2006

Predicting Third Grade Students' FCAT Reading Achievement and Oral Reading Fluency Using Student Demographic, Academic History, and Performance Indicators

Angela I. Canto
THE FLORIDA STATE UNIVERSITY
COLLEGE OF EDUCATION

PREDICTING THIRD GRADE STUDENTS’ FCAT READING ACHIEVEMENT AND
ORAL READING FLUENCY USING STUDENT DEMOGRAPHIC, ACADEMIC HISTORY,
AND PERFORMANCE INDICATORS

By

ANGELA I. CANTO

A Dissertation submitted to the
Department of Educational Psychology and Learning Systems
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Degree Awarded:
Spring Semester, 2006
The members of the Committee approve the dissertation of Angela I. Canto defended on February 24, 2006.

___________________________
Briley E. Proctor
Professor Directing Dissertation

___________________________
Stephanie Al Otaiba
Outside Committee Member

___________________________
Frances Prevatt
Committee Member

___________________________
Gary Peterson
Committee Member

Approved:

___________________________
Frances Prevatt, Ph.D., Chair,
Department of Educational Psychology and Learning Systems

The Office of Graduate Studies has verified and approved the above named committee members.
Philip: For all that you are and all that you have helped me to become. Without your love and undying support, none of this would have been possible. Thank you for your love and for our beautiful son. The two of you fulfill my life and complete me. I am left in breathless wonder, ever-thankful to share my life with you. “If you need a push, a pull, a tug, or just a hug…”

Mom and Dad: Thank you for all of your continued support in everything I try to accomplish. I am blessed to be your daughter and will always be your “little angel”, your “caboose”.

Kamela: Thank you for the time you gave me to work on this project. You are a wonderful supervisor and mentor and an even greater friend.
ACKNOWLEDGMENT

Thank you, Dr. Proctor, for the years of guidance and support you have graciously given to me. Your contribution to my education and this project is unparalleled and your mentorship is not forgotten.

Thank you to each committee member for you valuable insight and recommendations on this project. I appreciate all of your input and hard work.

Preliminary statistical assistance on this project was graciously provided by Dr. Richard Tate.
# TABLE OF CONTENTS

List of Tables .............................................................................................................vii  
Abstract ......................................................................................................................ix  

1. SOCIAL SIGNIFICANCE AND PROFESSIONAL RESPONSE ..............1  
   Social Significance.................................................................................................1  
   Professional Response ..........................................................................................2  
   Definition of Terms...............................................................................................3  

2. LITERATURE REVIEW ...............................................................................5  
   Introduction...........................................................................................................5  
   Reading ..................................................................................................................5  
   High-Stakes Testing..............................................................................................10  
   Florida Comprehensive Assessment Test (FCAT) ............................................21  
   Oral Reading Fluency .........................................................................................31  
   DIBELS Assessment............................................................................................35  
   Summary...............................................................................................................40  

3. METHODOLOGY .......................................................................................43  
   Research Questions.............................................................................................43  
   Participants...........................................................................................................43  
   Sampling...............................................................................................................44
LIST OF TABLES

1. Descriptive Statistics for Gender, Ethnicity, and SES for Third Grade, Fall 2004 ..........................................................50
2. Number of Grade Retentions Prior to 2004-2005..........................51
3. Descriptive Statistics for Participation Ratings .............................52
4. Descriptive Statistics for NWF, ORF and FCAT Scores....................52
5. Presence of Missing Data ..........................................................57
6. Correlations Among Missingness Indicators and Study Variables ........60
7. Comparison of Observation Replacement Strategies of Missing Participation Values .................................................................63
8. Reading Grades for 2004-2005 ....................................................64
9. Descriptive Statistics and Intercorrelations Among Variables .............66
10. Regression Results Summary for Hypothesis 1, Model 1—Final Model ....70
11. Stepwise Regression Results Summary for Hypothesis 1, Model 2—Complete Model ..........................................................74
12. Regression Results for Hypothesis 2 ............................................77
13. Frequency of Reading FCAT-Level scores ..................................79
14. Comparison of Proportion of Risk Group Failures on Reading FCAT-SSS ..........................................................80
15. Good et al. (2002) Risk Group Failures per Reading FCAT Level Score ....81
16. FCRR (2004) Risk Group Failures per Reading FCAT Level Score ........82
17. Reproduced Good et al. (2002) ORF Risk Level Classification .............112
18. Reproduced FCRR (2004) ORF Risk Level Classification
(2004-2005 School Year)........................................................................................................113
ABSTRACT

