.152173913
42.5	0.716666667	0.163043478
43.5	0.733333333	0.206521739
44.5	0.766666667	0.255434783
45.5	0.8	0.293478261
46.5	0.833333333	0.331521739
47.5	0.85	0.380434783

(table 12 continued)

48.5	0.916666667	0.445652174
49.5	0.916666667	0.47826087
50.5	0.916666667	0.510869565
51.5	0.933333333	0.559782609
52.5	0.933333333	0.597826087
53.5	0.933333333	0.614130435
54.5	0.95	0.635869565
55.5	0.95	0.668478261
56.5	0.966666667	0.711956522
57.5	0.966666667	0.755434783
58.5	0.966666667	0.777173913
59.5	0.966666667	0.831521739
61	1	0.97826087
63	1	0.983695652
64.5	1	0.989130435
66	1	1

Table 13— Third Grade CBM Cut Scores and Predictive Accuracy

CBM Measure	Cut Score	Percent at or above cut score with satisfactory score on criterion measure	Percent below cut score with satisfactory score on criterion measure
Paragraph Maze	19	94%	50%
Sentence Maze	26	94%	60%
Picture Word Fluency	45	92%	47%

Table 14—Third Grade Diagnostic Statistics for Cut Scores

CBM Measure	Cut Score	Positive Predictive Power	Negative Predictive Power	Sensitivity	Specificity	Kappa	Standard Error of Kappa	Phi Coefficient
Paragraph Maze	19	0.5	0.94	0.87	0.72	0.47	0.06	0.51
Sentence Maze	26	0.6	0.94	0.85	0.82	0.58	0.06	0.6
Picture Word Fluency	45	0.47	0.92	0.8	0.71	0.41	0.06	0.44

5.2.4 Fifth Grade Reliability and Validity Analyses

- Reliability. Of the 173 fifth grade students who participated in the study, 30 were randomly selected for inclusion in the reliability sample. Internal consistency reliability proceeded with split-half reliability analyses (Crocker & Algina 1986). For this, each response opportunity on the group-administered CBM assessments was dichotomously coded into a statistical database using "1" to indicate a correct response and "0" to indicate an incorrect response. The researcher then created two forms for the screener by assigning all odd-numbered items to form 1 and all even-numbered items to form 2. Correlation between forms was .99 for paragraph maze and picture word fluency. Correlation between forms was .98 for sentence maze. Guttman Split-Half coefficients were .99 for paragraph maze and picture word fluency and .98 for sentence maze. A comparison of both forms yielded Spearman-Brown coefficients of 1.0 for paragraph maze .99 for sentence maze, and 1.0 for picture word fluency.

- Concurrent Validity. Bivariate correlation coefficients, presented in Table 15, were computed between ORF, paragraph maze, sentence maze, picture word fluency, and MCT to show association among these variables. Pearson correlations among all CBM measures achieved statistical significance ($p \leq .01$). The researcher then employed the Steiger's Z-test for correlated-correlations (Meng, 1992) to examine whether the correlations between individual CBM assessments and MCT were stronger than others. Correlation between paragraph maze and MCT was significantly higher than the correlation between ORF and MCT ($Z = 2.30$, $p<.01$), between sentence maze and MCT ($Z = 2.95$, $p<.01$), and between picture word fluency and MCT ($Z = 2.34$, $p<.01$). Difference tests failed to achieve statistical

significance for ORF and sentence maze (Z = .45, p>.05), ORF and picture word fluency (Z = .18, p>.05), and sentence maze and picture word fluency (Z = .19, p>.05).

Table 15—Fifth Grade Bivariate Correlations Between Target Variables

		Variable			
Variable	N	2	3	4	5
1. ORF	173	0.75	0.73	0.592	0.59
2. Paragraph Maze	173		0.79	0.72	0.68
3. Sentence Maze	173			0.59	0.57
4. Picture Word Fluency	173				0.58
5. MCT	173				

Note. All coefficients are significant ($p < .01$)

- Predictive Validity. Individual regression coefficients were computed using the MCT as the criterion measure. Regression equations showed that all CBM measures accounted for a significant portion of the variance. Paragraph maze performance explained nearly half (47%) of the variance in student MCT performance (F (1, 171) = 150.57, $p < .01$). ORF and picture word fluency each explained 34% of the variance in student MCT performance (F (1, 171) = 89.03, $p < .01$; (F (1, 171) = 87.74, p .01). Sentence maze accounted for 32% of the variance in student MCT performance (F (1, 171) = 80.09, $p < .01$).

The relative contribution of the various CBM measures were then examined as predictors of MCT performance. The Simultaneous multiple regressions are presented in Table 16. In Part I, a simultaneous multiple regression utilizing ORF, paragraph maze, sentence maze, and picture word fluency was computed with the goal of evaluating the additional contribution of each group-administered CBM assessment over ORF alone. In part II, a simultaneous multiple regression with paragraph maze, sentence maze, and picture word fluency investigated which CBM assessment explained the most variance. Part I of Table 16 shows that all four variables accounted for 49% of the

variance F (4, 168) = 40.88, p <.01). In this analysis, paragraph maze accounted for 5% of the variance in MCT performance over and beyond ORF, sentence maze, or picture word fluency. Picture word fluency accounted for an additional 1% and achieved statistical significance. With all four predictors included, ORF and sentence maze failed to account for significant variance, as evidenced by their beta weights and squared semipartial correlations. Part II of Table 16 shows simultaneous multiple regressions directly comparing the unique influence of paragraph maze, sentence maze, and picture word fluency in predicting MCT performance. This analysis indicated that all three variables together accounted for 47% of the variance in student MCT performances F (3, 175) = 52.62, p <.01). In this analysis paragraph maze and picture word fluency uniquely accounted for significant variance; however, the contribution of sentence maze failed to achieve statistical significance.

Table 16—Fifth Grade Simultaneous Multiple Regression Analyses Comparing the Utility of CBM Assessments in Predicting MCT Performance

I. Predictors	Standard β	Sr^2
ORF	0.14	0.008
Paragraph Maze	0.45	0.05**
Sentence Maze	0.002	0
(table 16 continued)		
Picture Word Fluency	0.17	0.01*
		R^2 = .493
		Adjusted R^2 =.481
II. Predictors		
Paragraph Maze	0.52	0.08**
Sentence Maze	0.03	0
Picture Word Fluency	0.19	0.02*

(table 16 continued)

R^2 = .474

Adjusted R^2 =.465

*p<.05.
**p<.01.

Hierarchical multiple regression coefficients, presented in Table 17, were computed to evaluate the added predictability of the group-administered CBM measures compared to ORF, with MCT as the criterion measure. First, the researcher examined findings whereby group-administered CBM measures were used to predict MCT performance. ORF was then entered to see if it accounted for any additional variance after the group-administered CBM measures. Second, findings were examined where by ORF was used solely to predict performance on the MCT. After that, the group-administered CBM measures were entered to see if they explained and additional variance.

In the first analysis, paragraph maze was entered first, followed by picture word fluency, sentence maze, and ORF. Paragraph maze accounted for 47% of the variance; with the addition of picture word fluency, both variables accounted for 48% of the variance. This increase was statistically significant. The addition of sentence maze in Step 3 and ORF in Step 4 failed to yield significant increases in predictability. In the second set of analysis, ORF was entered first, followed by paragraph maze, picture word fluency, and sentence maze. In this analysis, ORF accounted for 34% of the variance in student MCT performances; with the addition of paragraph maze, both variables accounted for 48% of the variance. This increase was statistically significant. The addition of picture word fluency in Step 3 and sentence maze in Step 4 yielded negligible increases in predictability.

- Classification Accuracy. Individual logistic regression equations were computed to determine the usefulness of CBM assessments for categorical prediction. This analysis used a student's MCT performance to indicate a positive or negative problem. A student was categorized as having a problem

Table 17—Fifth Grade Hierarchical Multiple Regression Analyses Comparing the Utility of Group-administered CBM Measures Versus ORF in Predicting MCT Performance

Predictor		*R*	*R* Square	Adjusted *R* Square	Change Statistics R^2 Change	Change Statistics *F* Change
Step 1						
	Paragraph Maze	0.68	0.47	0.47	0.47	150.57**
Step 2						
	Picture Word Fluency	0.7	0.48	0.48	0.02	5.26*
Step 3						
	Sentence Maze	0.7	0.49	0.48	0.001	0.374
Step 4						
	ORF	0.7	0.49	0.48	0.008	2.63
Step 1						
	ORF	0.59	0.34	0.34	0.342	89.03**
Step 2						
	Paragraph Maze	0.693	0.48	0.47	0.138	44.97**
Step 3						
	Picture Word Fluency	0.7	0.49	0.48	0.013	4.43
Step 4						
	Sentence Maze	0.7	0.49	0.48	0	0.001

***p<.01*
**p<.05*

on the MCT when his/her performance was indicated to be below "minimal/basic." When a student's performance on the MCT indicated "minimal/basic" or above, the student was categorized as negative for a problem on MCT. Of the 164 students who entered into the logistic regression analysis, 11 student performances were below "minimal/basic" on the MCT. Table 18 presents classification accuracy statistics determined according to the below "minimal/basic" rates on the MCT as predicted by the student's scores on the CBM assessments.

Categorical prediction was examined by three sets of analyses. First, findings were examined in terms of variance accounted for by each individual predictor. Second, findings were examined in terms of goodness-of-fit between actual student performance on the MCT and predicted student performance per each CBM assessment. Third, findings were examined in terms of classification accuracy, with attention devoted to each CBM assessment's sensitivity, specificity, positive predictive power, and negative predictive power. In the first set of logistic analyses, all CBM assessments accounted for a significant portion of the variance in student MCT performances. ORF was the strongest predictor, accounting for 58% of the variance in student MCT performances. Paragraph maze accounted for 55% of the variance in student MCT performances. Picture word fluency accounted for 44% of the variance, while sentence maze accounted for 43% of the variance in student MCT performances. The second set of analysis utilized the Hosmer and Lameshow chi-square test showing the goodness-of-fit between actual and predicted student performance. Analysis indicated a Hosmer and Lameshow χ^2 (8, N = 173) = .753, $p > .05$ for ORF, Hosmer and Lameshow χ^2 (8, N = 173) = 10.527, $p > .05$ for paragraph maze, Hosmer and Lameshow χ^2 (8, N = 173) = 6.530, $p > .05$ for picture word fluency, and Hosmer and Lameshow χ^2 (8, N = 173) = 5.557, $p > .05$ for sentence maze. The third set of analysis investigated the classification accuracy of ORF, paragraph maze, sentence maze, and picture word fluency to evaluate the extent of correct categorical classification of students into two groups—those "at risk" and those "not at risk" on the MCT.

- Screening Accuracy. A two-step process for generating cut scores for screening accuracy was modeled after procedures described by Siberglitt and Hintze (2005) using a combination of logistic regression and ROC curve analysis. ORF was utilized as the criterion measure for this analysis using a cut score of 100 words correct per minute as proposed by Hasbrouk and Tindal, 1992. In Step 1, logistic regression was used for category prediction ("at risk" or "not at risk") to identify overall classification rate and the probability of various errors. Step 2 used ROC analysis, which yielded hit rate juxtaposed to the likelihood of false positives to determine the final cut scores. In Step 1, logistic regression

Table 18—Fifth Grade Logistic Regression for Classification Accuracy of CBM Assessments

CBM Measure	Positive Predictive Power	Negative Predictive Power	Sensitivity	Specificity	Kappa	Standard Error of Kappa	Phi Coefficient
ORF	0.44	0.96	0.36	0.97	0.36	0.14	0.37
Paragraph Maze	0.88	0.98	0.63	0.99	0.72	0.11	0.73
Sentence Maze	1	0.98	0.5	1	0.66	0.16	0.7
Picture Word Fluency	0.8	0.96	0.36	0.99	0.48	0.16	0.52

showed sensitivity and specificity for paragraph maze at 0.35 and 0.97 and 0.46 and 0.99 for sentence maze. Sensitivity and specificity for picture word fluency was 0.27 and 0.99, respectively. Table 19 shows the coordinates of the curve for Step 2, whereby the ROC analysis was used to generate the final set of decision rules for paragraph maze, sentence maze, and picture word fluency, with paramount importance geared for sensitivity. This analysis showed paragraph maze sensitivity and specificity at 0.78 and 0.69, respectively. Sensitivity and specificity for sentence maze was 0.92 and 0.68. For picture word fluency, sensitivity and specificity fell at 0.92 and 0.69. Table 20 shows the resultant cut scores for each measure accompanied by additional information about their predictive accuracy. Summary diagnostic statistics are presented in Table 21.

