11-4-2004

Relationship Between Curriculum-Based Measurement Reading and Statewide Achievement Test Mastery for Third Grade Students

Erin Elizabeth Ax
University of South Florida

Relationship Between Curriculum-Based Measurement Reading and Statewide Achievement Test Mastery for Third Grade Students

by

Erin Elizabeth Ax

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: Kathy L. Bradley-Klug, Ph.D.
Kelly A. Powell-Smith, Ph.D.
Lou M. Carey, Ph.D.

Date of Approval:
November 4, 2004

Keywords: oral reading fluency, Florida Comprehensive Assessment Test, Reading Assessment, high-stakes accountability tests, progress monitoring

Table of Contents

List of Tables iii

List of Figures iv

Abstract v

Chapter One 1
- Purpose of Study 4
- Research Questions 4
- Hypotheses 5
- Educational Significance 5

Chapter Two 7
- Introduction 7
- Reading 7
- No Child Left Behind Act of 2001 9
  - Florida Comprehensive Assessment Test 11
- Curriculum-Based Measurement 15
  - Criterion Validity 16
  - Oral Reading Fluency 19
- R-CBM and Statewide Tests of Achievement 20
- Purpose of the Current Study 32

Chapter Three 34
- Setting 34
- Participants 36
- Instruments 37
  - Measures of Oral Reading Fluency 38
  - Florida Comprehensive Assessment Test 39

Procedure 39
Data Collectors 41
R-CBM Administration and Scoring 43
FCAT Administration and Scoring 44
Research Design 45
Statistical Analyses 45
Missing Data 46

Chapter Four 47
Descriptive Statistics 47
Correlations 48
Scatterplots 49
Assumptions of Multiple Regression 53
Apriori Power Analysis 54
Multiple Regression Procedures 55

Chapter Five 58
Research Questions 59
Implications for Education/Educators 64
Implications for School Psychologists 64
Limitations 65
Directions for Future Research 67
Conclusion 68

References 69

Appendices 76
Appendix A: Test Scores by Ethnicity and SES 77
Appendix B: Scatterplot of the Relationship Between FCAT R-CBM Score and FCAT-SSS 79
Appendix C: Scatterplot of the Relationship Between Content R-CBM Score and FCAT-SSS 80
Appendix D: Scatterplot of the Relationship Between FCAT R-CBM Score and FCAT-NRT 81

List of Tables

Table 1	State by State Analysis of ORF/Statewide Achievement Test Studies	22
Table 2	Ethnic Group Membership of Sample	37
Table 3	Socioeconomic Status of Sample	37
Table 4	Spache Readability Indices by Grade	38
Table 5	FCAT Levels of Mastery and Corresponding Standard Scores for Third Grade	44
Table 6	Descriptive Information of Study Instruments	48
Table 7	Correlation Matrix for R-CBM Probe Scores and FCAT-SSS and FCAT-NRT	49
Table 8	Predicting FCAT-SSS Reading Score from Generic R-CBM Score	51
Table 9	Summary of Multiple Regression Analysis for Predictors of FCAT-SSS	56
Table 10	Summary of Multiple Regression Analysis for Predictors of FCAT-NRT	57

List of Figures

Figure 1	Scatterplot of the Relationship Between Generic R-CBM Probe Score and FCAT-SSS	51
Figure 2	Scatterplot of the Relationship Between Generic R-CBM Probe Score and FCAT-NRT	52
Figure 3	Scatterplot of the Relationship Between Content R-CBM Probe Score and FCAT-NRT	53

Relationship Between Curriculum-Based Measurement Reading and Statewide Achievement Test Mastery for Third-Grade Students

Erin Elizabeth Ax

ABSTRACT

The ability to read is highly valued in American society and important for social and economic advancement. One of the best strategies to prevent reading difficulties is to build basic literacy skills, thereby ensuring that all children are readers early in their educational careers. The purpose of this study was to determine the relationship between third-grade students' oral reading rate and scores on the Florida Comprehensive Assessment Test.

The present study examined the relationship between the independent variables of Curriculum-Based Measurement Reading (R-CBM), ethnicity and socioeconomic status and the dependent variable of performance on the reading portion of the Florida Comprehensive Assessment Test (FCAT) in 215 third-grade students. The data presented in this study were collected by the Florida Center for Reading Research (FCRR) as part of a larger assessment battery across three school districts and nine elementary schools in Florida. Student demographic variables as well as performance on three different types

of oral reading probes (generic, content, and FCAT passages) were investigated in relation to each student's performance on the reading portion of the FCAT.

Results of the current study were similar to investigations in other states; the correlations among the R-CBM probes and between all R-CBM probes and FCAT scores were high and statistically significant. These results indicate that student performance on any or all R-CBM probe types can be used to predict FCAT score. Ethnicity and SES were not significant predictors of FCAT score above R-CBM score.

Implications for educators and specifically school psychologists are discussed including opportunities for school psychologists to train educational personnel in the use of R-CBM. As evidenced by the current study, R-CBM may help identify students who are at-risk for reading failure and FCAT failure so that intensive interventions can be implemented early and student progress frequently monitored.

Chapter 1

Introduction

The ability to read is highly correlated to social and economic advancement, and thus failure to develop fundamental reading skills is detrimental to a child's likelihood of future success in life. Concern exists regarding the capabilities of American public schoolchildren to compete in the global market of the twenty-first century (National Research Council, 1998) and schools are being called upon to respond to increased expectations. One of the best strategies to address this concern is to prevent reading difficulties by building basic literacy skills thereby ensuring that all children are successful readers early in their educational careers.

Over the past two decades, sensing a discrepancy between what was expected and what was taught in our schools, legislators implemented assessment programs to ensure results. Standards-based reform, accountability, and high-stakes assessment entered the vocabulary of America's educators passed down from their governors (Linn, 2000; Thurlow & Thompson, 1999). With the implementation of the No Child Left Behind (NCLB) Act in 2003, America's schools entered the most stringent period of accountability assessment to date (U.S. Department of Education, 2003).

According to NCLB, statewide testing is mandatory in order to assess whether or not the state's public schoolchildren are meeting adequate yearly progress. In Florida, the Florida Comprehensive Assessment Test (FCAT) assesses student performance on the Sunshine State Standards (SSS) which is the state mandated level of achievement each student must master to be promoted to the next grade (FCAT Briefing Book, 2001). Students first take the FCAT in third grade to measure basic reading skills. In the assessment system described by the high-stakes accountability movement, results should be instructionally relevant and capable of forecasting educational change and student learning (Good, Simmons & Kame'enui, 2001). However, high-stakes outcome tests like the FCAT fail to provide teachers with data regarding ongoing progress toward educational goals, data tied to specific instructional goals and data useful for developing instructionally relevant interventions (Crawford, Tindal & Stieber, 2001). In addition, commercially developed norm-referenced achievement tests, such as the FCAT, are not based in the curriculum and lack curricular and instructional validity (Good & Salvia, 1988; Jenkins & Pany, 1978). Thus, though most states' primary tool to evaluate students' knowledge and understanding of content is some form of published or commercially available standardized achievement test, many educators have questioned whether these are the most appropriate assessment tools to catch students before they are left behind.

Due to the limitations of high stakes accountability tests, a need exists for additional measures sensitive to the curriculum or instructional outcomes and useful for ongoing monitoring to measure students' progress over time. Curriculum-Based Measurement Reading (R-CBM) fits this criterion. Curriculum-Based Measurement

Reading involves standardized procedures to directly monitor students' progress over time (Deno, 1985). These procedures are short in duration, have established reliability and validity, and are easy to administer and score (Deno, 1985; Deno, Mirkin & Chaing, 1982; Fuchs, Fuchs, Hosp & Jenkins, 2001; Fuchs, Fuchs & Maxwell, 1988). Many teachers and professionals have used R-CBM to document students' oral reading fluency (ORF) rate to inform educational decisions including progress monitoring, prereferral decisions and classification decisions, and are currently establishing a range of reading fluency scores that will predict students' scores on statewide achievement tests (Crawford, et al., 2001; Deno, 1985; Fuchs, Fuchs & Maxwell, 1988; Good et al., 2001).

A tool, such as R-CBM, which can help to predict performance on high-states tests, can also allow for intervention prior to students' failing accountability tests. Schools can use this tool to monitor student's progress toward long-range goals. In Florida, for example, third-grade students who failed the reading portion of the FCAT were retained and repeated third-grade. In total, 28, 028 third-graders were retained in 2003 because they received a score of 1 on the FCAT (Florida Department of Education, 2004). These students' reading difficulties could have been determined by measures of ORF as early as first grade and intensive interventions could have been put into place. These early intervention strategies may have prevented a majority of these students from being retained.

Through a series of investigations, R-CBM was found to be related to scores on statewide achievement tests across the country (Barger, 2003; Buck & Torgeson, 2003; Castillo, Torgeson, Powell-Smith & Al Otaiba, 2003; Crawford, Tindal, & Steiber, 2001; Good et al., 2001; Linner, 2001; McGlinchey & Hixon, 2004; Shaw & Shaw, 2002;

Sibley, Biwer & Hesch, 2001; Shapiro, Edwards, Lutz & Keller, 2004; Stage & Jacobson, 2001). Consistently throughout all regions of the country, R-CBM was highly correlated with scores on statewide achievement tests including a statistically significant correlation between R-CBM and scores on the FCAT. These results are encouraging considering that assessments in their current form fail to forecast attainment of high-stakes outcomes early enough to inform instruction and alter learning trajectories. However, additional research is needed that includes a more representative sample of students from Florida to generalize these results in order to conclude that oral reading fluency can be used to predict FCAT score.

Purpose of the Study

The purpose of the current study was to determine the relationship between third-grade students' oral reading rate (R-CBM) and scores on the Florida Comprehensive Assessment Test.

Research Questions

The following research questions will be addressed:

1. What is the relationship between third-grade students' oral reading rate and performance on the reading portion of the Florida Comprehensive Assessment Test?
2. What is the relationship between third-grade students' oral reading rate on three different passage types (FCAT passages, curriculum passages, content passages) and scores on the reading portion of the Florida Comprehensive Assessment Test?
3. What is the relationship between third-grade students' ethnicity, oral reading rate, and scores on the reading portion of the Florida Comprehensive Assessment Test?

4. What is the relationship between third-grade students' socioeconomic status, oral reading rate, and scores on the reading portion of the Florida Comprehensive Assessment Test?

Hypotheses

1. It is hypothesized that as third-grade students' oral reading rate increases, performance on the reading portion of the Florida Comprehensive Assessment Test will increase. More specifically, it is hypothesized that third-grade students' who are not at-risk as defined by levels of oral reading fluency mastery will pass the reading portion of the Florida Comprehensive Assessment Test.
2. It is hypothesized that third-grade students' will score similarly on three different passage types (FCAT passages, curriculum passages, content passages). More specifically, students who master basic early literacy skills will score at similar levels of mastery regardless of passage type.
3. It is hypothesized that there is no relationship between third-grade student's ethnicity, oral reading rate, and scores on the reading portion of the Florida Comprehensive Assessment Test.
4. It is hypothesized that there is no relationship between third-grade students' socioeconomic status, oral reading rate, and scores on the reading portion of the Florida Comprehensive Assessment Test.

Educational Significance

The results of this study should help inform teachers, school psychologists, and school personnel at the building and state level of the manner in which to best assess student achievement and prevent poor academic outcomes in the broad sense and prevent

failure on the FCAT specifically. The measures must be reliable, prevention oriented, and dynamic. Undoubtedly, students who are at-risk of failure on statewide achievement tests should be identified early in their educational careers and provided with intensive instruction and interventions. Curriculum-Based Measurement Reading may be a critical tool that can identify students who may fail the FCAT and prevent these students from potentially deleterious educational consequences such as retention.

Chapter 2

Literature Review

Introduction

This chapter presents a review of the literature on the relationship between a brief academic assessment measure and standardized accountability tests. First, the review will focus on current federal and state legislation. Second, Curriculum-Based Measurement Reading (R-CBM), psychometric properties of R-CBM and oral reading fluency as a measure of R-CBM will be discussed. Third, statewide achievement tests and their relationship to R-CBM follow. The chapter commences with a discussion of the purpose of the current study.

Reading

Reading is a survival skill necessary for success in today's society and yet, many children have difficulty acquiring this survival skill. It is not an innate skill; reading is taught and learned, requiring both direct instruction and practice. Learning to read is a lengthy and complex process that requires the fusing of exposure to written materials and reading practice through connections (Lyon, 1990). Large numbers of children from all social classes have always had difficulty learning to read (National Research Council,

1998). For example, as many as one in five children experience difficultly learning to read (Lyon, 1999).

Research on reading continues to conclude that students with poor reading skills early often have poor reading skills later (Juel, 1988). Stanovich (1986) described a spin-off effect in which problems in the early stages of learning to read negatively impact other reading processes. He termed this the Matthew Effect after the biblical passage Matthew 25:29 which states "To everyone who has will be given more, and they will have more than enough; but from those who have not, even what they have will be taken away." His suggestion is that children who have a difficult start will lag further and further behind in all aspects of reading.

According to the National Institute for Literacy (NIFL) many adults in the United States lack a sufficient foundation of basic reading skills to function successfully in society. Between 46% and 51% of adults have low literacy skills and lack the foundation they need to find and keep decent jobs, and actively participate in civic life (National Institute for Literacy, 2004).

Today's competitive economy requires increased levels of literacy than have been necessary in the past. Sadly, native-born adults in the United States ranked 10th out of 17 high-income countries for average literacy score (National Institute for Literacy, 2004). Nationally, scores of fourth graders on The Nation's Report Card (2000) have remained stable from 1992 through 2000. Sixty-three percent of fourth graders were considered at or above a basic reading level with only 32% at or above proficiency (Donahue, Finnegan, & Lutkus, 2001).