The Florida Comprehensive Assessment Test (FCAT) is Florida’s annual measure of student yearly academic progress with major implications for individual students and schools. The present study explored the inter-relationships among student background and reading performance indicators and the predictive utility of those variables as a means to identify students at risk for FCAT failure. The following indicators were incorporated into this study: student demographics (gender and socio-economic status), academic history (attendance rate and number of prior grade retentions), teacher-rated evaluation of student performance (student participation in Reading activities and Reading grade), and reading skill (decoding and oral reading fluency). Decoding and oral reading fluency were measured using the DIBELS Oral Reading Fluency (ORF) and Nonsense Word Fluency (NWF) subtests (Good and Kaminski, 2002b), three months prior to FCAT administration.

The student background and performance variables were first used in a multiple regression analysis to predict FCAT Reading achievement. Among these predictor variables, ORF, student’s third quarter Reading grade, and SES were found to be statistically significant predictors of Reading FCAT-SSS scores. The strong correlation found between ORF and FCAT achievement in this study was congruent with results in previous studies, whereas the addition of Reading grade and SES as significant predictors of Reading FCAT-SSS scores were new findings in this study. Next, it was of interest to determine whether any of the aforementioned indicators also predicted ORF ability. Reading grade and NWF scores were both statistically significant predictors of ORF performance.

Lastly, this study examined the utility of two risk models of Reading failure that use cutpoint guidelines for student performance on the DIBELS ORF subtest to determine risk level. These cutpoint models were applied to this data set to determine the usefulness of these models in identifying students who may be at risk for reading difficulty. Both models explored were moderately effective in identifying at-risk students. However, a three-category risk model (low,
moderate, high) was determined to be of greatest assistance to educators for efficiency and ease. In conclusion, this author identified limitations inherent in the present study and offered recommendations for future research directions.
In recent decades federal and state governments have become heavily involved in the assessment of student academic achievement, proposing the establishment of high academic standards for students and schools (Hiebert, 2002). Achievement tests are administered to students to ensure those standards are being met. With the federal “No Child Left Behind” (NCLB) initiative, it was expected that every child would be able to read by the end of the third grade (U.S. Department of Education [USDOE], 2002). With the implementation of the NCLB Act in 2001, USDOE required that states institute appropriate accountability systems to measure schools’ yearly progress with annual assessments for all students (Florida Association of School Psychologists [FASP], 2003; USDOE, 2002). Thus, state educational testing programs were increasingly utilized to communicate student performance results to parents, educators, policy makers, and the public in general (Feuer, Holland, Green, Bertenthal, & Hemphill, 1999).

The Florida Comprehensive Assessment test (FCAT) was developed as an integral part of educational reform in the state of Florida in an effort to raise standards for education, improve student education, and increase school accountability (Florida Department of Education [FDOE], 2001b). The FCAT meets the requirements of the NCLB Act and students’ scores on the FCAT impact individual students and schools at large. Based upon the aggregate student scores on the FCAT, the schools are assigned a grade, which is the state’s accountability and publicity measure to promote school improvement (Jefferson, 2003). Regarding individual students, FCAT scores are commonly used to determine promotion eligibility, unless just cause can be demonstrated that a student should or should not be promoted (FDOE, 2001b).
Reading achievement is a major focus on the FCAT due to its importance in students’ success in all aspects of education. It is important that by fourth grade students are able to read skillfully, with sufficient comprehension and fluency, to manage new text content and text types (Strickland, 2002). The primary reason for this is because for students in fourth grade, the focus of instruction shifts from “learning to read” to “reading to learn”. If a child does not learn to read with at least modest skill by third grade, the probability that the child will graduate high school is significantly reduced (Snow, Burns & Griffin, 1998). Failure to read and to graduate high school severely limits subsequent occupational and economic opportunities. Furthermore, a lack of reading skill can dramatically impact one’s skill in navigating an ever-increasing technological and literate society, potentially diminishing overall quality of life. Thirty-eight percent of Florida’s students failed the FCAT Reading test in 2003, 35% of students failed in 2004, and 33% of students failed in 2005 (FDOE, 2003b, 2004b, 2005a). Given the far-reaching consequences of FCAT scores, educators need accurate and efficient methods to predict student FCAT scores well in advance of the springtime examination to improve student outcomes.

Professional Response

The literature base regarding prediction of FCAT achievement is still in its infancy. It is increasingly important that professionals offer educators reliable data on the existing variables potentially impacting performance. Also, given that the FCAT is administered to students late in the third grade, early identification of reading skill deficits and potential for FCAT failure is desirable. Thus, educators are able to identify struggling students and implement interventions in effort to remediate deficits and decrease the likelihood of FCAT failure.

The Florida Center for Reading Research (FCRR) is one of several key agencies established in an effort to provide low-performing schools structure and guidance in completing these objectives in Florida. The “Just Read, Florida!” initiative of the FDOE was implemented to address early-reading instructional strategies, reading assessments, and reading intervention techniques (FCRR, 2004; FDOE, 2002a). FDOE provides assistance with screening and progress monitoring assessments in cooperation with FCRR using measures such as DIBELS (Dynamic Indicators of Basic Early Literacy Skills) (Good and Kaminski, 2002b). The DIBELS
assessment is useful in monitoring student achievement and identifying at-risk students (FCRR, 2004).

However, as mentioned previously, other variables such as student background and performance indicators may explain a portion of the predictive relationship. Research is needed to help elucidate further those relationships and to provide more complete information to educators, parents, and students. If, in fact, student background variables have a significant impact on FCAT scores, the presence of these characteristics may alert teachers to the need for closer monitoring of reading achievement.

Definition of Terms

**DIBELS:** Dynamic Indicators of Basic Early Literacy Skills; assessment battery developed by Good and Kaminski (2002b) to identify emergent reading skills and students who may be at risk for reading difficulty and in need of additional instruction.

**Direct performance assessment:** One-on-one assessment of students’ oral reading fluency and decoding skills.

**FCAT:** Florida Comprehensive Assessment Test

**FCAT-SSS:** Florida Comprehensive Assessment Test-Sunshine State Standards; Florida’s annual high-stakes criterion-referenced achievement test administered to students to evaluate the extent to which students are mastering the basic academic skills according to the Sunshine State Standards.

**High-Stakes Testing:** Any test used to make important decisions about students, educators, or schools based upon the results of that testing.

**Median score:** Score that is the median (or middle) score in a series of scores.

**NWF:** Nonsense Word Fluency; subtest on the DIBELS assessment battery that assesses decoding skill by requiring the student to read aloud novel consonant-vowel-consonant or vowel-consonant nonsense words by blending the phonemes of the words.
ORF: Oral reading fluency; subtest on the DIBELS assessment battery that assesses oral reading fluency by requiring the student to read aloud from a passage that is grade-level appropriate while the student’s speed and accuracy is recorded and used to calculate the fluency score.

Reading FCAT-SSS: Student’s score on the Reading portion of the FCAT-SSS.

Regular education student: Student that is not currently enrolled in any Exceptional Student Education services or classes for Reading.

Risk Category: Category of student risk for reading difficulty.

SES: Socio-economic status; measured by proxy variable, student economic disadvantage—whether or not the student currently receives free or reduced lunch benefits based upon the family’s income.

Sunshine State Standards: Curriculum frameworks in Florida’s public schools for language arts, mathematics, science, social studies, health and physical education.
CHAPTER 2
LITERATURE REVIEW

Introduction

Recent trends in assessing student achievement have led to the implementation of high-stakes testing programs across the nation to ensure that students are meeting standards of performance in academics. The Florida Comprehensive Assessment Test (FCAT) is used in Florida for this purpose. Students’ scores on the FCAT impact individual students and schools at large.

Reading achievement is a major focus on the FCAT given its importance in students’ success in all aspects of education. However, since the FCAT is first administered to students late in the third grade, educators seek alternative assessment methods to evaluate students’ reading skill prior to the FCAT. Assessment of an individual’s oral reading fluency (ORF) has been shown to be an efficient and effective method for predicting student achievement on the FCAT. However, the predictions are not perfect, possibly due to the omission of unique student characteristics, the omission of other important reading skills, or the type of analysis of available data. Therefore, educators may also seek to include such variables to provide a closer estimation of student success on the FCAT.

Reading

Reading Skills

Most researchers consider learning to read the most important task a student needs to accomplish in primary and elementary school (Strickland, 2002). According to the 2000 report of the National Reading Panel, students must acquire skills in five key areas to be successful
readers: phonemic awareness, phonics, fluency, vocabulary, and text comprehension (Armbruster et al., 2001). Each area will be discussed briefly below.

*Phonemic awareness*, a component of phonological awareness, involves a student’s ability to attend to, think about, and manipulate individual sounds (phonemes) in spoken language (Armbruster et al., 2001; Sindelar, