2005

The predictive validity of four reading fluency measures on a state's 'high-stakes' outcome assessment

Jose Michael Castillo
University of South Florida

The Predictive Validity of Four Reading Fluency Measures on a State's "High-Stakes" Outcome Assessment

by

Jose Michael Castillo, M.A.

A thesis submitted in partial fulfillment
of the requirements for the degree of
Education Specialist
Department of Psychological and Social Foundations
College of Education
University of South Florida

Major Professor: Kelly A. Powell-Smith, Ph.D.
Kathy L. Bradley-Klug, Ph.D.
Jeffrey Kromrey, Ph.D.

Date of Approval:
February 18, 2005

Keywords: statewide assessments, screening, accountability, decision-making, educational utility

Table of Contents

List of Tables

List of Figures

The Predictive Validity of Four Reading Fluency Measures on a State's "High-Stakes" Outcome Assessment

Jose Michael Castillo

ABSTRACT

This study examined the long-term predictive validity of four reading fluency measures on a statewide reading assessment. The reading fluency measures were administered to a group of first and second grade students and their utility for predicting outcomes on a state's reading assessment in third grade was examined. Specifically, the amount of variance accounted for on the outcome assessment by Curriculum-Based Measurement – Reading (R-CBM) probes from two sources, a list fluency measure, and a group measure of fluency was investigated. The extent to which each is an accurate predictor of long-term performance when considered individually and in combination with the other measures was included in the analyses, along with an examination of whether a difference exists in the obtained prediction between grade levels. Results of the analyses indicated that the R-CBM probes from both sources tended to account for the most variance on the state's outcome assessment, followed by the list fluency measure and then the group fluency measure. This pattern was evident regardless of whether a single predictor or a combination of predictors was considered. Results of the grade-level analysis indicated that no significant differences were obtained in the amount of variance

accounted for by the measures between grade levels. These results are discussed in terms of their potential implications for research and practice in the field of education.

Chapter I

Introduction

Background Information

Good reading skills have become a necessity in modern society. Decades of technological breakthroughs have changed the nature of the work place in the United States. Because of the use of complex machines in industries such as farming and automobile manufacturing, occupations that solely require manual labor are quickly disappearing (e.g., the number of farmers is expected to decrease by 328,000 between the years 2000 and 2010; Bureau of Labor Statistics, n.d.). In fact, according to the American Management Association Survey of 2001, 34.1% of applicants tested by respondent firms lacked the basic reading skills necessary for the jobs they sought. Of the firms that responded to this survey, 84.6% did not hire skill deficient applicants (American Management Association Survey, 2001).

Unfortunately, when one considers the literacy rates for adults in the United States, the seriousness of this trend is evident. According to the National Adult Literacy Survey (1992), 21 to 23% of Americans scored within the Level 1 range, thereby demonstrating the "lowest level of prose, document and quantitative proficiencies." Although many adults within this category could perform simple tasks involving uncomplicated text, others displayed such limited reading skills that they were unable to respond to many of the questions. Another 25 to 28% of Americans scored within the Level 2 range, thus

demonstrating basic reading skills. They could generally locate information within a text, make low-level inferences, and integrate easily identifiable pieces of information. These statistics indicate that approximately 50% of American adults were unable to read at a proficient level as of the early 1990's. More recent data on the illiteracy problem facing our nation will be available soon when the National Center for Education Statistics publishes the results of the National Assessment of Adult Literacy conducted in 2003 (National Center for Education Statistics, n.d.).

The literacy statistics for America's children are no less alarming. Fletcher and Lyon (1998) have reported that at least ten million school-age children in the United States are poor readers. Other researchers have found that as many as one in five children will have difficulty learning to read (Lyon, 1995; Shaywitz, Escobar, Shaywitz, Fletcher, & Makuch, 1992). These data were further supported by the National Assessment of Educational Progress (NAEP) conducted in 2003. The reading portion of the NAEP is comprised of a series of multiple choice and constructed-response questions. These questions are designed to ascertain students' ability to understand, integrate, and synthesize information both within and across texts. According to the NAEP, 69% of fourth graders and 68% of eighth graders scored below the proficient level (National Assessment of Educational Progress [NAEP], 2003). Thus, over two-thirds of the nation's fourth and eighth graders were unable to perform the aforementioned tasks at a level that would be considered proficient.

The data from Fletcher and Lyon (1998) and the NAEP (2003) are particularly disheartening when findings from recent research are considered. According to many researchers (e.g., Francis, Shaywitz, Stuebing, Shaywitz, & Fletcher, 1994; Good,

Simmons, & Smith, 1998; Juel, 1988; Torgesen & Burgess, 1998), those students who get off to a poor start in reading almost always continue to be poor readers. For example, Francis et al. (1994) showed that, on average, children who were poor readers in third grade were unable to catch up to their peers in terms of their reading skills. In this study, 74% of the students who were poor readers in third grade continued to be poor readers in ninth grade.

According to Stanovich (1986), children experience what he has termed a "Matthew Effect" in reading. The Matthew Effect refers to the idea that children with lesser initial reading abilities acquire subsequent skills at a slower rate than their peers with higher abilities. One explanation for this phenomenon is that struggling readers spend significantly less time practicing activities that involve reading and writing. Less engagement in reading related activities means that children who are already demonstrating deficiencies in reading ability spend less time developing skills in areas such as phonemic awareness, reading fluency, and vocabulary than their same-age peers. Thus, the children with lesser reading skills fall further and further behind their peers in the acquisition of higher order skills.

Based in large part on this research, legislation such as the No Child Left Behind Act (NCLB) of 2002 has been passed. NCLB mandates that all children read at grade-level by the end of third grade starting in 2013. States are required to develop goals that mandate the number of students who should be performing at grade-level each year. These goals must be raised a minimum of once every three years until 2013 to ensure that the states remain on track to meet the objective of 100% of third graders reading at grade-level. To monitor student progress toward NCLB, the use of statewide assessments is

mandated in grades three through eight beginning in 2005-2006 (No Child Left Behind [NCLB], 2002). Although these assessments vary from state to state, the majority measure reading comprehension skills, the most important, ultimate outcome of effective reading instruction (e.g., Torgesen, 1998; Adams, 1990).

"High-Stakes" Assessments

Retention. In addition to using these assessments to monitor progress toward NCLB, many states are using the data to make important decisions that effect schools and their students. One issue for which these assessments have been employed as a tool for decision making is retention. Students who do not achieve a predetermined criterion for performance on these tests can be automatically retained in some states. The level of performance that students needs to attain to be promoted (i.e., basic or proficient level) and the grade levels in which automatic retention occur vary from state to state. For example, in Florida, students who score at Level 1 (i.e., below basic) on the reading section of the Florida Comprehensive Assessment Test (FCAT), Sunshine State Standards (SSS) in third grade may be automatically retained. During the 2002-2003 school year, 23% of the 188, 414 third graders who took the assessment scored at Level 1. Of the approximately 40,000 students who scored at level 1, 28,028 were retained. The other 12,000 students were promoted for various reasons including demonstrated proficiency through an alternative assessment or portfolio, Limited English Proficiency (LEP), non-participation because of an Individual Education Plan (IEP), or previous retentions with remediation (Florida Department of Education, n.d.b).

States with a large number of retainees are forced to deal with serious logistical issues. Largely effected schools must find ways to accommodate additional students

within grade levels. Space, the availability of teachers, and supplies are issues requiring immediate attention in these instances. Because resources are limited, classroom locations and teacher assignments would need to be reorganized. The availability of supplies (e.g., textbooks) would be a consideredion as well. Because of an already constricted budget, schools have difficulty obtaining textbooks and other learning materials when needed. A large number of retainees would probably require additional spending to ensure that the necessary supplies are obtained.

Retention seems to have negative effects on students as well. Many researchers have demonstrated negative academic and behavioral outcomes for students who have been retained (e.g., Pagani, Tremblay, Vitaro, Boulerice, & McDuff, 2001; McCoy & Reynolds, 1999). For example, McCoy and Reynolds (1999) examined the long-term academic outcomes of students who graduated from a government funded kindergarten program for children in poverty. The authors found that retainees from this group performed significantly lower on a standardized test of reading and mathematics achievement in seventh grade than their same-age peers. With regard to the behavioral functioning of retainees, Jimerson, Carlson, Rotert, Egeland, & Stroufe (1997) examined the outcomes of students at-risk for social and emotional problems due to maternal characteristics. According to the authors, retained students in this sample displayed more maladaptive behaviors and lower levels of social emotional health than non-retainees.

The negative outcomes found by these researchers extend beyond the classroom. According to Jimerson (1999), the students retained in the aforementioned Jimerson et al. (1997) study displayed higher drop-out rates, a lower percentage of certificates of high-school completion, and a lower number of post-secondary education enrollments than

non-retainees. As adults, retainees were less likely to be employed full-time, enrolled as a full-time student, or involved in some combination of the two than non-retainees. They also earned lower wages and were rated as less competent within the workplace.

Interestingly, a few researchers have shown small academic gains for retainees (e.g., Mantzicopoulos, 1997); however, these gains appear to be temporary with most retainees falling below their same-grade and same-age regularly promoted peers within 2-3 years (e.g., Mantzicopoulos, 1997; Jimerson, 1999; Jimerson et al., 1997). Thus, even the literature on retention that shows small academic gains seems to reflect that the practice is a negative experience for students that may hinder their academic, behavioral, and life outcomes.

School Grades and Adequately Yearly Progress. Statewide assessments have been used to make other important decisions as well. NCLB mandates that each state establish standards for student knowledge and performance in the core subject areas (e.g., reading, math, and science). States are required to use their statewide assessment to monitor student progress toward achieving these benchmarks. NCLB further mandates that states use the data obtained from their assessments to grade schools based on their students' performance (NCLB, 2002). Some states employ predetermined formulas based on the benchmarks established that constitute Adequate Yearly Progress (AYP) toward NCLB (i.e., meeting the benchmarks established by the state for adequate progress toward NCLB), the percentage of students scoring at the proficient level, and other factors such as the number of students tested to calculate school grades (e.g., see Florida Department of Education, n.d.a). The exact methods used to determine school

grades vary from state to state; however, reading scores from the statewide assessment are a required component of the process (NCLB, 2002).

Grades are important to schools for many reasons. Students that are attending failing schools may become eligible for "supplemental services" such as tutoring, after-school help, and summer school. If a school does not meet the requirements for AYP for two years, and continues to be identified as in need of improvement after receiving special assistance and additional resources, its students become eligible to attend another school with transportation provided (NCLB, 2002).

NCLB also provides an option for monetary penalties to be enforced if school grades do not improve over time. For those schools that continue to fall short of the requirements for AYP, sources of funding may be removed. Thus, a school that incurs a financial penalty because of inadequate performance may be forced to remove programs and personnel that were paid for by the additional funding. On the other hand, schools that receive high grades or demonstrate significant improvement can be rewarded for their success. Additional funding may be awarded that may be used for various purposes such as purchasing extra supplies and increasing teacher salary (NCLB, 2002). The fundamental principle behind this movement is that financial penalties/rewards will provide motivation for teachers, administrators, and other educators to work harder to improve student performance.

A school's reputation is also tied to the grade that it receives. Although other factors such as the composition of the school's population (e.g., high versus low SES) are important considerations when deciding its effectiveness, the reality is that grades are the primary emphasis in an accountability driven system. Because the focus of the system is

on outcomes, schools that receive higher grades may be considered exemplary, whereas those that perform poorly may be viewed as inadequate. The resulting reputation can effect the decisions of parents, teachers, administrators, and other educators. Parents of children zoned for a "failing" school may decide to find an alternative site for their child's education (i.e., school choice; see NCLB, 2002). Qualified teachers and administrators may also target schools with a reputation for success over those that are underachieving.

Test Preparation. Because of the decisions that are being made with these "high-stakes" outcome assessments, many schools spend an inordinate amount of time preparing their students for these tests (i.e., "teaching to the test"). In many cases, individual teachers, grade levels, and even entire schools will stop regular instruction weeks before the assessment. Testing procedures, strategies, and practice problems are reviewed and drilled in an attempt to provide students with an opportunity to perform. These practices are particularly evident in schools/grades where automatic retention or previous poor performance is an issue. Although spending time to prepare students for these assessments may seem necessary in an accountability-based system, such procedures significantly reduce valuable academic instruction time.

While the long-term effects of reducing academic instruction time for test preparation are unknown, researchers have consistently demonstrated that spending more time teaching academic tasks leads to improved student performance (e.g., Fisher et al., 1978; Hawley & Rosenholtz, 1984). For example, Fisher et al. (1978) examined the achievement test scores of second grade students after they received varying amounts of reading instruction per week for five week intervals. When 100 minutes per week of

instruction in reading was received, student performance lagged behind entry-level percentiles. When the amount of time was increased to 573 minutes per week, percentiles on the test remained stable. However, when 1300 minutes per week was spent learning to read, the percentile scores increased significantly. Thus, as academic instruction time increased, performance on achievement tests improved. Although an explicit link has not been made between decreased academic instruction and student performance on statewide assessments, results from Fisher et al. suggest that the practice of "teaching to the test" may be detrimental to student learning.

Importance of Predicting Outcomes

The amount of time spent on test preparation, as well as issues surrounding retention and school grades make the prediction of outcomes on statewide assessments more important than ever. A study conducted by Buly and Valencia (2002) indicated that students fail statewide assessments for a myriad of reasons. The authors stated that students who do not meet the criterion for passing may be deficient in a number of critical areas such as decoding, vocabulary, fluency, and comprehension. Moreover, the data obtained by statewide assessments do not allow for in depth analysis of a student's strengths and weaknesses. Further assessment would be necessary to determine the skill areas in which the student is deficient, further delaying the targeted instruction the child requires.

If student performance could be estimated early, schools could employ preventative measures to ensure that those students who are identified as at-risk for reading failure receive additional assistance. School-wide screening systems could be developed that identify struggling students for early and intensive intervention, a key component in

helping struggling readers catch up to their peers (Fletcher & Lyon, 1998; Torgesen 2002). Thus, time could be reallocated from test preparation to early, intensive instruction on significant academic skills needed to pass the outcome assessment.