Table 19—Fifth Grade ROC Analysis Coordinates of the Curve

Paragraph Maze

Positive if Less Than or Equal To	Sensitivity	1 - Specificity
1	0	0
3	0.07407407	0

(table 19 continued)

4.5	0.18518519	0
5.5	0.22222222	0
7	0.22222222	0.00613497
8.5	0.25925926	0.00613497
9.5	0.2962963	0.00613497
10.5	0.2962963	0.01226994
11.5	0.33333333	0.01840491
13	0.33333333	0.02453988
14.5	0.37037037	0.04294479
15.5	0.37037037	0.04907975
16.5	0.44444444	0.05521472
17.5	0.44444444	0.0797546
18.5	0.55555556	0.10429448
19.5	0.7037037	0.1595092
20.5	0.7037037	0.17177914
21.5	0.74074074	0.24539877
22.5	0.74074074	0.26993865
23.5	0.77777778	0.31288344
24.5	0.85185185	0.33742331
25.5	0.85185185	0.35582822
26.5	0.85185185	0.3803681
27.5	0.92592593	0.44785276
28.5	0.92592593	0.49693252
29.5	0.92592593	0.53374233
30.5	0.92592593	0.61349693
31.5	0.92592593	0.65644172
32.5	0.92592593	0.6993865
33.5	0.92592593	0.72392638

(table 19 continued)

34.5	0.92592593	0.77300613
35.5	0.92592593	0.81595092
36.5	0.92592593	0.85276074
37.5	0.92592593	0.87116564
38.5	0.92592593	0.88343558
39.5	0.96296296	0.91411043
40.5	0.96296296	0.93251534
41.5	0.96296296	0.95092025
43	1	0.97546012
45	1	0.98159509
47	1	0.98773006
48.5	1	0.99386503
50	1	1

Coordinates of the Curve

Sentence Maze

Positive if Less Than or Equal To	Sensitivity	1 - Specificity
-1	0	0
1	0	0.00343643
2.5	0.02702703	0.00343643
5.5	0.05405405	0.00343643
8.5	0.10810811	0.00343643
9.5	0.16216216	0.00343643
10.5	0.21621622	0.00687285

(table 19 continued)

11.5	0.24324324	0.00687285
12.5	0.2972973	0.00687285
13.5	0.35135135	0.01030928
14.5	0.43243243	0.01718213
15.5	0.59459459	0.03436426
16.5	0.64864865	0.04810997
17.5	0.72972973	0.0652921
18.5	0.75675676	0.09278351
19.5	0.75675676	0.12371134
20.5	0.75675676	0.14089347
21.5	0.83783784	0.20618557
22.5	0.86486486	0.25085911
23.5	0.89189189	0.28522337
24.5	0.91891892	0.31958763
25.5	0.94594595	0.45360825
26.5	0.94594595	0.50515464
27.5	0.97297297	0.56013746
28.5	0.97297297	0.62199313
29.5	0.97297297	0.65979381
30.5	0.97297297	0.69072165
31.5	0.97297297	0.73883162
32.5	0.97297297	0.79037801
33.5	0.97297297	0.83848797
34.5	0.97297297	0.86254296
35.5	0.97297297	0.89003436
36.5	0.97297297	0.89690722
37.5	0.97297297	0.90721649

(table 19 continued)

38.5	0.97297297	0.91408935
39.5	0.97297297	0.92783505
40.5	0.97297297	0.95876289
41.5	0.97297297	0.96219931
42.5	1	0.96563574
43.5	1	0.96907216
45	1	0.97594502
46.5	1	0.98281787
47.5	1	0.9862543
49	1	0.98969072
51	1	1

Coordinates of the Curve

Picture Word Fluency

Positive if Less Than or Equal To	Sensitivity	1 - Specificity
-1	0	0
3	0	0.00355872
7.5	0.02777778	0.00355872
11.5	0.05555556	0.00355872
14.5	0.08333333	0.00355872
18	0.11111111	0.00355872
21.5	0.16666667	0.00355872
23	0.19444444	0.00355872
25	0.25	0.00355872
26.5	0.27777778	0.00711744

(table 19 continued)

27.5	0.30555556	0.00711744
28.5	0.30555556	0.01779359
29.5	0.30555556	0.02135231
30.5	0.30555556	0.02491103
31.5	0.33333333	0.02491103
32.5	0.36111111	0.03202847
33.5	0.38888889	0.03558719
34.5	0.52777778	0.04982206
35.5	0.52777778	0.0569395
36.5	0.55555556	0.07829181
37.5	0.58333333	0.08896797
38.5	0.63888889	0.10676157
39.5	0.69444444	0.12811388
40.5	0.72222222	0.16014235
41.5	0.72222222	0.17437722
42.5	0.72222222	0.19928826
43.5	0.77777778	0.23131673
44.5	0.83333333	0.24199288
45.5	0.86111111	0.27402135
46.5	0.91666667	0.30960854
47.5	0.91666667	0.34163701
48.5	0.94444444	0.38434164
49.5	0.94444444	0.40569395
50.5	0.94444444	0.43772242
51.5	0.94444444	0.47330961
52.5	0.94444444	0.54448399
53.5	0.94444444	0.60498221
54.5	0.94444444	0.65836299

(table 19 continued)

55.5	0.97222222	0.69395018
56.5	1	0.76156584
57.5	1	0.79003559
58.5	1	0.82562278
59.5	1	0.85053381
60.5	1	0.87544484
61.5	1	0.88967972
62.5	1	0.89323843
63.5	1	0.92170819
64.5	1	0.95373665
66	1	1

Table 20—CBM Cut Scores Predictive Accuracy

CBM Measure	Cut Score	Percent at or above cut score passing the criterion measure	Percent below cut score on criterion measure failing the criterion measure
Paragraph Maze	22	94%	31%
Sentence Maze	22	98%	30%
Picture Word Fluency	44	97%	31%

Table 21—Fifth Grade Diagnostic Statistics for Cut Scores

CBM Measure	Cut Score	Positive Predictive Power	Negative Predictive Power	Sensitivity	Specificity	Kappa	Standard Error of Kappa	Phi Coefficient
Paragraph Maze	22	0.31	0.94	0.74	0.73	0.3	0.07	0.35
Sentence Maze	22	0.3	0.98	0.86	0.75	0.34	0.05	0.39
Picture Word Fluency	44	0.31	0.97	0.83	0.76	0.3	0.05	0.4

CHAPTER 6. EXPERIMENT 2

Experiment 2 was designed to accomplish several goals. First, this study allowed for an examination of whether the results obtained in Experiment 1 generalized to another population in another state using an alternative form of CBM. This experiment also explored the utility of paragraph maze, sentence maze, and picture word fluency for screening purposes. Regression analysis examined paragraph maze, sentence maze, picture word fluency and ORF as predictors of performance on criterion measures. Second, predictive power estimates compared the classification accuracy of paragraph maze, sentence maze, and picture word fluency for purposes of validating cut scores established in Experiment 1. The primary criterion measure was GRADE for first grade and the Integrated Louisiana Assessment Program (*i*LEAP) for third and fifth grades. DIBELS ORF served as a secondary criterion variable in some analyses for first grade and third grade. ORF served as a secondary criterion variable for fifth grade.

6.1 Method

6.1.1 Participants and Setting

A total of 307 students participated in the second experiment; they were enrolled in first, third and fifth grade regular education classrooms in elementary schools in the Southeastern United States. Of the participants, 59% received free/reduced lunch, 41% were African American, 1% were Hispanic, 58% were Caucasian and <1% were Asian/Pacific Islander.

6.1.2 Measures

This study examined the generalizeability and predictive utility of paragraph maze, sentence maze, picture word fluency and DIBELS ORF (DORF) against performance on group achievement testing. The author established decision rules (benchmark cut scores) for each variable, based on his assessment of the appropriate balance among degrees of error for screening purposes.

- Dynamic Indicators of Basic Early Literacy Skills oral reading fluency (DIBELS –ORF).

DIBELS oral reading fluency (DORF) is a well-researched, individually administered measure of

reading performance and achievement for school-aged students. DORF test-retest reliability for elementary students ranges from .92 to .97 and alternate form reliability ranges from .89 to .94. DORF criterion-related validity ranges from .52 to .91 (Good, Simmons, & Kame'enui, 2001). DORF requires students to read a passage aloud for one minute while simultaneously being scored by an examiner. Words omitted, word substitutes, and hesitations of more than three seconds are scored as errors. Words self-corrected within three seconds are scored as accurate. The number of words read per minute, minus the number of errors, equals the student's score on DORF.

- Paragraph Maze. Paragraph maze is a paper-and-pencil measure of basic reading performance and comprehension. Administration is individual or group. The paragraph maze requires a student to read related passages in which words have been selectively omitted. Administration of paragraph maze asks the student to select, in a multiple-choice format, the one word that best completes the sentence. Students have three minutes to complete the paragraph maze. A student's paragraph maze score comprises the total correct words marked.

- Sentence Maze. Sentence maze is a variation of paragraph maze. Sentence maze is a paper-and-pencil measure of basic reading performance and comprehension, which is administered to students individually or in groups. Sentence maze requires a student to read unconnected sentences wherein the last word is omitted. For sentence maze the student is asked to select, in a multiple-choice format, the one word that best completes the sentence. Students have three minutes to complete the sentence maze. A student's sentence maze score comprises the total correct words marked.

- Picture Word Fluency. Picture word fluency, which is similar to WIF, is a measure of sight word reading. This procedure was developed using words from frequently used word lists in children's basal reading series including the Fry and Dale-Chall word lists. The student is asked to select, in a multiple-choice format, the word that best matches the picture. Administration is individual or in groups. Students have three minutes to complete picture word fluency. A student's picture word fluency score comprises the total correct words marked.

- Integrated Louisiana Educational Assessment Program (*i*LEAP). The *i*LEAP is a group administered, norm and criterion referenced achievement test published by the Louisiana Department of Education. In Louisiana elementary schools, the *i*LEAP is administered to students in the third and fifth grades for comparison of their achievement to national and state performance standards. Tests in English language arts, math, science, and social studies yield valuable information about the development of the students' skills. This study will use *i*LEAP English Language Arts Tests as criterion for comparison against the experimental screening probes for third and fifth grade students. The English Language Arts section of the *i*LEAP yields a scaled score measuring content standards including: read, comprehend, and respond; read, analyze, and respond to literature; apply reasoning and problem solving skills; write competently; use conventions of language; and locate, select and synthesize information. Data Recognition Corporation (2006) reported Fled-Raju stratified alpha and Cronbach reliability coefficients of .93 for third grade and .93 to .92 respectively for fifth grade. Test content items for the *i*LEAP test were selected from two sources: the Iowa Test of Basic Skills (ITBS) and new test items specifically developed to measure certain content standards, benchmarks, and grade level expectations. For the content items coming from the ITBS, the Northwest Evaluation Association (NWEA) reported concurrent validity at .77 for third grade and .84 for fifth grade. For the new test items on the augmented *i*LEAP, the Data Recognition Corporation reported that items would be considered valid in terms of content if score points for each content unit were similar to those set by the test's blueprint. In these terms, direct correspondence was reported for both third and fifth grade.

6.1.3 Procedure

The curriculum-based screening probes and the criterion measures were administered within one month of each other. The author administered the GRADE and the group-administered CBM measures while school personnel administered the *i*LEAP and DORF. First grade students were administered DIBELS oral reading fluency, sentence maze, picture word fluency, and the GRADE. Third grade students were administered DIBELS oral reading fluency, paragraph maze, sentence maze, picture word

fluency, and the *i*LEAP. Fifth grade students were administered one ORF, paragraph maze, sentence maze, picture word fluency, and the *i*LEAP. Administration of paragraph maze, sentence maze, and picture word fluency were counterbalanced by the author to control for order effects. Administration took approximately 15 minutes per class including startup time.

6.1.4 Assessment Administration

The same administration procedures used for oral reading fluency, paragraph maze, sentence maze, and picture word fluency in experiment 1 were followed for experiment 2.

- Integrated Louisiana Educational Assessment Program (iLEAP*).* School personnel administered the Integrated Louisiana Educational Assessment Program according to the prescribed procedures routinely used in that district. Scoring also followed these procedures.

6.1.5 Data Collection and Scoring

- Dynamic Indicators of Basic Early Literacy Skills oral reading fluency (DIBELS –ORF). DORF was administered to first and third grade students. Scoring of DORF followed the procedures described by Good, Simmons, & Kame'enui (2001). The number of words read per minute, subtracting the number of errors equaled the student's score on DORF.

- Oral Reading Fluency (ORF)*.* The author administered ORF to fifth grade students following the procedures described by Shinn (1989). In doing so, assessors asked students to read individually for one minute from grade appropriate reading probes.

- Paragraph Maze*.* The author scored maze passages by tallying the number of correct choices a student made during a three-minute timed assessment. A student's story paragraph score equaled the total correct words circled.

- Sentence Maze*.* The author scored sentence maze by tallying the number of correct choices a student made during a three-minute timed assessment. A student's sentence maze score equaled the total correct words circled.

- Picture Word Fluency. A student's picture word fluency score equaled the total correct words marked during a three-minute timed assessment.
- Integrated Louisiana Educational Assessment Program (iLEAP). The *i*LEAP was scored by the test manufacturers. The school shared the results with the author for entry into a statistical database.
- Group Reading Assessment and Diagnostic Evaluation (GRADE). The author scored GRADE protocols for first grade students, according to the procedures described in the administration manual.

6.2 Results

The purpose of Experiment 2 was to examine whether the results obtained in Experiment 1 generalized to another population in another state with different criterion measures. Specifically, the author examined the utility of paragraph maze, sentence maze, and picture word fluency as reliable and valid CBM measures of reading performance. Descriptive statistics are provided in the first section for the CBM assessments and criterion variables in the grades studied. In the second section, organized by grade, are data pertaining to reliability and validity of the group screening measures. Specifically, split-half coefficients were used for each grade to examine internal consistency reliability. In the following section, correlation and regression analyses are presented for each grade to examine the extent to which the CBM assessments are associated with and are predictors of performance on criterion assessments (i.e., GRADE and *i*LEAP). In the final section, predictive and screening accuracy statistics are presented.

6.2.1 Descriptive Statistics

Descriptive statistics for all variables are presented by grade in Table 22. According to Kline (1988), skewness values greater than an absolute value of three and kurtosis values greater than an absolute value of eight indicate normality problems. Absolute values of skewness and kurtosis levels in this study were acceptable (skewness less than 3 and kurtosis less than 8) for all variables thus satisfying the assumption of normality (Kline 1988). Additionally, following the power analysis guidelines

recommended in Cohen and Cohen (1983), sample size was sufficient for multiple correlation and regression analysis (Faul, Erdfelder, Lang, & Buchner, 2007).

Table 22—Descriptive Statistics of Target Variables

Variable	Mean (SD)	Skewness	Kurtosis
First Grade			
GRADE	67.78 (15)	0.69	0.44
DORF	57.2 (21.60)	0.32	0.5
Sentence Maze	13.3 (6.75)	0.09	0.82
Picture Word Fluency	30.88 (8.9)	0.03	0.04
Third Grade			
*i*LEAP	303.1 (48.8)	0.75	1.92
DORF	99.24 (28.30)	0.33	0.58
Paragraph Maze	17.77 (6.34)	0.16	0.18
Sentence Maze	26.87 (7.27)	0.46	1.24
Picture Word Fluency	42.36 (9.95)	0.49	1.3
Fifth Grade			
*i*LEAP	298.97 (40.88)	0.44	0.53
ORF	129.59 (34.20)	0.46	0.96
Paragraph Maze	23.99 (8.89)	0.833	1.78
Sentence Maze	24.31 (7.71)	1.1	1.93
Picture Word Fluency	42.27 (10.67)	0.3	0.05

6.2.2 First Grade Reliability and Validity Analyses

- Reliability. Of the 81 first grade students who participated in the study, 30 were randomly selected for inclusion in the reliability sample. Internal consistency reliability proceeded with split-half reliability analysis (Crocker & Algina 1986). For this, each response opportunity on the group-administered CBM assessments was dichotomously coded into a statistical database using "1" to indicate a correct response and "0" to indicate an incorrect response. The author then created two forms on the screener by assigning all odd-numbered items to form 1 and all even-numbered items to form 2. Correlation between forms was .97 for sentence maze and .99 for picture word fluency. Guttmant Split-Half coefficients were .98 and .99, respectively. A comparison of both forms yielded Spearman-Brown of .99 for sentence maze and .99 for picture word fluency.

- Concurrent Validity. Bivariate correlation coefficients, presented in Table 23, were computed between DORF, sentence maze, picture word fluency, and GRADE to show association among these variables. Pearson correlations among all CBM measures achieved statistical significance ($p \leq .01$). The author then employed the Steiger's Z-test for correlated-correlations (Meng, 1992) to examine whether one or more of the correlations between individual CBM assessments and GRADE were significantly stronger than others. Difference tests suggested that correlations between DORF and GRADE were significantly greater than sentence maze and GRADE ($Z=2.19$, $p<.05$) and picture word fluency and GRADE ($Z=4.46$, $p<.01$). The correlation coefficient for sentence maze and GRADE was significantly greater than the coefficient for picture word fluency and GRADE ($Z= 2.63$, $p>.01$). A similar pattern was seen in Experiment 1 with CBM-ORF showing the strongest association with GRADE performance. In Experiment 2, however, the second strongest correlation was seen between sentence maze and GRADE, whereas in Experiment 1 the second strongest correlation was seen between picture word fluency and GRADE.

Table 23—First Grade Bivariate Correlations Between Independent Variables

Variable	Variable n	2	3	4
1. DORF	81	0.76**	0.59**	0.68**
2. Sentence Maze	81		0.56**	0.54**
3. Picture Word Fluency	81			0.30**
4. GRADE	81			

**$p < .01$.

- Predictive Validity. Individual regression coefficients were computed using the GRADE as the criterion measure. Regression equations showed that the percentage of variance accounted for by all CBM measures was statistically significant. ORF performance explained 46% of the variance (F (1, 79) = 65.99, $p < .01$), followed by sentence maze, which explained 30% of the variance F (1, 79) = 33.21, $p < .01$), and picture word fluency, which explained 9% of the variance in student GRADE performance F (1, 79) = 7.70, $p < .01$). The predictive strength of all CBM measures in Experiment 2 was approximately the same as that shown in Experiment 1. In Experiment 1, however, picture word fluency explained markedly more variance in GRADE performance than it did in Experiment 2.

The relative contributions of the various CBM measurement scores in explaining variability in GRADE performance were examined. Simultaneous multiple regressions are presented in Table 24. In Part I, a simultaneous multiple regression utilizing DORF, sentence maze and picture word fluency was computed with the goal of evaluating the additional contribution of each group-administered CBM assessment over ORF alone. In Part II, a simultaneous multiple regression with sentence maze and picture word fluency investigated which of these CBM assessments explained the most variance. Part I of Table 24 shows simultaneous multiple regressions comparing the relative contributions of ORF, sentence maze and picture word fluency in explaining GRADE performance. This analysis showed that

all three variables together accounted for 48% of the variance, but DORF uniquely accounted for 18% of the variance in student GRADE performance over and beyond sentence maze and picture word fluency alone. Sentence maze and picture word fluency had little impact in explaining GRADE performance, as both yielded insignificant beta weights and squared semipartial correlation coefficients. Part II of Table 24 shows a simultaneous multiple regression analysis directly comparing the relative influence of sentence maze and picture word fluency together in predicting GRADE performance. This analysis indicated that together both predictors simultaneously accounted for 30% of the variance in student GRADE performance (F (2, 78) = 16.402, $p < .01$). In this analysis, only sentence maze yielded a statistically significant beta weight and squared semipartial correlation, thus uniquely accounting for 21% of the variance. Similar findings were seen in Experiment 1. In Experiment 1, however, of the group-administered CBM measures it was picture word fluency, which alone accounted for the most variance.

Table 24— First Grade Simultaneous Multiple Regression Analysis Comparing the Utility of CBM Assessments in Predicting GRADE Performance

Part I	Standard β	sr^2
DORF	0.69	0.18**
Sentence Maze	0.124	0.006
Picture Word Fluency	-0.18	0.02
		R^2 = .48
		Adjusted R^2 =.46
Part II		
Sentence Maze	0.55	0.21**
Picture Word Fluency	-0.01	0

(table 24 continued)

R^2 = .30

Adjusted R^2 =.28

**p<.01.

Hierarchical multiple regression coefficients, presented in Table 25, were computed to evaluate the added predictive value of sentence maze and picture word fluency scores compared to DORF with GRADE as the criterion measure. First, the author examined findings whereby group-administered CBM measures were used to predict GRADE performance. DORF was then entered into the equation last to determine if it contributed significantly to explained variation in GRADE. Second, the author examined findings whereby DORF was entered first in the equation to predict performance on the GRADE and was then followed by the group-administered CBM assessments to determine the degree to which the latter explained variation over and above DORF. In the first analysis, sentence maze entered the equation followed by picture word fluency. Sentence maze accounted for 30% of the variance. The addition of picture word fluency in Step 2 was insignificant. In Step 3 DORF added significantly to prediction with an R^2 change value of 0.18. In the second set of analysis, DORF was entered at Step 1 and accounted for 46% of the variance in GRADE performance. In Steps 2 and 3, the inclusion of sentence maze and picture word fluency did not contribute significantly to understanding the variance in GRADE. The same general findings in Experiment 2 were seen in Experiment 1 in that the ORF/DORF predictive relationship with the GRADE was consistently stronger that that of the group-administered CBM measures. In Experiment 1, however, the combined influence of both picture word fluency and sentence maze did significantly increase variance accounted for in GRADE performance.

Table 25—First Grade Hierarchical Multiple Regression Analyses Comparing the Utility of Group-administered CBM Versus DORF in Predicting GRADE Performance

Predictor		*R*	*R* Square	Adjusted *R* Square	R^2 Change	*F* Change
	First Analysis					
Step 1						
	Sentence Maze	0.54	0.3	0.29	0.3	33.21**
Step 2						
	Picture Word Fluency	0.54	0.3	0.28	0	0.01
Step 3						
	DORF	0.69	0.48	0.46	0.18	26.76**
	Second Analysis					
Step 1						
	DORF	0.68	0.46	0.45	0.46	65.99**
Step 2						
	Sentence Maze	0.68	0.46	0.44	0	0.36
Step 3						
	Picture Word Fluency	0.69	0.48	0.46	0.02	2.94

***p<.01*

- Classification Accuracy. Individual logistic regression equations were computed to determine the usefulness of the CBM assessments for categorical prediction. This analysis used a student's GRADE performance to indicate a positive or negative problem. A student was categorized as having a problem on the GRADE when his/her performance yielded a standard score at or below 85, which coincides with performance at or below the 16th percentile, indicating a significant risk status. When a student's performance yielded a standard score above 85, the student was categorized as negative for a problem on the GRADE. Of the 85 students who entered into the logistic regression analyses only nine

student performances were indicated to be at risk on the GRADE. Table 26 presents classification accuracy statistics determined according to the "pass/fail" rates on the GRADE as predicted by the student's scores on the CBM assessments.

Categorical prediction was examined by three sets of analyses. First, findings were examined in terms of variance accounted for by each individual predictor. Second, findings were examined in terms of goodness-of-fit between a student's actual performance on the GRADE and predicted student performance per each CBM assessment. Third, findings were examined in terms of classification accuracy, with attention devoted to each CBM assessment's sensitivity, specificity, positive predictive power, and negative predictive power. In the first set of logistic analyses only DORF accounted for a significant portion of the variance in student GRADE performance. That is, DORF accounted for 30% of the variance in student GRADE performance, whereas sentence maze and picture word fluency accounted for only 8% and 7%, respectively. The second set of analysis utilized the Hosmer and Lameshow chi-square test showing the goodness-of-fit between actual and predicted student performance. Analysis indicated Hosmer and Lameshow χ^2 (8, N = 81) = 5.61, p = .691 for DORF, Hosmer and Lameshow χ^2 (8, N = 81) = 4.48, p = .811 for sentence maze, followed by Hosmer and Lameshow χ^2 (8, N = 81) = 7.89, p= 34 for picture word fluency. The third set of analysis investigated the classification accuracy of ORF, sentence maze and picture word fluency to evaluate the extent of correct categorical classification of students into two groups; i.e., those at risk and those not at risk on the GRADE. Sensitivity for all three CBM measures fell at 0.0, while specificity was. 97 for DORF and 1.0 for sentence maze and picture word fluency. Positive predictive power for all three measures was 0.0. Negative predictive power was .92 for DORF and .90 for sentence maze and picture word fluency. Similar findings were seen in Experiment 1 with regard to the CBM assessments' specificity and negative predictive power. Specificity and negative predictive power was higher in Experiment 1.

Table 26—First Grade Logistic Regression for Classification Accuracy of CBM Assessments

CBM Measure	Positive Predictive Power	Negative Predictive Power	Sensitivity	Specificity	Kappa	Standard Error of Kappa	Phi Coefficient
DORF	0	0.92	0	0.97	-0.04	0.02	0.05
Sentence Maze	0	0.9	0	1	0	0	0
Picture Word Fluency	0	0.9	0	1	0	0	0

- Screening Accuracy. A primary question in this Experiment pertains to the degree to which group-administered screening measures would identify the same students as DORF. A two-step process for generating cut scores for screening accuracy was modeled after the procedures described by Siberglitt and Hintze (2005) using a combination of logistic regression and ROC curve analysis. DORF was utilized as the criterion measure for this analysis using a cut score of 40 words correct per minute as proposed by DIBELS' end-of-year benchmarks. In Step 1, logistic regression was used for category prediction ("at risk" or "not at risk") to identify overall classification rate and the probability of various errors. Step 2 used ROC analysis, which yielded hit rate juxtaposed to the likelihood of false positives to determine the final cut scores. In Step 1, logistic regression showed sensitivity for sentence maze and picture word fluency at 0.48 and 0.20, respectively. Specificity was 0.93 for sentence maze, and 0.96 for picture word fluency. Table 27 shows the coordinates of the curve for Step 2, whereby the ROC analysis was used to generate the final set of decision rules for sentence maze and picture word fluency with paramount importance geared for sensitivity. This analysis showed sentence maze sensitivity and specificity at 0.87 and 0.77, respectively. Sensitivity and specificity for picture word fluency was 0.80 and 0.73. Table 28 shows the resultant cut scores for each measure accompanied by additional information about their predictive accuracy. Diagnostic statistics are presented in Table 29.

The cut scores that were generated in Experiment 2 were similar to those generated in Experiment 1, particularly cut scores for sentence maze. Table 30 presents cut scores generated in Experiment 1 but applied to the data set used in Experiment 2. Juxtaposed to these findings are cut

scores generated in Experiment 2 but applied to the data set in Experiment 1. These findings show that sentence maze and picture word fluency will identify many of the same students that were identified as "at risk" on ORF/DORF.

Table 27—First Grade ROC Analysis

Coordinates of the Curve

Sentence Maze

First Positive if Less Than or Equal To	Sensitivity	1 – Specificity
1	0	0
2.5	0.2	0.01449275
3.5	0.3333333	0.01449275
4.5	0.4666667	0.07246377
5.5	0.5333333	0.08695652
6.5	0.6	0.10144928
7.5	0.6666667	0.14492754
8.5	0.7333333	0.17391304
9.5	0.8	0.2173913
10.5	0.8666667	0.23188406
11.5	0.8666667	0.33333333
12.5	0.8666667	0.34782609
13.5	0.9333333	0.42028986
14.5	1	0.46376812
15.5	1	0.49275362
16.5	1	0.60869565
17.5	1	0.62318841

(table 27 continued)

18.5	1	0.66666667
19.5	1	0.75362319
20.5	1	0.8115942
21.5	1	0.89855072
24	1	0.92753623
26.5	1	0.97101449
27.5	1	0.98550725
29	1	1

Coordinates of the Curve

Picture Word Fluency

Positive if Less Than or Equal To	Sensitivity	1 - Specificity
11	0	0
12.5	0	0.01428571
14	0.0666667	0.01428571
15.5	0.0666667	0.02857143
16.5	0.0666667	0.04285714
17.5	0.2	0.04285714
18.5	0.4	0.05714286
19.5	0.5333333	0.05714286
20.5	0.5333333	0.08571429
22	0.5333333	0.11428571
23.5	0.6	0.14285714
25	0.6	0.15714286
26.5	0.7333333	0.21428571

(table 27 continued)

27.5	0.8	0.22857143
28.5	0.8	0.27142857
29.5	0.8	0.32857143
30.5	0.8666667	0.38571429
31.5	0.8666667	0.47142857
32.5	0.8666667	0.48571429
33.5	0.8666667	0.54285714
34.5	0.8666667	0.61428571
35.5	0.8666667	0.65714286
36.5	1	0.72857143
38	1	0.75714286
39.5	1	0.77142857
40.5	1	0.81428571
42	1	0.85714286
43.5	1	0.88571429
45	1	0.92857143
46.5	1	0.94285714
47.5	1	0.95714286
49	1	0.98571429
51	1	1

Table 28—First Grade CBM Cut Scores and Predictive Accuracy

CBM Measure	Cut Score	Percent at or above cut score passing the criterion measure	Percent below cut score on criterion measure failing the criterion measure
Picture Word Fluency	28	94%	39%
Sentence Maze	10	96%	45%

Table 29—First Grade Diagnostic Statistics

	Diagnostic Statistics for Cut Scores							
CBM Measure	Cut Score	Positive Predictive Power	Negative Predictive Power	Sensitivity	Specificity	Kappa	Standard Error of Kappa	Phi Coefficient
Picture Word Fluency	28	0.39	0.94	0.8	0.73	0.37	0.1	0.42
Sentence Maze	10	0.45	0.96	0.87	0.77	0.47	0.1	0.51

Table 30—First Grade Generalization Statistics: Experiment 1 Cut Scores Applied to Experiment 2 Data Set

CBM Measure	Cut Score	Positive Predictive Power	Negative Predictive Power	Sensitivity	Specificity	Kappa	Standard Error of Kappa	Phi Coefficient
Sentence Maze	11	0.36	0.96	0.87	0.67	0.35	0.09	0.41
Picture Word Fluency	24	0.45	0.91	0.6	0.84	0.39	0.12	0.4

First Grade Generalization Statistics: Experiment 2 Cut Scores Applied to Experiment 1 Data Set

CBM Measure	Cut Score	Positive Predictive Power	Negative Predictive Power	Sensitivity	Specificity	Kappa	Standard Error of Kappa	Phi Coefficient
Sentence Maze	10	0.84	0.77	0.8	0.81	0.61	0.09	0.62
Picture Word Fluency	28	0.81	0.92	0.95	0.73	0.69	0.08	0.7

6.2.3 Third Grade Reliability and Validity Analysis

- Reliability. Of the 109 third grade students who participated in the study, 30 were selected at random for inclusion in the reliability sample. Internal consistency reliability proceeded with split-half reliability analysis (Crocker & Algina 1986). For this, each response opportunity on the group-administered CBM assessments was dichotomously coded into a statistical database using "1" to indicate a correct response and "0" to indicate an incorrect response. The author then created two forms for the screener by assigning all odd-numbered items to form 1 and all even-numbered items to form 2.

Correlation between forms was .97 for paragraph maze, .98 for sentence maze and .99 for picture word fluency. Guttman Split-Half coefficient were .98 for paragraph maze and .99 for sentence maze and picture word fluency. A comparison of both forms yielded correlation coefficients of .983 for paragraph maze, .991 for sentence maze and .996 for picture word fluency.

- Concurrent Validity. Bivariate correlation coefficients, presented in Table 31, were computed between DORF, paragraph maze, sentence maze, picture word fluency, and *i*LEAP to show association among these variables. Pearson correlations among all CBM measures achieved statistical significance ($p < .01$). The author then employed the Steiger's Z-test for correlated-correlations (Meng, 1992) to examine whether the correlations between individual CBM assessments and *i*LEAP were stronger than others. Paragraph maze showed the strongest correlation with *i*LEAP, surpassing both sentence maze and iLEAP ($Z = 3.01$, $p<.01$) and picture word fluency and *i*LEAP ($Z = 1.88$, $p<.05$) but similar to the association seen between DORF and *i*LEAP ($Z = 0.48$, $p>.05$). The association between DORF and *i*LEAP was stronger than sentence maze and *i*LEAP ($Z = 2.96$, $p<.01$) but not picture word fluency and *i*LEAP ($Z = 1.53$, $p>.05$). Correlations between sentence maze and *i*LEAP and picture word fluency and *i*LEAP were similar ($Z=.92$, $p>.05$). The same relationship among variables was seen in Experiment 1.

Table 31—Third Grade Bivariate Correlations Between Target Variables

Variable	Variable n	2	3	4	5
1. DORF	109	0.75	0.71	0.55	0.53
2. Paragraph Maze	109		0.63	0.54	0.56
3. Sentence Maze	109			0.61	0.34
(table 31 continued)					
4. Picture Word Fluency	109				0.41
5. *i*LEAP	109				

Note. All coefficients are significant ($p < .01$)

- Predictive Validity. Individual analyses were computed using the *i*LEAP as the criterion measure. Regression equations showed that all CBM(s) accounted for a significant portion of the variance in *i*LEAP performance. Paragraph maze and DORF explained the most variance in student's *i*LEAP performance at 31% (F (1, 107) = 48.04, $p < .01$) and 28% (F (1, 107) = 41.75, $p < .01$). Picture word fluency accounted for 17% of the variance in student's MCT performance (F (1, 107) = 21.55, $p < .01$) and sentence maze accounted for 11% (F (1, 107) = 13.67, $p < .01$). The CBM measures in Experiment 2 accounted for the same portion of variances as they did in Experiment 1.

The relative contributions of the various CBM measures were then examined as predictors of *i*LEAP performance. Simultaneous multiple regressions are presented in Table 32 showing the relative influence of the various CBM measures for predicting *i*LEAP performance. In part I, a simultaneous multiple regression utilizing DORF, paragraph maze, sentence maze, and picture word fluency was computed with the goal of evaluating the additional contribution of each group-administered CBM assessment over ORF alone. In part II, a simultaneous multiple regression with paragraph maze, sentence maze, and picture word fluency investigated which CBM assessment explained the most variance. Part I of Table 32 shows that all four variables together accounted for 37% of the variance F (4, 104) = 15.21, $p<.01$). Paragraph maze alone accounted for 5% of the variance in student's *i*LEAP performance. DORF accounted for 4% of the variance in student's *i*LEAP performance. Sentence maze and picture word fluency each accounted for an additional 2% of the variance in student's *i*LEAP performance, yet failed to yield significant beta weights and squared semipartial correlations. Part II of Table 32 shows simultaneous multiple regression analyses directly comparing the relative influence of paragraph maze, sentence maze, and picture word fluency together in predicting *i*LEAP performance. This analysis indicated that used together all three predictors accounted for 33% of the variance in student's *i*LEAP performance F (3, 105) = 17.43, $p<.01$). Paragraph maze alone accounted for 15% of the variance in student's *i*LEAP performance. Contribution from sentence maze and picture word fluency failed to achieve statistical significance, each yielding insignificant beta weights and squared

semipartial correlations. These findings were similar to those obtained in Experiment 1, particularly with respect to student performance on paragraph maze and ORF. In Experiment 1, with all four measures utilized, picture word fluency also contributed significantly to prediction, a phenomenon not seen in Experiment 2. Furthermore, in Experiment 1, when only group-administered CBM measures were used, paragraph maze and picture word fluency contributed significantly. Only paragraph maze and DORF contributed significantly in Experiment 2. When only group-administered CBM measures were used, only paragraph maze contributed significantly to prediction, with the other group-administered CBM measures accounting for no additional unique variance.

Table 32—Third Grade Simultaneous Multiple Regression Analyses Comparing the Utility of CBM Assessments in Predicting *i*LEAP Performance

	GRADE 3	
	Standard	
I. Predictors	β	sr^2
DORF	0.336	0.04*
Paragraph Maze	0.36	0.05**
Sentence Maze	-2.27	0.02
Picture Word Fluency	0.17	0.02
		R^2 = .37
		Adjusted R^2 = .35

(table 32 continued)

II. Predictors		
Paragraph Maze	0.52	0.15**
Sentence Maze	-0.11	0.01
Picture Word Fluency	0.2	0.02
		R^2 = .33
		Adjusted R^2 =.31

*p<.05.

**p<.01.

Hierarchical multiple regression coefficients, presented in Table 33, were computed to evaluate the added predictability of the group-administered CBM measures compared to DORF with *i*LEAP as the criterion measure. First, findings were examined whereby group-administered CBM measures alone were used to predict *i*LEAP performance. ORF was then entered to see if it accounted for any additional variance after the group-administered CBM measures. Second, findings were examined where by ORF was used solely to predict performance on the *i*LEAP. After that, the group-administered CBM measures were entered to see if they explained any additional variance. In the first analysis, paragraph maze was entered first, followed by picture word fluency, sentence maze, and DORF. Paragraph maze accounted for 31% of the variance. The addition of picture word fluency and sentence maze in Steps 2

and 3 showed only negligible increases in predictability. The addition of DORF in Step 4, however, was significant, with all four predictors accounting for 37% of the variance in the student's *i*LEAP performance. In the second set of analysis, DORF was entered at Step 1 and accounted for 28% of the variance in a student's *i*LEAP performance. Paragraph maze was entered in Step 2, resulting in a significant increase, with both variables combined accounting for 58% of the variance. The addition of picture word fluency and sentence maze in Steps 3 and 4 failed to add significantly to the variance accounted for. The same trend in Experiment 2 was seen in Experiment 1 in terms of the unique variance accounted for by paragraph maze and ORF. In Experiment 1, however, both paragraph maze and picture word fluency were significant contributors among the group-administered CBM measures. Of the group-administered CBM assessments in Experiment 2, no other variables accounted for significant variance after paragraph maze entered the equation. In Experiment 1, after the inclusion of ORF, additional variance was accounted for by both paragraph maze and picture word fluency. In Experiment 2, only paragraph maze contributed significantly to prediction after DORF.

Table 33—Third Grade Hierarchical Multiple Regressions Analyses Comparing the Utility of Group-Administered CBM Versus ORF in Predicting *i*LEAP Performance

Predictor		*R*	*R* Square	Adjusted *R* Square	R^2 Change	*F* Change
	Paragraph Maze	0.56	0.31	0.3	0.31	48.04**
Step 2						
	Picture Word Fluency	0.57	0.33	0.31	0.02	2.67
Step 3						
	Sentence Maze	0.57	0.33	0.31	0.01	0.88
Step 4						
	DORF	0.61	0.37	0.35	0.04	6.03*

(table 33 continued)

Step 1						
	DORF	0.53	0.28	0.27	0.28	41.71**
Step 2						
	Paragraph Maze	0.58	0.34	0.33	0.06	9.47**
Step 3						
	Picture Word Fluency	0.59	0.35	0.33	0.01	1.25
Step 4						
	Sentence Maze	0.61	0.37	0.35	0.02	3.56

*$p<.05$
**$p<.01$

- Classification Accuracy. Individual logistic regression equations were computed to determine the usefulness of the CBM assessments for categorical prediction. This analysis used student's *i*LEAP performance to indicate a positive or a negative problem. A student was categorized as having a problem on the *i*LEAP when his/her performance was indicated to be below "basic." When a student's performance on the *i*LEAP indicated "basic" or above, the student was categorized as negative for a problem on *i*LEAP. Of the 109 students who entered into the logistic regression analysis, 36 students were below "basic" on the *i*LEAP. Table 34 presents classification accuracy statistics determined according to the number of students whose performance was below "basic" according to their *i*LEAP performance as predicted by the student's scores on the CBM assessments.

Categorical prediction was examined by three sets of analyses. First, findings were examined in terms of variance accounted for by each individual predictor. Second, findings were examined in terms of goodness-of-fit between a student's actual performance on the *i*LEAP and predicted student performance per each CBM assessment. Third, findings were examined in terms of classification accuracy with attention devoted to each CBM assessment's sensitivity, specificity, positive predictive

power, and negative predictive power. In the first set of logistic analysis, all CBM assessments, with the exception of sentence maze, accounted for a significant portion of the variance in the student's *i*LEAP performance. Paragraph maze was the strongest predictor, accounting for 25% of the variance in student *i*LEAP performance. DORF accounted for 22%, while picture word fluency accounted for 7% of the variance in *i*LEAP performance. Sentence maze accounted for 4% of the variance in student *i*LEAP performance. The second set of analysis utilized the Hosmer and Lameshow chi-square test showing the goodness-of-fit between the actual and predicted student performance. All CBM measures achieved a good fit with the exception of sentence maze. This analysis showed paragraph maze achieving Hosmer and Lameshow (χ^2 (8, N = 109) = 8.64, p > .37). DORF achieved [Hosmer and Lameshow χ^2 (8, N = 109) = 4.26, p > 0.83], and picture word fluency achieved [Hosmer and Lameshow χ^2 (8, N = 109) = .149, p > .149]. Sentence maze achieved a significant (poor fit) Hosmer and Lameshow χ^2, showing that there was a significant difference between the predicted and observed number of cases whose performance was low on the *i*LEAP.

The third set of analyses investigated the classification accuracy of ORF, paragraph maze, sentence maze, and picture word fluency to evaluate the extent of correct categorical classification of students into two groups; i.e., those at risk and those not at risk on the iLEAP ORF sensitivity was .36 and specificity was .89. Positive and negative predictive power for ORF were .62 and .74, respectively. Paragraph maze sensitivity was .50 and specificity was .88. Paragraph maze positive and negative predictive power was .67 and .78. For sentence maze sensitivity and specificity were .11 and 1.00, respectively, whereas positive and negative predictive power were 1.0 and .69. Picture word fluency sensitivity was .14 and specificity was .98. Positive and negative predictive power for picture word fluency were .83 and .71. Classification accuracy in Experiment 2 was similar to that shown in Experiment 1; that is, all CBM measures tended to show greater specificity than sensitivity. Similar to Experiment 1, paragraph maze in Experiment 2 was one of the top 2 predictors. Whereas picture word

fluency accounted for the most variance among the predictors in Experiment 1, the percentage of variance accounted for in Experiment 2 was lower.

Table 34—Third Grade Logistic Regression for Classification Accuracy of CBM Assessments

Diagnostic Statistics for Cut Scores

CBM Measure	Positive Predictive Power	Negative Predictive Power	Sensitivity	Specificity	Kappa	Standard Error of Kappa	Phi Coefficient
ORF	0.62	0.73	0.36	0.89	0.28	0.1	0.3
Paragraph Maze	0.67	0.78	0.5	0.88	0.4	0.09	0.41
Sentence Maze	1	0.7	0.11	1	0.14	0.07	0.28
Picture Word Fluency	0.83	0.71	0.15	0.99	0.17	0.08	0.27

- Screening Accuracy. A two-step process for generating cut scores for screening accuracy was modeled after procedures described by Siberglitt and Hintze (2005) using a combination of logistic regression and ROC curve analysis. DORF was utilized as the criterion measure for this analysis using a cut score of 110 words correct per minute as proposed by DIBELS end-of-year benchmarks. In Step 1, logistic regression was used for category prediction ("at risk" or "not at risk") to identify overall classification rate and the probability of various errors. Step 2 used ROC analysis, which yielded hit rate juxtaposed to the likelihood of false positives to determine the final cut scores. In Step 1, the logistic regression showed sensitivity for paragraph maze and picture word fluency at 0.77. Sensitivity for sentence maze was 0.83. Specificity was 0.76 for paragraph maze and 0.51 for picture word fluency. Sentence maze specificity was 0.58. Table 35 shows the coordinates of the curve for Step 2, whereby the ROC analysis was used to generate the final set of decision rules for paragraph maze, sentence maze, and picture word fluency, with paramount importance geared for sensitivity. This analysis showed that sensitivity and specificity for paragraph maze was 0.85 and 0.71, respectively; sensitivity and specificity for sentence maze was 0.76 and 0.67, respectively whereas sensitivity and specificity for picture word fluency was 0.77 and 0.51, respectivly. Table 36 shows the resultant cut scores for each measure,

accompanied by additional information about their predictive accuracy. Summary diagnostic statistics are presented in Table 37.

The cut scores that were generated in Experiment 2 were similar to those generated in Experiment 1 for all CBM measures. Table # 38 presents the cut scores generated in Experiment 1 but applied to the data set used in Experiment 2. Juxtaposed to these findings are the cut scores generated in Experiment 2 but applied to the data set in Experiment 1. These findings show that group-administered CBM assessments will identify many of the same students that were identified as at risk on ORF/DORF. Paragraph maze or sentence maze appears to be the most useful of the group-administered CBM measures for this purpose according to Table 38.

Table 35—Third Grade ROC Analysis

Coordinates of the Curve

Paragraph Maze

Positive if Less Than or Equal To	Sensitivity	1 – Specificity
3	0	0
5	0.015152	0
6.5	0.045455	0
7.5	0.106061	0
8.5	0.121212	0
9.5	0.166667	0
10.5	0.227273	0
11.5	0.287879	0
12.5	0.409091	0.022222
13.5	0.439394	0.022222
14.5	0.5	0.022222

(table 35 continued)

15.5	0.575758	0.066667
16.5	0.666667	0.066667
17.5	0.712121	0.155556
18.5	0.742424	0.222222
19.5	0.772727	0.244444
20.5	0.848485	0.288889
21.5	0.924242	0.444444
22.5	0.984848	0.555556
23.5	0.984848	0.622222
24.5	0.984848	0.688889
25.5	1	0.733333
26.5	1	0.777778
27.5	1	0.822222
28.5	1	0.844444
29.5	1	0.933333
33.5	1	0.977778
38	1	1

Coordinates of the Curve

Sentence Maze

Positive if Less Than or Equal To	Sensitivity	1 – Specificity
8	0	0
10	0.015152	0
11.5	0.030303	0
12.5	0.045455	0

(table 35 continued)

13.5	0.060606	0
14.5	0.075758	0
15.5	0.090909	0
16.5	0.136364	0
17.5	0.151515	0
18.5	0.166667	0
19.5	0.212121	0
20.5	0.242424	0.022222
21.5	0.30303	0.044444
22.5	0.378788	0.066667
23.5	0.424242	0.066667
24.5	0.439394	0.066667
25.5	0.606061	0.222222
26.5	0.681818	0.222222
27.5	0.757576	0.333333
28.5	0.833333	0.422222
29.5	0.878788	0.488889
30.5	0.924242	0.488889
31.5	0.969697	0.533333
32.5	0.969697	0.555556
33.5	0.969697	0.622222
34.5	0.984848	0.688889
35.5	1	0.755556
36.5	1	0.8
37.5	1	0.822222
38.5	1	0.866667
40	1	0.888889
43	1	0.933333

(table 35 continued)

47.5	1	0.955556
51	1	1

Coordinates of the Curve

Picture Word Fluency

Positive if Less Than or Equal To	Sensitivity	1 – Specificity
1	0	0
12	0.015152	0
23	0.030303	0
25	0.045455	0
27	0.075758	0
28.5	0.121212	0
29.5	0.136364	0
30.5	0.151515	0
31.5	0.227273	0
32.5	0.227273	0.022222
33.5	0.242424	0.066667
34.5	0.318182	0.088889
35.5	0.363636	0.133333
36.5	0.454545	0.133333
37.5	0.469697	0.133333
38.5	0.484848	0.133333
39.5	0.515152	0.133333
40.5	0.560606	0.177778
41.5	0.560606	0.244444

(table 35 continued)

42.5	0.575758	0.266667
43.5	0.651515	0.288889
44.5	0.666667	0.466667
46	0.772727	0.488889
47.5	0.818182	0.577778
48.5	0.848485	0.577778
49.5	0.863636	0.622222
50.5	0.878788	0.644444
51.5	0.893939	0.733333
52.5	0.909091	0.755556
53.5	0.924242	0.755556
54.5	0.939394	0.777778
55.5	0.969697	0.777778
56.5	0.984848	0.8
57.5	1	0.8
58.5	1	0.866667
59.5	1	0.888889
61	1	1

Table 36—Third Grade CBM Cut Scores and Predictive Accuracy

CBM Measure	Cut Score	Percent at or above cut score with satisfactory score on criterion measure	Percent below cut score with satisfactory score on criterion measure
Paragraph Maze	20	76%	81%
Sentence Maze	27	65%	77%
Picture Word Fluency	45	61%	70%

Table 37—Third Grade Diagnostic Statistics for Cut Scores

CBM Measure	Cut Score	Positive Predictive Power	Negative Predictive Power	Sensitivity	Specificity	Kappa	Standard Error of Kappa	Phi Coefficient
Paragraph Maze	20	0.81	0.76	0.85	0.71	0.57	0.08	0.57
Sentence Maze	27	0.77	0.65	0.76	0.67	0.42	0.09	0.42
Picture Word Fluency	45	0.7	0.61	0.77	0.51	0.29	0.09	0.29

Table 38—Third Grade Generalization Statistics: Experiment 1 Cut Scores Applied to Experiment 2 Data Set

CBM Measure	Cut Score	Positive Predictive Power	Negative Predictive Power	Sensitivity	Specificity	Kappa	Standard Error of Kappa	Phi Coefficient
Paragraph Maze	19	0.82	0.69	0.77	0.76	0.52	0.08	0.52
Sentence Maze	26	0.81	0.63	0.68	0.78	0.44	0.08	0.45
Picture Word Fluency	45	0.7	0.61	0.77	0.51	0.29	0.09	0.29

Third Grade Generalization Statistics: Experiment 2 Cut Scores Applied to Experiment 1 Data Set

CBM Measure	Cut Score	Positive Predictive Power	Negative Predictive Power	Sensitivity	Specificity	Kappa	Standard Error of Kappa	Phi Coefficient
Paragraph Maze	20	0.45	0.95	0.9	0.64	0.4	0.05	0.47
Sentence Maze	27	0.54	0.96	0.92	0.74	0.54	0.05	0.58
Picture Word Fluency	45	0.47	0.92	0.8	0.71	0.41	0.06	0.44

6.2.4 Fifth Grade Reliability and Validity Analysis

- Reliability. Of the 117 fifth grade students who participated in the study, 30 were randomly selected for inclusion in the reliability sample. Internal consistency reliability proceeded with split-half reliability analysis (Crocker & Algina 1986). For this, each response opportunity on the group-administered CBM assessments was dichotomously coded into a statistical database using "1" to indicate a correct response and "0" to indicate an incorrect response. The author then created two forms for the screener by assigning all odd-numbered items to form 1 and all even-numbered items to form 2.

Correlation between forms was .97 for paragraph maze and .95 for sentence maze and .99 for picture word fluency. Guttman Split-Half coefficients were .98 for paragraph maze and .99 for sentence maze and picture word fluency.

- Concurrent Validity. Bivariate correlation coefficients, presented in Table 39, were computed between ORF, paragraph maze, sentence maze, picture word fluency, and *i*LEAP to show association among these variables. Pearson correlations among all CBM measures achieved statistical significance ($p \leq .01$). The author then employed the Steiger's Z-test for correlated-correlations (Meng, 1992) to examine whether the correlations between individual CBM assessments and *i*LEAP were stronger than others. Correlations between ORF and *i*LEAP and paragraph maze and *i*LEAP were significantly stronger than those between sentence maze and *i*LEAP and picture word fluency and *i*LEAP. Difference tests showed that correlation between ORF and iLEAP and paragraph maze and iLEAP ($Z = 1.31$, $p>.01$) failed to achieve statistical significance. Difference tests also showed that correlation between sentence maze and *i*LEAP and picture word fluency and *i*LEAP ($Z = .88$, $p>.05$) failed to achieve statistical significance. Experiment 2 again showed paragraph maze to have a strong relationship to reading performance as measured by the state accountability test.

Table 39—Fifth Grade Bivariate Correlations Between Target Variables

Variable	n	2	3	4	5
1. ORF	117	0.8	0.77	0.59	0.68
2. Paragraph maze	117		0.79	0.6	0.63
3. Sentence Maze	117			0.55	0.52
4. Picture Word Fluency	117				0.46
5. *i*LEAP	117				

Note. All coefficients are significant ($p < .01$)

- Predictive Validity. Individual regression coefficients were computed using the *i*LEAP as the criterion measure. Regression equations showed that all CBM measures accounted for a significant portion of the variance in *i*LEAP performance. ORF performance explained nearly half (47%) of the variance in *i*LEAP performance (F (1, 115) = 100.17, $p < .01$). Paragraph maze explain 39% of the variance in *i*LEAP performance (F (1, 115) = 74.12, $p < .01$). Sentence maze accounted for 27% of the variance in *i*LEAP performance (F (1, 115) = 42.08, $p < .01$), whereas picture word fluency explained 21% (F (1, 115) = 30.26, $p < .01$).

The relative contribution of the various CBM measures were then examined as predictors of iLEAP performance. The simultaneous multiple regressions are presented in Table 40. In Part I a simultaneous multiple regression utilizing ORF, paragraph maze, sentence maze, and picture word fluency was computed with the goal evaluating the additional contribution of each group-administered assessment over ORF alone. In Part II, a simultaneous multiple regression with paragraph, sentence maze, and picture word fluency investigated which CBM assessment explained the most variance. Part I of Table 40 shows that all four variables accounted for 49% of the variance (F (4, 112) = 26.57. $p<.01$). In this analysis, ORF uniquely accounted for 8% of the variance in predicting *i*LEAP performance and achieved statistical significance. Paragraph maze accounted for an additional 2% of the variance over and beyond that of ORF but yielded insignificant beta weights and squared semipartial correlations. Neither sentence maze nor picture word fluency added significantly to explained variation of the criterion. Part II of Table 40 shows simultaneous multiple regressions directly comparing the unique influence of paragraph maze, sentence maze and picture word fluency in predicting *i*LEAP performance. This analysis indicated that all three variables together accounted for 40% of the variance in *i*LEAP performance (F (3, 113) = 25.58. $p<.01$). In this analysis, paragraph maze exclusively accounted for 9% of the variance in *i*LEAP performance. Picture word fluency uniquely accounted for an additional 1% of the variance but yielded insignificant beta weights and squared semipartial correlations. Sentence maze added little to prediction. In Experiment 2, with all variables entered simultaneously, ORF

exclusively accounted for the most variance in *i*LEAP score over and beyond that of the other CBM measures. This was in contrast to Experiment 1, where paragraph maze showed as the major predictor of the MCT score. Of the group-administered CBM measures in Experiment 2, paragraph maze alone accounted for the majority of the variance, whereas in Experiment 1 predictability was partially shared with picture word fluency.

Table 40—Fifth Grade Simultaneous Multiple Regression Analyses Comparing the Utility of CBM Assessments in Predicting *i*LEAP Performance

	GRADE 3	
I. Predictors	Standard β	sr^2
ORF	0.54	0.08**
Paragraph Maze	0.24	0.02
Sentence Maze	-0.1	0
Picture Word Fluency	0.48	0
		R^2 = .49
		Adjusted R^2 =.47
II. Predictors		
Paragraph Maze	0.5	0.09**
Sentence Maze	0.08	0
Picture Word Fluency	0.36	0.01
		R^2 = .40
		Adjusted R^2 =.39

*p<.05.
**p<.01.

Hierarchical multiple regression coefficients, presented in Table 41, were computed to evaluate the added predictability of the group-administered CBM measures compared to ORF with *i*LEAP as the criterion measure. First, the author examined findings whereby group-administered CBM measures were used to predict *i*LEAP performance. DORF was then entered into the equation last to determine if it contributed significantly to explained variation in *i*LEAP over and above the group-administered CBM measures. Second, the author examined findings whereby ORF was used solely to predict performance on *i*LEAP. The group-administered CBM measures were then entered into the equation to determine if the would contribute significantly to explained variation in the *i*LEAP. In the first analysis, paragraph maze was entered first, followed by sentence maze, picture word fluency, and ORF. Paragraph maze accounted for 39% of the variance. The addition of sentence maze and picture word fluency in Steps 2 and 3 was insignificant; however, the addition of ORF in Step 4 significantly improved prediction, with all four variables accounting for nearly half of the variance. In the second set of analysis, ORF was entered first followed by paragraph maze, sentence maze, and picture word fluency. ORF accounted for 47% of the variance in *i*LEAP performance. The addition of paragraph maze, sentence maze, and picture word fluency failed to significantly increase prediction. This differed from the results shown in Experiment 1, where, after utilizing paragraph maze and picture word fluency, sentence maze and ORF failed to significantly change predictability. In Experiment 2, paragraph maze alone accounted for 47% of the variance. Picture word fluency and sentence maze produced negligible changes in R^2 but the addition of ORF resulted in a significant increase in proportion of variance accounted for. Overall, a notable difference in the results for Experiment 2 was that when ORF entered the equation first, other variables subsequently entered had a negligible effect in accounting for explained variance.