In order to meet the demands of an increasingly educated society, many policymakers have recommended a 100% literacy rate (Improving America's School's Act of 1994, Goals 2000, No Child Left Behind Act of 2001). Policymakers desire to ensure that students are receiving the literacy education they need, but they are removed from direct observation and instruction in the classroom. As a result, large scale assessment has become the standard for policymakers to measure progress (Linn, 2000).

Assessment and accountability have historically appealed to policymakers as agents of reform for a number of reasons (Linn, 2000). First, assessments are relatively inexpensive relative to programmatic or instructional change such as increasing instructional time, reducing class size, hiring more aides, or additional professional development for teachers. Second, assessment can be externally mandated, which may be easier than changing what happens inside the classroom. Third, testing can be rapidly implemented, particularly within the term of the elected official. Fourth, results of assessments are visible in that they can be reported to the press. An increase in scores over the first few years inevitably results from a large scale system of assessment (Linn, 2000). The appeal of assessment and accountability as the sole measure of reading achievement continues today with the policy of the current presidential administration, the No Child Left Behind Act (2001).

No Child Left Behind Act of 2001

A Bush Administration prescribed overhaul of the educational system was ratified January 8, 2002 when the President signed the No Child Left Behind Act (NCLB) of 2001 into law. No Child Left Behind constituted the most sweeping reauthorization of the Elementary and Secondary Education Act of 1965, defining the federal government's role

in education (U.S. Department of Education, 2003). No Child Left Behind is an act "to close the achievement gap with accountability, flexibility, and choice so that no child is left behind" (P.L. 107-110, 2002).

According to NCLB rhetoric, there are four pillars of the system: accountability for results, emphasis on doing what works based on scientific research, expanded parental options, and expanded local control and flexibility (U.S. Department of Education, 2003). No Child Left Behind requires each state to measure every public school student's annual progress in reading and math in grades three through eight and at least one time during tenth through twelfth grades. Those measurements must be aligned with state academic content and achievement standards (U.S. Department of Education, 2003).

The hallmark of NCLB is accountability. Under the law, each state is responsible for creating their own standards for what a child should know and learn for every grade. Each state, school district, and school is expected to make adequate yearly progress (AYP) toward meeting state standards. Yearly progress is measured for all students regardless of socioeconomic status, race, and language factors. Locally and nationally, school and district performance is publicly reported, and if a district or a school fails to make progress, they will be held "accountable" (U.S. Department of Education, 2003).

Inherent in accountability is yearly progress monitoring. State-wide tests of achievement are required to monitor each student's progress. Each state must to define AYP for each district and individual schools within the parameters set by Title I (U.S. Department of Education, 2003). In Florida, the Florida Comprehensive Assessment Test (FCAT) measures student achievement on the Sunshine State Standards (SSS), which are grade level standards of achievement that students are expected to meet for grade

promotion (Florida Department if Education, 2003). The results of the FCAT form the basis of Florida's system of school improvement and accountability.

Florida Comprehensive Assessment Test

The FCAT is part of Florida's plan to increase student achievement by implementing higher standards for public school students. As such, there are two components to the test. The first component is a criterion-referenced test (CRT) where scores can be measured against benchmarks in reading, writing, and mathematics from the Sunshine State Standards (SSS). The second component is a norm-referenced test (NRT) which measures each student's performance against national norms (Florida Department of Education, 2003).

The history of standardized testing in Florida follows a lengthy trajectory resulting with the FCAT. In the early 1970s, statewide assessment was first authorized by the Florida legislature to measure student's acquisition of minimum competency skills (Linn, 2000). By 1976, the legislature approved competency assessment in third, fifth, eighth, and eleventh grades and the nation's first high school graduation test (Florida Department of Education, 2003).

The conceptualization for the FCAT began in 1995 including development of the SSS. The FCAT in its current form has been administered each year since 1998 assessing students in fourth, fifth, eighth, and tenth grades. In 1999, sensing a need to raise educational expectations in order to give students necessary skills to compete in the job market, Governor Jeb Bush introduced his A+ plan. It was adopted by the legislature who amended section 229.57 of the Florida Statutes to expand achievement assessment to

include grades 3-10 (FCAT Briefing Book, 2001; Florida Department of Education, 2002).

Development of the FCAT took place over a number of years. In May of 1996, the Florida Department of Education (DOE) contracted with McGraw-Hill Education for the development of FCAT tests for grades four, five, eight, and ten. After their contract expired in 1999, the DOE contracted with the Harcourt Educational Measurement Company expanding the FCAT to third through tenth grades (Florida Department of Education, 2003).

According to the Florida Department of Education, the FCAT was developed by the Department of Education with the assistance of commercial testing companies and validated by committees of practicing teachers and curriculum specialists. FCAT questions draw from a variety of topic and subject areas. The test uses graphic displays and illustrations, and incorporates thinking and problem solving skills that match the complexity of the standards being assessed. The FCAT involves a variety of item types including multiple-choice items, and performance items which require the student to write-in answers. Performance items are not used in the third-grade FCAT (Florida Department of Education, 2003). Florida Comprehensive Assessment Test scores are reported on a scale of 100 to 500 and are assigned a number from 1-5 based on level of material mastery (FACT Briefing Book, 2001). Scores at levels one or two are considered below level and levels three through five represent passing scores on the FCAT.

The overall FCAT is purported to have good technical adequacy. Reliability indices from the 2000 administration are above .90 for fourth, fifth, eighth, and tenth grades however, information is not specified for grade level or content (reading or

mathematics). Concurrent validity has been reported between .70 and .81 for the FCAT (including the NRT) and the Stanford Achievement Test, ninth edition (SAT-9) for fourth, fifth, eighth, and tenth grades (FCAT Briefing Book, 2001). Of consideration, however, is that Harcourt Educational Measurement Company makes both the FCAT and the SAT-9 (Harcourt Educational Measurement Company, 2003).

The FCAT is reported to have content validity because of its development procedures. Content validity refers to the match between items on an achievement test and content covered by the curriculum. The content covered by the test should be representative of the content of instruction. However, commercially developed norm-referenced achievement tests have been shown to be a one-time measure, not based in the curriculum and thus lacking curricular and instructional validity (Good & Salvia, 1988; Jenkins & Pany, 1978). A preliminary study by Jenkins and Pany (1978) and replication study by Good and Salvia (1988) found that significant discrepancies exist between standardized achievement tests and the curriculum in that achievement tests do not measure what they purport to measure (namely the curriculum). Student achievement in a particular curriculum may not be reflected in an achievement score. Therein lies the problem, for if there is not a match, then it is not possible to interpret a student's score since it is unknown whether the test score represents the student's score or the test's content validity.

Since 2001, students' scores on the reading portion of the FCAT have increased each year. In 2001, the first year the FCAT was administered to third-graders, 57% achieved a passing score. The percentage increased to 60% in 2002 (Florida Department of Education, 2003).

Sixty-three percent of 188,107 third-grade students received passing scores on the reading portion of the FCAT in 2003 (Florida Department of Education, 2003). Beginning in 2003, students who achieved level one, the lowest level on the reading portion of the FCAT, were retained and repeated third-grade. In total, 28,028 third-graders were retained (Florida Department of Education, 2003).

It is the position of both the National Association of School Psychologists and the American Psychological Association that large scale assessment for high-stakes testing should be used cautiously (APA, 2001; NASP, 2003). Tests are considered "large scale assessments" when they assess all students within a given population or geographic region on the attainment of high academic standards (NASP, 2003), and they are considered high stakes for students when the results of the tests are used to make decisions about promotion or retention or high school graduation. Both governing bodies urge caution when using any single measure, such as the FCAT, as the sole determinant for making high-stakes decisions about a single student such as grade promotion or retention, receipt of a diploma, or access to educational opportunity (APA, 2001; NASP, 2003).

The FCAT and other achievement tests continue to be supported by policymakers because of face validity in that they seem to be representative of what students are learning or tasks that they are performing in school. However, the extent to which achievement tests including the FCAT actually test the curriculum is unknown. A high-stakes test administered one time at the end of the school year does not give the teachers enough time to deliver intensive interventions to students who do not meet standards. An alternative or complement to such high-stakes outcome tests would be to monitor student

progress in the curriculum periodically throughout the school year. Such monitoring would provide a more effective tool to use in determining instructional changes and interventions. Curriculum-Based Measurement reading is one such instrument.

Curriculum-Based Measurement

Curriculum-Based Measurement reading (R-CBM) is a one-minute timed assessment of oral reading fluency (ORF). Students read a grade level passage from their curriculum or a generic curriculum while an examiner notes words read correct and errors per minute. In its current form, R-CBM is a standardized procedure where students' results can be compared to national or local norms as well as established grade level benchmarks. These procedures have been developed over a number of years and can be used to make a variety of instructional decisions.

Research on R-CBM began over thirty years ago under the direction of Stan Deno at the Institute for Research on Learning Disabilities at the University of Minnesota. Sensing the apparent discrepancy between measurement procedures and instructional decisions, the developers of R-CBM sought to create reliable and valid, standardized procedures. The goal was to aid teachers in routinely and directly monitoring student achievement in the curriculum over time in order to make decisions about instructional change (Deno, 1985).

The early R-CBM research focus was threefold. First, the measures had to be quick and efficient so that students could be assessed frequently, even daily. Second, the measures had to be inexpensive and easy to create with comparable alternate forms. Third, the measures had to be easy to teach teachers, aides, instructional assistants and other educational personnel as well as reliable among testers (Deno, 1985).

Initial research supported the reliability and validity of R-CBM. Criterion validity was initially established with high correlations between CBM and other widely used, well-known tests of achievement such as the Stanford Achievement Test, Woodcock-Johnson Test of Reading Mastery, and the Peabody Individual Achievement Test (Deno, 1985; Deno, Mirkin & Chaing, 1982; Fuchs & Deno, 1992; Fuchs, Fuchs & Maxwell, 1988).

Criterion Validity

In an initial review of the literature by Deno, Mirkin and Chiang (1982), reading aloud was determined as a behavior that might index reading progress. To test their hypotheses, they conducted three studies to determine the best procedures synonymous with the tenets of Curriculum-Based Measurement being quick and easy in order to measure student's progress daily, inexpensive to produce, unobtrusive, and simple to teach teachers.

Deno, Mirkin and Chiang's (1982) exploratory study revealed that many of their initial assumptions regarding reading behavior were accurate. In their studies, they confirmed criterion and concurrent validity for R-CBM. In the first study, researchers gave 18 students in general education and 15 students in special education in grades one through five a published norm-referenced, achievement test (Stanford Diagnostic Reading Test, 1975) as well as reading measures of curricular achievement (words in isolation, words in context, oral reading, cloze comprehension, word meanings). The results of study one showed that oral reading, words in isolation, and words in context correlated with Stanford Diagnostic Reading Test (r=.73-.91). In addition, oral reading rate was correlated at a higher level than words in isolation and words in context.

The second study sought to determine whether the grade level of the measure or the length of the measure impacted the correlations with published norm-referenced reading and comprehension measures. Forty-five students in first through sixth-grades including twenty-seven students in general education and 18 students in special education participated in the study. The study employed an alternate form of each of the three measures of reading aloud from study one taken from third and sixth-grade basal readers. Researchers administered two 30- and two 60-second parallel forms tests to each student. Results of study two found similarly high correlations in the .80s and .90s on the three measures of word recognition for both third and sixth-grade materials. In addition, the 30-second tests correlated very highly with the 60-second tests (r=.95-.97 for oral reading).

The third study sought to replicate studies one and two to determine concurrent validity. In this study, researchers assessed 43 students in general education and 23 students in special education in first through sixth grades using a third-grade word list, a sixth-grade word list, a 300 word third-grade passage, a 300 word sixth-grade passage, a sixth-grade cloze passage, the Stanford Diagnostic Reading Test, and the Peabody Test of Reading Comprehension. The results of the study were that virtually all coefficients for oral reading were high and significant for each sample, both individual and combined.

Overall findings of all three studies indicated that reading aloud from a reader (oral reading fluency or ORF), reading lists of words, and cloze were all related to performance on published norm-referenced reading tests. In terms of psychometric properties, validity coefficients were all high and reliable. Interestingly, the authors found that correlations between reading aloud for 30 seconds and one minute were .90 or

higher. However, recommendations were to use the one minute interval because it was a common reporting response by teachers, and one with which they were comfortable. In summary, the authors' conclusions based on the data were that any of the informal reading procedures including oral reading can be used to estimate proficiency in decoding and comprehension (Deno et al., 1982).

The criterion validity of R-CBM with respect to commercially available standardized tests of achievement was determined in the first published study by Deno and his colleagues. More specifically, correlations between reading aloud from texts correlated with the Stanford reading achievement test at .78 and .80, and the Woodcock Johnson reading test at .93 (Deno, 1985). Criterion validity was additionally strengthened as reading aloud discriminated between students in general education and special education. In response to the significant findings of Deno, Mirkin and Chaing (1982), additional research was conducted with children from diverse geographic locations over the United States to determine if the developmental patterns of growth that occur with reading would affect reading aloud. Findings were consistent with developmental growth in that first graders read only a small number of words, which increased in second and third-grades with a negatively accelerating trend from third to sixth-grades. Overall, the number of words read correctly from a basal reading series reliably and validly discriminated growth in reading proficiency in the elementary school years. In summary, student performance in the curriculum generally and reading aloud (ORF) specifically was determined to reliably and validly measure student reading achievement (Deno, 1985). However, further research for better understanding of ORF was necessary.