The prediction of failure on statewide reading measures would not only benefit students, but also the schools they attend. Because at-risk students could be targeted for early and intense intervention, many will receive the assistance they require to perform proficiently. More children scoring at a level that meets the criterion for passing means less retentions in schools. Such a decrease in the number of retentions would reduce the need to find space, reorganize teaching assignments, and obtain additional supplies (e.g., textbooks).

In addition to reducing the number of retentions, the prediction of performance on statewide assessments has the potential to help improve school grades. Through targeting students identified as at-risk, schools would be presented with an opportunity to improve student performance, thus improving their grades. Such improvements have a tremendous impact for both low and high performing schools. Traditionally low performing sites could avoid sanctions such as financial penalties and move closer to the goal of NCLB. Those sites that traditionally obtain higher grades would have the opportunity to improve student scores and obtain financial rewards. Improvements in school grades would have the potential to lead to a better reputation within the community as well. Parents, teachers, and administrators might be more likely to involve themselves with schools that are demonstrating improvement.

Perhaps the most important objective that predicting success on statewide outcome assessments could accomplish is the estimation of student comprehension skills. Good

comprehension skills are necessary for academic success across subjects (i.e., math, English, science, the social sciences, etc.). The early identification and prevention of failure in the most important, ultimate outcome of reading instruction (i.e., comprehension) could translate to more students successfully making their way through the education system. Thus, more students would graduate with proficient reading skills, meaning more qualified workers might be found within the job market.

Graduating with proficient reading skills is beneficial to students as well. The development of adequate reading skills lays the foundation necessary for students to succeed in elementary school and beyond. Students that succeed academically have the opportunity to attend college and obtain a degree. In an era dominated by technological breakthroughs, higher education is needed to be competitive within the labor market.

In addition to their increasing importance in the labor market, proficient reading skills may help students avoid the negative life consequences associated with illiteracy. The Orton Dyslexia Society (1986) reported that illiteracy is found in 75% of the unemployed, 85% of juveniles who appear in court, and 60% of prison inmates. More recent statistics indicate that as many as 7 out of 10 prison inmates and 68% of women on welfare are unable to read at the proficient level (National Adult Literacy Surven, 1992). Although a direct, causal link between illiteracy and these outcomes cannot be made, a strong relationship between the two is evident. Thus, when one considers both the positive and negative outcomes associated with the development of proficient reading skills, the importance of early identification of struggling and at-risk students is strikingly clear.

Review of Research on Prediction of Outcomes

A few methods have been investigated for predicting outcomes on statewide assessments. One method that has received extensive consideration with regard to predicting performance is frequently referred to as Curriculum-Based Measurement – Reading (R-CBM) (Deno, 1985; Marston, 1989; Shinn, 1989, 2002). R-CBM is an oral reading fluency measure that typically requires a child to read aloud from a passage for 1-3 minutes. The score is the number of words read correctly. Research on R-CBM has demonstrated adequate reliability (Marston, 1989) and a strong relationship with reading comprehension (e.g., Deno, Mirkin, & Chiang, 1982; Fuchs, Fuchs, & Maxwell, 1988; Shinn, Good, Knutson, Tilly, & Collins, 1992). An examination of its predictive utility has led researchers to conclude that it is an accurate predictor of short-term (i.e., within-year) outcomes on statewide reading assessments (e.g., Shaw & Shaw, 2002; Buck & Torgesen, 2003; Barger, 2003; Good, Simmons, & Kame'enui, 2001). However, little research exists on the utility of R-CBM as a predictor of long-term (i.e., across years) outcomes on these measures.

In addition to R-CBM, a few other measures have been developed that may have utility as predictors of performance on statewide assessments. One such measure involves the use of word lists (i.e., list fluency). List fluency is typically measured by prompting a child to read aloud from a printed list of words for a short period of time (i.e., 45 seconds to 5 minutes). The score is the number of words read correctly. Adequate reliability (e.g., see Torgesen, Wagner, & Rashotte, 1999) and a moderate (r=.53; Jenkins, Fuchs, van der Broek, Espin, & Deno, 2003) to strong (e.g., r=.50 to .87; Torgesen, Wagner, & Rashotte, 1999) relationship with reading comprehension have

been demonstrated. Group measures of reading fluency are a recently developed method for measuring fluency that may have some utility as well. One example of a group measure currently being developed is the Test of Critical Early Reading Skills (TOCERS; Torgesen, Wagner, Lonigan, & DeGraff, 2002). One of the subtests from the TOCERS requires students to identify printed real words on a page by placing a slash through them. The score is based a formula involving the number of words identified correctly and the ratio of real to non-real words in each column (a detailed description of this measure is located in the methods section). Adequate reliability and correlations with reading comprehension ranging from low to moderate have been demonstrated for the subtest (Castillo, Torgesen, Powell-Smith, Al Otaiba, 2004). No research was evident that examined the predictive validity (within or across years) of list or group measures of reading fluency on statewide assessments.

Thus, various types of reading fluency measures have been examined with regard to their relationship with reading comprehension. However, with the exception of R-CBM, little research has been conducted on their predictive validity on statewide reading assessments. Therefore, additional research is needed to determine if alternate measures of reading fluency are accurate predictors of outcomes. Additional research is also needed to determine if reading fluency measures, including R-CBM, can reliably predict long-term outcomes.

Rationale for Current Study

Many states are using their mandated statewide assessments to make "high-stakes" decisions. In Florida, the reading section of the Florida Comprehensive Assessment Test (FCAT) is being used to retain students in third grade who score below the basic level

and is instrumental in determining school grades. The reading section of the FCAT is comprised of two subtests, the Sunshine State Standards (SSS) and the Norm-Referenced Test (NRT). The FCAT SSS is a criterion-referenced test that is used to determine if students are meeting the standards established by the Florida Department of Education for a particular grade level. The FCAT SSS is the subtest of the reading section that is primarily used to make the "high-stakes" decisions mentioned previously. The FCAT NRT is a norm-referenced test that is used to compare students to a national standardization sample. Both subtests primarily measure reading comprehension skills.

Because scores on the FCAT are being used for such important decisions, the prediction of student performance would be an invaluable tool for schools. To investigate this issue, Buck and Torgesen (2003) examined the predictive validity of R-CBM probes on the FCAT SSS. Three R-CBM probes were administered to 1102 third grade students approximately one month after the FCAT was administered. Results indicated that 91% of the students who obtained a median score of at least 110 words per minute performed at the proficient level or better on the FCAT. Eighty-one percent of the students who obtained a median score of 80 words per minute or less scored below proficient. These results suggest that the median score from R-CBM probes in third grade can accurately predict performance on the FCAT SSS.

Buck and Torgesen (2003) established the validity of using R-CBM to predict outcomes on the FCAT SSS. However, in light of research regarding the outcomes of those students who get off to a poor start in reading (e.g., Francis, Shaywitz, Stuebing, Shaywitz, & Fletcher, 1994; Good, Simmons, & Smith, 1998; Juel, 1988; Torgesen & Burgess, 1998), identifying a child who is at-risk for reading failure in third grade may be

too late. Unfortunately, little is known about the long-term predictive validity of R-CBM probes on either subtest of the FCAT. In addition, little information is available on the predictive utility of alternate reading fluency measures (i.e., list fluency and group measures). If any of these measures could predict outcomes in first and/or second grade, schools would be able to target those students at-risk for failure. Thus, at-risk students would receive the intensive intervention they require in time to make a meaningful difference in their reading performance.

Purpose

The purpose of this study was three-fold. The first purpose was to investigate the long-term predictive validity of various reading fluency measures on the reading section of the FCAT. To investigate this issue, this study examined three types of reading fluency measures (i.e., R-CBM, list fluency, and group measures) to determine the extent to which each is an accurate predictor of performance on the FCAT SSS and FCAT NRT in third grade. The second purpose was to determine if there are significant differences in the predictive utility of these reading fluency measures across years (i.e., grade 1 versus grade 2). The third and final purpose was to determine if the various reading fluency measures predicted outcomes differently across the FCAT SSS and FCAT NRT. Thus, the research questions that were targeted in this study were:

1) To what extent is each measure of reading fluency an accurate predictor of FCAT performance?
2) What is the difference in the predictive utility of reading fluency measures across grade levels (i.e., grade 1 versus grade 2)?

3) What is the difference in the prediction obtained by the various reading fluency measures across the FCAT SSS and FCAT NRT?

Chapter II

Literature Review

Theoretical Basis for Examining Reading Fluency

The importance that has been placed on outcome assessments at both the state and federal levels makes the development of reliable and valid measures for early identification and prevention of reading difficulties critical. Because reading comprehension measures are time consuming and often unreliable in young children with reading skill deficits, other assessments of reading proficiency have been developed. One group of assessments that have received attention examines reading fluency. Reading fluency is defined as rate and accuracy in reading. Many theories (e.g., LaBerge & Samuels, 1974; Posner & Snyder, 1975) have purported that fluent reading is as an integral part of the ability to comprehend text.

According to LaBerge and Samuels (1974), reading is a complex skill that requires the coordination of many component processes in a short time frame. If each of these component processes required attention, the performance of the complex skill would exceed attentional capacity and therefore become impossible. However, successful performance could be achieved if enough of the components were executed automatically, making the attentional load tolerable. The authors assumed that comprehension processes require attention and therefore are not good targets for automaticity. However, they considered lower-level lexical processes such as

phonological coding prime candidates for automatic processing. In essence, the authors believed that proficient reading involves reallocating attentional capacity from lower-level word recognition to more demanding comprehension functions. Put simply, if a child is focusing too much attention on decoding words, little is left over for comprehension. In contrast, a child who is able to read fluently can focus more attention on comprehending the text.

Other theories, such as one proposed by Posner and Snyder (1975), purport a more interactive process. According to Posner and Snyder, word recognition is affected by semantic context via two independently acting processes. The first is an automatic-activation process in which a stimulus activates a memory location and spreads automatically to nearby semantically related locations in the network. This process requires no attentional capacity. The second is a conscious-attention mechanism that relies on context to make predictions about the upcoming word and directs the limited capacity processor to the location of the expected information. This slow-acting process is optional and utilizes valuable attentional capacity. Good rapid word recognition skills can short-circuit the conscious-attention mechanism (i.e., the spreading-activation component dominates), freeing up attenional resources for integrative comprehension processes. In other words, readers with poor word recognition skills are forced to rely on prior contextual knowledge to aid in identifying unknown words. Because the attention of poor readers is focused on the use of prior knowledge to identify words, little is left for integrating what is read with existing knowledge structures. In contrast, readers with strong word recognition skills are able to focus attention on constructing and integrating new knowledge.

Both theories share the assumption that efficient word-recognition frees up attentional capacity for comprehension. If this assumption is valid, then a child's ability to read fluently should be a valid indicator of his/her comprehension skills. In this review, studies that examine how different types of reading fluency measures relate with tests of reading comprehension are discussed. More specifically, studies that examined how oral reading fluency, list fluency, and group measures of fluency relate with reading comprehension are reviewed.

Relationship between Oral Reading Fluency and Reading Comprehension

Oral Reading Fluency (ORF) is defined as rate and accuracy in oral reading. One of the most common ways to measure ORF is frequently referred to as Curriculum-Based Measurement Reading (R-CBM) (Deno, 1985; Marston, 1989; Shinn, 1989, 2002). R-CBM typically requires a child to read aloud from a passage for one to three minutes. The score is the number of words read correctly. Researchers have concluded that R-CBM demonstrates adequate reliability. Test-retest reliability estimates range from .82 to .97, with most correlations above .90 (Marston, 1989). Alternate-form estimates range from .84 to .96, again with most estimates above .90 (Marston, 1989). Inter-rater agreement of .98 or better has been demonstrated as well (Marston, 1989). Because R-CBM is reliable and can be administered efficiently, many researchers have examined R-CBM to determine if the measure has a significant relationship with reading comprehension.

In a recent review of the literature, Hosp (2004) found 21 studies with 33 correlations that addressed the relationship between R-CBM and reading comprehension. The studies included in the analysis met the following criteria: (a) participants were in grades one

through eight, (b) measures used real words, (c) word reading was conducted orally, (d) a measure of reading comprehension was used as the criterion measure, and (e) correlations were provided between the R-CBM and reading comprehension measures. The average correlation found in these studies was .69 (SD = .11). Although the average correlation found by Hosp provides an overall picture of a moderately strong relationship between R-CBM and reading comprehension, a more detailed discussion is required. Earlier research on the relationship between R-CBM and reading comprehension yielded some stronger correlations. Additionally, more recent research has focused on the relationship between R-CBM and outcomes on high-stakes statewide assessments. It is these two areas that this literature review will now address.

Relationship with Reading Comprehension. Deno, Mirkin, and Chiang (1982) conducted one of the seminal studies that examined the relationship between R-CBM and reading comprehension. The participants were 18 general education students and 15 students with learning disabilities who were randomly selected from grades one through five. The researchers administered five formative measures, three fluency measures (one list fluency, one R-CBM, and one oral reading), and two informal tests of reading comprehension (one cloze and one word meaning). Two published norm-referenced tests (PNTs) of reading comprehension, the reading comprehension subtest of the Stanford Diagnostic Reading Test (SDRT), and the word identification and word comprehension subtests of the Woodcock Reading Mastery Test (WRMT) were administered as well. The scores from the formative measures were correlated with the results from the PNTs. The correlations for the three reading fluency measures and the PNTs ranged from .73 to .91, with most falling in the .80's. Interestingly, these correlations were stronger than the

correlations between the informal and PNTs of reading comprehension (r's ranged from .60 to .83). The researchers concluded that the results provided strong evidence that a simple measure of reading aloud would be a valid index of student reading proficiency.

In a slightly different approach, Shinn, Good, Knutson, Tilly, and Collins (1992) examined the relation of ORF to reading from a theoretical perspective. One-hundred-fourteen third grade and one-hundred-twenty-four fifth grade general education students participated in the study. Participants were administered two measures of decoding skills (i.e., Test of Written Spelling and the word attack subtest of the Woodcock Reading Mastery Test), three comprehension measures (i.e., written retell, cloze, and the SDRT comprehension subtest), and two R-CBM passages. Confirmatory factor analysis was employed to determine whether ORF constituted a significant role in a single-factor model of reading or whether it should be defined as a decoding construct, a comprehension construct, or a separate construct. In third grade, a single factor model of reading, labeled Reading Competence, could not be rejected. Each measure that was administered contributed significantly to the model. Interestingly, the two R-CBM passages correlated more highly with the model (*r's* = .88 and .90) than the measures of reading comprehension. For fifth grade, a unitary model was rejected. Reading competence was composed of two separate constructs, decoding and comprehension (ORF was loaded on the decoding construct). Although the two constructs could be differentiated, they were still highly correlated (*r* = .83). In fact, the R-CBM passages correlated as high or higher with reading comprehension (*r* = .74 and .75) than the SDRT measures (*r* = .73 and .75). Only the cloze measure was correlated more highly with reading comprehension (*r* = .86). The authors concluded that regardless of the model

employed, the results provide further evidence that R-CBM is a valid measure of reading competence, including comprehension.