Table 41—Fifth Grade Hierarchical Multiple Regression Analyses Comparing the Utility of Group-Administered CBM Versus ORF in Predicting *i*LEAP Performance

					Change Statistics	
Predictor		*R*	*R* Square	Adjusted *R* Square	R^2 Change	*F* Change
Step 1						
	Paragraph Maze	0.63	0.39	0.39	0.39	74.12**
Step 2						
	Sentence Maze	0.63	0.4	0.39	0.01	0.954
Step 3						
	Picture Word Fluency	64	0.4	0.39	0.01	1.42
Step 4						
	ORF	0.7	0.49	0.47	0.08	17.99**
Step 1						
	ORF	0.68	0.47	0.46	0.47	100.17**
Step 2						
	Paragraph Maze	0.7	0.48	0.47	0.02	3.72
Step 3						
	Sentence Maze	0.7	0.49	0.47	0.003	0.69
Step 4						
	Picture Word Fluency	0.7	0.49	0.47	0.001	0.3

***p<.01*
**p<.05*

- Classification Accuracy. Individual logistic regression equations were computed to determine the usefulness of the CBM assessments for categorical prediction. This analysis used student's *i*LEAP performance to indicate a positive or negative problem. A student was categorized as having a problem on the *i*LEAP when his/her performance was indicated to be below "basic." When a student's

performance on the *i*LEAP indicated "basic" or above, the student was categorized as negative for a problem on *i*LEAP. Of the 117 students who entered into the logistic regression analysis, 39 student performances were below "basic" on the *i*LEAP. Table 42 summarizes the number of students whose performance was below basic according to their *i*LEAP performance as predicted by the student's scores on the CBM assessment.

Categorical prediction was examined by three sets of analyses. First, findings were examined in terms of variance accounted for by each individual predictor. Second, findings were examined in terms of goodness-of-fit between a student's actual performance on the *i*LEAP and their predicted performance per each CBM assessment. Third, findings were examined in terms of classification accuracy, with attention devoted to each CBM assessment's sensitivity, specificity, positive predictive power, and negative predictive power. In the first set of logistic analysis, all CBM assessments accounted for a significant portion of the variance in *i*LEAP performance. ORF accounted for 31% of the variance in *i*LEAP performance. Paragraph maze accounted for 25% of the variance in *i*LEAP performance. Sentence maze accounted for 18%, while picture word fluency accounted for 12% of the variance in *i*LEAP performance. The second set of analysis utilized the Hosmer and Lameshow chi-square test showing the goodness-of-fit between actual and predicted student performance. Analysis indicated that ORF achieved a Hosmer and Lameshow (χ^2 (8, N = 117) = 5.98, p > .05), paragraph maze achieved Hosmer and Lameshow (χ^2 (8, N = 117) = 10.78, p > .05), sentence maze achieved Hosmer and Lameshow (χ^2 (8, N = 117) = 4.77, p > .05), and picture word fluency achieved a Hosmer and Lameshow (χ^2 (7, N = 117) = 15.33, p > .05) of the variance in the students *i*LEAP performance.

The third set of analysis investigated the classification accuracy of ORF, paragraph maze, sentence maze, and picture word fluency to evaluate the extent of correct categorical classification of students into two groups; i.e., those at risk and those not at risk on the *i*LEAP. ORF sensitivity was .49 and specificity was .87. Positive and negative predictive power for ORF were .66 and .77, respectively. Paragraph maze sensitivity was .46 and specificity was .85. Paragraph maze positive and negative

predictive power was .60 and .76. For sentence maze, sensitivity and specificity were .28 and .90, respectively, whereas positive and negative predictive power were .58 and .77. Picture word fluency sensitivity was .26 and specificity was .99. Positive and negative predictive power for picture word fluency were .71 and .72. The same general pattern was also seen in Experiment 1 in terms of each CBM measures utility for prediction.

Table 42—Third Grade Logistic Regression for Classification Accuracy of CBM Assessments

CBM Measure	Positive Predictive Power	Negative Predictive Power	Sensitivity	Specificity	Kappa	Standard Error of Kappa	Phi Coefficient
ORF	0.66	0.77	0.49	0.87	0.38	0.09	0.39
Paragraph Maze	0.6	0.76	0.46	0.85	0.33	0.09	0.33
Sentence Maze	0.58	0.71	0.28	0.9	0.21	0.09	0.23
Picture Word Fluency	0.71	0.72	0.26	0.95	0.24	0.09	0.3

- Screening Accuracy. A two-step process for generating cut scores for screening accuracy was modeled after procedures described by Siberglitt and Hintze (2005) using a combination of logistic regression and ROC curve analysis. ORF was utilized as the criterion measure for this analysis using a cut score of 100 words correct per minute as proposed by Hasbrouk and Tindal, (1992). Logistic regression was used for category prediction ("at risk" or "not at risk") to identify overall classification rate and the probability of various errors. Step 2 used ROC analysis, which yielded hit rate juxtaposed to the likelihood of false positives to determine the final cut scores. In Step 1, logistic regression showed sensitivity and specificity for paragraph maze at 0.35 and 0.96. Sensitivity for sentence maze and picture word fluency was 0.39, whereas specificity corresponded to 0.94 and 0.97. Table 43 shows the coordinates of the curve for Step 2, whereby the ROC analysis was used to generate the final set of decision rules for paragraph maze, sentence maze and picture word fluency with paramount importance geared for sensitivity. This analysis showed paragraph maze sensitivity and specificity at 0.83. Sensitivity and specificity for sentence maze was 0.83 and 0.79. For picture word fluency sensitivity and specificity was 0.83 and 0.71. Table 44 shows the resultant cut scores for each measure

accompanied by additional information about their predictive accuracy. Useful diagnostic statistics are presented in Table 45.

The cut scores that were generated in Experiment 2 were similar to those generated in Experiment 1, particularly for sentence maze. Table 46 presents the cut scores generated in Experiment 1 but applied to the data set used in Experiment 2. Juxtaposed to these findings are the cut scores generated in Experiment 2 but applied to the data set in Experiment 1. These findings both paragraph maze or sentence maze may be the most suitable of the group-administered CBM measures for this purpose according to Table 46.

Table 43—Fifth Grade ROC Analysis

Coordinates of the Curve

Paragraph Maze

Positive if Less Than or Equal To	Sensitivity	1 - Specificity
1	0	0
4.5	0.043478	0
7.5	0.086957	0
8.5	0.130435	0
10	0.173913	0
11.5	0.26087	0
12.5	0.26087	0.009524
13.5	0.26087	0.028571
14.5	0.347826	0.038095
15.5	0.434783	0.066667
16.5	0.565217	0.085714
17.5	0.652174	0.142857
18.5	0.826087	0.171429

(table 43 continued)

19.5	0.826087	0.2
20.5	0.913043	0.257143
21.5	0.913043	0.32381
22.5	0.913043	0.361905
23.5	0.956522	0.438095
24.5	1	0.533333
25.5	1	0.580952
26.5	1	0.638095
27.5	1	0.666667
28.5	1	0.67619
29.5	1	0.714286
30.5	1	0.733333
31.5	1	0.742857
32.5	1	0.819048
34	1	0.828571
35.5	1	0.847619
36.5	1	0.885714
37.5	1	0.904762
38.5	1	0.933333
40	1	0.952381
42.5	1	0.961905
45.5	1	0.980952
54	1	0.990476
62	1	1

(table 43 continued)
Coordinates of the Curve

Sentence Maze

Positive if Less Than or Equal To	Sensitivity	1 - Specificity
7	0	0
8.5	0.043478	0
9.5	0.086957	0
11.5	0.130435	0
14	0.173913	0
15.5	0.304348	0.019231
16.5	0.391304	0.057692
17.5	0.521739	0.067308
18.5	0.608696	0.096154
19.5	0.652174	0.163462
20.5	0.826087	0.211538
21.5	0.869565	0.307692
22.5	0.956522	0.346154
23.5	1	0.432692
24.5	1	0.509615
25.5	1	0.605769
26.5	1	0.625
27.5	1	0.673077
28.5	1	0.711538
29.5	1	0.75
30.5	1	0.788462
31.5	1	0.836538
32.5	1	0.865385

(table 43 continued)

33.5	1	0.875
35	1	0.894231
36.5	1	0.923077
38.5	1	0.932692
42.5	1	0.951923
46	1	0.961538
48.5	1	0.980769
51	1	1

Coordinates of the Curve

Picture Word Fluency

Positive if Less Than or Equal To	Sensitivity	1 - Specificity
12	0	0
15	0.043478	0
18	0.086957	0
19.5	0.173913	0
22.5	0.26087	0.009524
25.5	0.26087	0.019048
26.5	0.304348	0.019048
27.5	0.347826	0.028571
28.5	0.391304	0.028571
29.5	0.434783	0.047619
30.5	0.478261	0.057143
31.5	0.478261	0.07619
32.5	0.652174	0.095238
33.5	0.652174	0.104762

(table 43 continued)

34.5	0.695652	0.142857
35.5	0.695652	0.161905
36.5	0.695652	0.180952
37.5	0.73913	0.2
38.5	0.73913	0.238095
39.5	0.826087	0.285714
40.5	0.869565	0.314286
41.5	0.869565	0.342857
42.5	0.869565	0.390476
43.5	0.869565	0.419048
44.5	0.869565	0.466667
45.5	0.869565	0.52381
46.5	0.869565	0.6
47.5	0.913043	0.647619
48.5	0.956522	0.685714
49.5	0.956522	0.72381
50.5	1	0.72381
51.5	1	0.752381
52.5	1	0.819048
53.5	1	0.828571
54.5	1	0.857143
55.5	1	0.87619
56.5	1	0.885714
57.5	1	0.904762
59	1	0.942857
61	1	0.952381
63.5	1	0.980952
66	1	1

Table 44—Fifth Grade CBM Cut Scores and Predictive Accuracy

CBM Measure	Cut Score	Percent at or above cut score passing the criterion measure	Percent below cut score on criterion measure failing the criterion measure
paragraph maze	18	98%	44%
sentence maze	20	95%	46%
picture word fluency	39	95%	39%

Table 45—Fifth Grade Diagnostic Statistics for Cut Scores

CBM Measure	Cut Score	Positive Predictive Power	Negative Predictive Power	Sensitivity	Specificity	Kappa	Standard Error of Kappa	Phi Coefficient
Paragraph Maze	18	0.51	0.96	0.83	0.83	0.53	0.09	0.55
Sentence Maze	20	0.46	0.95	0.83	0.79	0.47	0.08	0.51
Picture Word Fluency	39	0.39	0.95	0.83	0.71	0.37	0.08	0.43

Table 46—Fifth Grade Generalization Statistics: Experiment 1 Cut Scores Applied to Experiment 2 Data Set

CBM Measure	Cut Score	Positive Predictive Power	Negative Predictive Power	Sensitivity	Specificity	Kappa	Standard Error of Kappa	Phi Coefficient
Paragraph Maze	22	0.36	0.97	0.91	0.64	0.34	0.07	0.42
Sentence Maze	22	0.38	0.99	0.96	0.65	0.38	0.07	0.47
Picture Word Fluency	44	0.29	0.95	0.87	0.53	0.22	0.06	0.31

Fifth Grade Generalization Statistics: Experiment 2 Cut Scores Applied to Experiment 1 Data Set

CBM Measure	Cut Score	Positive Predictive Power	Negative Predictive Power	Sensitivity	Specificity	Kappa	Standard Error of Kappa	Phi Coefficient
Paragraph Maze	18	0.47	0.92	0.56	0.9	0.42	0.09	0.42
Sentence Maze	20	0.41	0.97	0.76	0.86	0.45	0.06	0.48
Picture Word Fluency	39	0.41	0.96	0.69	0.87	0.43	0.07	0.46

CHAPTER 7. DISCUSSION

The aim of this study was to examine the efficacy of screening methods in reading, which, because they are group administered, hold promise for being more efficient than those that are administered individually. In doing so, this study investigated the reliability and validity of three group-administered procedures for the screening of reading problems. Specifically studied were paragraph maze, sentence maze, and picture word fluency. Analyses sought both to establish the technical adequacy of these group-administered measures and to demonstrate their predictive utility. The study also sought to determine whether the results obtained in one population of study would generalize to another population of study using alternative criterion measures.

To evaluate the psychometric adequacy of the group screening measures, this study examined the properties of the assessments, which derive from classical measurement theory, as well as the utility and accuracy of the assessments for decision-making in the context of the requirements for universal screening. Overall, the data reported here corroborate findings from others sources (Fuchs & Fuchs, 1992; Ritchey & Speece, 2004), which suggest that group-administered screening measures show promise and, in most cases, produce outcomes which are roughly equivalent to individually administered ORF assessments. Two groups of validity analyses are reported here: (a) those pertaining to concurrent/predictive validity, and (b) those pertaining to classification accuracy.

Although the results from both types of analyses reported here are somewhat positive, the results of the classification accuracy analyses are more germane to the evaluation of screening measures. In a recent and incisive review of research on screening measures for reading, Jenkins, Hudson, and Johnson (2007) indicated:

> The key feature of a screening measure is its ability to accurately classify students as at risk or not at risk…Whereas criterion validity studies are informative in identifying measures that hold potential as screens, classification studies are the sine qua non of screening research.

It is thus possible for a screening measure to correlate very highly with an important criterion but be completely inadequate as a screening measure because of poor classification accuracy. For example,

Jenkins, et al. (2007) noted that DIBELS correlates well with state achievement tests but has a high false positive rate and identified an average of 51% of all students as at risk with some subtests identifying 71% of all students tested as at risk. A well-rounded screening measure, however, measures what it supposed to measure while maintaining respectable rates of classification accuracy. The importance of classification accuracy and the minimization of false negatives are especially important when screening at-risk students, because a false negative will result in an at-risk student not receiving necessary remedial educational services.

With this in mind, the classification accuracy results will be considered first, giving particular attention to the specificity of the measures, which refers to the accuracy with which the measures correctly identify a student as not at risk. In this study, specificity was consistently in the .70's and .80's across all three group-administered measures, used across all three grades in two separate states. The specificity of the group-administered screening measures are consistent with many studies of individually administered screening measures in reading reviewed by Jenkins et al (2007) and considerably better than the accuracy of rates of DIBELS reported by Hintz, Ryan and Stoner (2003). Sensitivity data, which refers to the accuracy with which a measures correctly indicate that a student is at risk, were also considered. These analyses yielded results across all measures and populations ranging form the upper .60's to upper .80's. These results are consistent with other studies involving individually administered screening measures.

In addition to the results pertaining to classification accuracy, the results were also examined with respect to the psychometric tradition of classical measurement theory. Split-half reliability coefficients for paragraph maze, sentence maze and picture word fluency were acceptable, and met commonly accepted standards for reliability in two states (Croker & Algina 1986). Reliability coefficients ranged from .95 to .99.