Oral Reading Fluency

Oral reading fluency (ORF) is the oral translation of text with speed and accuracy and is a performance indicator of overall basic reading competence which includes comprehension. It usually develops during the elementary school years and involves direct measure of phonological segmentation and decoding skill as well as rapid word recognition. As part of a system to monitor student achievement, ORF is best used within a normative framework so that performance levels can be compared between individuals, and so that gains represented by performance slopes can track the development of reading competence within individuals (Fuchs, Fuchs, Hosp, & Jenkins, 2001). Students who read fewer than 80 words correct in a one-minute time frame at their curricula level are said to be in the "high-risk" range for reading failure and in need of intensive intervention. Eighty-one to 110 words read correct in one-minute is in the "some-risk" range where students are at an appropriate yet challenging level. Students who read more than 110 words correct per minute are in the "low-risk" for that level (Good et al., 2001).

Fuchs, Fuchs and Maxwell (1988) assessed and contrasted the validity of informal strategies of reading comprehension measurement. The reading comprehension measures they used were question answering, which is the most commonly used strategy in classrooms, recall procedures, oral reading measures, and cloze procedures. Oral reading fluency is not typically viewed as a reading comprehension measure, thus face validity is low. Seventy middle and junior high school boys ages nine through fifteen in fourth through eighth grade with mild to moderate handicaps participated in the study. A subgroup of 35 students were selected for the oral reading portion using stratified random sampling. Besides the oral reading test, students were given a comprehension question

test, passage recall test, and cloze test, as well as The Word Study Skills and Reading Comprehension subtests of the Stanford Achievement Test 7th edition as global achievement tests.

Findings revealed oral reading to be the best measure of reading comprehension. All correlations with the Stanford Achievement Test ranged from .70 (passage recall) to .91 (oral reading) which was significantly higher than any other measures. Questions answered, which is the most commonly employed classroom test of reading comprehension, including statewide achievement tests such as the FCAT, correlated at .82 (Fuchs, Fuchs, & Maxwell, 1988). Ultimately, with ORF the best measure of reading comprehension researchers have begun to explore the relationship between ORF and statewide achievement tests. Several studies highlighted the importance of using R-CBM as a complement to state-wide standardized tests of achievement (Buck & Torgeson, 2003; Crawford, Tindal & Stieber, 2001; Good, Simmons & Kame'enui, 2001; Nolet & McLaughlin, 1997).

R-CBM and Statewide-Tests of Achievement

Though accountability through testing is not a new concept, high-stakes accountability has become a constant warning to the public school system and every district, school, and educator therein. Statewide accountability tests are external tests imposed by the state or national legislature. By design, statewide assessments are "blunt instruments" in that they lack sensitivity over time or intra-student variables (Nolet & McLaughlin, 1997). Curriculum-Based Measurement Reading on the other hand, is sensitive, and progress-monitoring data from R-CBM would not replace statewide tests, but rather empower teachers to understand and utilize the outcomes from statewide tests.

In response to a growing concern among school psychologists and educators regarding the single "snap-shot" high-stakes nature of statewide tests, a number of studies have examined the relationship between oral reading fluency and statewide tests of achievement (Barger, 2003; Buck & Torgeson, 2003; Castillo, Torgeson, Powell-Smith & Al Otaiba, 2003; Crawford, Tindal, & Steiber, 2001; Good et al., 2001; Linner, 2001; McGlinchey & Hixon, 2004; Shaw & Shaw, 2002; Sibley, Biwer & Hesch, 2001; Shapiro, Edwards, Lutz & Keller, 2004; Stage & Jacobson, 2001) (see Table 1 for a summary of these results). These studies were conducted in ten states throughout the country and represent four geographic regions (West, Midwest, Northeast, and Southeast). Each of these studies include the same independent variables (R-CBM) and the dependent variable (a state's high-stakes achievement test) but differ as to the grade of the participants, number of participants, geographic region of participants, demographic make-up of the sample, number of probes administered, as well as reported oral reading score (single probe versus the median score of three probes). Overall, all studies found scores on R-CBM statistically significantly or highly correlated with scores on statewide achievement test, thus showing R-CBM to be an important tool for early identification and intervention. The following section provides a review of these studies. This review is organized by geographic region of the country and will end with a discussion of the southeast and Florida.

Table 1

State by State Analysis of ORF/Statewide Achievement Test Studies

State	Authors	Score Used	Sample Size	Grade	Correlation	Cut Scores	Accuracy in Predicting Passing Scores
AK	Linner (2001)	NA	NA	3rd	NA*	110 WCPM	Unavailable
WA	Stage & Jacobson (2001)	Single Probe	173	4th	.43-.51***	100 WCPM	90%
OR	Crawford, Tindal, & Steiber (2001)	Median of 3 probes	51	2nd-3rd	NA	119 WCPM	94%
	Good, Simmons, & Kame'enui (2001)	NA	364	3rd	.67*	110 WCPM	99%
CO	Shaw & Shaw (2002)	Median of 3 probes	52	3rd	.73-.80*	110 WCPM	90%
IL	Sibley, Biwer, & Hesch, (2001)	NA	82	3rd	.79*	110 WCPM	99%
MI	McGlinchey, & Hixon (2004)	single/median	1362	4th	.49-.81 (M=.67***)	100 WCPM	72%
OH	Stoller (2004)	Median of 3 probes	332	4th	.59**	110 WCPM	NA
PA	Shapiro, Edwards, Lutz & Keller (2004)	NA	185	3rd	.65-.67**	114 WCPM	93%
NC	Barger (2003)	Median of 3 probes	38	3rd	.73*	110 WCPM	100%
FL	Castillo, Torgeson, Powell-Smith & Al Otaiba (2003)	Single Probe	101	3rd	.60-.65**	110 WCPM	NA
	Buck & Torgeson (2003)	Median of 3 probes	1102	3rd	.70-.74***	110 WCPM	91%

**=.01
***=.0001
*=not reported

Analyses of the relationship between R-CBM and statewide achievement tests have been performed to varying complexity in the West. In Washington state and Oregon complex analysis of large, published studies supported this relationship. Presentations and technical papers in Alaska and Colorado also supported this relationship.

Stage and Jacobsen (2001) examined the relationship between R-CBM and the Washington Assessment of Student Learning (WASL) in a sample of 173 fourth graders. Examiners administered three oral reading probes from students' basal reading series in September, January, and May. Each probe came from materials considered to be mid-year level of difficulty. Examiners reported the median probe score for each student. The students took the WASL in May of the same year.

Results supported a positive relationship between R-CBM and scores on the WASL. Curriculum-Based Measurement Reading was correlated with WASL at all three data collection points (r=.43-.51). Researchers found that using students' September ORF scores was a better predictor than growth in ORF across the year. Students whose scores fell below mastery level in September were at risk for failing the WASL. Limitations of the study included that participating students were from a high performing school (80% passed the WASL). In addition, 90% of the students identified themselves as Caucasian and just 15% of the sample received free and reduced lunch. The generalization of this study beyond the school or district is questionable (Stage & Jacobson, 2001).

Crawford, Tindal, and Stieber (2001) studied the strength of the relationship between R-CBM and future performance on the Oregon statewide reading and math achievement tests as well as the levels of oral reading rate that best predicted student's scores on the Oregon statewide reading and math achievement tests. This longitudinal

study followed 51 students through second and third-grades. For each of the two years, students were administered three oral reading passages selected from Houghton Mifflin Basal Reading Series (1989). Passages for the first year of the study (second grade) were selected randomly from the second grade basal reader. Passages for the second year of the study (third-grade) were randomly selected from the third-grade basal reader. Curriculum-Based Measurement Reading data were collected on one day in January in each of the years. In March of the second year, third-graders were tested on the Oregon statewide math and reading assessments.

Results were powerful and significant. Second graders read an average of 62.3 words correctly per minute and the same students in third-grade jumped to an average of 103.8 words read correctly per minute. The mean gain in words read correct over the year was 42 per minute. Scores on the statewide reading achievement test ranged from 172-235, with the average third-grade reading assessment score at 202.5 (a passing score was 201 or above). Out of the 51 students followed from second through third-grade, 65% passed the reading assessment in third-grade.

Using the norms established by Hasbrouck and Tindal (1992) for the winter of third-grade, students falling below the 25th percentile read between 0 - 70 words correct per minute, students in the second group read 71-92 words correct per minute, students in the third group read 93-122 words correct per minute, and students in the fourth group read 123 or more words correct per minute. Eighty-one percent of students reading at the third and fourth group levels passed the statewide reading achievement test. In addition, 94% of third-graders reading at least 119 words correct per minute passed the statewide

test, thus 119 was the critical rate needed to pass the test. A chi-square demonstrated statistical significance between reading rate and statewide reading test scores.

Conclusions of this study focused on the stability of R-CBM as well as its utility of predicting a passing score on accountability tests. Overall, the student's rate of oral reading increased by an average of 42 words from second to third-grades. A strong correlation existed between rates of oral reading in second grade and third-grade, which confirmed the stability of R-CBM. In addition, nonparametric analysis revealed a significant relationship between reading rate and performance on the state achievement test. In fact, second graders who read at least 72 words correct per minute passed the statewide test in third-grade and 81% of third-grade students reading at the 50th percentile and above passed the statewide reading test. Reading at least 119 words correctly per minute in third-grade virtually assured a passing score (94%). Curriculum-Based Measurement Reading was sensitive enough to detect growth in 50 of the 51 students studied over the two year period. In the wake of ever increasingly high stakes decisions being made about students based on their test score, continual progress should be used in order to intervene when students are not reading at least 90 words correct per minute (Crawford, Tindal & Stieber, 2001).

Although the study by Crawford, Tindal and Stieber (2001) demonstrated the advantage of R-CBM, there were limitations of the study**.** First, as reported by the authors, the student's classroom teachers were used as testers, and no reliability checks were implemented by the researchers. Second, the sample size was small and represented only one district in Oregon. Third, the levels of ORF used by Crawford et al. are not commonly cited in the literature related to R-CBM. To fully explore the research

questions, a future study should include reliability checks, a larger and more representative sample, and common levels of fluency grounded in the R-CBM literature.

A study by Good, Simmons and Kame'enui (2001) discussed the importance of decision-making utility of oral reading fluency for third-grade high-stakes testing in Oregon. In this study, four cohorts of students in Kindergarten through third-grade from six elementary schools participated. Five of the six schools qualified for Title I services with the percentage of students receiving free and reduced lunch ranging from 37% - 63%. Within the district, 10% of students were identified as minority students and 18% fell at or below poverty level.

The measures employed in this study were threefold. Students were given Dynamic Indicators of Basic Early Literacy Skills (DIBELS) which are fluency-based measures of early literacy. Students were also given R-CBM and the Oregon statewide achievement assessment (OSAT), a high stakes measure of comprehensive reading achievement. The three oral reading fluency passages selected were from third-grade screening and level C progress monitoring passages of the Tests of Reading Fluency (TORF) a generic source of R-CBM materials.

The results of this study were insightful for both R-CBM and statewide assessment. First, correlations between earlier and later oral reading skills ranged from .34-.82. In addition, of the 98 students that reached first grade R-CBM benchmarks, 97% attained second grade benchmarks. Ninety-six percent of those who attained the May of third-grade R-CBM goal of at least 110 words correct in one minute were rated as "meets expectation" or "exceeds expectation" which are the passing levels on the OSAT. For students reading 70-110 WCM, the likelihood of meeting expectations on the statewide

achievement test was less clear and predictions of passing rates could not be made with precision. Twenty-eight percent of students who scored below 70 words correct per minute scored "meets expectation" on the OSAT.

Good, Simmons and Kame'enui (2001) found that DIBELS benchmarks were related to meeting later benchmarks. The DIBELS measures can be administered as early as preschool to monitor students' reading progress and assess students who may be at risk for reading failure. The results of R-CBM supported accuracy and fluency with connected text as an important foundation for reading competence. Students who read at grade level (110 words correct per minute or better) were more likely to meet or exceed expectation, and students who read less than 70 words correct per minute were not likely to pass the Oregon statewide achievement test.

Shaw and Shaw (2002) studied DIBELS ORF with a small sample of third-graders in one elementary school in Colorado. They, too, found ORF correlated with the Colorado achievement tests. Shaw and Shaw (2002) collected DIBELS ORF data in September, January and April which were correlated with the April administration of the CSAP (r=.73-.80).

The study found high scores on ORF to be predictive of passing the state tests however predictions were less clear for students who did not read at least 90 WCM. For example, Shaw and Shaw (2002) found a 91% likelihood that students who read 90 WCM or above would receive a passing score at or above proficiency level on the CSAP. Seventy-three percent of students who read fewer than 90 WCM scored partially proficient or unsatisfactory which are failing scores. Though this study supported the relationship between ORF and statewide tests, there were many limitations. First, a small

sample of third-graders was used (n=52). Second, no demographic information was available on the sample. Third, the study sampled only students from one elementary school.

In the Midwest (Ohio, Michigan, Illinois) R-CBM scores were correlated with statewide achievement scores. In Ohio, 332 fourth graders were administered three R-CBM probes in October and the Ohio Proficiency Test in March of the same year (Stoller, VanderMeer, & Lentz, 2004). Overall, a statistically significant correlation was found between the median of three Houghton-Mifflin R-CBM probes and OPT (r=.59).

In one school district in Illinois, 99% of third-graders who scored at or above 110 words correct in one-minute scored "meets standards" or "Exceeds standards" on the Illinois State Assessment Test (ISAT) (Siblet et al., 2001). Correlations between TORF and ISAP was high (r=.79) for the 82 third-grade students that participated in the study. Further information on the methods used was unavailable.

McGlinchey and Hixon (2004) replicated the findings of Stage and Jacobson (2001) with fourth-graders in Michigan across an eight year time span. Researchers examined fourth graders in one elementary school for seven of the eight years and all fourth-graders in one school district for one of the eight years. In total, 1,362 fourth-graders participated across eight years. Across the school district, 52% of students were identified as non-Caucasian and 60% received free or reduced lunch. Each year for eight years examiners administered ORF probes to fourth graders two week before the Michigan Educational Assessment Program (MEAP). In years one through five, a single probe was administered and recorded and in years six through eight, the median score of three probes was recorded.

Correlations in the Michigan study were higher than those from the Washington study, ranging from .49-.81 with an average correlation of .67 between ORF and the MEAP over eight years (McGlinchey & Hixon, 2004). The differences in results could have been due to differences in economic status or racial make-up. Some methodological issues of the study include comparing different samples each year (variation in sample size, free and reduced lunch status, ethnicity) as well as the entire district being included just one year. Overall however, both studies found ORF scores for fourth graders in one elementary school and one district significantly correlated with scores on statewide achievement tests.

In the Northeast, a study by Shapiro, Edwards, Lutz, and Keller (2004) demonstrated further evidence for the use of CBM General Outcomes Measures as predictors of the Pennsylvania System of School Assessment (PSSA). Oral reading probes were administered to 185 third-graders in the Fall, Winter, and Spring and the PSSA was administered in the Spring. The third-graders were selected from among eight elementary schools in one mixed urban and suburban school district in which 32.8% of the students were considered low income. Oral reading fluency was significantly correlated with PSSA at all three times. However, ORF was correlated highest with PSSA at the Spring administration (r=.67), followed by Winter (r=.66), and Fall (.65). In addition, 93% of students who read 114 words correct per minute or above were considered to be "successful" on the PSSA. Curriculum-Based Measurement Reading was a strong predictor of performance on the state standardized test in Pennsylvania.

Finally, in the Southeast (North Carolina, Florida) R-CBM was found to not only predict scores on statewide achievement tests, but also was the best predictor from among

other reading comprehension measures (Castillo et al., 2003). Barger (2003) found a high correlation between student scores of the median of three DIBELS ORF probes with the North Carolina End of Grade reading assessment administered one week later (r=.73). As in Colorado, high scores on ORF were predictive of passing the North Carolina End of Grade reading assessment but other predictions were unclear. Barger found all students who read 100 WCM or better passed the North Carolina End of Grade reading assessment but half of the students who did not read 100 WCM also passed. Though this study supported the relationship between ORF and statewide tests, there were many limitations including small sample size (n=38), no available sample demographic information, and a limited sample (one elementary school).

Castillo et al. (2003) further explored the relationship between two forms of ORF probes as well as other individually and group administered reading fluency measures and the Florida Comprehensive Assessment Test (FCAT) in a small, rural district in Florida. Students in two elementary schools in first, second and third grades were administered three ORF probes developed from the DIBELS and three ORF probes from the Monitoring Basic Skills Program (MBSP). They were also administered the Test of Word Reading Efficiency (TOWRE; Torgeson, Wagner, & Rashotte, 1999) which is an individually administered measure designed to assess fluency with reading word lists rather than a complete passage. To addresses the lengthy amount of time teachers can spend administering ORF probes to their entire class, two group administered word lists currently being developed were administered including the Test of Critical Early Reading Skills (TOCERS; Torgeson, Wagner, Lonigan, & DeGraff, 2002) and the Test of Silent Word Reading Fluency (TOSWRF; Mather, Hammill, Allen, & Roberts, 2004).

All fluency measures were administered to participants in two sessions one month after the students took the statewide achievement tests including the FCAT for third-graders. Overall, ORF was correlated highest with the FCAT-SSS (for reading) score in 102 third-grade students (r=.60-.65). Both probes types were equally related to the FCAT score, but there was significant variability in the mean scores across the three passages at each grade level. In Florida, ORF was found to be the highest correlated brief assessment with the FCAT (SSS or NRT) from among four assessments (Castillo et al., 2003).

In response to the study by Good and colleagues, Buck and Torgeson (2003) replicated their study with a sample from Florida. Thirteen schools from one Florida district provided data that included Curriculum-Based Measures of oral reading fluency and scores from the Florida Comprehensive Assessment Test for third-grade students (n=1102). Forty-nine percent of the students were female. The ethnic composition of the sample included 83% Caucasian, 7% African American, and 6% Hispanic. Forty-six percent of the students qualified for free or reduced lunch.

Similar to the Good, Simmons and Kame'neui (2001) study, Buck and Torgeson (2003) found a significant correlation between ORF and reading scores on the FCAT-SSS (r=.70). Ninety-one percent of students who read at or above 110 words correct per minute passed the FCAT. Of those who read below 80 words correct per minute, 81% did not pass. As in the above study by Good et al. (2001), midrange (between 80 - 110 words correct per minute) was unpredictable and students were equally likely to pass or not pass.

Minority representation also offered insight into using oral reading fluency as a predictive measure. Hispanic students' oral reading fluency scores were correlated

highest with FCAT score (r=.78), followed by Caucasian students' scores (r=.70) and African American students' whose scores were correlated lowest (r=.62). In addition, scores above 110 words correct per minute were slightly less predictive of success for minority groups while scores below 80 words correct per minute were more predictive of failure for these groups. Multiway frequency analyses were conducted to determine whether the interaction between racial/ethnic background and predictive accuracy for oral reading fluency scores was statistically reliable. The interactions were not significant indicating that predictive relationship was not significantly different for Caucasian or African American students. However, the racial/ethnic make-up of this study was not representative of either the population of the United States or Florida with an over representation of Caucasian students and an under representation of African American and Hispanic students. This is one limitation of the Buck and Torgeson (2003) study.

Overall, R-CBM has been found to be correlated with statewide tests of achievement across the country. However, few studies compared socioeconomic status, free and reduced lunch status, and ethnicity among or between districts in the state. To this point, none of the studies correlating R-CBM with statewide achievement tests can be generalized across Florida let alone nationally.

Purpose of the Current Study

The current study will attempt to add to the existing body of literature regarding the relationship between R-CBM and statewide achievement tests. Buck and Torgeson (2003) and Castillo et al. (2003) explored the relationship between R-CBM and mastery on the Florida Comprehensive Assessment Test (FCAT), a criterion-referenced statewide achievement test. The current study proposes to further study the relationship between R-

CBM and performance on the Florida Comprehensive Assessment Test for third-grade students across Florida. The limitations such as sample homogeneity and representation will be addressed. Finally, the current study will explore the relationship between three different oral reading passage types (i.e. curriculum, content area, FCAT) and their relationship to FCAT mastery.

Chapter 3

Methods

This chapter outlines the procedures and instruments that were utilized to determine the relationship between third-grade students' oral reading fluency and scores on the reading portion of the Florida Comprehensive Assessment Test (FCAT). First, a description of the project grant, from which the data for this study were obtained, is presented followed by a discussion of the setting and research participants. Next, the instruments, data collection procedures, and research design are presented. Finally, a description of the data analysis and the study limitations are discussed.

Setting

The Florida Center for Reading Research (FCRR) has as its mission conducting basic and applied research to impact policy and practices of literacy instruction as well as reading assessment (Florida Center for Reading Research, 2003). To that end, FCRR received a nationally funded grant entitled *Individual Differences in FCAT performance* (FCAT grant) in order to study the cognitive and reading profiles of third-grade, seventh-grade, and tenth-grade students in Florida who took the FCAT.

The FCRR staff selected three sites across Florida to act as regional representatives for the purpose of collecting data for the FCAT grant. Broward,

Hillsborough, and Leon Counties were selected as representatives of Southern, Central, and Northern Florida. These counties were selected based on location, demographic makeup, as well as proximity to universities to facilitate data collection.

At the time of data collection, Broward County ranked as the nation's fifth largest school district and the largest fully-accredited public school district. In Florida, it ranked as the second largest school district behind Miami-Dade County. For the 2002-2003 school year, there were more than 266,000 students enrolled in kindergarten through twelfth grade representing over 155 countries and 57 different languages. Average per pupil expenditure in Broward County was $4,383 (Broward County School District, 2003). Approximately 46% of elementary school children received free or reduced lunch as a measure of socioeconomic status (Florida Department of Education, 2004). There were 136 elementary schools in the district which enrolled approximately 122,162 students. The racial composition of public school student enrollment was 37% Caucasian, 36% African American, 22% Hispanic, 3% Asian, less than 1% Native American and 2% classified as multi-racial (Broward County School District, 2003).

Hillsborough County is located in west central Florida. For the 2002-2003 school year there were approximately 165,164 students enrolled in kindergarten through twelfth grade. Average per pupil expenditure in Hillsborough County was $4,080 (Hillsborough County School District, 2003). Approximately 54% of elementary school children received free or reduced lunch (Florida Department of Education, 2004). There were 121 elementary schools in the district which enrolled 78,919 students excluding charter schools. The racial composition of all elementary school students enrolled was approximately 45% Caucasian, 22% African American, 25% Hispanic, 2% Asian, less

than 1% Native American, and 5% classified as Multi-racial (Hillsborough County School District, 2003).

Florida's capital, Tallahassee, is in Leon County. Leon County enrolled 31,752 students in the 2002-2003 school year. For the 2001-2002 school year, average per pupil expenditure in Leon County was $4,252 (Leon County School District, 2003). Approximately 44% of elementary school children received free or reduced lunch (Florida Department of Education, 2003). The 25 elementary schools enrolled 15,445 students for the 2002-2003 school year. The racial composition was 52% Caucasian, 42% African American, 2% Hispanic, 2% Asian, less than 1% Native American, and 2% multi-racial (Leon County School District, 2003).

Participants

Participants in this study include a subset of students from the FCAT grant. In 2003, all third-grade students enrolled in a public school in the state of Florida were required to take the Florida Comprehensive Assessment Test in the spring of third grade. A total of 188,107 third-grade students took the FCAT in 2003 including 36,285 third graders from Broward, Hillsborough, and Leon Counties. A representative sample of 215 third graders enrolled in 9 elementary schools from these three Florida counties was selected to participate in the current study. The specific schools were selected based on the geographic region and socioeconomic makeup of the children they serve. See Table 2 for a description of ethnicity and Table 3 for a description of socioeconomic status of the students in the sample. To be included in the present study, participants' parents must have given written informed consent prior to study participation and the individual

student must have given written informed assent immediately before study participation. Also, each of the students had to be eligible to take the FCAT.

Table 2

Ethnic Group Membership of Sample

	Ethnicity						
	Black	Caucasian	Hispanic	Asian	Mixed	Total	Missing
n	90	83	32	3	5	213	5
%	42.3	39.0	15.0	1.4	2.3	100.0	

Table 3

Socioeconomic Status of Sample

	Meal Status				
	Full Priced Lunch	Reduced-Price Lunch	Free Lunch	Total	Missing
N	92	10	85	187	28
%	49.2	5.3	45.5	100.0	

Instruments

Instruments for this study were selected by the principle investigator from among all of the instruments administered as part of the FCAT grant for the purpose of answering the research questions. The larger test battery included instruments measuring general knowledge, listening comprehension, vocabulary, non-verbal reasoning, working memory, reading fluency, decoding, reading comprehension, motivation and exposure to print, as well as teacher ratings. For the purpose of this study, measures of oral reading fluency and scores on the FCAT were selected.

Measures of Oral Reading Fluency

Measures of R-CBM consisted of nine probes selected from three different sources. Three of the R-CBM probes came from text books on the state adoption list for third graders. The texts were selected by FCRR staff for the content contained within the passage, as well as third-grade level. Three probes came from AIMSweb/Edformation, which is a database of R-CBM probes (AIMSweb, 2003). These probes were selected at random from among all third grade level probes available on the website. The final three probes came from published FCAT practice passages from the 2001-2002 school year. Each probe was approximately 250 words in length and was retyped onto separate pieces of paper which matched the print in standard basal text books. As part of this current study, the readability of each of the R-CBM passages was determined using the Spache Readability Formula (Spache, 1953). The Spache formula assesses the difficulty of a passage by computing two values of the text. The first is the average number of words per sentence. The second is the percentage of words not found on the Spache revised word list which is a list of accepted and common words for students through third grade. The average of the three probes were all in the third-grade level for each of the three probe types (see Table 4).

Table 4

Spache Readability Indices by Grade

Probe Type	Probe 1	Probe 2	Probe 3	Mean Probe Level
FCAT R-CBM	2.4	3.6	3.8	3.3
Generic R-CBM	3.5	3.6	3.7	3.6
Content R-CBM	3.3	3.5	4.8	3.9

The technical adequacy of R-CBM has been well documented over the past two decades. Oral Reading Fluency was developed by Deno and his colleagues at the University of Minnesota Institute for Research on Learning Diabilities (Deno, 1985). Studies of test-retest reliability yielded coefficients ranging from .82-.97, with parallel forms ranging from .84 to .96 with most correlations above .90. In addition, interrater reliability has been found to be .99 (Marston, 1989). Studies investigating criterion related and construct validity with published norm-referenced tests of achievement have been moderate to high ranging from .63-.90 with most correlations above .80 (Deno, Mirkin & Chaing, 1982; Fuchs, Fuchs & Maxwell, 1988; Marston, 1989; Shinn, Good, Knutson, Tilly & Collins, 1992).

Florida Comprehensive Assessment Test

The FCAT is a criterion-referenced test containing multiple choice questions. Results of the items are reported in standardized scores which are then converted into levels of Mastery rated 1-5 (Florida Department of Education, 2003).

The technical adequacy of the FCAT is described by the Florida Department of Education as being excellent (FCAT Briefing Book, 2003). Reliability was reported to be above .90. In addition, the FCAT is reported to have content validity. Criterion-referenced validity was reported between .70 and .81 when correlated with scores on the SAT-9 (FCAT Briefing Book, 2003). Unfortunately, more precise information was unavailable.

Procedure

The data for this study were collected through the FCRR as part of the FCAT grant. To adequately sample public school students from across Florida, three regional,

university based representative sites were nominated by FCRR. At the nominated sites, FCRR invited faculty from Florida State University, University of South Florida and Florida Atlantic University representing Leon, Hillsborough, and Broward counties, respectively, to participate in the study. Regional representatives were responsible for obtaining Institutional Review Board approval at each university site and each participating school district, recruiting and training data collectors as well as soliciting school involvement and arranging data collection at the individual schools.

To account for socioeconomic variance, the schools selected were to fall into low, middle, or high socioeconomic categories. The selection process for these schools varied across counties. For example, there were three elementary schools selected in Hillsborough County (high, middle, low socioeconomic status) to assess seventy third graders. To narrow the list of all schools in Hillsborough County to determine socioeconomic status, the principal investigators consulted with two experts who were familiar with the schools in the county. They narrowed the list to 18 elementary schools, 16 middle schools, and 16 high schools. The study administrator entered those fifty schools into a database (Great Schools, 2003) to further categorize the schools. The final list resulted in three schools at each socioeconomic level, for a total of nine schools with eight back-up schools. After receiving IRB approval, the investigators verbally asked each school principal to participate and in some cases emailed a study summary. Eight school principals (two elementary schools, three middle schools, three high schools) agreed for their school to participate. Since one elementary school did not agree to participate, the principal of the back-up school was called and agreed to participate.

In Leon and Broward Counties, schools were selected primarily based on ethnicity. Schools were secondarily selected based on the socioeconomic make-up of the school. All schools selected agreed to participate. After the schools agreed, Institutional Review Board Approval was obtained through Florida State University and Leon County Public School district and Florida Atlantic University and Broward County Public School district for Leon and Broward Counties respectively.

At each school, the building principal was given the choice as to how to recruit participants. All third grade students at all nine elementary schools across the study were given informed consent forms to participate in the current study. Teachers sent home the informed consent forms with their students and those whose parents agreed for their children to participate were eligible for the study.

Data Collectors

At the University of South Florida, data collectors were solicited by email and flyer in the College of Education and Department of Psychology. Those who responded were interviewed by the regional representatives to determine their level of experience assessing students in the schools as well as their flexibility in terms of scheduling to participate in data collection. The resulting data collectors were two undergraduate students majoring in psychology, one graduate student enrolled in a Ph.D. program in Curriculum and Instruction with a full time program Emphasis in Special Education, five graduate students enrolled in an Applied Behavior Analysis Master's Program, and six graduate students enrolled in a Ph.D. program in School Psychology. Data collectors were compensated for their time by the FCRR. Training for the data collectors occurred in two sessions for two groups of data collectors. Each training session lasted six hours

and training was given by the head research faculty member of the FCRR. At each of the trainings, data collectors were instructed on each assessment tool, practiced each tool, and asked questions. They received a tenth grade protocol to use during the training and an additional tenth grade protocol to practice with after the training session. In addition, each data collector received a packet of all testing materials.

Each data collector was assigned to one or more schools by the regional representatives. Each week, the regional representatives confirmed the assignments of the data collectors based upon data collection need. Though data collectors were assigned one school, most also collected data at another school site due to the fact that assessments were conducted on 215 students and assessments lasted approximately two hours each.

Similar procedures were followed at both the Leon and Broward County sites. Procedures included similar data collector background, training, assignment, and compensation. Regional representatives followed procedures delineated by FCRR for the purpose of data collection.

Once at their assigned school, the individual data collectors decided which students to test based on various factors including their attendance that day and their class schedules. The entire test battery lasted approximately two hours and the data collectors decided to test for the entire two hours or split the session into two one-hour sessions. Though there was a prescribed sequence of tests, the data collectors were allowed to give them in any order.

Typically, the data collector went to the students' classroom and escorted the child to a quiet testing room. Though the students' parents gave informed consent for the students to participate, the data collector first explained the voluntary nature of the study

and asked for student assent. Once assent was given, the data collector proceeded with the assessment. After asking various demographic questions, the data collector administered the assessments and walked the student back to his or her classroom.

Once per week, data collectors turned completed protocols in to the regional representatives. The regional representatives maintained a database for the purpose of keeping track of the data. Every two weeks, the regional representatives mailed the completed protocols to FCRR where the protocols remained. Two independent researchers at the FCRR completed inter-rater checks on 100% of the protocols. A senior researcher at the FCRR settled any discrepancy between the site rated protocol and the researcher rated protocol by making final decisions.

R-CBM Administration and Scoring

Following standard R-CBM procedures, students were asked to read nine passages for one-minute each. These passages were interspersed among other assessment tools. However, no two R-CBM passages were administered together and there was at least one other instrument in between the R-CBM probes. Standardized instructions were given to each student prior to administering each probe. The instructions were as follows:

> When I say start, begin reading aloud at the top of the page. Read across the page (demonstrate by pointing). Try to read each word. If you come to a word you don't know, I'll tell it to you. Be sure to do your best reading…Start.

The examiner timed each student for one minute noting any errors on his or her copy of the R-CBM probe. An error was defined as a mispronunciation, substitution, omission, or if a participant struggled with a word for more than

three seconds. At that point, the examiner scored the word as incorrect and supplied the student with the word (Shinn, 1989). At the end of one minute, the examiner noted the total number of words read correct by the student.

Words read correct per minute were calculated by individual data collectors at the completion of each testing session. These scores consisted of the number of words read correct in one minute by probe. Following Shinn's (1989) scoring model, an error was defined as a mispronunciation, substitution, omission, or if a participant struggled with a word for more than three seconds. Scores were checked in the same manner as described above.

FCAT Administration & Scoring

The Florida Comprehensive Assessment Test was administered over one week in March to third grade students across Florida. Individual FCAT scores were obtained for each participant from district records. Results of FCAT items are reported in standardized scores which are then converted into levels of Mastery rated 1-5 (Florida Department of Education, 2003). For the purpose of the current study, standardized scores were utilized. Table 4 delineates levels of Mastery from standard scores (Florida Department of Education, 2003).

Table 5

FCAT Levels of Mastery and Corresponding Standard Scores for Third Grade

Level	1	2	3	4	5
Standard Scores	100-258	259-283	284-331	332-393	394-500

Research Design

A quasi-experimental research design was used in the current study. This design does not include the use of random assignments because this study does not involve grouping students at the time of the study. The lack of random assignment is a limitation to the quasi-experimental design as it threatens internal validity (Frankel & Wallen, 2003).

Statistical Analyses

A multiple regression was used to correlate variables in the current study. In a multiple regression, a dependent variable is predicted from a set of predictors (Stevens, 1999). The analyses that were used to test each research question are described below.

1. What is the relationship between third-grade students' oral reading rate and performance on the reading portion of the Florida Comprehensive Assessment Test?

   Descriptive statistics including mean, and standard deviation were calculated to describe the characteristics of the R-CBM and FCAT scores. The average R-CBM score for each participant was correlated with FCAT score in a multiple regression equation.

2. What is the relationship between third-grade students' oral reading rate on three different R-CBM passage types (FCAT passages, curriculum passages, content passages) and scores on the reading portion of the Florida Comprehensive Assessment Test?

   Descriptive statistics including mean, and standard deviation, were calculated to describe the characteristics of each passage type. Each R-CBM passage type for

each participant was correlated with FCAT score in a multiple regression equation.

3. What is the relationship between third-grade students' ethnicity, oral reading rate, and scores on the reading portion of the Florida Comprehensive Assessment Test?

   Descriptive statistics including mean, and standard deviation were calculated to describe the characteristics of participants' ethnicity. Ethnicity was correlated with FCAT scores in a multiple regression equation.

4. What is the relationship between third-grade students' socioeconomic status, oral reading rate, and scores on the reading portion of the Florida Comprehensive Assessment Test?

   Descriptive statistics including mean, and standard deviation were calculated to describe the characteristics of participants' socioeconomic status. Socioeconomic status was correlated with FCAT scores in a multiple regression equation.

Missing Data

Each of the variables had cases of missing data. In the case of missing data among the variables in the regression analyses, cases were excluded pairwise. In other words, pairs with complete data were used to compute the correlation coefficient on which the regression analyses were based.

Chapter 4

Results

This chapter presents the results on the relationship between R-CBM and scores on the reading portion of the Florida Comprehensive Assessment Test. The analyses used to address each research question are described in detail.

Descriptive Statistics

Descriptive information for the study instruments can be found in Table 5. Study instruments include FCAT Sunshine State Standards Reading Scale Score (FCAT-SSS), FCAT Norm-Reference Test Reading Scale Score (FCAT-NRT), FCAT R-CBM probe score, generic R-CBM probe score, and content R-CBM probe score. For each measure, the number of cases, mean or median, minimum and maximum scores and standard deviation are included. See Appendix A for a table of scores by ethnic group and socioeconomic make-up.

The scores reported on each of the instruments are consistent with score reported in the literature. For the FCAT-SSS and FCAT-NRT, individual student scores were used in the analyses. The median score of the three R-CBM probes by type was used. For example, three generic R-CBM probes were administered to each student. The

median of the three scores was then used in the subsequent analyses. The use of the median score is consistent with research in the area of CBM.

Table 6

Descriptive Information of Study Instruments

Variable	N	Mean	Minimum	Maximum	Standard Deviation
FCAT-SSS	207	310.31	100.00	500.00	64.63
FCAT-NRT	210	638.00	526.00	765.00	47.13
FCAT R-CBM	215	95.58	8.00	207.00	39.72
Generic R-CBM	215	104.60	7.00	213.00	41.92
Content R-CBM	215	92.55	7.00	195.00	41.35

Note. Scores reported are scale scores. For FCAT R-CBM, Generic R-CBM, and Content R-CBM, the median score of three probes was used.

Correlations

To assess the relationship between R-CBM probes and FCAT, correlations were computed among and between the variables. Significant correlations were found between all R-CBM probe scores and FCAT SSS reading scale scores and FCAT NRT reading scale scores (see Table 6). Correlations ranged from .74 to .76 for all R-CBM probe scores and FCAT scores.

Table 7

Correlation Matrix for R-CBM Probe Scores and FCAT-SSS and FCAT-NRT

Variable	1	2	3	4	5
1. FCAT-SSS	--	.85**	.75**	.75**	.74**
2. FCAT-NRT		--	.74**	.75**	.76**
3. FCAT R-CBM			--	.96**	.94**
4. Generic R-CBM				--	.93**
5. Content R-CBM					--

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

Scatterplots

Figure 1 shows the relationship between Generic R-CBM scores and FCAT-SSS using the benchmarks described by Good, Simmons, Kame'enui, Kaminski, and Wallin (2002). In this figure, the horizontal line represents the passing score on the FCAT-SSS (standard score at or above 284). Thus, students scoring at or above the horizontal lines were at or above the acceptable range on the test, or grade level, as determined by the FCAT-SSS. The vertical lines represent the cut scores determined by Good et al. (2002). According to these benchmarks, students who read over 110 words correct in one minute are considered to be at low risk. Students who read 80-110 words correct in one minute are considered to be a some risk and students who read fewer than 80 words correct in one minute are at-risk and in need of intensive intervention.

For a breakdown of Generic R-CBM prediction of FCAT-SSS scores, see Table 7. Of the students who were in the "low-risk" range reading at or above 110 words correct per minute, 96% received passing scores on the FCAT. Fifty-three percent of student's

whose R-CBM scores fell in the "some risk" range passed the FCAT. Interestingly, 22% of students who score in the "at-risk range" for number of words read correct in one minute passed the FCAT.

Figures 2 and 3 display scatterplots of the relationship between generic R-CBM and content R-CBM scores and FCAT-NRT scores. When analyzing the data, the researcher created scatterplots for each of the R-CBM probe types and both FCAT-SSS and FCAT-NRT. The scatterplots shown here were selected because they represent the R-CBM probe types that came out statistically significant in the multiple regression analyses explained later. Due to the fact that the R-CBM probe types were highly correlated, the examiner felt it would be redundant to graphically represent the relationship between each probe type and both FCAT-SSS and FCAT-NRT score.