In a review of the literature conducted by Fuchs, Fuchs, Hosp, and Jenkins (2001), the authors discussed a study in which four measures (one R-CBM and three informal reading comprehension), were correlated with the Stanford Achievement Test (SAT). The study discussed by the authors was conducted by Fuchs, Fuchs, and Maxwell (1988). In the Fuchs et al. (1988) study, R-CBM, question answering, passage recall, and cloze measures were administered to each subject. Seventy middle and junior high school students with a reading disability participated in the study. Criterion validity correlation coefficients were calculated between the SAT and the other four measures. Coefficients for the question answer, passage recall, cloze, and R-CBM measures were .82, .70, .72, and .91 respectively. Interestingly, R-CBM was more strongly correlated with the SAT than any of the informal measures of reading comprehension. Tests for differences between correlations demonstrated that the correlation between R-CBM and the SAT was statistically significantly higher than the correlation between the other three methods and the SAT. Fuchs et al. (2001) concluded that the findings in the Fuchs, Fuchs, and Maxwell (1988) study were consistent with other researchers' claims that ORF appears to reflect individual differences in reading competence.

The relationship between R-CBM and published norm-referenced reading tests was also examined by Jenkins and Jewell (1993). The PNTs used in the study were the Gates-MacGinitie Reading Tests (GMRT) and the Metropolitan Achievement Test (MAT). Both tests measure vocabulary and reading comprehension. The total reading score was used in the analysis for the GMRT and MAT. Three R-CBM probes were

administered as well. The median words read correctly were used in the analysis of the data. Cross-grade correlations were all above .80 and statistically significant at the $p<.01$ level. Within-grade correlations between words read correctly and the GMRT total reading score ranged from .67 to .88 with a median of .83 obtained in second grade. Within-grade correlations between words read correctly and the MAT total reading score ranged from .60 to .87 with a median of .70 obtained in third grade. Examination of these correlations revealed a negative trend across grade levels. For the GMRT, correlations of .83, .88, and .86 were found for grades two, three, and four respectively and declined to .67 by grade six. Correlations for the MAT dropped from .87 for second grade to .60 for sixth grade. These results suggest that the relationship between R-CBM and reading comprehension declines as the grade-level increases.

In an attempt to explore the effect of the source of R-CBM probes, Hintze, Shapiro, Conte, and Basile (1997) examined the relationship between survey-level R-CBM probes and a reading comprehension measure. The R-CBM passages were taken from literature-based basals and authentic reading material ranging from a first through fifth grade difficulty level. The reading comprehension measure used was the Degrees of Reading Power (DRP). The DRP uses a modified cloze technique in which reading passages are ordered by increasing difficulty. Each passage has seven deleted words and five options are provided for each deletion. Raw scores on the DRP were converted into DRP index scores to allow for cross-grade comparisons. The number of words read correctly was used as the data for the R-CBM probes. Participants were 57 students in grades two, three, and four. Tests of differences between the correlations revealed that there was no significant difference between type and material across any of the difficulty levels.

Averaged across difficulty level, the mean correlation for the authentic book series with the DRP was .665. The mean correlation across difficulty level for the literature-based basal series was .655. No statistically significant difference was found between the two correlations. In addition, R-CBM accounted for a significant amount of the variability in DRP scores. Thee authentic reading material accounted for 49 percent of the variance in DRP scores, while the literature-based basal series accounted for 52 percent. Interestingly, R-CBM predicted reading comprehension skills in two-thirds of the cases on the DRP. The researchers concluded that these results provide evidence that R-CBM and reading comprehension are similarly correlated regardless of the source of the R-CBM probes. They also stated that their findings support R-CBM as a valid measure of overall reading competence.

The aforementioned studies provide substantial evidence for a relationship between R-CBM and reading comprehension. A myriad of researchers have found moderately strong to strong correlations between the two constructs across setting (e.g., general education, special education), grade, and difficulty level. In addition, this relationship has been demonstrated regardless of the source of R-CBM probes (e.g., literature-based, authentic material, generic materials) and type of reading comprehension measure used (e.g., standardized test, cloze, passage recall).

Prediction of "High-Stakes" Outcomes. In one of the first studies that utilized a state's outcome assessment as the criterion measure, Stage and Jacobson (2001) used growth curve analysis to examine the relationship between a student's slope in R-CBM and the Washington Assessment of Student Learning (WASL). The WASL is an un-timed performance-based test that utilizes norm-referenced and criterion-referenced

scoring procedures. Participants consisted of 173 fourth grade students in one elementary school. R-CBM benchmarks were developed for the entire school for three intervals (September, January, and May). During each interval, students were given a 250 word R-CBM passage from the Silver Burdette and Ginn curriculum and given one minute to read. The score was the number of words read correctly. The WASL was administered to the students in May. Attainment of the September R-CBM benchmarks increased the predictive power of student failure and success on the WASL by 30% over the base rate levels. The increase in slope of R-CBM scores across the year did not explain as much variance in WASL scores as attainment or non-attainment of the benchmarks in September, January, and May. Interestingly, correlations between R-CBM and WASL performance were moderate at best (r's ranged from .43 to .51). However, the sample came from one elementary school that outperformed the average student's WASL score across the state (i.e., 80% of the students passed in the sample, 59% passed statewide). Therefore, the demographics of the sample may have had an influence on the results. The researchers concluded that R-CBM scores are good predictors of WASL performance.

In a related study, Good, Simmons, and Kame'enui (2001) examined the relationship between R-CBM and the Oregon Statewide Assessment (OSA) -Reading/Literature high stakes outcome measure. The OSA in reading and literature is a standardized achievement test that is used to assess the achievement level of individual students based on standards established by the Oregon State Board of Education. Participants were 364 students in third grade. During the fall, winter, and spring intervals, students were administered three R-CBM passages from the Grade 3 level of the Test of Reading

Fluency (generic R-CBM materials). The OSA was administered in the spring to each of the participants. Results indicated that 96% of students who reached the benchmark goal for May of 110 words per minute met or exceeded expectations on the OSA. For those students who read below 70 words per minute, only 28% met expectations on the OSA. The outcome for students who read between 70 and 110 words per minute on the OSA was less clear. The authors concluded that these results provide further evidence of the utility of R-CBM as a good predictor of reading comprehension skills.

McGlinchey and Hixson (2004) examined the relationship between R-CBM and the Michigan Educational Assessment Program's (MEAP) fourth grade reading assessment. The MEAP is the statewide outcome assessment used in Michigan and is based on Michigan's Essential Goals and Objectives for Education. Across the eight year duration of the study, a total of 1362 fourth grade students from one urban school participated. All students were administered R-CBM probe(s) from the MacMillan Connections Reading Program in the two weeks prior to the administration of the MEAP each year. A cut-off score of 100 words read correct per minute (WCPM) was used. Those students who read over 100 WCPM were predicted to pass the MEAP. Those that read less than 100 WCPM were predicted to fail. The results indicated that R-CBM probe(s) correctly identified students who would pass or fail the MEAP 74% of the time. This rate of correct classification was 48% above chance (kappa = .48). The authors concluded that R-CBM is a valid predictor of performance on the MEAP.

Barger (2003) invesitaged the predictive power of R-CBM passages for performance on the North Carolina End of Grade reading assessment in third grade. Thirty-eight students were administered three R-CBM probes from the Dynamic Indicators of Basic

Early Literacy Skills (DIBELS) one week before the administration of the outcome assessment. One-hundred percent of the 26 students who scored 100 correct words per minute (CWPM) passed the North Carolina End of Grade assessment. Of the six students who read 70-99 CWPM, 50% achieved a passing score. Only 33% of the 6 students that read below 70 CWPM passed. The authors concluded that R-CBM may be an accurate predictor of student achievement on the North Carolina End of Grade assessment.

Shaw and Shaw (2002) examined the relationship between R-CBM probes and the Colorado State Assessment Program (CSAP). The CSAP is a standards-based reading comprehension assessment that is administered statewide each year. Fifty-eight third grade students were administered R-CBM probes from the DIBELS in September, January, and April of the 2001-02 school year. The CSAP was administered in April of the 2001-02 academic year. A cut-off score of 90 words read correctly (WRC) in the spring was used. Thirty-nine of the forty-three students (91%) who scored 90 WRC or above received a passing grade on the CSAP. Only 4 out of the 15 (27%) students who scored less than 90 WRC were proficient. Thus, the use of R-CBM probes resulted in classifying 50 of the 58 (86%) students correctly with regard to receiving a passing score on the CSAP.

Buck and Torgesen (2003) investigated the efficacy of using R-CBM passages to predict outcomes on the Florida Comprehensive Assessment Test (FCAT) Sunshine State Standards (SSS). The FCAT SSS is based on the standards developed for Florida students by the Florida Department of Education and is administered statewide each year. R-CBM passages from The Standard Reading Passage: Measures for Screening and Progress Monitoring were administered to 1,102 students in May of 2002. The FCAT

SSS was administered in April of the same year. Ninety-one percent of the students who obtained a median score of 110 words per minute on three R-CBM passages scored above proficient. Of those students who scored below 80 words per minute, 81% scored below proficient. The authors concluded that R-CBM passages are accurate predictors of performance on the FCAT SSS.

The aforementioned studies demonstrate that performance on R-CBM probes is a good predictor of success on state-mandated outcome assessments. However, more research needs to be done in order to determine if R-CBM can be used to predict reading performance on these outcome measures across grades. The available studies examine the short-term predictive validity of R-CBM probes. The R-CBM probes in the studies were administered in the same year (typically third or fourth grade) as the statewide assessment. In light of research regarding the outcomes of those students who get off to a poor start in reading (Francis, Shaywitz, Stuebing, Shaywitz, & Fletcher, 1994; Good, Simmons, & Smith, 1998; Juel, 1988; Torgesen & Burgess, 1998), identifying a child who is at-risk for reading failure in third or fourth grade may be too late.

Unfortunately, little is known about the long-term predictive validity of R-CBM. In one study that examined this issue, Crawford, Tindal, and Stieber (2001) administered three R-CBM probes from the Houghton Mifflin Basal Reading Series (the average score was used in the analysis) to 51 students in second and third grades in a rural school district in Oregon. The students were also administered the statewide reading assessment used in Oregon (i.e., the OSA) in third grade. This assessment is criterion-referenced, and consists of multiple-choice questions and performance tasks. In the within year analysis, 81% of the students who read at the 50th percentile or above (as established by

Hasbrouk & Tindal's, 1992 norms) passed the statewide reading test. Of those students that read at least 119 words per minute, 94% passed the assessment (i.e., 16 out of 17). Only 37% of students that read below the 50th percentile passed. In the across year analysis, 29 of the 37 (78%) students reading in the top three quartiles (i.e., 47 words per minute and above) passed the statewide assessment in third grade. One-hundred percent of students that read at least 72 words per minute in second grade passed the statewide test in third grade. Only 29% (i.e., 4 out of 14) of those in the first quartile passed the test. Thus, the researchers concluded that R-CBM can be used to effectively predict future performance on statewide assessments.

The findings by Crawford et al. (2001) provide evidence for the long-term predictive validity of R-CBM; however, more research is needed. The size and geographic location of the sample highlights the need for replication studies. In addition, the study examined the predictive validity of R-CBM in second and third grades, but did not address the utility of using the measure in first grade. Therefore, future research should examine the efficacy of using R-CBM probes in grades one and two to predict outcomes on "high-stakes" assessments across states.

Relationship between List Fluency and Reading Comprehension

R-CBM is not the only measure of reading fluency that has been examined in regard to its relationship with reading comprehension. Several researchers have examined whether list fluency correlates with reading comprehension as well. List fluency measures examine the fluency with which children read words in lists. It is typically measured by prompting a child to read aloud from a printed list of words for a short

period of time (can be from 45 seconds to five minutes). Adequate reliability has been demonstrated for list fluency as well (e.g., see Torgesen et al., 1999).

In a recent review of the literature, Hosp (2004) discussed several studies that examined the relationship between list fluency and reading comprehension. The studies included in the analysis met the following criteria: (a) participants were in grades one through eight, (b) measures used real words, (c) word reading was conducted orally, (d) a measure of reading comprehension was used as the criterion measure, and (e) correlations were provided between the list fluency and reading comprehension measure. Seven studies were found that contained seven correlations with a mean of .60 (SD = .08). The review by Hosp gives an overall picture of a moderate relationship between list fluency and reading comprehension, however a more detailed discussion is required. Research on the relationship between list fluency and reading comprehension is limited and has yielded conflicting results. Because no research was found that focuses on the relationship between list fluency and outcomes on high-stakes statewide tests, this review primarily focuses on list fluency's relationship with reading comprehension.

One of the seminal studies that examined the utility of list fluency measures was mentioned above. In the aforementioned study, Deno, Mirkin, and Chiang (1982) examined the relationship between reading aloud measures (i.e., one list fluency, one R-CBM, and one oral reading), informal measures of reading comprehension (i.e., one cloze and one word meaning) and published norm-referenced reading tests (i.e., the SDRT and WRMT). The participants were 18 general education students and 15 students with learning disabilities who were randomly selected from grades one through five. Correlations between the reading aloud measures and the PNTs ranged from .73 to .91

with most falling in the .80's. These correlations were higher than the correlations between the informal measures and standardized tests of reading comprehension (*r's* ranged from .60 to .83). The authors concluded that the results provided evidence for the validity of using words read in isolation as an index of reading performance.

In a review written by Fuchs, Fuchs, Hosp, and Jenkins (2001), the authors discuss a study that examined the criterion validity of R-CBM and list fluency. In the study conducted by Jenkins, Fuchs, van der Broek, Espin, and Deno (2003), one 400 word R-CBM passage, one list of randomly ordered words from the R-CBM passage, and the reading comprehension subtest of the Iowa Test of Basic Skills (ITBS) were administered. The sample consisted of 113 fourth graders. Within the sample, 85 of the students were classified as skilled readers, 21 were classified as below average but without reading disabilities, and 7 were classified as having a reading disability. The participants were selected in this manner to approximate the normal distribution of fourth graders. The results indicated that R-CBM (r = .83) had a higher correlation with reading comprehension than list fluency (r = .53). In addition, R-CBM accounted for much more variance in ITBS scores than list fluency. Together, R-CBM and list fluency accounted for 70% of the variance in the criterion measure. R-CBM uniquely accounted for 42% of the variance, whereas list fluency uniquely explained only 1%. Jenkins et al. also examined R-CBM as the criterion measure to determine how reading comprehension and list fluency contribute to R-CBM performance. Surprisingly, the ITBS uniquely explained 28% of the variance in R-CBM performance, whereas list fluency only uniquely accounted for 11%. The authors concluded that R-CBM appears to have a stronger relationship with reading comprehension than list fluency.

The results obtained in the aforementioned study lend credibility to a theory proposed by Perfetti (1995). In this theory, other reading subcomponents besides word-recognition can become automatized (e.g., integrating propositions within text and with background knowledge, inferencing, etc.). As they become more automatic, additional attentional resources are released (beyond the resources freed by word-recognition skills) for constructing a deeper text model. The findings in Jenkins et al. (2003) suggest that R-CBM indexes individual differences in verbal efficiency for the subcomponents of reading beyond those at the word level. Therefore, R-CBM would be a more valid measure of reading comprehension than list fluency.

However, findings from research conducted by Torgesen, Wagner, Rashotte, Alexander, and Conway (1997) suggest that list fluency may be a better indicator of reading comprehension skills than was purported by Jenkins et al. (2003). In this study, Torgesen et al. administered a measure of list fluency (i.e., the Test of Word Reading Efficiency (TOWRE)) and two PNTs of reading comprehension (i.e., the Gray Oral Reading Test -3 (GORT-3) and the Woodcock Reading Mastery Test – Revised (WRMT-R)) to fourth and fifth graders with severe reading disabilities. Correlations of .75 and .87 were obtained between the TOWRE and GORT-3 and the TOWRE and WRMT-R respectively. These correlations are more similar to those reported for R-CBM. One hypothesis for the differences in the correlations between the studies is that the words used in the Jenkins et al. (2003) study were all at the same difficulty level, whereas the words used in the Torgesen et al. (1997) study increased gradually in difficulty. It is possible that administering a list with words that are of the same difficulty level may artificially limit the range of individual differences.

In a similar study, Wagner et al. (1997) administered the TOWRE, the GORT-3, and the WRMT-R to a randomly selected group of 201 fifth grade students with average reading skills. Correlations between the TOWRE and the GORT-3 and the TOWRE and WRMT-R were .50 and .76 respectively. The correlation between the TOWRE and GORT-3 fell within the moderate range, while the relationship between the TOWRE and the WRMT-R was stronger. Much like the data that have been reported within the literature, these correlations provide mixed results with regard to list fluency's relationship with reading comprehension.

The literature on list fluency is not nearly as extensive as the literature on R-CBM. The studies that have been conducted have produced evidence of a moderate to moderately strong relationship between list fluency and reading comprehension. However, many of the correlations reported have not been as high as those obtained between R-CBM and reading comprehension measures. In addition, in one study, list fluency accounted for significantly less variance in reading comprehension than R-CBM. These findings indicate that although a relationship between list fluency and reading comprehension exists, it is not as strong as the relationship between R-CBM and reading comprehension. However, the limited number of studies on the topic and the findings reported by Torgesen et al. (1997) necessitate that more research be conducted before any definitive conclusions can be reached. In addition, no research was evident that examines the predictive validity of list fluency measures on statewide outcome measures.

Relationship between Group Measures and Reading Comprehension

Because R-CBM and list fluency measures are administered individually to each student, they may require upwards of two hours to administer to an entire class.

Concerns over time have lead to the development of several group measures of reading fluency. These tests can be administered to an entire class at once, thus requiring less time than R-CBM and list fluency measures. Two examples of group measures that have been developed include the Test of Critical Early Reading Skills (TOCERS; Torgesen, Wagner, Lonigan, & DeGraff, 2002) and the Test of Silent Word Reading Fluency (TOSWRF; Mather, Hammill, Allen & Roberts, 2004).

The TOCERS is a group measure that is currently underdevelopment. It is composed of several subtests that can be used to obtain an index of a student's reading ability. When administering the subtest that measures reading fluency, each student receives a form with five lists composed of real and non-real words. The student's task is to go through each list and mark as many real words as possible in 90 seconds. The score is the number of real words marked correctly minus non-words marked incorrectly. The score for the TOCERS was calculated by examining each column separately. The score in each column is the total number of real words marked correctly minus non-words marked incorrectly times the ratio of real to non-real words in that column. The scores for each column are then added to together to get the total score.

The TOSWRF is a recently published group measure of reading fluency. When administering the TOSWRF, each student is given a form that has lines of real words on it, but there are no spaces between the words. The class is given three minutes to separate as many real words as possible by placing a slash between each word. The score is the number of words separated correctly.

Because the development of measures such as the TOCERS and TOSWRF is fairly recent, little research on group measures exists in the literature. One study by Castillo,

Torgesen, Powell-Smith, and Al-Otaiba (2004) examined the relationship between the TOCERS and school-administered measures of reading comprehension. In this study, the TOCERS, two sets of R-CBM probes (i.e., probes from the Dynamic Indicators of Early Literacy Skills (DIBELS), and the Monitoring Basic Skills Progress (MBSP) program) and a list fluency measure (i.e., the TOWRE) were administered to 282 students in first through third grades. In grades one and two, the students also received the Stanford Achievement Test – 9 (SAT-9), a norm-referenced test of reading comprehension. In grade three, the Florida Comprehensive Assessment Test (FCAT), Florida's statewide outcome assessment, was administered. Two subtests from the FCAT were administered, the FCAT SSS (Sunshine State Standards) and the FCAT NRT (Norm-Referenced Test). Correlations between the TOCERS and the SAT-9 in first and second grades were .49 and .53 respectively. Correlations between the TOCERS and FCAT SSS and FCAT NRT were .24 and .23 respectively. In first and third grades, both sets of R-CBM (*r's* ranged from .64 to .82) probes and the TOWRE (*r's* ranged from .45 in third grade to .78 in first grade) were significantly more correlated with reading comprehension than was the TOCERS. In second grade, R-CBM (r=.78 for both sources) had a stronger relationship with reading comprehension than the TOWRE (r=.62) and TOCERS. There was no significant difference between the TOWRE and TOCERS in terms of their relationship with the SAT-9. Based on these results, the authors concluded that the low to moderate relationship between the TOCERS and reading comprehension was the weakest of the three types of reading fluency measures.

In the same study conducted by Castillo et al. (2004), correlations were obtained between the TOSWRF and two school administered tests of reading comprehension (i.e.,

the SAT-9 and the FCAT). Thirty-eight first graders, 78 second graders, and 101 third graders were administered the TOSWRF. The SAT-9 was also administered to the first and second grade students. The 3rd grade students received the FCAT SSS and FCAT NRT. Correlations of .72 and .60 were found between the TOSWRF and the SAT-9 in grades one and two respectively. Correlations between the TOSWRF and the FCAT SSS and FCAT NRT were .41 and .46 respectively. These correlations indicate that the relationship between reading comprehension and the TOSWRF ranged from moderate to moderately strong, with the correlations declining in magnitude as the grade level increased.

Thus, the limited information available on group measures revealed a wide range of correlations with reading comprehension (i.e., low to moderately strong). In a study comparing one of the group tests (i.e., the TOCERS) with other types of reading fluency measures, the authors found that the TOCERS had the weakest relationship with reading comprehension. In the same study, another group measure (i.e, the TOSWRF) demonstrated correlations with measures of reading comprehension that ranged from moderate to moderately strong. The conflicting results from this study necessitate that more research be conducted in this area before any definitive statements regarding the efficacy of group measures can be made. Additional studies involving other group measures of reading fluency may yield different results. In addition, studies examining the predictive utility of these tests on statewide assessments are warranted.

Conclusions

Many states are using statewide outcome assessments to make "high-stakes" decisions. Because of the importance of these tests, a prediction of performance would

be invaluable. The use of reading fluency measures is one avenue that has been examined. R-CBM, list fluency, and group measures of fluency have been studied with regard to their relationship with reading comprehension. Strong correlations have been found between R-CBM probes and measures of reading comprehension. The relationship between list fluency and reading comprehension was slightly lower, with more moderate correlations evident in the literature. The limited data available on group measures revealed correlations that ranged from low to moderately strong. More data are needed before any definitive statements can be made about group measures of fluency and their relationship with reading comprehension. With regard to predicting outcomes on statewide assessments, researchers have demonstrated that R-CBM is an accurate predictor of performance across many states; however, more research is needed on long-term predictive validity. No research was evident that examined the predictive utility (within or across years) of list fluency or group measures of fluency.

Chapter III

Method

Participants

Participants consisted of 42 students from two elementary schools in northern Florida. These students previously participated in a study conducted in the spring of the 2001-02 school year in which several reading fluency measures were administered to 182 first (n = 91) and second graders (n = 91). Letters of consent were sent to the parents of these students to ask permission to obtain their child's third grade FCAT reading scores. Only the participants for whom signed parental consent was received were included in the study. Of the 42 participants for whom consent was received, 19 took the FCAT during the 2002-03 school year (i.e., 19 students were in second grade during the previous study), while 23 took the statewide assessment during the 2003-04 school year (i.e., 23 students were in first grade during the previous study).

The two schools varied in ethnic and socio-economic diversity during the 2001-02 school year. Approximately 5% of the students at school one were Caucasian, while the majority of the remaining 95% were African-American. The exact composition of the school's population in terms of ethnicity was unavailable. Eighty-three percent of the school's students were eligible for the free or reduced lunch plan. Conversely, approximately 79% of the students at school two were Caucasian, while the remaining 21% were of Asian-American, Hispanic, and African-American backgrounds. Again, an

exact breakdown of the percentage of students in terms of their ethnic background was not available. Nine percent of school two's students were eligible for the free or reduced lunch plan.

In the original study, 81 of the 182 students were from school one, while 101 of the participants were from school two. Of the 42 participants in the current study, five were from school one and thirty-seven were from school two. Therefore, a significant percentage of the students in the current sample were from the school with a low proportion of students on the free or reduced lunch plan (i.e., school two). As was the case with the composition of the schools' populations, the exact composition of both the original and current samples in terms of ethnic background was unavailable. The socio-economic status of the participants across both samples was unavailable as well. However, it should be noted that only one out of forty-two participants failed the FCAT in the current sample (i.e., scored at Level 1). The percentage of students who failed the statewide assessment in this sample is well below the average for the state of Florida (i.e., 23% of third grade students failed the FCAT during the 2002-03 school year). In fact, 40 out of the 42 participants in the sample scored at Level 3 or greater. A score at or above Level 3 corresponds with proficient performance on grade level reading material. Thus, these data suggest that the sample is somewhat homogenous and high performing.

Measures

R-CBM probes. R-CBM probes from the DIBELS (Good & Kaminski, 2001) and the MBSP (Fuchs & Fuchs, 1992) program were administered to each student individually. Both the DIBELS and MBSP probes were administered to determine if probes drawn from different sources would have significantly different relationships with reading

comprehension. The first grade probes from each source were administered to first graders, while the second grade probes were administered to second graders. The measures were administered and scored according to standardized R-CBM procedures (Shinn, 1989). Previous research on the reliability for this measure found high reliability coefficients. Both test-retest and alternate-form estimates ranged from the low .80's to the mid .90's, with most correlations above .90 (Marston, 1989). Research on the construct validity of this measure has produced high correlations as well. Most correlations that have been found between R-CBM and reading comprehension measures range from .73 to .91 (Fuchs, Fuchs, Hosp, & Jenkins, 2001; Fuchs, Fuchs, & Maxwell, 1988; Hintze, Shapiro, Conte, & Basile, 1997; Shinn et al., 1992) in elementary school.

TOWRE - Sight Word Efficiency. The TOWRE Sight Word Efficiency subtest consisted of a list of real printed words that is administered to each student individually. Standardized administration and scoring procedures were used (Torgesen, et al., 1999). Each student was provided with a practice list and asked to read the words. The practice words consisted of several real words. After completing this list, each student had 45 seconds to read as many words as possible aloud from the test list. A valid response included any word read correctly. Any word that was pronounced incorrectly was marked with a slash. In addition, a hesitation of three seconds was scored as incorrect. This measure was scored by counting the total number of words read correctly in 45 seconds. The research on the reliability for this test also found high coefficients. For ages 6-9 years, test-retest reliability was .97 and alternate-form reliability ranged from .93 and .97. Correlations between the TOWRE and measures of reading comprehension ranged from moderate to high (*r's* ranged from .50 to .87; Torgesen et al., 1999).

TOCERS - Real and Non-Real Word Lists. The TOCERS Real and Non-Real Word Lists subtest was a group-administered measure that consisted of a form with five lists of both real and non-real words. Standardized administration and scoring procedures were used during the study (Torgesen et al., 2002). Each student received a practice form and was asked to complete it. The class received thirty seconds to mark as many real words as possible on this short list. After this brief time period, the class was shown the correct responses. Following this demonstration, the students were told to begin the test. Ninety seconds were allotted to mark as many real words as possible on the form. A correct response included any real words marked with a slash. The score for the TOCERS was calculated by examining each column separately. The score in each column was the total number of real words marked correctly minus non-words marked incorrectly times the ratio of real to non-real words in that column. The scores for each column were then added to together to get the total score. Relatively high reliability coefficients have been found for this measure. Alternate-form reliability of .85 in first grade, .91 in second grade, and .93 in third grade have been demonstrated (Torgesen et al., 2002).

FCAT Reading Section. The FCAT is composed of two subtests for reading, the FCAT Sunshine State Standards (SSS) and the FCAT Norm-Referenced Test (NRT). The FCAT SSS is a criterion-referenced measure that is used to help determine the extent to which students are meeting the reading objectives set forth by the Florida Department of Education. Scores are reported in terms of standard scores and five achievement levels (1-low to 5-high). The FCAT NRT is used to compare Florida's students with a national standardization sample. Both subtests measure reading comprehension (Florida Comprehensive Assessment Test, n.d.). Reliability indices above .90 have been found

for the FCAT. Concurrent validity estimates range from .78 to .85 (Florida Department of Education, 2004).

Procedure

Participant recruitment. Several steps were taken to obtain permission from the participants. First, permission was obtained from the school district in which the original study occurred. Next, a request for consent from the two schools was made. Finally, a letter was sent home to the parent(s) of the participants requesting permission to obtain their child's FCAT scores. Only data from those participants who returned a signed copy of the consent form were included in the study.

Administration and scoring of measures. All subjects were administered four reading fluency measures during the spring of the 2001-02 school year. The measures consisted of three individually administered tests and one group-administered test. All measures were administered and scored using standardized procedures.

The participants also took the FCAT during their third grade year. The first graders in the original sample took the FCAT during the spring of the 2003-04 school year. The second graders took the FCAT during the spring of the 2002-03 school year. Their scores were obtained from the program monitoring and evaluation office in the district where the two schools were located.

Examiners. The five examiners who administered the individual measures consisted of four upper-division undergraduate psychology majors and one upper-division undergraduate education major. The examiners were chosen through an interview process. Each student was asked their GPA, major, year in school, experience with

children, and experience administering reading measures. The five most qualified students were chosen based on the answers provided to these questions.

Two training sessions were held to train the examiners to administer and score the individual measures. The first training session was held two weeks before testing began. During the first session, examiners were trained on testing procedures, and administration and scoring rules by a university professional versed in the administration and scoring of curriculum-based measures. The examiners also practiced administering and scoring the measures across three different modeled passages until all examiners met a 100 percent criterion of accuracy. In other words, the university professional read the three passages at different fluency rates (i.e., a disfluent, average, and very fluent reader) until all examiners' scoring of the passages matched the models. The second training session was held a week before testing began. Testing procedures, and administration and scoring rules were reviewed. Also, the examiners received additional practice with the measures. Completion of the second session took one hour.

One examiner was also trained to administer and score the TOCERS. During the 30-minute training session, standardized administration and scoring procedures were reviewed. The examiner also received copies of the scoring templates at the conclusion of the session.

Data collection. Forms A and B of the TOWRE and TOCERS were used in this study. Standardized administration and scoring procedures were used for both measures. The R-CBM passages were chosen from two sources, the DIBELS and MBSP probes. The R-CBM probes from the DIBELS were selected from the oral reading fluency benchmark passages used for end of the year monitoring in each grade level. The probes

from the MBSP were chosen by selecting the 2nd, 12th, and 22nd passage for each grade level. Standardized R-CBM administration procedures were used for both sources of probes.

Testing occurred in two sessions. During the group testing, each class was administered both forms A and B of the TOCERS. During individual testing, the students were administered the fluency measures in the following order: TOWRE form A, 3 R-CBM passages, TOWRE form B, 3 R-CBM passages. In half of the packets, the R-CBM probes from the DIBELS were placed first. In the other half, the R-CBM probes from the MBSP were placed first. The order of the DIBELS and MSBP passages was varied to ensure that the participants did not perform better on any particular set of passages due to fatigue. The order of the passages within each source was not counterbalanced. Individual testing occurred within two empty classrooms at school one. At school two, individual testing occurred in the media center. Inter-scorer reliability was obtained by having one examiner score along with another examiner for approximately ten percent of the participants. For each R-CBM probe and TOWRE form, the number of words read correctly was used to calculate inter-scorer reliability. Inter-scorer reliability for both types of R-CBM probes and the TOWRE was extremely high, with reliability coefficients of .99 obtained across all grades. No reliability data were collected for the TOCERS.

Data analysis. Several data analyses were required to answer the three research questions proposed in the study. Recall that the first purpose was to determine which reading fluency measures are the most accurate predictors of student performance on the FCAT reading sections. To determine the accuracy of a single predictor, linear regression

was used for both the FCAT SSS and FCAT NRT. R^2 values were calculated for each predictor; however, because of the relatively small sample size in the study (i.e., limited power), tests for significant differences between regression coefficients were not conducted. Therefore, effect sizes for regression coefficients were calculated based on Cohen's (1988) formula. The effect sizes for the measures were then compared to one another and ranked according to Cohen's (1988) criteria for interpreting the magnitude of regression effect sizes.

In addition to linear regression, a dominance analysis, a method for comparing the relative importance of predictors in multiple regression (Budescu, 2003), was used. In a dominance analysis, all R^2 values for all possible subset models are examined. These R^2 values are used to measure the relative importance of the predictors in a pairwise fashion (i.e., each possible pair is examined and compared for all possible subset models). After each model is examined for a particular pair of predictors, 3 qualitative values of "dominance" can be assigned: complete, conditional, and general (Budescu, 2003).

The first level of dominance is complete dominance. A predictor is said to completely dominate if its additional contribution to each of the subset models that form the basis for comparison is greater than that of the other predictor. Complete dominance cannot be established between all pairs in all possible subset models. The next level of dominance is conditional dominance. A predictor is said to conditionally dominate if the average additional contribution within each model size is greater for one predictor than the other. As with complete dominance, conditional dominance cannot be established between all pairs of predictors. The last level of dominance (i.e., general dominance)

summarizes the additional contributions of each predictor to all subset models by averaging all the conditional values. If the overall averaged additional contribution is greater for one predictor than another, the predictor is said to generally dominate. General dominance is always possible to establish unless the general measure is identical for a pair of predictors. Thus, through the use of a dominance analysis, it was possible to determine which reading fluency measures completely, conditionally, or generally dominated in terms of their predictive validity on the FCAT reading section (Budescu, 2003).

As was mentioned previously, the second purpose of this study was to determine if there are significant differences in the prediction of FCAT reading scores across grade levels. To determine if the four reading fluency measures predict FCAT performance differently across years, multiple regression with continuous and categorical variables was used. One predictor was grade level (dummy coded, 0 = first grade & 1 = second grade). There was also a predictor for each reading fluency measure (i.e., DIBELS, MBSP, TOWRE, & TOCERS). Finally, 4 interaction variables were included. Each interaction variable was obtained by computing the product of grade level and one of the reading fluency measures. To determine if any of these interactions contributed to the R^2 values, the model with all interactions was compared to a model with no interactions and the change in R^2 was tested for statistical significance. The regression coefficients of any interaction variables that lead to a significant change in R^2 were examined to determine the grade level in which a better prediction was achieved. Regression coefficients with positive values indicated a better prediction in second grade, whereas coefficients with negative values indicated that a better prediction was achieved in first

grade. This analysis was used to determine if a main effect or any differences in individual measures existed.

Finally, to answer the third research question posed in this study (i.e., the predictive validity of the various measures across subtests of the FCAT), the effect sizes and dominance values obtained to answer the first research question were used. The effect sizes calculated for both the FCAT SSS and FCAT NRT were compared to each other using the same procedure outlined above (i.e., Cohen's (1988) criteria for interpreting the magnitude of regression effect sizes). The researcher conducted this analysis to determine if there was a difference in the variance accounted for by each individual fluency measure across subtests. The dominance values derived for both subtests were used to determine if there were any differences between the predictive utility of multiple measures. These values were compared to determine if there were any differences in the dominance levels assigned.

Chapter IV

Results

Data on four reading fluency measures and a state's outcome assessment were collected over a two-year period. Forty-two first and second grade students were administered R-CBM probes from two sources, a list fluency measure, and a group measure of fluency in the spring of the 2001-02 school year. The students also received the reading section of the FCAT (i.e., the FCAT SSS and FCAT NRT) during their third grade year. Linear regression was used to examine the predictive utility of the fluency measures on the FCAT when considered as single predictors. Next, a dominance analysis was used to investigate the relative utility of the measures when considered as multiple predictors. Finally, multiple regression with continuous and categorical variables was used to determine if the measures predicted outcomes differently across grade levels.

Descriptive Statistics

Table 1 provides the mean and standard deviations of each measure by grade level. The participants' scores on the R-CBM probes indicate that as a group, the students in this sample scored above average. For example, participants in this study averaged approximately 95 (SD = 35.32) and 137 (SD = 41.87) correct words per minute (cwpm) on the DIBELS probes in first and second grades respectively. These averages are well above the DIBELS benchmark levels for the spring interval of 40 cwpm in first grade and 90 cwpm in second grade (DIBELS Benchmark Levels, n.d.). In fact, if the standard

deviations for the DIBELS probes in first and second grades are subtracted from the participants' average scores, the resulting scores would still exceed the benchmarks for each grade level. When combined with the distribution of FCAT scores in the sample (e.g., only one student failed the FCAT, 40 out of 42 scored at or above Level 3), this information suggests that the current sample is somewhat homogeneous and high performing. See Figure 1 below for a complete distribution of the students FCAT levels.

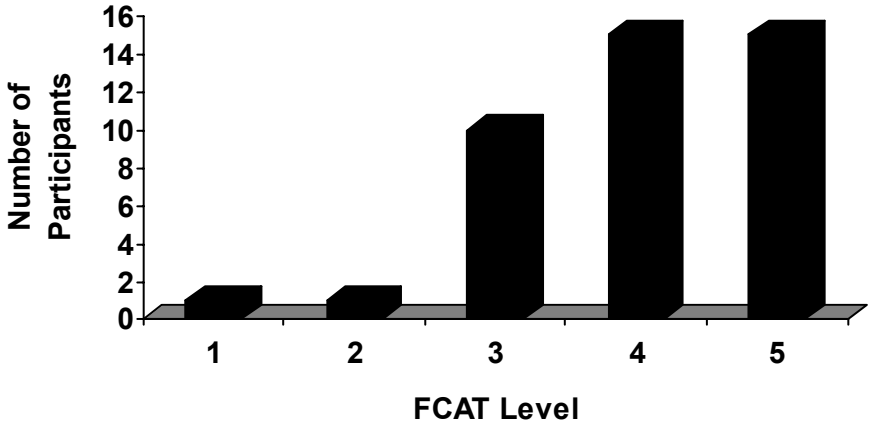
Table 1

Mean and standard deviation for each measure by grade level

Measure	Mean	Standard Deviation	*n*
DIBELS			
Grade 1	94.91	35.32	23
Grade 2	136.79	41.87	19
MBSP			
Grade 1	90.78	37.25	23
Grade 2	133.26	47.40	19
TOWRE			
Grade 1	50.00	11.80	23
Grade 2	61.87	8.95	19
TOCERS			
Grade 1	47.20	10.08	23
Grade 2	50.90	13.50	19
FCAT SSS	367.90	51.60	42
FCAT NRT	688.19	41.47	42

Note. DIBELS = *Dynamic Indicators of Basic Early Literacy Skills* (Good & Kaminski, 2001); MBSP = *Monitoring Basic Skills Progress* (Fuchs & Fuchs, 1992); TOWRE = *Test of Word Reading Efficiency* (Torgesen et al., 1999); TOCERS = *Test of Critical Early Reading Skills* (Torgesen et al., 2002); FCAT SSS = *Florida Comprehensive Assessment Test – Sunshine State Standards* (FCAT, 2004); FCAT NRT = *Florida Comprehensive Assessment Test – Norm-referenced Test* (FCAT 2004).

Figure 1. Frequency polygon depicting distribution of FCAT levels in the sample.



Correlations with 95% confidence intervals among each reading fluency measure and the FCAT SSS and FCAT NRT are shown in Table 2. For each subtest of the FCAT (the correlation between the two subtests was .76 [$p < .01$], suggesting a moderately strong to strong relationship), the R-CBM probes were the most highly correlated, followed by the TOWRE and then the TOCERS. Although the magnitude of the differences between these correlations is somewhat sizeable, the overlap in the 95% confidence intervals for each measure and the FCAT subtests reveal that these differences may not be significant. In other words, because there is a 95% chance that the true population correlations fall anywhere within the obtained confidence intervals, the overlap indicates that the true correlation coefficients between each of the measures and the FCAT subtests could be close to, if not identical to each other. Tests of statistical significance between correlation coefficients were not conducted because of limited power due to the sample size.

Table 2

Correlations among fluency measures and the FCAT SSS and FCAT NRT including 95% confidence intervals

Measure	FCAT SSS	FCAT NRT
DIBELS	.51** (.25, .71)	.68** (.48, .82)
MBSP	.55** (.30, .73)	.67** (.46, .81)
TOWRE	.48** (.21, .69)	.64** (.42, .79)
TOCERS	.24 (-.07, .51)	.32* (.02, .57)

Note. Values enclosed within parentheses represent the lower and upper bound limits of the 95% confidence intervals. DIBELS = *Dynamic Indicators of Basic Early Literacy Skills* (Good & Kaminski, 2001); MBSP = *Monitoring Basic Skills Progress* (Fuchs & Fuchs, 1992); TOWRE = *Test of Word Reading Efficiency* (Torgesen et al., 1999); TOCERS = *Test of Critical Early Reading Skills* (Torgesen et al., 2002); FCAT SSS = *Florida Comprehensive Assessment Test – Sunshine State Standards* (FCAT, 2004); FCAT NRT = *Florida Comprehensive Assessment Test – Norm-referenced Test* (FCAT 2004).

** Correlation is significant at the .01 level (2-tailed).

* Correlation is significant at the .05 level (2-tailed).

Accuracy of Each Fluency Measure in Predicting Outcomes

Linear regression and a dominance analysis were used to address the first research question. First, the accuracy of each individual predictor was examined using linear regression and Cohen's (1988) criteria for calculating and interpreting effect size. Next, the relative accuracy of the predictors when various model sizes and combinations were considered was examined using a dominance analysis.

When investigating the measures as single predictors, R^2 values and effect sizes were calculated, as well as adjusted R^2 values. Effect sizes were calculated by dividing the R^2 value for each measure by one minus the R^2 value (i.e., R^2 / (1-R^2)). Effect sizes were included to provide a qualitative interpretation of the accuracy of each predictor based on Cohen's effect sizes for regression (Cohen, 1988). Small effect sizes

ranged from .02 to .14, medium effect sizes ranged from .15 to .34, and large effect sizes were at or above .35. These interpretations were necessary because tests of significant differences between linear regression coefficients would not have been meaningful due to the small sample size in the study. Adjusted R^2 values were included to adjust for sampling error and the small sample size. Refer to Table 3 for R^2 values, adjusted R^2 values, and effect sizes for each predictor for both the FCAT SSS and FCAT NRT. Because the differences between the R^2 and adjusted R^2 values were minimal, and the effect sizes were calculated from the R^2 values, the adjusted R^2 values are not discussed here.