The analyses pertaining to criterion and/or concurrent validity suggested that group-administered measures hold some promise for universal screening. Overall, correlational analyses in two states with

different criterion measures suggested a moderate to strong association both between the group measures and ORF and between the group measures and the various criterion achievement tests. When ORF was classified as an independent variable along with the group measures, the results varied by grade.

In both first grade experiments, the bivariate correlations between all measures evaluated (ORF/DORF, sentence maze, and picture word fluency) and the criterion (GRADE) were all statistically significant. At first grade, the ORF/DORF correlations with the criterion measures were significantly higher. Furthermore, ORF/DORF accounted for unique variance in GRADE over and above the variance accounted for by sentence maze and picture word fluency, individually or in combination with each other. The correlations between ORF/DORF and the group measures at first grade were moderate to strong.

At third grade, the results of the predictive/concurrent validity analyses indicated a slightly different pattern of results. Although group measures (paragraph maze, sentence maze, and picture word fluency) and ORF had moderate to strong and statistically significant relationships with the criterion measures in both experiments, the percentage of variance accounted for by the various measures differed from the first grade data. Specifically, all variables in Experiment I, with the exception of sentence maze, made a unique contribution to explained variance in the criterion. In Experiment II, paragraph maze had the strongest degree of association with the criterion measure and also accounted for the most unique variance. DORF was the only other predictor, which accounted for unique variance beyond the variance accounted for by paragraph maze.

At fifth grade, all bivariate validity coefficients were statistically significant in both experiments. In Experiment I, paragraph maze was the best predictor of the criterion measure, and only ORF contributed significantly beyond paragraph maze to explained variance. In Experiment II, ORF was the best predictor of the criterion and none of the group-administered measures contributed significantly to explained variance over and above ORF.

Overall, this study extended previous research by others (e.g., Fuchs and Fuchs 1992) by adding to the literature pertaining to the validity and reliability of group-administered CBM for the measurement of reading performance. Although additional research will be needed, this study suggests that group-administered measures hold promise as potential substitutes for ORF. While this study also strongly supports the use of ORF for screening, in some cases the use of group screening measures may be more efficient than ORF (Fuchs & Maxwell, 1988; Wesson, King, & Deno, 1984). This study adds to previous research that has investigated alternative screening methods for reading that are easy to use and valid for their intended use (Fuchs & Maxwell, 1988). This study continued this line of research by examining the three group-administered CBM alternatives by first examining their reliability and validity and then examining their predictive and screening accuracy, both necessary factors for greater acceptance and use (Jenkins, et. al, 2007; Ritchey & Speece, 2004).

Maze reliability estimates in this study were acceptable and similar to those shown in previous studies (Fuchs & Fuchs, 1992; Parker & Hasbrouck, 1992). Reliability estimates were also quite strong for sentence maze, and picture word fluency. Concurrent validity coefficients for these measures were also similar to those shown in previous research studies (Fuchs, Fuchs, Hamlett, & Ferguson, 1990; Jenkins & Jewell 1993) and in preliminary studies conducted with the present measures.

Although the question originally posed by this study was whether the group measures performed significantly different or better than ORF, one interesting aspect of the findings is the degree of similarity across the various measures. When examined from the perspective of classical measurement theory, the validity data for ORF and paragraph maze appeared to be slightly better than the other measures. However, in terms of classification accuracy, it is difficult to draw a clear distinction among the measures. This finding deserves additional research on the construct validity of various reading measures. Other studies have observed that ORF has a high degree of correlation with measures of reading comprehension (e.g. Fuchs, Fuchs, & Maxwell, 1988; Jenkins & Jewell 1993). Even studies of simple measures of reading such as reading a word list have yielded excellent validity coefficients with

more complex reading tasks. This raises the question, which perhaps can be addressed with future research, regarding the extent to which various types of reading assessment actually measure different constructs or components of reading.

Some aspects of the study delimit the findings and are noted. First, the classification accuracy aspect of the study depends upon known groups of students. In this case, there are students whose performance was at risk and those who were not at risk. However, there were a relatively low number of students who were truly low achieving that participated in this study. That is, there were very few at-risk students, as indicated by achievement testing. This limitation was especially apparent in first grade for both Experiment 1 and Experiment 2 and in third and fifth grades in Experiment 1. Future research is needed where samples have a larger proportion of students who are at risk.

A second limitation concerned the criterion variables used in third and fifth grade. The data collected in Experiment 1 utilized the MCT as the criterion variable. The data collected in Experiment 2 utilized *i*LEAP as the criterion variable. In an ideal generalization study, only the student population, and not the criterion, changes. It is possible that the results generalize across populations and measures, but future research will need to examine this more closely.

Much research remains to be done to fully understand the validity and efficiency of various screening measures. Researchers may want to replicate this study recognizing and improving upon the limitations noted above. Future researchers may also want to improve upon the criterion used in this study, as low scores on the GRADE, MCT and *i*LEAP do not represent an ideal definition of "at risk." Using additional measures and, perhaps, a student's response to intervention would provide a stronger definition of students who are truly at risk as opposed to those who are not.

A premise of this study was that group-administered measures hold the potential to be both psychometrically acceptable and more efficient than individually administered measures. This study focused on the technical properties of the group measures but did not evaluate their efficiency. Additional research will be needed to examine whether they are truly more efficient and whether they

are perceived by school based professionals as more efficient. Similarly, the "face validity" of paragraph maze and sentence maze for teachers can be evaluated.

In conclusion, early screening for the identification of reading problems plays a valuable role in identifying students who may be at risk for reading problems. If screening is conducted more efficiently without losing accuracy, it may increase the overall prevalence of screening, thereby maximizing prevention efforts while potentially minimizing the overhead associated with its conduct. This study provides preliminary evidence that paragraph maze, sentence maze and picture word fluency offer these benefits for screening purposes and warrant future consideration (Fuchs et al, 1988; Jenkins & Jewell, 1993; Shinn, Good, Knutson, Tilly, & Colline, 1992; Wesson, King, & Deno 1984). This study supports continued use of ORF as a valid measure of general reading performance for universal screening (Fewster & Macmillan, 2002; Hintz, 1998; Hintz, Conte, & Kristin, 1997; Shinn, Good & Roland, 1992). However, because ORF is administered individually, thus requiring considerably more time for data collection (Wesson, Fuchs, Tindal, Mirkin, & Deno, 1986) group-administered assessments, such as those studied here, may prove to be an alternative to ORF. In addition, sentence maze and paragraph maze may have additional face validity with teachers because they appear to be broader than ORF and represent an index of reading comprehension (Fuchs & Maxwell, 1988).

REFERENCES

Algozzine, B., Ysseldyke, J.E. & Christenson, S. (1983). An analysis of the incidence of special class placement: The masses are burgeoning. *The Journal of Special Education, 17 (2),* 141-147.

Allinder, R. M., & Oates, R.G. (1997). Effects of acceptability on teachers' implementation of curriculum based measurement and student achievement in mathematics computation. *Remedial and Special Education, 18.* 113-120.

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

Bocian, K., Beebe, M., MacMillan, D., & Gresham, F.M. (1999). Competing paradigms in leaning disabilities classification by schools and variations in meaning of discrepant achievement. *Learning Disabilities Research and Practice, 14,* 1-14.

Bormuth, J. R. (1968). Cloze test readability criterion referenced scores. *Journal of Educational Measurement, 5,* 189-196.

Bradley, J., Ackerson, G., & Ames, W. (1978). The reliability of the maze procedure for classroom assessment. *Journal of Educational Measurement, 5,* 291-296.

Bruning, R. (1985). Review of degrees of reading power. In J.V. MItchell (Ed.), *The ninth mental measurements yearbook* (pp. 442-444). Lincoln, NE: Buros Institute of Mental Measurements.

Carlberg, C., & Kavale, K. A. (1980). The efficacy of special versus regular education placement for exceptional children: A meta-analysis. *Journal of Special Education, 14,* 296-309.

Clay, M. (1987). Learning to be learning disabled. *New Zealand Journal of Educational Studies, 22,* 155-173.

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

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

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* New York: Harcourt Brace.

Data Recognition Corporation (2006). *2006 Operational Technical Report.* Submitted to the Louisiana Department of Education, *October, 2006.*

Deno, S. L. (1985). Curriculum-based measurement: the emerging alternative. *Exceptional Children, 52,* 219-231.

Deno, S.L. (2003). Developments in Curriculum-Based Measurement. *The Journal of Special Education*, 37(3), 184-192.

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

Deno, S. L., Mirken, P. K., & Chiang, B. (1982). Identifying valid measures of reading. *Exceptional Children, 49,* 86-45.

Donovan, M., & Cross, C. (2002). *Minority students in special and gifted education.* Washington, DC: National Academy Press.

DuBay, W. H. (2004). The principles of readability. *Online submission;* William H. DuBay; August 25, 2004.

Elliot, S. N., & Fuchs, L. S. (1997). The utility of curriculum-based measurement and performance assessment as alternative to traditional intelligence and achievement test. *School Psychology Review, 26*(2).

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

Fewster, S., McMillian, P. D., & Peter, D. (2002). School-based evidence for the validity of curriculum-based measurement of reading and writing. *Remedial and Special Education, 23*(3).

Fletcher, J., & Satz, P. (1984). Test-based versus teacher-based predictions of academic achievement: A three-year longitudinal study. *Journal of Pediatric Psychology, 9,* 193-203.

Fuchs, L. S., & Fuchs, D. (1998). Treatment Validity: A unifying concept for reconceptualizing the identification of learning disabilities. *Learning Disabilities Research and Practice, 13,* (4), 204-219.

Fuchs, L. S. (1986). Monitoring progress among mildly handicapped pupils: Review of current practice and research. *Remedial and Special Education, 7,* 5-12.

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

Fuchs, L. S., Fuchs, D., Campton, D. L. (2004). Monitoring Early Reading Development in First Grade: Word Identification Fluency Versus Nonsense Word Fluency. *Exceptional Children*, *71*, 7-21.

Fuchs, L. S., Fuchs, D., & Hamlett, C. L. (1989). Monitoring reading growth using student recalls: Effects of two teacher feedback *systems. Journal of Educational Research, 83,* 103-111.

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

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

Fuchs, L. S., Fuchs, D., Hosp, M. K., & Jenkins, J. R. (2001). Oral reading fluency as an indicator of reading competence: a theoretical, empirical, and historical analysis. *Scientific Studies of Reading, 5*(3), 239-256.

Fuchs, L. S., Fuchs, D., & Maxwell, L. (1988). The validity of informal reading comprehension measures. *Remedial and Special Education, 9,* 20-29.

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.

Gerber, M., & Semel, M. (1984). Teacher as imperfect test: Reconceptualizing the referral process. *Educational Psychologist*, *14,* 137-146.

Good , R. H., Simmons, D. C., & Kam'enui, E. J. (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*(3), 257-288.

Gresham, F.M. (2002). Response to Treatment. In Bradley, R., Danielson, L., and Hallahan, D., *Identification of Learning Disabilities: Research to Practice.* Washington, DC: USED.

Gresham, F.M., MacMillan, D., & Bocian, K. (1997). Teachers as "tests": Differential validity of teacher judgments in identifying students at-risk for learning difficulties. *School Psychology Review, 26,* 47-60.

Gresham, F.M., Reschly, D., & Carey, M. (1987). Teachers as "tests": Classification accuracy and concurrent validation in the identification of learning disabled children. *School Psychology Review. 16,* 543-563.

Gresham, F.M. Vanderheyden, A., & Witt, J.C. (2005) Response to intervention in the identification of learning disabilities: Empirical support ad future challenges. *Unpublished manuscript. Available online at* http://www.joewitt.org/downloads/responce

Gresham, F.M., & Witt, J.C. (1997). Utility of intelligence tests for treatment planning, classification, and placement decisions: Recent empirical findings and future directions. *School Psychology Quarterly, 12,*(3), 249-267.

Gutherie, J. T. (1973). Reading comprehension and syntactic responses in good and poor readers. *Journal of Educational Psychology, 65,* 294-300.

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

Hintze, J. M., Conte, K. L., Shapiro, E. S., & Basile, I. M. (1997). Oral reading fluency and authentic reading material: Criterion validity of the technical features of CBM survey-level assessment. *School Psychology Review, 26,* 535–553.

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.

Jenkins, J. R., Hudson, R. F., & Johnson, E. S. (2007). Screening for at-risk readers in a response to intervention framework. *School Psychology Review, 36*, 582-600.

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

Jongsma, E. (1971). *The cloze procedure as a teaching technique.* Newwark, DE: International Reading Association.

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

Kline, R. B. (1988). *Principles and practice of structural equation modeling.* New York: Guilford Press.

Langenberg, D., N. (1999). *Testimony before the U.S. Senate Appropriations Committee's Subcommittee on Labor, Health & Human Services, and Education,* Washington, DC.

Lyon, G.R. (1985). Learning Disabilities Research: False starts and broken promises. In S. Vaughn & C. Bos (Eds.), *Research in learning disabilities: Issues and future directions* (pp. 69-83). Boston: Little Brown.

Lyon, G. R., Fletcher, J. M., Shaywitz, S. E., Shaywitz, B. A., Torgenson, J. K., Wood, F. B., Schulte, A., & Olson, R. (2002). Rethinking learning disabilities. In *Rethinking special education for a new century.* C.E. Finn, A.J. Rotherham & C.R. Hokanson (Eds.) Washington, D. C.: Thomos B. Fordham and the Progressive Policy Institute.

MacMillan, D. L., Gresham, F. M., Siperstein, G. N., & Bocian, K. M. (1996). The labyrinth of I.D.E.A.: School decisions on referred students with subaverage general intelligence. American Journal on Mental Retardation, 101, 161-174.

MacMillan, D., Gresham, F.M., Bocian, K., & Siperstein, G. (1997). The role of assessment in qualifying students as eligible for special education: What is and what's supposed to be. *Focus on Exceptional Children, 33,* 83-94.

MacMillan, D., & Siperstien, G. (2002). Learning disabilities as operationally defined by schools. In Bradley, L. Danielson, & D. Hallahan (Eds.). *Identification of learning disabilities: Research to practice* (pp. 287-333). Mahwah, N.J.: Lawrence Erlbaum.

MacMillan, D., Siperstein, G., & Gresham, F. M. (1996), Mild mental retardation: A challenge to its viability as a diagnostic category. *Exceptional Children, 62,* 356-371.

Maheady, L., Algozzine, B., & Ysseldyke, J.E. (1984). Minority overrepresentation in special education: A functional assessment perspective. *Special Services in the Schools, 1(2)*, 5-19.Marshall, N. (1983). Using story grammar to assess reading comprehension. *The Reading Teacher, 36,* 616-620.

Marshall et al., v. Georgia. U.S. District Court for the Southern District of Georgia, CV 482-233, June 28, 1984.

McGlinchey, M. T., & Hixon, M. D. (2004). Using curriculum-based measurement to predict performance on state assessments in reading. *School Psychology Review, 33*(2), 193-204.

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

Mississippi Department of Education (2002). *Mississippi Curriculum Test: Summary of Technical Information*. Jackson, MS: Office of Research and Statistics, Mississippi Department of Education.

National Reading Panel, (2000). *Report of the National Reading Pane. Teaching children to Read: An evidenced-based assessment of the scientific research literature on reading and its implications for reading instruction (NIH Publications No. 00-4754.* Washington, DC U. S. Government Printing Office.

National Research Council. (2002) *Division of Behavioral and Social Sciences and Education, Minority Students in special and gifted education.* Washington, DC: National Academy Press.

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

Parker, R., & Hasbrouck, J. E. (1992). The maze as a classroom-based reading measure: Construction methods, reliability, and validity. *Journal of Special Education, 26*(2), 195-219.

Parker, R., Tindal, G., & Hasbrouk, J. (1989). Initial validation of two classroom-based measures of reading comprehension. *Diagnostique, 14(4),* 222-240.

Parker, R., Hasbrouck, J., & Tindal, G. (1989). *Combining informal teacher judgment and objective test scores to make class rooms reading group placement* (Resource Consultant Training Program research Rep. No. 4). Eugene, OR: University of Oregon, College of Education.

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

President's Commission on Excellence in Special Education. (2002). *A new era: revitalizing special education for children and their families.* Jessup, MD: ED Pubs.

Ritchey K. D. & Speece, D. L. (2004). Early identification of reading disabilities: current status and new directions. *Assessment for Effective Interventions.* 29 (4), 13-24.

Scarborough, H.S. (1998). Early identification of children at risk for reading disabilities: Phonological awareness and some promising predictors. In B. K. Shapiro, P. J. Accardo, & A. J. Capute (Eds.), *Specific reading disability: A view of the spectrum.*(pp. 75-119). Timonium, MD: New York Press.

Siegel, L. S. (1989). IQ is irrelevant to the definition of learning disabilities. *Journal of Learning Disabilities, 22,* 469−478.

Shaywitz, S.E. Fletcher, J.M, Holahan, J.M., Scheider, A.E., Marchione, K.E., Stuebing, K.K.,. Francis, D.J., & Shaywitz, B.A.(1999) "Persistency of dyslexia: The Conneaut Longitudinal Study at Adolescents," *Pediatrics 104*, 1357-1359.

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

Shinn, J., Deno, S. L., & Espin, C. (2000). Technical adequacy of the maze task for curriculum based measurement of reading growth. *The Journal of Special Education, 3*(34), 164-172.

Shinn, M. R., & Good III, R. H. (1992). Curriculum-based measurement of oral reading fluency: a confirmatory analysis or its relation to reading. *School Psychology Review, 21*(3).

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

Shinn, M. R., Tindal, G. A., & Spira, D. A. (1987). Special education referrals as an index of teacher tolerance: Are teachers imperfect tests? Exceptional Children, 54, 32-40.

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.

Simmons, D. C., Kame'ennui, E. J., Good III, R. H., Harn, B. A., Cole, C., Braun, D. Building, Implementing, and Sustaining a Beginning Reading Improvement Model: Lessons learned School by School. In: M R.Shinn, H. M. Walker, and G. Stoner *Interventions for Academic and Behavior Problems II: Preventative and Remedial Approaches (pp. 537-569).* NASP Publication, Bethesda, MD.

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

Stage, S. A., Jacobsen, M. D., & Michael, D. (2001). Predicting student success on a state-mandated performance based assessment using oral reading fluency. *School Psychology Review, 30*(3), 407-420.

Tomkowicz, J., & Schaeffer, G.A. (2002, April). Vertical scaling for custom criterion-referenced tests. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans.

Torgesen, J.K. (2000). Individual responses to reasons to early interventions in reading: the lingering problem of treatment resisters. *Learning Disabilities Research & Practice, 15, 55-64.*

Torgesen, J.K. & Burgess, S.R. (1998). Consistency of reading-related phonological processes throughout early childhood: Evidence from longitudinal-correlational studies. In J. L. Metsala & L. C. Ehri (Eds.), *Word recognition in beginning literacy (*pp. 161-188). Mahwah, NJ: Erlbaum.

U. S. Department of Education. (1998). *To assure the free appropriate education of all children with disabilities: Twentieth annual report to Congress on the implementation of the Individuals with Disabilities Education Act.* Washington, DC: U.S. Government Printing Office.

U. S. Department of Education. (2002). *Twenty-fourth annual report to congress on the implementation of the individuals with disabilities educational act.* Washington, DC: U.S. Government Printing Office.

U. S. Department of Education. (2005).*Twenty-fifth annual report to congress on the implementation of the individuals with disabilities educational act.* Washington, DC: U.S. Government Printing Office.

Vaughn, S. (2003). Response to instruction as means of identifying students with reading/learning disabilities. *Exceptional Children, 69,* 391-409.

Vellutino, F. R., Scanlon, D. M., & Lyon, G. R. (2000). Differentiating between difficult-to-remediate and readily remediated poor readers: More evidence against the IQ-achievement discrepancy of reading disability. *Journal of Learning Disabilities, 33,* 223-238.

Whaley, D.L., & Surratt, S.L. (1968). *Attitudes of science.* Kalamazoo, MI: Behaviordelia.

Wesson, C., Fuchs, L.S., Tindal, G., Mirkin, P.K., and Deno, S.L. (1986). Facilitating the efficiency of ongoing curriculum-based measurement. *Teacher Education and Special Education, 9,* 166-172.

Wesson, C., King, R., & Deno, L. S. (1984). Direct and frequent measurement: if it's so good for us, why don't we use it? *Learning Disability Quarterly, 7*(45-48).

Wiley, H. &Deno, S. (2005). Oral reading and maze measures as predictors of success for English Language Learners on state standard assessments. *Remedial & Special Education,* 26 (4), 207-14.

Williams, K., T. (2001). *Group Reading Assessment and Diagnostic Evaluation.* Cirde Pines, MN: American Guidance Service

Witt, J.C. (2005). *Sentence Maze Assessment*. Unpublished.

Witt, J.C. (2005*). Picture Word Fluency Assessment*. Unpublished.

Woodcock, R, W. (1987). *Woodcock Reading Mastery. Tests* (Rev. ed.). Cirde Pines, MN: American Guidance Service

Ysseldyke, J. (1979). Psychoeducational decision-making. In J. E. Ysseldyke & P. K. Mirkin (Eds.), *Proceedings of the Minnesota Roundtable Conference on Assessment of Learning Disabled Children* (Monograph No. 8). Minneapolis: University of Minnesota Institute for Research on Learning Disabilities.

Ysseldyke, J.E., Algozzine, B., Shinn, M.R. & McGue, M. (1982). Similarities and differences between low achievers and students classified as learning disabled. *The Journal of Special Education, 16* 73-85.

Ysseldyke, J.E., & Thurlow, M.L. (1984). Assessment practices in special education: Adequacy and appropriateness. *Educational Psychologist, 9,(3),* 123-136.

APPENDIX A: INSTRUCTIONS FOR PARAGRAPH MAZE

Instructions for Paragraph Maze

1. Write the following sentence on the board:

   When it is cold, I wear a warm _____ (book, tree, coat).

2. Say to Students: **"Take out a sharpened pencil. I will be passing out a piece of paper face down. Please do not turn the paper face up until you are instructed to do so."** Insure that all students have a sharpened pencil and are ready to take part in the assessment.

3. Tell Students "**Please write your teacher's name on the back of the page. If your homeroom teacher is different, then write your homeroom teacher's name instead. Now write your first name and your last name on the back of the paper.** Watch students to insure they are following the directions and successfully complete this task.

4. Tell Students: **"We will be doing some reading today. You will be asked to read a story. In the story there are sentences with words missing. Beside the missing word, you will see a blank with three words next to it. Read the sentence and choose the best word to complete the sentence. CIRCLE the word that best fits the sentence." The word you circle should be the word that causes the sentence to make the most sense.**

5. Using the example sentence on the board, Say: **Now look at the example sentence. It says, When it is BLANK, I wear a warm coat. After the blank there are three words: (warm, cold, late). The word that makes the most sense is COLD. When it is cold, I wear a warm coat. So you would circle the word cold. Cold makes this sentence make the most sense.**

6. Say, **"Work quickly without making errors."**

7. Say, **"Are there any questions?"** If asked, you may tell the students they have 3 minutes to work but indicate they should working quickly and accurately.

8. Say: **When I tell you to start, you will turn the paper over and, for each sentence, circle the best word.**

9. Set timer: **three minutes**. Say **"Start."** And THEN begin timer to allow students a moment to turn the paper over.

10. When the timer rings, say, **"Stop working. Hold your papers in the air now please."**

11. Circulate and collect all worksheets prior to thanking the students and exiting.

APPENDIX B: INSTRUCTIONS FOR SENTENCE MAZE

Instructions for Sentence Maze

1. Write the following sentence on the board:
   When it is cold, I wear a warm _____ (book, tree, coat).

2. Say to the students: **"Take out a sharpened pencil. I will be passing a piece of paper face down. Please do not turn the paper face up until you are instructed to do so."** Insure that all students have a sharpened pencil and are ready to take part in the assessment.

3. Tell students "**Please write your teachers name. If your homeroom teacher is different, then write our homeroom teacher's name instead. Now write your first and last name on the paper."** Watch students to insure they are following the directions and successfully complete this task.

4. Tell students: "**We will be doing some reading today. You will be asked to read some sentences. For each sentence, the last word is missing, for the missing word, you will see a blank with three words next to it. Read the sentence and choose the best word to complete the sentence. CIRCLE the word that best fits the sentence. The word you circle should be the word that makes the sentence makes the most sense."**

5. Using the example sentence on the board, Say: **"Now look at the example sentence. If says – When it is cold, I wear a warm_____ (book, tree, coat). The word that makes the most sense is COAT. When it is cold I wear a warm coat. So you would circle the word coat.**

6. Say: **"when I tell you to start, you will turn the paper over and, for each sentence, circle the best word."**

7. Say **"Work quickly without making any errors. If you finish a page, turn the page and go on to the next page."**

8. Say: "are **there any questions."**

9. Set timer for **three minutes.** Say "**Start".** And THEN begin the timer to allow students a moment to turn their papers over.

10. When the timer rings, say, "**Stop working. Hold you papers in the air now please."**

11. Circulate and collect all worksheets prior to exiting.

APPENDIX C: INSTRUCTIONS FOR PICTURE WORD FLUENCY

Directions for Picture Word Fluency

1. Say to Students: **"Take out a sharpened pencil. I will be passing out a paper face up and some papers face down. Please do not turn the papers over until you are instructed to do so."** Insure that all students have a sharpened pencil and are ready to take part in the assessment.

2. Tell Students "**Please write your teacher's name on the back of the page which is face down. If your homeroom teacher is different, then write your homeroom teacher's name instead. Now write your first name and your last name on the back of the paper.** Watch students to insure they are following the directions and successfully complete this task.

3. Tell Students: **"We will be doing some reading today. You will be asked to match a word with a picture.**

4. Using the example picture word match on the board, Say: "**Now look at the example picture word match. This is a picture of a boy. Beneath the picture are four words: (boy, run, dog, day). The word that best matches the picture is BOY. So you would circle the word BOY. The word BOY best matches the picture."**

5. Say, **"For each picture, you will have four word choices. Look carefully at each picture and the words beneath it. Choose the best word that matches the picture and CIRCLE it. " The word you circle should be the one word that best matches the picture. Work quickly without making errors. If you finish a page, turn the page and go on to the next page."**

6. Say, **"Are there any questions?"** If asked, you may tell the students they have 3 minutes to work but indicate they should working quickly and accurately.

7. Say: **When I tell you to start, you will turn the paper over and, for each picture, circle the best word.**

8. Set timer: **three minutes**. Say **"Start."** And THEN begin timer to allow students a moment to turn the paper over.

9. When the timer rings, say, **"Stop working. Hold your papers in the air now please."**

10. Circulate and collect all worksheets prior to thanking the students and exiting.

VITA

James A. Van Hook, III, is a candidate for the degree of Doctor of Philosophy in the psychology program at Louisiana State University. Mr. Van Hook has a bachelor's degree in psychology from Centenary College, and a Master of Science degree in clinical psychology from Northwestern State University. Before enrolling full-time at LSU, he worked for several years as a psychological and neuropsychological assistant under the supervision of Thomas E. Staats, Ph.D. During this time, he gained experience in the administration, scoring and interpretation of numerous psychological and neuropsychological instruments. Mr. Van Hook moved to Washington, D.C,, where he worked for over two years as a School Psychologist for the District of Columbia public school system, where he also worked as a member of a Rapid Response Team, which was charged with the evaluation of time-sensitive special education cases. Also while in Washington, he co-chaired the Preliminary Educational Review committee, whose purpose was to oversee the needs of special educational services. Mr. Van Hook returned to his native Louisiana in 2001 and took a position as a School Psychologist for East Baton Rouge Pupil Appraisal Services. He entered the LSU doctoral program in August 2003 under the supervision of Joseph C. Witt, Ph.D., where he continued to pursue his interests in assessment and measurement—particularly the screening of "at risk" students in reading. He is to complete his Doctor of Philosophy in psychology in May 2008.