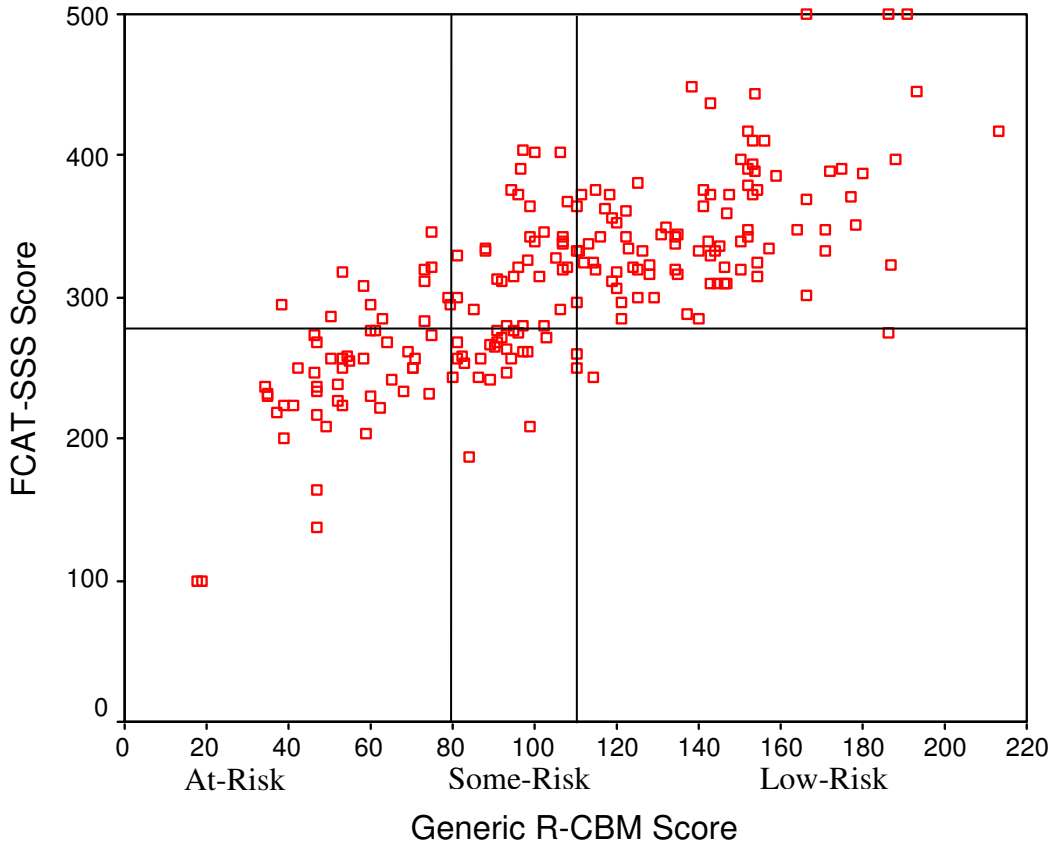


Figure 1. Scatterplot of the Relationship between Generic R-CBM Probe Score and FCAT-SSS

Table 8

Predicting FCAT-SSS Reading Score from Generic R-CBM Scores

	Oral Reading Fluency Classification		
FCAT Performance	High-Risk (<80 WRCM)	Some-Risk (80-109 WRCM)	Low-Risk (>110 WRCM)
Passing (FCAT-SSS Scores 284-500)	22%	53%	96%
Not Passing (FCAT-SSS Scores 100-283)	78%	47%	4%

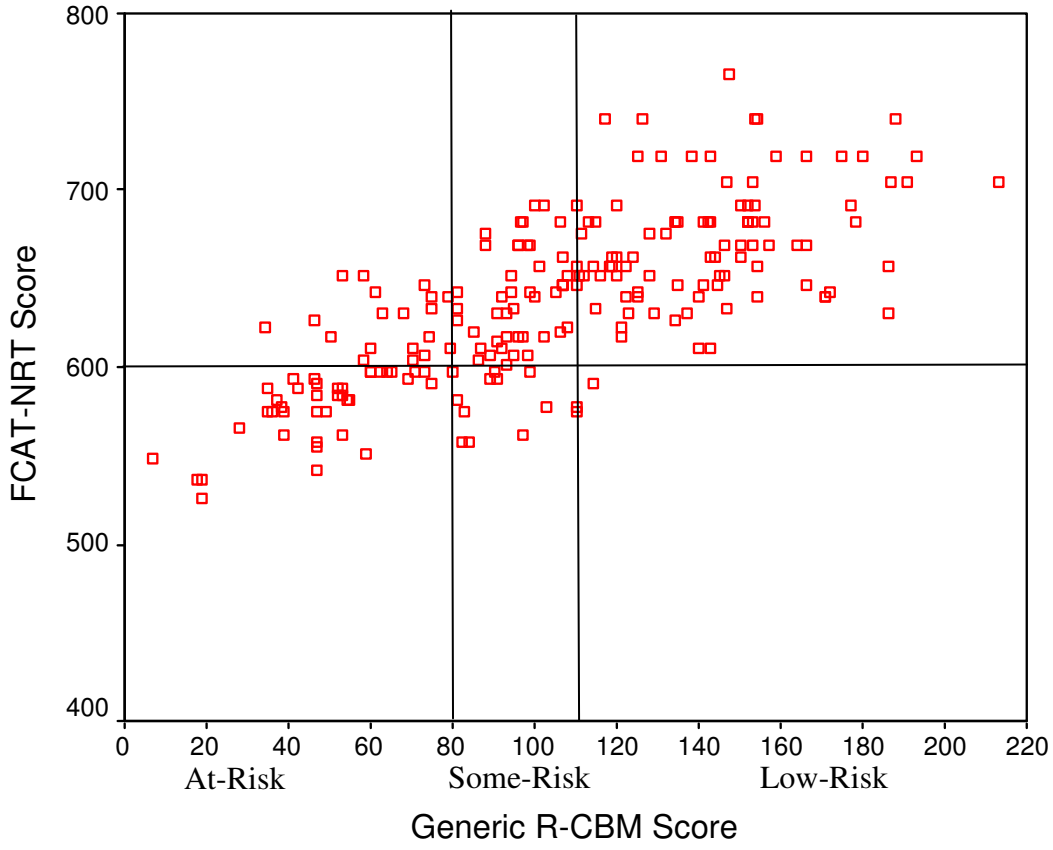


Figure 2. Scatterplot of the Relationship Between Generic R-CBM Probe Score and FCAT-NRT

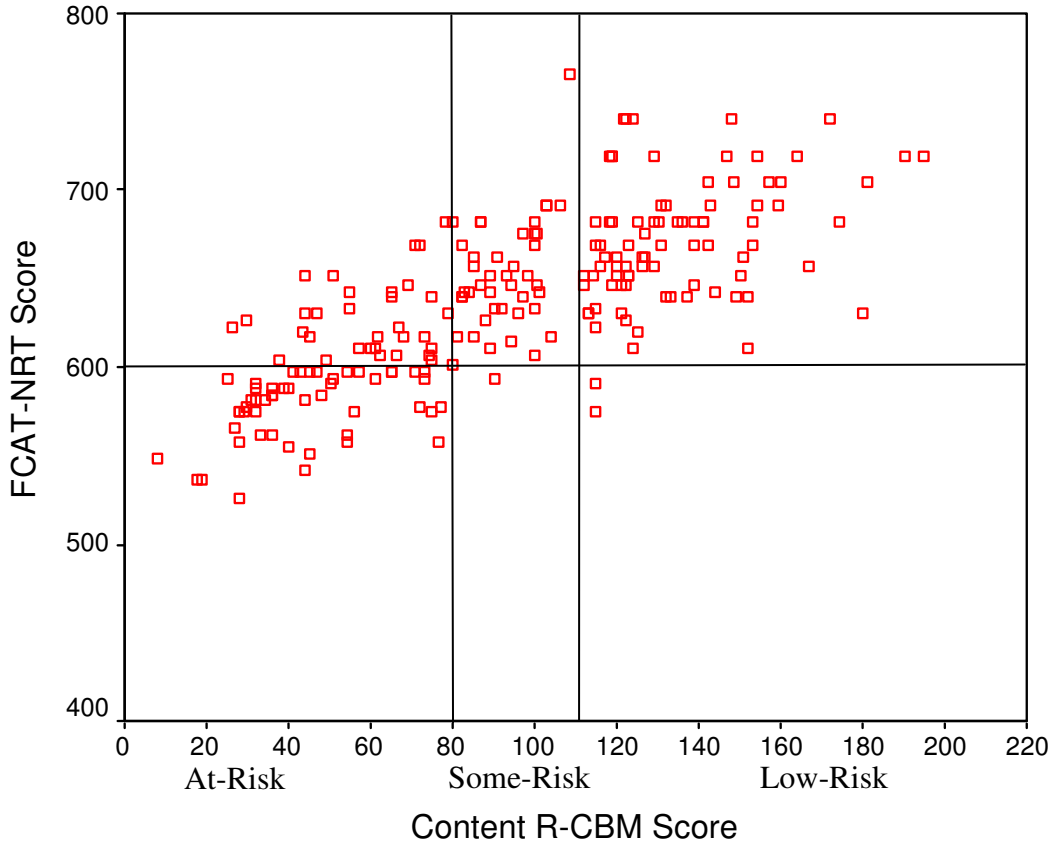


Figure 3. Scatterplot of the Relationship Between Content R-CBM Probe Score and FCAT-NRT

Assumptions of Multiple Regression

Prior to answering the research questions, the underlying assumptions for regressions were examined. For regressions to provide a good index of the association between two or more variables, these assumptions must be met. Overall, the data for this study did not violate the assumptions and thus the obtained results are considered valid.

Multicollinearity and singularity, or the relationships among the independent variables were examined. Multicollinearity was examined for the R-CBM probes because they were highly correlated (above .90). Due to the fact the probes were taken from different passages, this assumption was not violated. Singularity occurs when one

variable is a combination of other variables. In order to not violate the assumption of singularity, the three different probe scores were considered in one multiple regression. The average of the three probe scores was not entered into the same regression as any of the other three probes scores which it comprised.

Multiple regression analyses are sensitive to outliers; thus extreme scores were examined for all of the variables in the regression analyses. Outliers were visually identified using the standardized residual plot for each regression, and there were no apparent outliers which Tabachnick and Fidell (1996) define as those with standardized residual values less than -3.3 or above 3.3.

To assess the assumptions of normality, linearity, homoscedasticity and independence of residuals, the residuals scatterplots and normal probability plots were examined. The normal probability plots suggested linearity with no major deviations from normality in that the points were in a relatively straight diagonal line from bottom left to top right. The assumption of homoscedasticity did not appear to be violated upon examination of the scatterplots of the standardized residuals with most of the scores centered but with the same variance throughout the predictor (Pallant, 2001).

Apriori Power Analysis

The number of participants required for multiple regression analyses is 15 observations per predictor variable (Stevens, 1996). Because there were a maximum of five predictors included in each of the analyses, 75 participants were necessary to achieve adequate power for the regressions. The current study satisfied the condition by including 178 participants at minimum in the regression analysis.

Multiple Regression Procedures

Multiple regression analyses were performed to determine whether ethnicity, socioeconomic status, median FCAT R-CBM score, median generic R-CBM score, and/or median content R-CBM score significantly predicted scores on the FCAT. Two regression analyses were performed using the variables to predict FCAT-SSS and FCAT-NRT for the purpose of cross validation.

The multiple regression analysis predicting FCAT-SSS by ethnicity, socioeconomic status, median FCAT R-CBM score, median generic R-CBM score, and median content R-CBM score was statistically significant ($p<.0005$). The predictor variables accounted for 60% of the variance in the criterion variable (FCAT-SSS), ($R^2=.60$, $F(8,171)=199.23$, $p<.001$). The adjusted R^2 was .58, indicating little shrinkage of the true value in the population. To determine which of the predictors contributed to the prediction of FCAT-SSS, the contributions of each of the independent variables were compared (Table 8). The generic R-CBM score yielded the largest beta coefficient, .41, and made the strongest unique contribution to explaining FCAT-SSS when the variance explained by ethnicity, socioeconomic status, FCAT R-CBM and content R-CBM were controlled for. In addition, generic R-CBM score made a statistically significant unique contribution to the equation. Ethnicity, socioeconomic status, FCAT R-CBM and content R-CBM did not make any contributions to FCAT-SSS over and above generic R-CBM score.

The multiple regression analysis predicting FCAT-NRT from ethnicity, socioeconomic status, median FCAT R-CBM score, median generic R-CBM score, and median content R-CBM score was statistically significant ($p<.0005$). The R^2 was .62,

which indicates that the predictor variables accounted for 62% of the variance in the criterion variable (FCAT-NRT) (R^2=.62, F(8, 174)=558.88, p<.001). The adjusted R^2 was .60, indicating little shrinkage of the true value in the population. To determine which of the predictors contributed to the prediction of FCAT-NRT, the contributions of each of the independent variables were compared (Table 9). The median Generic R-CBM score yielded the largest beta coefficient, .48, and made the strongest unique contribution to explaining FCAT-NRT when the variance explained by ethnicity, socioeconomic status, FCAT R-CBM score and content R-CBM score were controlled for.

Table 9

Summary of Multiple Regression Analysis for Predictors of FCAT-SSS

Variable	B	β
Black	-11.26	-.09
Asian	.68	.00
Hispanic	-9.13	-.05
Mixed	-21.57	-.05
Socioeconomic Status	4.08	.06
FCAT R-CBM	.31	.19
Generic R-CBM	.63	.41*
Content R-CBM	.19	.12

*p<.05

The beta coefficient for median content R-CBM probe was .31. Thus, both the median generic R-CBM probe score and median content R-CBM probe score made statistically significant contributions to the equation. Ethnicity, socioeconomic status, and FCAT R-CBM probe score, however, did not make any contributions to FCAT-NRT over and above generic R-CBM score and content R-CBM score.

Table 10

Summary of Multiple Regression Analysis for Predictors of FCAT-NRT

Variable	B	β
Black	-10.03	-.11
Asian	-7.08	-.01
Hispanic	-6.01	-.05
Mixed	-15.59	-.05
Socioeconomic Status	5.16	.11
FCAT R-CBM	-.10	-.09
Generic R-CBM	.53	.48*
Content R-CBM	.36	.31*

*p<.05

Chapter 5

Discussion

The purpose of this study was to determine the relationship between oral reading fluency (ORF) and reading scores on the Florida Comprehensive Assessment Test (FCAT) for third-grade students. Several studies conducted in states in all regions of the country have found a positive, and in most cases, a statistically significant relationship between ORF and statewide tests of achievement (Barger, 2003; Buck & Torgeson, 2003; Castillo, Torgeson, Powell-Smith & Al Otaiba, 2003; Crawford, Tindal, & Steiber, 2001; Good et al., 2001; Linner, 2001; McGlinchey & Hixon, 2004; Shaw & Shaw, 2002; Sibley, Biwer & Hesch, 2001; Shapiro, Edwards, Lutz & Keller, 2004; Stage & Jacobson, 2001). This study extended the literature of ORF and statewide achievement tests by including a large sample of minority students. The current chapter discusses the results of this study in light of the proposed research questions. Limitations of the study are presented along with implications for educators, school psychologists, and future research.

Research Questions

Research Question 1: What is the relationship between third-grade students' oral reading rate and performance on the reading portion of the Florida Comprehensive Assessment Test?

Median Curriculum-based Measurement-Reading (R-CBM) probe score was strongly related to third-grade students' performance on the Florida Comprehensive Assessment Test (FCAT). Correlations between R-CBM and FCAT were high and statistically significant. Consistent with prior research, R-CBM score was found to be highly predictive of a passing score on the FCAT (Barger, 2003; Buck & Torgeson, 2003; Good, Simmons, & Kame'enui, 2001; Shapiro, Edwards, Lutz, & Keller, 2004; Shaw & Shaw, 2002; Sibley, Biwer, & Hesch, 2001; Stage & Jacobson, 2001). Specifically, students whose oral reading fluency fell in the low-risk range (over 110 words read correct in one minute) as determined by DIBELS benchmarks were virtually assured a passing score (Level 3) on the FCAT.

The relationship between students' R-CBM scores in the some-risk or at-risk categories (80-110 words read correct in one minute and 0-80 words read correct in one minute, respectively) was less clear. These results are consistent with the findings by Good et al. (2001) and Buck and Torgeson (2003) who also reported that for students reading 80-110 words correct in one minute, the relationship to statewide achievement test scores was unpredictable and students were equally likely to pass or fail. Synonymous with their findings, data from this study showed that 53% of the students in the some-risk range passed the FCAT and 47% in the same range failed the FCAT. Students were equally likely to pass or not pass the FCAT. Further research might

explore additional correlates such as comprehension, vocabulary knowledge, motivation to read or other related variables to further understand the reading abilities and passing abilities of students falling in the some-risk group.

Students whose R-CBM scores fell in the at-risk range were more likely to receive a failing score on the FCAT (Level 1 or 2). Specifically, 75% of students whose scores fell in the at-risk range failed the FCAT. Although R-CBM was found to be more sensitive at predicting passing than failing scores on the FCAT, it was sensitive to passing at both extremes. In other words, high R-CBM scores (over 110 wcpm) were related to passing the FCAT and low R-CBM scores (below 80 wcpm) were related to failing the FCAT. Overall, results of this study found students who read in the low-risk range as defined by DIBELS benchmarks were more likely to pass the FCAT and students who were at-risk for reading failure as defined by DIBELS benchmarks were not likely to pass the FCAT. Findings were similar for both the FCAT-SSS and FCAT-NRT.

Research Question 2: What is the relationship between third-grade students' oral reading rate on three different R-CBM passage types (FCAT passages, curriculum passages, content passages) and scores on the reading portion of the Florida Comprehensive Assessment Test?

In the current study, the relationship between R-CBM score and FCAT score was statistically significant. There were no statistical differences among probe type when correlated with FCAT score. However, when ethnicity, socioeconomic status, and the three probe types (FCAT R-CBM, generic R-CBM, and content R-CBM) were entered into a prediction equation, only generic R-CBM significantly predicted FCAT-SSS score.

For the FCAT-NRT score, both generic R-CBM and content R-CBM significantly predicted this score.

These results should be interpreted cautiously, however, since the three R-CBM probes were so highly correlated. Because the three probes were highly correlated, this has potentially decreased the amount of variance in FCAT score that is accounted for uniquely by the individual probe types. The result of the redundancy of R-CBM probes is a net decrease in the total amount of variance that is accounted for by the linear combination of individual R-CBM probes and FCAT score (Hatcher & Stepanski, 1994). Future research should consider the relationship between individual R-CBM probe types, ethnicity, and socioeconomic status in order to determine if a higher proportion of variance would be accounted for.

Due to the fact that the R-CBM probes were highly related to each other, for practical reasons, they should not be separated from each other and considered independently. In other words, though the generic R-CBM probe was found to be statistically significant, it is not practically significant and all three probes were equally related to the FCAT score. Because the R-CBM probes were highly correlated with each other, the assumption of multicollinearity might again be addressed. Multicollinearity can cause regression coefficient estimates to fail to demonstrate statistical significance (Hatcher & Stepanski, 1994). This could have been the case in the current study.

Because all three R-CBM probe scores were highly correlated, it may not have been necessary to run analyses on each of the probe types separately. The data used in the current study was archival, thus the researcher was not involved in the selection of the probes used in the study. However, the granting institute (FCRR) called for the

researcher to compare the three different probes (as this has not before been considered in relation to statewide testing). In order for the researcher to answer questions pertinent to the grant, it was necessary to proceed with the subsequent analyses.

The finding that all R-CBM probe were highly correlated support previous research that suggests that it is not essential for R-CBM reading passages to be "curriculum-based" (e.g., Bradley-Klug, Shapiro, Lutz, & DuPaul 1998; Fuchs & Deno; 1994; Hintze & Shapiro, 1997, Powell-Smith & Bradley-Klug, 2001). These results suggest that either generic, content (curriculum), or FCAT practice passage probe type could be used to monitor students' progress over time.

Overall, this is the first study to examine the relationship between three different R-CBM probe types and a statewide achievement test. The findings of the current study were consistent with the findings of Castillo et al. (2003) who found that two forms of generic probes, DIBELS and Monitoring Basic Skills Program (MSBP) probes, were equally related to FCAT scores. Besides Castillo et al. (2003), all other studies considering the relationship between R-CBM and statewide achievement tests used either a single probe or the median of three probes (generic probes or basal series probes). Prior research did not include passages taken directly from the statewide achievement test, therefore, this study offers a unique contribution to the literature.

Research Question 3: What is the relationship between third-grade students' ethnicity, oral reading rate, and scores on the reading portion of the Florida Comprehensive Assessment Test?

In the current study, a significant relationship was not found between ethnicity and FCAT score. Compared with the other studies reviewed, this study included the

largest number of minority students. In fact, there were more minority students than Caucasian students included in this sample. The finding that ethnicity is not related to FCAT score is an interesting result when compared with the findings of Buck and Torgeson (2003). These authors found that the relationship between African American students' R-CBM scores was lower than the Caucasian students' scores. Hispanic students' R-CBM scores were correlated highest with FCAT score; the study by Buck and Torgeson (2003) included only 7% African American students and 6% Hispanic students compared with 42% African American students and 1% Hispanic students in the current study. Therefore, with this large number of minority students included in the sample ethnicity was not a predictor of FCAT score over and above R-CBM scores. Reading CBM scores are better predictors of students' FCAT score than ethnicity.

Research Question 4: What is the relationship between third-grade students' socioeconomic status, oral reading rate, and scores on the reading portion of the Florida Comprehensive Assessment Test?

The results demonstrated that socioeconomic status was not significantly correlated with FCAT score. Socioeconomic status was determined by free and/or reduced lunch status. Approximately half of the students in the sample received free and/or reduced lunch and half did not receive assistance. A few previous studies considered the relationship between free and/or reduced lunch status, R-CBM scores, and statewide tests of achievement (Buck & Torgeson, 2003; Good, Simmons, & Kame'enui, 2001; McGlinchey & Hixon, 2004; Shapiro, Edwards, Lutz, & Keller, 2004; Stage & Jacobsen, 2001). None of the studies that considered free and/or reduced lunch status reported any differences between groups with respect to statewide achievement-test

score. Consistent with prior research, this study also found no significant differences between SES groups' scores on the FCAT. Free and/or reduced lunch status was not a predictor of FCAT score over and above R-CBM score.

Implications for Education/Educators

The results of this study have implications for educators across Florida. First, this study found that third-grade students' R-CBM scores were significantly correlated with third-grade students' FCAT score. Because R-CBM is sensitive to students' growth over time and can be used as a tool to monitor progress, students' level of risk for failure can be determined as early as the beginning of third-grade. From that assessment, teachers can implement intensive interventions for those students found to be at-risk. Good et al. (2001) reported that third-grade R-CBM scores are related to second grade scores, first grade scores, and kindergarten scores. Students can thus be identified in Kindergarten and monitored yearly or more frequently so that they are prepared with the skills tested by the FCAT by the time they reach third-grade.

Another implication for educators is that R-CBM probes for monitoring can be taken directly from generic curricula, text book passages or FCAT materials. Generic R-CBM probes are available to educators on-line and free of charge which saves such important resources as time and money.

Implications for School Psychology

School psychologists are in unique positions to influence change within schools. Many school psychologists in Florida have been trained to assess students using R-CBM. School psychologists can train educators and educational personnel to use R-CBM for continual progress monitoring. Many schools in Florida are concerned with receiving a

passing grade on the FCAT. School psychologists can encourage schools to use R-CBM to determine which students are at risk for failing the FCAT in order to develop intensive interventions and ensure that more students are successful on the FCAT and in reading in general.

Besides No Child Left Behind (NCLB) (of which FCAT is a component), a second piece of national legislation is being considered with wide ranging implications for students. The soon to be reauthorized Individuals with Disabilities Education Act (IDEA), will require responsiveness to intervention (RTI) to be part of Learning Disability identification. Specifically, in the problem-solving model using RTI, students are provided effective instruction which is monitored and those who don't respond to the instruction get additional or different instruction which is also monitored. Only if the students fail to respond do they qualify for special education evaluation (Fuchs, Mock, Morgan, & Young, 2003). Reading Curriculum-based Measurement is one excellent tool for monitoring RTI.

In addition, for students requiring an Academic Improvement Plan (AIP), due to academic skills below level, R-CBM can be used to set specific goals and to monitor attainment of those goals.

Limitations

Several threats to internal and external validity limit the interpretation of the results. Internal validity can be described as the stipulation that the observed differences on the dependent variable are the result of the independent variable and not something else (Gay & Airasian, 2000). Consequently, internal validity is threatened when rival hypotheses can not be eliminated. Several potential threats to the internal validity of this

study exist including instrumentation, differential selection of participants or selection bias, implementation bias, and order bias. External validity, by contrast, is the extent to which study findings can be generalized to and across populations, settings, and times (Johnson & Christensen, 2000). Threats to external validity in the current study include population validity, ecological validity, and specificity of variables.

Threats to internal validity limit interpretation and generalizability of the results. First, instrumentation is a threat to internal validity. Specific measures were selected to determine oral reading fluency and achievement from among many. Second, differential selection of participants or selection bias is a threat to internal validity. Participants were self-selected for the current study based on parent and student interest, parent consent and student assent; only a small subset of the population of each school (and consequently district and state of Florida) participated. Third, implementation bias is a threat to internal validity due to the number and different backgrounds of data collectors (Onweugbuzie, 2003). There were many data collectors and though all data collectors were trained in the same manner, it is not clear if the specific training procedures were followed by all data collectors because there was no systematic observation of data collectors. In addition, some students many have felt more comfortable with the gender or ethnicity of one data collector than another. In either case, one data collector many have elicited a more representative sample of behavior than another. Fourth, order bias is a potential threat to internal validity due to the fact that though there was a recommended sequence of assessments, data collectors could administer the assessments in any order. For the most part, students were exposed to R-CBM probes in the same order. Practice effects might

have contributed to the results. Conclusions of the current study must be interpreted and extrapolated with caution outside of the measures given and individuals assessed.

Threats to external validity include population validity, ecological validity, and specificity of variables. Population and ecological validity refer to the extent to which results are generalizable from the sample of participants to the larger population, as well as across settings, contexts, and conditions (Onwuegbuzie, 2003). Due to the fact that this study was conducted with third graders across three counties in Florida, results should be generalized cautiously to the larger national population. Research findings will also be less generalizable due to specificity of variables or the combination of specific variables (e.g., participants, time, context, conditions, and variables).

Directions for Future Research

Future research should address the limitations of the current study. First, in order to evaluate the ability of R-CBM to predict performance on statewide achievement tests, R-CBM data should be collected prior to FCAT administration.

Second, additional research comparing different R-CBM probes and statewide achievement tests must be examined before a conclusive argument for use of any or all of the R-CBM probe types can be made. This line of research has been relatively unexplored.

Third, to generalize better the results of this study, a truly representative geographic, ethnic, and socioeconomically diverse sample of students rather than a convenience sample of students should be used. Though the current study sampled students from three districts across Florida, results are not generalizable outside of

Florida. A replication of the current study using a broad population base to determine if similar results would be found is recommended.

Finally, longitudinal R-CBM data collected on a sample of students from kindergarten through third-grade when they take the statewide achievement test would be beneficial. Through these data, a more complete picture of the long term identification effectiveness of R-CBM could be found. In addition, these data would provide information on best timing for interventions to predict passing scores.

Conclusion

In conclusion, the current study contributed to the research base by examining the relationship between Curriculum-based Measurement reading (R-CBM) and the Florida Comprehensive Assessment Test (FCAT). Consistent with previous research, R-CBM probe scores were highly and statistically significantly correlated with third-grade students' scores on the reading portion of the Florida Comprehensive Assessment Test. It appears that all three probe types (FCAT R-CBM, generic R-CBM, and content R-CBM) explored in this study can be used to monitor students' performance in relation to FCAT outcomes. Ethnicity and socioeconomic status were not significant predictors of FCAT scores above student R-CBM score. These data have several implications but must be interpreted with caution due to the limitations of the study. School psychologists can advocate for R-CBM and train educators in the benefits and administration of R-CBM. Research to date supports the use of R-CBM (a sensitive, brief, inexpensive measure) to identify students who might be at-risk for reading failure in order to provide them with intensive interventions to avoid failing a high-stakes achievement test.

References

American Psychological Association. (2001, May). *Appropriate use of high-stakes testing in our nation's schools.* Retrieved July 17, 2003, from http://www.apa.org/pubinfo/testing.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.

Bradley-Klug, K.L., Shapiro, E.S., Lutz, J.G., & DuPaul, G.J. (1998). Evaluation of oral reading rate as a curriculum-based measure within literature-based curriculum. Journal of School Psychology, 36, 183–197.

Broward County School District (n.d.) Retrieved on September 30, 2003 from http://www.browardschools.com.

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 No. 1). Tallahassee, FL: Florida Center for Reading Research.

Castillo, J. M., Torgeson, J. K., Powell-Smith, K. A., & Al Otaiba, S. (2003). Relationships of five reading fluency measures to reading comprehension in first through third grade. Manuscript in preparation.

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

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

Deno, S. L., Mirkin, P. K., & Chiang, B. (1982). Identifying valid measures of reading, *Exceptional Children,* 46, 36-45.

Donahue, P. L., Finnegan, A. D., & Lutkus, N. L. (2001). The nation's report card: fourth-grade reading 2001, U.S. Department of Education: Washington, DC.

Florida Center for Reading Research (n.d.). *The Florida center for reading research four -part mission*. Retrieved December 10, 2003, from http://www.fcrr.org/aboutFCRR/mission.htm.

Florida Comprehensive Assessment Test Briefing Book. (2001). Florida Department of Education: Tallahassee.

Florida Department of Education. (2002). *Florida Statutes and State Boards of Education Rules Excerpts for Special Programs* (Volume 1-B). Tallahassee, Florida: Florida Department of Education.

Florida Department of Education. (n.d). *Florida Comprehensive Assessment Test*. Retrieved March 18, 2003, from http://www.firn.edu.

Fraenkel, J. R. & Wallen, N. E. (2003). *How to design and evaluate research in education.* New York: McGraw-Hill Higher Education.

Fuchs, L. S. & Deno, S. L. (1992). Effects of curriculum within curriculum-based measurement. *Exceptional Children*, 58, 232-242.

Fuchs, L.S., & Deno, S.L. (1994). Must instructionally useful performance assessment be based in the curriculum? Exceptional Children, 61, 15–24.

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

Fuchs, L. S., Fuchs, D., & Maxwell, L. (1988). The validity of informal reading Comprehension measures. *Remedial and Special Education,* 92, 20-28.

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

Good, R. H., & Salvia, J. (1988). Curriculum bias in published norm-referenced reading tests: Demonstrable effects. *School Psychology Review,* 17, 51-60.

Good, 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,* 257-288.

Great Schools Data Base (n.d.). Retrieved December 9, 2003, from http://www.greatschools.net.

Harcourt Educational Measurement Company (n.d.). Retrieved March 20, 2003 from http://www.hemweb.com.

Hatcher, L. & Stepanski, E. J. (1994). *A step-by-step approach to using the SAS system for univariate and multivariate statistics.* Cary, NC: SAS Institute Inc.

Hillsborough County School District. (n.d.). Retrieved on September 30, 2003 from http://www.sdhc.k12.fl.us.

Hintze, J.M., & Shapiro, E.S. (1997). Curriculum-based measurement and literature-based reading: Is curriculum-based measurement meeting the needs of changing reading curricula? Journal of School Psychology, 35, 351–357.

Jenkins, J. R. & Pany, D. (1978). Standardized achievement tests: How useful for special education? *Exceptional Children*, *44,* 448-453.

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

Leon County School District (n.d.) Retrieved on September 30, 2003 from http://www.leon.k12.fl.us/.

Linn, R. L. (2000). Assessments and Accountability. *Educational Researcher, 29,* 4-16.

Linner, S. (2001, January). *Curriculum based assessment in reading used as a predictor for the Alaska Benchmark Test*. Paper presented at the Alaska Special Education Conference, Anchorage, AK.

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

Mather, N., Hammill, D. D., Allen, E. A., & Roberts, R. (2004). *Test of silent word reading fluency*. Austin, TX: PRO-ED.

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

National Association of School Psychologists (2003, April, 12). *Position statement on using large scale assessment for high stakes decisions.* Retrieved July 17, 2003, from http://www.nasponline.org/information/pospaper_largescale.html.

National Institute for Literacy (n.d.). Retrieved March 18, 2004 from http://www.nifl.gov.

National Research Council (1998). *Preventing reading difficulties in young children.* Washington, D.C: National Academy Press.

Nolet, V & McLaughlin, M. (1997). Using CBM to explore a consequential basis for the validity of a state-wide performance assessment. *Diagnostique, 23*, 147-163.

Onwuegbuzie, A. J. (2003). Expanding the framework of internal and external validity in quantitative research. *Research in the Schools, 10,* 71-89.

Powell-Smith, K.A., & Bradley-Klug, K.L. (2001). Another Look at the "C" in CBM: Does it really matter if curriculum-based measurement reading probes are curriculum-based? *Psychology in the Schools, 38*, 299-312.

Shapiro, E.S. (1996). *Academic skills problems* (2nd ed.). New York: The Guilford Press.

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.

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

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

Sibley, D., Biwer, D., & Hesch, A. (2001). *Unpublished data*. Arlington Heights, IL: Arlington Heights School District 25.

Spache, G. (1953). A new readability formula for primary-grade reading materials. *The Elementary School Journal, 55,* 410-413.

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

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

Stevens, J. P. (1999). *Intermediate statistics a modern approach (2nd Ed.).* Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Tabachnick, B. G., & Fidell, L. S. (1996). *Using multivariate statistics* (3rd edition). New York: HarperCollins.

Thurlow, M. L., & Thompson, S. J. (1999). District and state standards and assessments. *Journal of Special Education Leadership, 12,* 3-10.

Torgeson, J. K., Wagner, R. K., Lonigan, C. J., & DeGraff, A. (2002). *Test of critical early reading skills.* Unpublished Manuscript, Florida State University.

Torgeson, J. K., Wagner, R. K., & Rashotte, C. A. (1999). *Test of word reading efficiency.* PRO-ED inc.

United States Department of Education. (n.d.). *No Child Left Behind Act.* Retrieved March 15, 2003 from http://www.nochildleftbehind.gov.

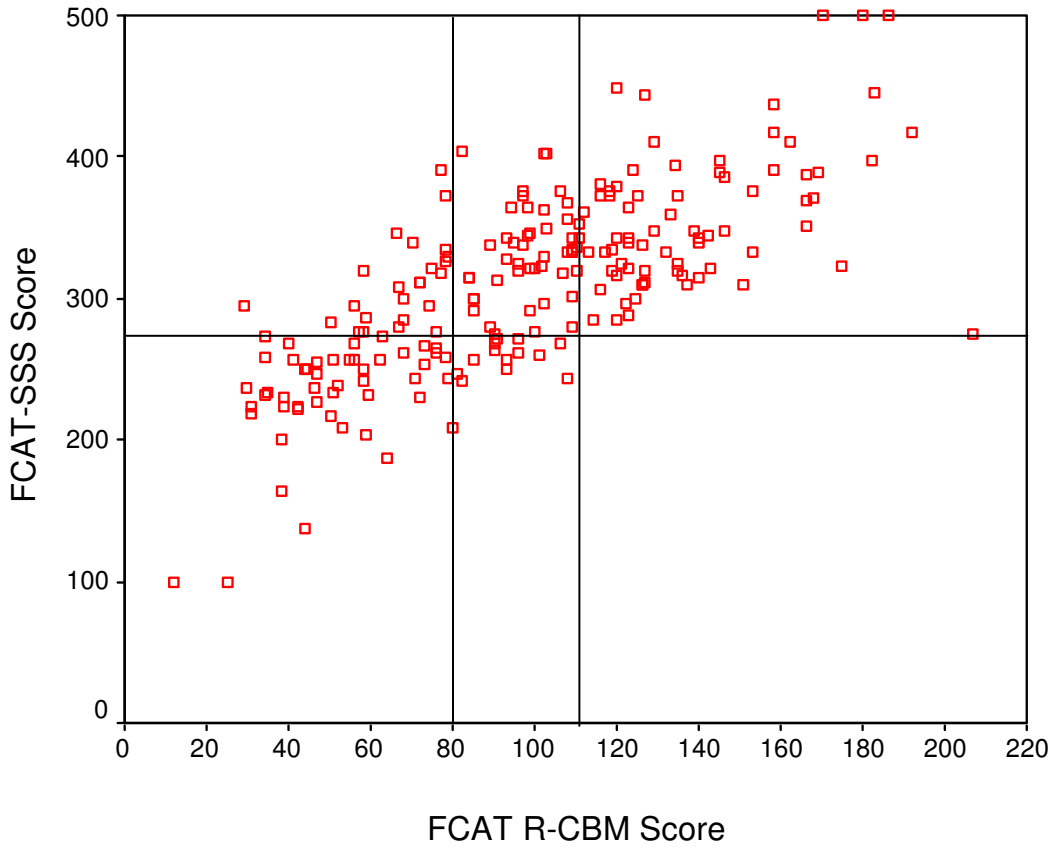
Appendices

Appendix A: Test Scores by Ethnicity and SES

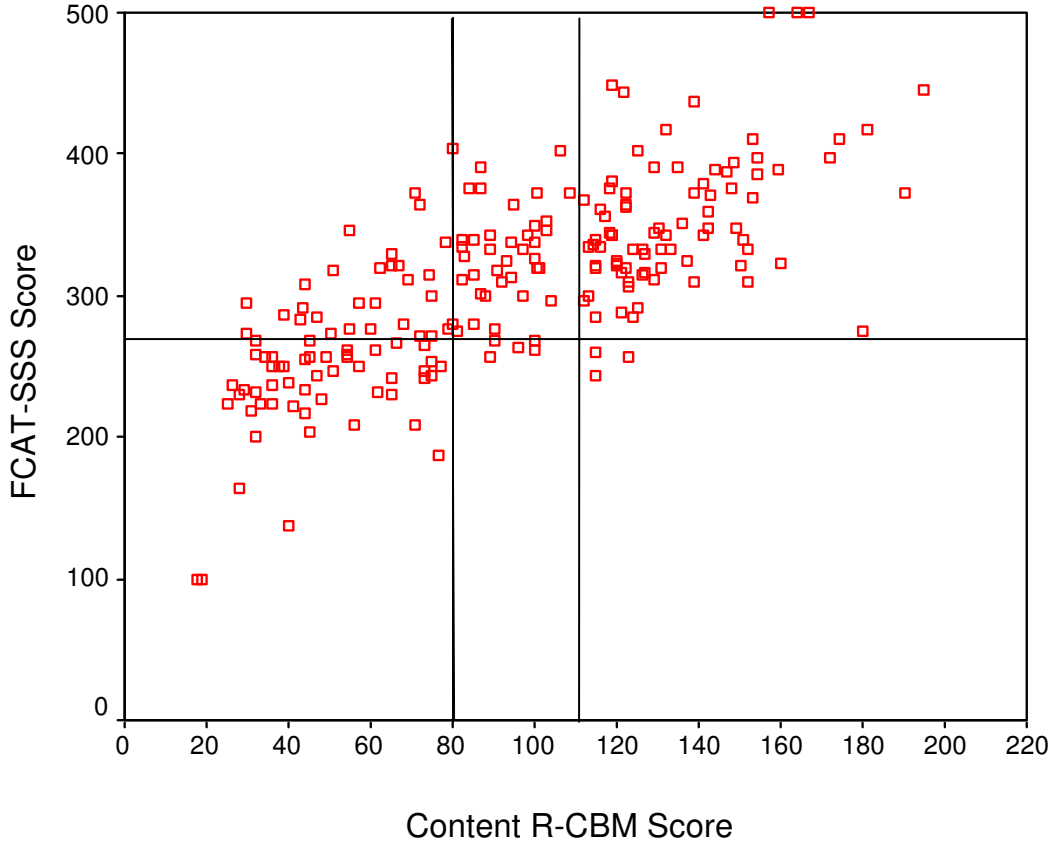
	Black	Caucasian	Asian	Hispanic	Mixed	Free Lunch	Reduced-Price Lunch	Full-Priced Lunch
FCAT-SSS								
n	84	82	3	31	5	80	9	91
Mean	285.57	336.51	349.00	302.42	309.60	284.91	311.67	335.36
Standard Deviation	51.18	59.96	13.89	68.47	135.108	57.32	55.21	63.90
FCAT-NRT								
n	87	83	3	30	5	82	10	91
Mean	618.01	659.06	660.33	633.57	636.00	617.09	630.40	658.84
Standard Deviation	42.93	44.80	18.93	38.37	65.84	42.50	52.16	43.40

FCAT R-CBM								
n	90	83	3	32	5	85	10	92
Mean	80.22	111.64	118.45	91.50	113.60	79.29	90.60	109.58
Standard Deviation	35.47	33.98	7.88	41.29	70.13	35.30	40.01	36.12
Generic R-CBM								
n	90	83	3	32	5	85	10	92
Mean	90.65	119.79	133.11	99.81	113.80	88.59	93.80	119.45
Standard Deviation	38.90	37.19	14.77	40.69	76.18	39.11	45.38	37.66
Content R-CBM								
n	90	83	3	32	5	85	10	92
Mean	74.66	111.38	111.79	89.34	103.80	74.94	87.60	108.75
Standard Deviation	36.81	35.74	12.14	39.49	64.76	37.77	44.03	37.42

Appendix B: Scatterplot of the Relationship Between FCAT R-CBM Probe Score and FCAT-SSS



Appendix C: Scatterplot of the Relationship Between Content R-CBM Probe Score and FCAT-SSS



Appendix D: Scatterplot of the Relationship Between FCAT R-CBM Probe Score and FCAT-NRT