Table 3

R^2 values with effect sizes and adjusted R^2 values by fluency measure on the FCAT SSS and FCAT NRT

Measure	FCAT SSS	FCAT NRT
DIBELS		
R^2	.26** (.35)	.46** (.85)
Adj. R^2	.24**	.44**
MBSP		
R^2	.30** (.43)	.44** (.79)
Adj. R^2	.29**	.43**
TOWRE		
R^2	.23** (.30)	.40** (.67)
Adj. R^2	.21**	.39**
TOCERS		
R^2	.06 (.06)	.10* (.11)
Adj. R^2	.03	.08*

Note. Values enclosed within parentheses represent effect size. DIBELS = *Dynamic Indicators of Basic Early Literacy Skills* (Good & Kaminski, 2001); MBSP = *Monitoring Basic Skills Progress* (Fuchs & Fuchs, 1992); TOWRE = *Test of Word Reading Efficiency* (Torgesen et al., 1999); TOCERS = *Test of Critical Early Reading Skills* (Torgesen et al., 2002); FCAT SSS = *Florida Comprehensive Assessment Test – Sunshine State Standards* (FCAT, 2004); FCAT NRT = *Florida Comprehensive Assessment Test – Norm-referenced Test* (FCAT 2004).

** Significant at the .01 level (2-tailed).

* Significant at the .05 level (2-tailed).

For the FCAT SSS, the R-CBM probes produced the largest R^2 values and effect sizes. R^2 values and effect sizes for the DIBELS and MBSP were .26 (*ES* = .35) and .30 (*ES* = .43) respectively (*p's* <.01). Both of the effect sizes were large according to Cohen's (1988) criteria. The TOWRE produced the next largest values, with an R^2

value of .23 (p<.01) and a medium effect size of .30. The TOCERS produced a non-significant R^2 value of .06 (p>.05) with an effect size of .06, which was small according to Cohen's effect size criteria.

For the FCAT NRT, both the R-CBM probes and the TOWRE produced R^2 values with effect sizes in the large range. R^2 values for the DIBELS, MBSP, and TOWRE were .46 (*ES* = .85), .44 (*ES* = .79), and .40 (*ES* = .67) respectively (p's < .01). The TOCERS produced an R^2 value of .10 (p<.05) and an effect size of .11, which fell within the small range. Thus, across subtests, administration of the individually administered measures resulted in higher R^2 values and effect sizes, while the TOCERS consistently produced the lowest values.

In addition to the examining the predictive utility of each measure singly, the relative utility of each measure was investigated across multiple model sizes and combinations using a dominance analysis. Based on a pairwise comparison of the additional contribution of each predictor across all possible model combinations and sizes, three qualitative levels of dominance were derived; complete, conditional, and general. Complete dominance was established when one measure's additional contribution to all possible models was greater than another measure's additional contribution. Conditional dominance was established when one measure's average additional contribution to every possible model size was greater than another measure's additional contribution. Finally, general dominance was established when neither complete nor conditional dominance was found. General dominance occurred when one measure's overall average additional contribution was greater than another measure's overall additional contribution across all possible models.

Table 4 contains all of the possible models (including R^2 values) for predicting FCAT SSS outcomes and the additional contribution to R^2 that would occur if one of the four predictors was added to a particular model. More specifically, the column on the left of the table contains all the possible models, the second column contains the R^2 value for each model, and the four columns on the right provide the additional contribution to R^2 that would be gained by adding the corresponding predictor. The values in the four columns to the right were used to establish dominance for each of the predictors. For example, when the MBSP probes were added to the single predictor models of the DIBELS, TOWRE, and TOCERS, its additional contributions of .060, .079, and .262 respectively were greater than the additional contributions of any of the other measures to those models. In other words, when compared in a pairwise fashion to the DIBELS, TOWRE and TOCERS' additional contributions to those models, the MBSP probes additional contribution was greater in each instance. This pattern occurred when the MBSP probes were added to the models containing two and three predictors as well. Therefore, complete dominance was established for the MBSP probes when compared with each of the other predictors.

However, complete dominance was not established for every pairwise comparison for all of the measures. For example, when the TOWRE was compared to the TOCERS across the models, the additional contribution of the TOWRE when added to the one predictor models was greater than the TOCERS, but the opposite occurred when the two measures were added to the two predictor models (e.g., the additional contribution of the TOCERS of .019 to the model including the DIBELS and MBSP probes was greater than the TOWRE's additional contribution of .000). Because complete dominance could not

be established, the next step was to determine if conditional dominance could be established by examining the average additional contribution of the TOWRE and TOCERS to each of the model sizes. These values are found in the four rows that contain $k = 0$, $k = 1$, $k = 2$, and $k = 3$. The symbol k represents the model size prior to adding another predictor to the model. Thus, before adding any predictors to the models or in the one predictor models, the TOWRE's average additional contributions of .231 and .061 across the model sizes were greater than the TOCERS average additional contributions of .059 and .009. However, the opposite relationship occurred when the two and three predictor models were examined. The TOCERS' average additional contributions when the number of predictors in the models equaled two and three were .013 and .020 respectively, while the TOWRE's were .001 in both instances. Therefore, conditional dominance could not be established, leading the researcher to examine which measure was generally dominant. General dominance was investigated by examining the overall average additional contributions of the two measures across all model sizes. The overall average additional contributions of the measures can be found in the last row of the table containing the label overall average. When the additional contributions of the TOWRE and TOCERS were averaged across the various model sizes, the overall average additional contribution of the TOWRE of .074 was greater than the TOCERS' value of .025, indicating that the TOWRE generally dominated the TOCERS. The reader interested in a more detailed explanation of a dominance analysis is referred to Budescu (2003). A summary of the results for the FCAT SSS and FCAT NRT for the current study is provided below.

For the FCAT SSS, the MBSP probes completely dominated the DIBELS, TOWRE, and TOCERS. Thus, the additional contribution of the MBSP probes was greater for every possible model when compared to each of the three other measures. Next, the DIBELS completely dominated the TOWRE, but only generally dominated the TOCERS. This means that the additional contribution of the DIBELS was greater that the TOWRE for all possible models, but only the average additional contribution was greater than the TOCERS. Finally, the TOWRE generally dominated the TOCERS, indicating that the TOWRE's average additional contribution to each model was greater than that of the TOCERS. Refer to Table 4 for the dominance analysis numerical values for the FCAT SSS.

Table 4

Dominance analysis for fluency measures prediction of FCAT SSS scores

		Additional contribution of:			
Subset Model	P^2yx	DIBELS	MBSP	TOWRE	TOCERS
k=0 average	0	.260	.304	.231	.059
DIBELS	.260		.060	.002	.007
MBSP	.304	.016		.006	.017
TOWRE	.231	.031	.079		.002
TOCERS	.059	.208	.262	.175	
k=1 average		.085	.134	.061	.009
DIBELS MBSP	.320			.000	.019
DIBELS TOWRE	.262		.058		.007
DIBELS TOCERS	.267		.072	.002	
MBSP TOWRE	.310	.010			.012
MBSP TOCERS	.321	.018		.001	
TOWRE TOCERS	.233	.036	.089		
k=2 average		.021	.073	.001	.013
DIBELS MBSP TOWRE	.320				.020
DIBELS MBSP TOCERS	.339			.001	
DIBELS TOWRE TOCERS	.269		.071		
MBSP TOWRE TOCERS	.322	.018			
k=3 average		.018	.071	.001	.020
DIBELS MBSP TOWRE TOCERS	.340				
Overall average		.096	.146	.074	.025

Note. The column labeled P^2yx represents the variance in *Y* explained by the model appearing in the corresponding row. Columns labeled with fluency measures contain the additional contributions to the explained variance gained by adding the column variable (i.e. fluency measure) to the row model. Blank cells indicate that data are not applicable. DIBELS = *Dynamic Indicators of Basic Early Literacy Skills* (Good & Kaminski, 2001); MBSP = *Monitoring Basic Skills Progress* (Fuchs & Fuchs, 1992); TOWRE = *Test of Word Reading Efficiency* (Torgesen et al., 1999); TOCERS = *Test of Critical Early Reading Skills* (Torgesen et al., 2002); FCAT SSS = *Florida Comprehensive Assessment Test – Sunshine State Standards* (FCAT, 2004); *k* = the number of predictors in the subset model.

For the FCAT NRT, the DIBELS probes completely dominated the MBSP probes and the TOWRE, but only generally dominated the TOCERS. Thus, the additional contribution of the DIBELS was greater for every possible model when compared to the MBSP and TOWRE, but only the average additional contribution was larger than the TOCERS' contribution. Next, the MBSP probes generally dominated both the TOWRE and TOCERS, indicating that the average additional contribution across the different models for the MBSP was greater than the other two measures. Finally, the TOWRE generally dominated the TOCERS, once again indicating that the average additional contribution across the different models was greater for the TOWRE. Refer to Table 5 for the dominance analysis numerical values for the FCAT NRT.

A general pattern emerged regarding the dominance levels for the measures across the FCAT SSS and FCAT NRT. The R-CBM probes were the most "dominant", followed by the TOWRE and then the TOCERS. In other words, the R-CBM probes tended to add more variance to the models when compared with the alternate measures, while the TOWRE tended to add more variance than the TOCERS.

Table 5

Dominance analysis for fluency measures prediction of FCAT NRT scores

		Additional contribution of:			
Subset Model	P^2yx	DIBELS	MBSP	TOWRE	TOCERS
k=0 average	0	.459	.444	.403	.100
DIBELS	.459		.000	.002	.012
MBSP	.444	.015		.006	.016
TOWRE	.403	.058	.047		.005
TOCERS	.100	.371	.360	.308	
k=1 average		.148	.136	.105	.011
DIBELS MBSP	.459			.002	.015
DIBELS TOWRE	.461		.000		.014
DIBELS TOCERS	.471		.003	.004	
MBSP TOWRE	.450	.011			.018
MBSP TOCERS	.460	.014		.008	
TOWRE TOCERS	.408	.067	.060		
k=2 average		.031	.021	.005	.016
DIBELS MBSP TOWRE	.461				.016
DIBELS MBSP TOCERS	.474			.003	
DIBELS TOWRE TOCERS	.475		.002		
MBSP TOWRE TOCERS	.468	.009			
k=3 average		.009	.002	.003	.016
DIBELS MBSP TOWRE TOCERS	.477				
Overall average		.162	.151	.129	.036

Note. The column labeled P^2yx represents the variance in *Y* explained by the model appearing in the corresponding row. Columns labeled with fluency measures contain the additional contributions to the explained variance gained by adding the column variable (i.e. fluency measure) to the row model. Blank cells indicate that data are not applicable. DIBELS = *Dynamic Indicators of Basic Early Literacy Skills* (Good & Kaminski, 2001); MBSP = *Monitoring Basic Skills Progress* (Fuchs & Fuchs, 1992); TOWRE = *Test of Word Reading Efficiency* (Torgesen et al., 1999); TOCERS = *Test of Critical Early Reading Skills* (Torgesen et al., 2002); FCAT NRT = *Florida Comprehensive Assessment Test – Norm Referenced Test* (FCAT, 2004); *k* = the number of predictors in the model.

The Predictive Utility of Fluency Measures Across Grade Levels

Multiple regression with continuous and categorical variables was used to examine the issue of prediction across grade levels. Both main and individual effects for grade level were examined. Positive and negative values indicated the direction in which a better prediction was obtained. In other words, a positive value meant a better prediction in second grade, while a negative value meant a better prediction was achieved in first grade. No main effect was found for the FCAT SSS (F = .814, p>.05) or the FCAT NRT (F = 1.69, p>.05). No individual effects were found for grade level for either subtest as well. For the FCAT SSS, the t values were 1.37, -0.75, -0.58, and 0.42 (p's > .05) for the DIBELS, MBSP, TOWRE, and TOCERS respectively. For the FCAT NRT, the t values were 1.93, -1.06, -0.87, and 0.72 (p's > .05) for the DIBELS, MBSP, TOWRE, and TOCERS respectively.

Differences in the Predictive Utility of the Fluency Measures Across FCAT subtests

Both linear regression and a dominance analysis were used to address the third research question. First, the predictive utility for a single predictor across the FCAT SSS and FCAT NRT was compared using linear regression and Cohen's (1988) criteria for calculating and interpreting effect size. Next, the relative predictive utility of the measures when multiple predictors were considered was compared across the subtests using a dominance analysis.

Linear regression was used to determine if the individual reading fluency measures predicted differently across the FCAT SSS and FCAT NRT. R^2 values and effect sizes were calculated for each measure. The effect sizes were used to compare the predictive utility of each measure across the FCAT subtests because statistically significant

differences among regression coefficients would have been difficult to detect due to limited power. The effect sizes were compared in the same manner described previously (i.e., using Cohen's (1988) criteria for comparing effect sizes of regression coefficients). Refer back to Table 3 for R^2 values and effect sizes for each measure for both subtests of the FCAT.

The accuracy of the predictors was only slightly different between the two subtests according to Cohen's (1988) criteria. For the FCAT SSS, the MBSP and DIBELS probes obtained the highest R^2 values and effect sizes (i.e., large effect sizes), followed by the TOWRE (i.e., medium effect size), and TOCERS. For the FCAT NRT, both sets of R-CBM probes and the TOWRE obtained the highest R^2 value and effect sizes (i.e., large effect sizes), followed by the TOCERS. Thus, only the TOWRE produced a different prediction across subtests, producing a medium effect size on the FCAT SSS and a large effect size on the FCAT NRT.

However, the magnitudes of the effect sizes were somewhat larger for the FCAT NRT than the FCAT SSS. Across all the predictors, effect sizes were calculated for the FCAT NRT that were approximately twice as large as those calculated for the FCAT SSS. Thus, it would appear that the ratio of the amount of variance accounted for by the measures to error was twice as large on the FCAT NRT than on the FCAT SSS.

Results from the dominance analyses mentioned previously were used to determine if the relative predictive utility of the measures was different across subtests when multiple predictors were considered. Refer back to Tables 4 and 5 for the dominance analyses numerical values. Across subtests, the R-CBM probes were the most dominant, although the MBSP probes tended to be more dominant on the FCAT SSS and the DIBELS probes

tended to be more dominant on the FCAT NRT. The TOWRE was generally dominant over the TOCERS across both subtests as well. One difference across subtests was that the additional contribution of the TOWRE and TOCERS to some of the models was greater on the FCAT NRT. On the FCAT SSS, the TOWRE was completely dominated by both sets of R-CBM probes, but was only completely dominated by the DIBELS on the FCAT NRT (the TOWRE was generally dominated by the MBSP). With regard to the TOCERS, the group measure was completely dominated by the MBSP and generally dominated by the other two individually administered measures for the FCAT SSS, while it was only generally dominated by all three of the individually administered measures on the FCAT NRT.

Chapter V

Discussion

This study examined the utility of various reading fluency measures as across grade predictors of performance on the reading subtests of the FCAT in third grade. Several analyses were conducted to determine the following: the extent to which each measure is an accurate predictor, both in isolation and in combination with the other measures, whether administration of the measures in first or second grade affects the accuracy of the prediction, and whether a difference exists in the accuracy of the predictors across the FCAT SSS and FCAT NRT. A summary of the findings and implications for research and practice follows.

Summary of Findings

Recall that linear regression was used to determine the accuracy of each measure when a sole predictor was considered. Due to the relatively small sample size, tests for significant differences between regression coefficients were not conducted. Therefore, Cohen's (1988) formula for calculating effect sizes from regression coefficients was used to compare the measures, along with Cohen's criteria for interpreting the magnitude of the effect sizes.

The results of this analysis for the FCAT SSS indicated that the R-CBM measures had the largest effect sizes. The DIBELS and MBSP probes accounted for 26% and 30% of the variance respectively, resulting in large effect sizes for both measures. The

TOWRE accounted for 23% of the variance, resulting in a medium effect size. Finally, the TOCERS had the least predictive power of the four measures, resulting in a small effect size.

The results of the analysis for the FCAT NRT were similar to those from the FCAT SSS. The individually administered measures (i.e., the DIBELS, MBSP, and TOWRE) accounted for 40-46% of the variance, resulting in large effect sizes for all three tests. The TOCERS accounted for the least amount of variance, resulting in a small effect size. Thus, the only difference between the analyses for the FCAT SSS and FCAT NRT in terms of Cohen's (1988) criteria for interpreting the magnitude of effect sizes was the TOWRE's large effect size on the FCAT NRT and medium effect size on the FCAT SSS.

Although similar results were observed across subtests in terms of Cohen's (1988) criteria, an interesting finding was noted regarding the magnitude of the effect sizes. The effect size for each of the measures was approximately twice as large for the FCAT NRT than the FCAT SSS. For example, the effect size for the DIBELS probes more than doubled from .35 on the FCAT SSS to .85 on the FCAT NRT. Although both effect sizes were considered large according to Cohen's criteria, the signal to noise ratio (i.e., the ratio of predictive power to error) was more than twice as large on the FCAT NRT. The effect size for the TOWRE more than doubled on the FCAT NRT as well, increasing from .30 on the FCAT SSS to .67 on the FCAT NRT. Thus, the effect size increased from medium on the former subtest to large on the latter subtest. Interestingly, these results are consistent with correlational research that has been conducted on both subtests of the FCAT. In an analysis of the relationship between various reading fluency measures and the FCAT, Castillo et al. (2004) reported higher correlations between the fluency

measures and the FCAT NRT than between the measures and the FCAT SSS. No research is currently available that examines the reasons for these observed differences.

To examine the relative utility of the four reading fluency measures when considered as multiple predictors, a dominance analysis was used. Each of the measures was compared in a pairwise fashion across all possible model combinations and sizes. Based on the pairwise comparisons, one of three levels of dominance was assigned for each pair of measures, complete, conditional, or general.

Results for the FCAT SSS indicated that the R-CBM measures tended to be the most "dominant" measures. In other words, their additional contribution to the various models tended to be greater than the alternate fluency measures. The TOWRE was the next most dominant measure, followed by the TOCERS. The additional contribution of the TOWRE tended to be slightly better than the TOCERS across the various model combinations and sizes.

Results for the FCAT NRT were somewhat similar to those found for the FCAT SSS. The additional contribution of the R-CBM measures across the models tended to be greater than the alternate fluency measures, although less so than for the FCAT SSS. Both the TOWRE and TOCERS accounted for a greater amount of variance across some of the model sizes than the R-CBM measures. Once again, the additional contribution of the TOWRE tended to be slightly greater than the TOCERS across the various models. Thus, the main difference between the fluency measures across the FCAT SSS and FCAT NRT was that the additional contribution of the alternate measures was slightly greater on the FCAT NRT than the FCAT SSS.

However, it is important to note that one limitation of the dominance analysis may be that the size of the R^2 values are not considered. In the current analyses, the majority of the increases in R^2 that occurred after adding a predictor to a model were less than .1 and some were below .01. Although the dominance analyses provided information regarding which measures tend to produce greater increases in the R^2 values, it provides little information regarding the practical importance of those increases without examining the R^2 values themselves. Therefore, when making decisions regarding the practical importance of adding a predictor to a model, it is important to examine the size of the increase in the R^2 value as well as the dominance label assigned to the measure. Examining both sources of information will increase the probability of making a practically meaningful decision when choosing from multiple predictors.

Finally, multiple regression with continuous and categorical variables was used to determine if the measures predicted differently across grade levels. Both main effects and differences within individual measures were examined. In addition, the direction of the differences were examined to determine the grade level in which the better prediction was obtained. Results indicated that there was no main effect or differences in predictive power for the individual measures between grade levels. The lack of differences across the measures in first and second grade indicates that no statistically significant differences were evident in their predictive power across grades. However, these results must be interpreted cautiously because of the difficulty in detecting interaction effects in field research. McClelland and Judd (1988) demonstrated that limited statistical power in field research due to factors such as smaller sample sizes and increased measurement error (due to less stringent experimental control) makes interaction effects more difficult

to detect. A potential example of this phenomenon was noted when examining the difference in prediction across grade levels for the DIBELS probes. The prediction obtained in second grade for the DIBELS probes was somewhat better than in first grade (t = 1.37, p = .06). Although these results are not technically statistically significant, they suggest that future research with a larger sample (i.e., more power) may find a difference in the prediction obtained between grade levels.

Overall, the results of the aforementioned analyses are consistent with a long line of research that has demonstrated a relationship between R-CBM probes and reading comprehension (e.g., Deno, Mirkin, & Chiang, 1982; Fuchs, Fuchs, & Maxwell, 1988; Shinn, Good, Knutson, Tilly, & Collins, 1992) as well as the efficacy of R-CBM probes as within-year predictors of performance on statewide assessments, including the FCAT (e.g., Good, Simmons, & Kame'enui, 2001; Buck & Torgesen, 2003; McGlinchey & Hixson, 2004). However, little research has been conducted on the validity of R-CBM as a predictor of across year performance. The results of the current study are consistent with the one study that was found in the literature examining across year prediction. In a study examining the Oregon Statewide Assessment (OSA), Crawford, Tindal, and Steiber (2001) showed that R-CBM probes administered in second grade were accurate predictors of OSA performance in third grade. The current study provides preliminary evidence for extending the findings from the Crawford et al. (2001) study to first grade as well. Both sources of R-CBM probes produced large effect sizes on both subtests of the FCAT and no significant differences were found between grade level prediction for either measure.

The results of the study also provide preliminary evidence for the predictive validity of alternate measures of reading fluency. The medium and large effect sizes found for the TOWRE suggest that the measure may have some predictive power. Although no research is evident on the predictive utility of list fluency measures, this finding is consistent with research that has demonstrated a moderate to moderately strong relationship between measures of list fluency and comprehension measures (e.g., Jenkins, et al., 2003; Torgesen et al., 1999).

In addition, small effect sizes were noted for the TOCERS across subtests. Again, no research was evident examining the predictive utility of group fluency measures, however, the small effect sizes that were calculated for the TOCERS are consistent with a previous study conducted by Castillo et al., (2004) in which a small to moderate relationship between the TOCERS and measures of reading comprehension was found. Although small effect sizes may be reason for skepticism, the current study used a single subtest from an experimental group measure. Future research on the predictive validity of group measures could yield different results.

Implications for Research and Practice

Results of this study may help practitioners determine which assessments have the potential to accurately predict performance on statewide assessments. The measures selected to screen for at-risk readers would depend on whether schools are interested in using a single or multiple predictors. If educators are interested in administering a single measure, then either R-CBM probes or the TOWRE should be considered. The large effect sizes found for the R-CBM probes across both subtests of the FCAT indicated that the measures resulted in good predictive power on both the FCAT SSS and FCAT NRT.

A medium and large effect size was calculated for the TOWRE across the FCAT SSS and FCAT NRT respectively. The smaller effect size for the TOWRE on the former subtest suggest that the R-CBM probes may have slightly better predictive power, but educators should be cautious when interpreting these results because of the size and composition of the sample (e.g., tests for significant differences could not be conducted, the vast majority of the sample passed the FCAT).

When considering multiple predictors, educators must weigh their options regarding which measures to include. Results of the dominance analysis indicated R-CBM probes should be one of the measures included in a screening battery. Both the DIBELS and MBSP probes were the most "dominant" measures across both subtests of the FCAT. The selection of which additional measures to include is a decision that educators should make based on practical considerations. Neither alternate measure's additional contribution to the models including the R-CBM probes was substantial. Although the TOWRE's additional contribution tended to be greater across both subtests, one would need to consider if the additional time that administering the measure individually to each student would take is worth the relatively small increase in predictive power. Conversely, the additional contribution of the TOCERS to the various models tended to be the smallest. However, the amount of time saved by administering such a group measure may be worth considering. The particular needs and resources at the disposal of schools would need to be considered when selecting the measures within a screening battery.

The results of this study also suggest the possibility of screening for at-risk readers in first or second grade using reading fluency measures. Findings from research on reading have consistently demonstrated that early intervention is necessary to improve the

outcomes of struggling readers (Francis, Shaywitz, Stuebing, Shaywitz, & Fletcher, 1994; Good, Simmons, & Smith, 1998; Juel, 1988; Torgesen & Burgess, 1998). The results of the current study provide preliminary evidence that measures of reading fluency could be administered in first and/or second grade to identify at-risk readers. Screening in first grade would allow educators to intervene early and provide students with additional time to close the gap with their peers. The measures could also be administered in second grade to screen for those students who were not identified in first grade (e.g., students who are new to the school, were not identified by the initial screening). Thus, if future research were to extend the findings from the current study and screening cut-scores were developed, interventions could be provided to at-risk readers in both grades to improve their performance on statewide assessments and overall reading outcomes.

However, whether considering the use of a single screening measure or a battery of measures, it is important for educators to remember that error is involved in the assessment process. Because students must master skills in domains such as phonological awareness, phonics, fluency, vocabulary, and comprehension, a myriad of variables contribute to the variability in the reading scores obtained from statewide assessments (Buly & Valencia, 2002). Despite the medium and large effect sizes calculated for both sources of R-CBM probes and the TOWRE, a significant proportion of the variance in FCAT reading scores was unaccounted for by the measures.

The amount of error involved highlights the need for screening in other areas that are associated with reading outcomes (e.g., vocabulary) as well as continuous progress monitoring. The addition of other screeners with adequate predictive utility would reduce the amount of error involved in identifying at-risk readers, but some false positives and

false negatives would still occur. Therefore, it is important for schools to continue to monitor the reading progress of all students (e.g., using R-CBM measures). Students identified as low-risk for reading failure might be assessed three to four times per year to ensure adequate progress, while those who are flagged as at-risk readers could be monitored more frequently to evaluate the effects of interventions that are implemented. Thus, should schools decide to use reading fluency measures to screen for at-risk readers, educators would need to keep in mind that the examination of other variables (e.g., vocabulary) and continuous progress monitoring would reduce the chances of providing services that are not commensurate with the student's need.

In fact, the amount of error introduced by other variables is one of several potential contributors to the apparent differences in the predictive utility of the reading fluency measures across the FCAT SSS and FCAT NRT. Although definitive statements cannot be made regarding the difference in prediction across subtests, the results from the current study suggest that one could expect a better prediction of performance on the FCAT NRT than the FCAT SSS using reading fluency measures. Because little research has been conducted examining these differences, the possible explanations offered for this finding are theoretical.

One such explanation for the lower correlations and predictive power of the fluency measures on the FCAT SSS is the open-ended response format of some of the items. Perhaps the written response requirement leads to more variability in scores than a format that consists of solely multiple-choice questions. Another possibility related to the open-ended response format is that scoring error is introduced when examiners are asked to score written responses. Finally, another proposed explanation for the observed

differences between the subtests is that the reading fluency measures are not sensitive to vocabulary differences in students taking the FCAT SSS. In other words, the measures are not sensitive to the fact that students with less vocabulary development would perform worse on the FCAT SSS. Such a difference highlights the need for interventions with a broad focus or alternatively highlights the concerns with focusing interventions solely on reading fluency. Regardless of the measure used to screen for at-risk readers, interventions implemented to improve outcomes often will need to target decoding and comprehension skills, as well as skills within the areas of language and vocabulary. However, more research is needed on reasons for the observed differences between the two subtests before any definitive statements can be made.

This study may have broader implications for educators attempting to improve the reading outcomes of at-risk students as well. Because reading problems begin early and are associated with negative academic and life outcomes (e.g., Stanovich, 1986; Orton Dyslexia Society, 1986; National Adult Literacy Survey, 1992), a significant need exists to identify those at-risk students and intervene early to prevent the persistent difficulty associated with early reading struggles. In addition, educators are under pressure from external sources (e.g., laws such as NCLB, state departments of education) in charge of student performance criteria, retention policy, schools grades, etc. Such high-stakes decisions have created an increased need for tools that can reliably identify at-risk students. Reliable screening measures might help prevent a significant proportion of the reading difficulties reported by the NAEP (2003), thereby meeting many of the demands of the external sources.

Although predicting performance on statewide assessments will not accomplish all of these goals in and of itself, the accurate identification of at-risk readers is an important step in the right direction. This study provides further evidence that R-CBM probes can provide a common metric for both general and special education for predicting performance and making instructional changes when needed. Providing a common metric for general and special education is important because laws such as NCLB and the Individuals with Disabilities Education Improvement Act (IDEIA) require that both groups of students perform adequately in relation to the general curriculum (NCLB 2002; Individuals with Disabilities Education Improvement Act [IDEIA], 2004). For example, NCLB mandates that *all* students read at grade level by the end of third grade, including both general and special education students. Not only is the progress of the entire school tracked from year to year, but the data are disaggregated to ensure that both groups of students are improving. In other words, schools cannot rely on the scores for general education students to demonstrate that they are making progress (i.e., AYP). Schools also must ensure that students receiving services under IDEIA meet grade-level standards as well. Ensuring that special education students meet general education standards is consistent with the spirit of IDEIA, which calls for judging students performance in relation to the general education curriculum. Thus, having a common metric for both groups of students allows educators to examine the performance of their special education students in relation to general education students.

Despite the apparent usefulness of such measures, they do not tell educators specifically how to intervene with a struggling reader. Therefore, it is imperative that educators do not begin teaching to the screening measures. Although improving reading

fluency is a worthwhile goal, educators need to ensure that skills needed to improve fluency (e.g., phonemic awareness, decoding) are instructed, not strategies for taking a screening test. Such instruction in test-taking strategies essentially would defeat the purpose of the screening measures (i.e., the identification of at-risk readers who need supplemental instruction and/or intense intervention). Through screening and additional diagnostic assessment, the focus of instruction can be reallocated from test preparation to intensive instruction in the academic skills needed to pass the statewide assessment and improve reading outcomes.

Limitations

The results of the current study must be interpreted in light of several limitations. One limitation was the relatively small sample size that resulted from the low parental consent letter return rate. Because of the low sample size, tests for significant differences between the regression coefficients were not conducted. In addition, the low sample size may have made the detection of differences in the fluency measures' prediction across grade levels more difficult.

Another limitation of the current study involves the homogeneity of the sample. As was previously mentioned, a significant proportion of the participants were from a high-SES school. The homogeneity of the sample does not allow the results of the study to be generalized to the population within the state of Florida or other regions of the country. In addition, the vast majority of the participants passed the FCAT, eliminating the possibility of examining the accuracy of screening decisions derived from the measures.

Finally, the time frame in which the measures were administered may be another limitation. The fluency measures were only administered in the spring of the 2001-02

school year, following the administration of the FCAT in the schools. It is possible that administration of the fluency measures at different points of the year (e.g., fall, winter) could lead to different results. In other words, administration of these measures during the fall, winter, or spring could lead to different screening decisions, thereby affecting the reliability of the results.

Directions for Future Research

Future research should examine a variety of issues involving reading fluency measures and statewide assessments such as the FCAT. Future studies examining the long-term predictive validity of reading fluency measures should include a larger and more diverse sample. Such a sample would allow for more statistically sound analyses as well as generalization of the results. Including a larger and more diverse sample would also allow for the development of specific cut-points used for making screening decisions.

Further investigation of the alternate measures of reading fluency is warranted as well. Both the TOWRE and the TOCERS accounted for variance in FCAT reading scores in the current study that resulted in small to large effect sizes. However, because of the limitations of the sample, more research is needed before definitive statements regarding their utility can be made. Finally, the differences between the FCAT SSS and FCAT NRT should be examined. Researchers need to determine if the differences are due to vocabulary demands, response demands, scoring error, and/or other variables. An examination of these proposed explanations could impact the nature of the screening measures used as well as policy decisions.

Conclusion

The current accountability climate has forced educators to rethink the manner in which students are taught reading. Schools and their students are being held accountable for their performance on statewide assessments, exacerbating the need for early identification and intervention. A myriad of researchers have demonstrated that R-CBM is a reliable and valid within-year predictor of performance on statewide assessments. However, in light of the research on reading intervention, the identification of struggling students in third or fourth grade may be too late. The current study provides preliminary evidence for R-CBM probes as predictors of performance in first and second grades. Preliminary evidence for the long-term predictive validity of list fluency measures such as the TOWRE is provided as well. The need for future research on these measures, including group measures of reading fluency, is also highlighted.

References

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

American Management Survey (2001). *National Institute for Literacy Facts Sheets Overview Page.* Retrieved January 25, 2005, from http://www.nifl.gov/nifl/facts/facts_overview.html

Barger, J. (2003). *Comparing the DIBELS Oral Reading Fluency indicator and the North Carolina end of grade reading assessment* (Technical Report). Asheville, NC: North Carolina Teacher Academy.

Buck, J., & Torgeson, J. (2003). *The relationship between performance on a measure of oral reading fluency and performance on the Florida Comprehensive Assessment Test* (Technical Report 1). Tallahassee, FL: Florida Center for Reading Research.

Budescu, D.V. (2003). The dominance analysis approach for comparing predictors in multiple regression. *Psychological Methods, 8 (2)*, 129-148.

Buly, M.R., & Valencia, S.W. (2002). Below the bar: Profiles of students who fail state reading assessments. *Educational Evaluation and Policy Analysis, 24 (3)*, 219-239.

Bureau of Labor Statistics (n.d.). *Bureau of Labor Statistics Tomorrow's Jobs Page.* Retrieved March 15, 2004, from http://www.bls.gov/oco/oco2003.htm

Castillo, J. M., Torgeson, J. K., Powell-Smith, K. A., & Al Otaiba, S. (2004). Relationships of five reading fluency measures to reading comprehension in first through third grade. Manuscript submitted for review.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale: Lawrence Erlbaum Associates, Inc.

Crawford, C., Tindal, G., & Stieber, S. (2001). Using oral reading rate to predict student performance on statewide achievement tests. *Educational Assessment, 7 (4)*, 303-323.

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

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

DIBELS Benchmark Levels (n.d.). Retrieved January 27, 2005, from http://dibels.uoregon.edu/benchmark.php

Fisher, C.W., Berliner, Filby, Marliave, Cahen, Dishaw, & Moore (1978). *Teaching behaviors, academic learning time and student achievement: Final report of phase III-B, beginning teachers evaluation study* (Technical Report No. V-1). San Francisco, CA: Far West Laboratory for Educational Research and Development. (ERIC Document Reproduction Service No. ED183525)

Fletcher, J.M., & Lyon, G.R. (1998). Reading: A researched-based approach. In W.M. Evers (Ed.), *What's gone wrong in America's classrooms* (pp. 50-72). Stanford: Hoover Institution Press.

Florida Comprehensive Assessment Test (n.d.). *Florida's Comprehensive Assessment Test Home Page*. Retrieved March 24, 2004, from http://www.firn.edu/doe/sas/fcat.htm

Florida Department of Education (n.d.a). *Fact Sheet: NCLB and Adequately Yearly Progress.* Retrieved January 25, 2005 from http://www.fldoe.org/NCLB/FactSheet-AYP.pdf

Florida Department of Education (n.d.b). *Third Grade Reading 2003 State Profile.* Retrieved May 9, 2004 from http://www.firn.edu/doe/commhome/pdf/3statesum.pdf

Florida Department of Education (2004). *2003-04 assessment & accountability briefing book.* Retrieved January 26, 2005, from http://www.firn.edu/doe/sas/fcat/pdf/fcataabb.pdf

Francis, D.J., Shaywitz, S.E., Stuebing, K.K., Shaywitz, B.A., & Fletcher, J.M. (1994). Measurement of change: Assessing behavior over time and within a developmental context. In G.R. Lyon (Ed.), *Frames of reference for the assessment of learning disabilities: New views on measurement issues* (pp. 29-58). Baltimore: Brookes.

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

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

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

Good, R.H., & Kaminski, R.A. (2001). *Dynamic Indicators of Basic Early Literacy Skills* (5th ed.). Eugene, OR: Institute for the Development of Educational Achievement.

Good III, R.H., Simmons, D.C., & Kame'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)*, 257-88.

Good, R.H., Simmons, D.C., & Smith, S.B. (1998). Effective academic interventions in the United States: Evaluating and enhancing the acquisition of early reading skills. *School Psychology Review, 27 (1)*, 45-56.

Hasbrouck, J.E., & Tindal, G. (1992). Curriculum-based oral reading fluency norms for students in grades 2 through 5. *Teaching Exceptional Children, 24 (3)*, 41-44.

Hawley, W.D., & Rosenholtz, S.J. (1984). Effective Teaching. *Peabody Journal of Education, 61 (4),* 15-52.

Hintze, J.M., Shapiro, E.S., Conte, K.L, & 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* (4), 535-553.

Hosp, M.K. (2004). *Variables that affect the correlation between fluency and accuracy with a measure of reading comprehension*. Manuscript in preparation.

Individuals with Disabilities Education Improvement Act, U.S.C. H.R. 1350 (2004).

Jenkins, J.R., Fuchs, L.S., van der Broek, P., Espin, C., & Deno, S.L., (2003). Sources of individual differences in reading comprehension and reading fluency. *Journal of Educational Psychology, 95 (4)*, 719-729.

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

Jimerson, S. (1999). On the failure of failure: Examining the association between early grade retention and education and employment outcomes during late adolescence. *Journal of School Psychology*, 37(3), 243-272.

Jimerson, S., Carlson, E., Rotert, M., Egeland, B., & Sroufe, A. (1997). A prospective, longitudinal study of the correlates and consequences of early grade retention. *Journal of School Psychology*, 35(1), 3-25.

Juel, C. (1988). Learning to read and write: A longitudinal study of 54 children from first through fourth grades. *Journal of Educational Psychology, 80 (4)*, 437-447.

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

Lyon, G.R. (1995). Research in learning disability: Contributions from scientists supported by the National Institute of Child Health and Human Development. *Journal of Child Neurology, 10,* 120-126.

Mantzicopoulos, P. (1997). Do certain groups of children profit from early retention? A follow-up study of kindergartners with attention problems. *Psychology in the Schools*, 34, 115-127.

Marston, D.B. (1989). Curriculum-based measurement: What is it and why do it? In M.R. Shinn (Ed.), *Curriculum-based measurement: Assessing special children* (pp. 18-78). New York: Guiliford Press.

Mather, N., Hammill, D. D., Allen, E. A., & Roberts, R. (2004). *Test of Silent Word Reading Fluency*. Austin, TX: PRO-ED.

McClelland, G.H., & Judd, C.M. (1993). Statistical difficulties of detecting interactions and moderator effects. *Psychological Bulletin, 114 (2)*, 376-390.

McCoy, A. R., & Reynolds, A. J. (1999). Grade retention and school performance: An extended investigation. *Journal of School Psychology*, 37(3), 273-298.

McGlinchey, M.T., & Hixson, M.D. (2004). Using curriculum-based measurement to predict performance on state assessments in reading. *School Psychology Review, 33 (2)*, 193-203.

National Adult Literacy Survey (1992). *National Center for Educational Statistics Publications & Resources Page.* Retrieved March 23, 2004 from http://nces.ed.gov//naal/resources/execsumm.asp

National Center for Educational Statistics (n.d.). *National Center for Educational Statistics Publications & Resources Page.* Retrieved March 23, 2004 from http://nces.ed.gov//naal/resources/execsumm.asp

National Assessment of Educational Progress (2003). *National Center for Educational Statistics the Nation's Report Card Reading Page*. Retrieved November 30, 2003, from http://nces.ed.gov/nationsreportcard/reading/results2003/natachieve-g4.asp

No Child Left Behind Act, U.S.C. 115 STAT. 1426 (2002).

Orton Dyslexia Society (1986). Some facts about illiteracy in America. *Perspectives on dyslexia, 13(4)*, 1-13.

Pagani, L., Tremblay, R., Vitaro, F., Boulerice, B., & McDuff, P. (2001). Effects of grade retention on academic performance and behavioral development. *Development & Psychopathology*, 13(2), 297-315.

Perfetti, C.A. (1995). Cognitive research can inform reading education. *Journal of Research in Reading, 18*(2), 106-115.

Posner, M.I., & Snyder, C.R. (1975). Attention and cognitive control. In R. Solso (Ed.), *Information processing and cognition: The Loyola Symposium* (pp. 55-85). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Shaw, R., & Shaw, D. (2002). DIBELS Oral Reading Fluency-Based Indicators of Third Grade Reading Skills for Colorado State Assessment Program (CSAP). (Technical Report) Eugene, OR: University of Oregon.

Shaywitz, S.E., Escobar, M.D., Shaywitz, B.A., Fletcher, J.M., & Makuch, R. (1992). Evidence that dyslexia may represent the lower tail of the normal distribution of reading ability. *New England Journal of Medicine, 326,* 145-150.

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

Shinn, M.R. (2002). Best practices in using curriculum-based measurement in a problem-solving model. *Best Practices in School Psychology IV, 1*, 671-697.

Shinn, M.R., Good, R.H., Knutson, N., Tilly, W.D., and Collins, V.L. (1992). Curriculum-based measurement of oral reading fluency: A confirmatory analysis of its relation to reading. *School Psychology Review,* 21, 459-79.

Stage, S.A., & Jacobsen, M.D. (2001). Predicting student success on a state-mandated performance-based assessment using oral reading fluency. *School Psychology Review, 30 (3)*, 407-19.

Stanovich, K.E. (1986). Matthew effects in reading: Some consequences in individual differences in the acquisition of literacy. *Reading Research Quarterly, 21(4)*, 360-406.

Torgesen, J.K. (2002). *Personal Communication*. Tallahassee, FL. February.

Torgesen, J.K. (1998). Catch them before they fall. *American Educator,* 32-39.

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

Torgesen, J.K., Wagner, R.K., Lonigan, C.J., & DeGraff, A. (2002). *Test of Critical Early Reading Skills*. Unpublished Manuscript, Florida State University.

Torgesen, J.K., Wagner, R.K., & Rashotte, C.A. (1999). *Test of Word Reading Efficiency*. PRO-ED inc.

Torgesen, J.K., Wagner, R.K., Rashotte, C.A., Alexander, A.W., & Conway, T. (1997). Preventive and remedial interventions for children with severe reading disabilities. *Learning Disabilities: An Interdisciplinary Journal, 8,* 51-62.

Wagner, R.K., Torgesen, J.K., Rashotte, C.A., Hecht, S.A., Barker, T.A., Burgess, S.R., Donahue, J., & Garon, T. (1997). Changing causal relations between phonological processing abilities and word-level reading as children develop from beginning to fluent readers: A five-year longitudinal study. *Developmental Psychology, 33,* 468-479.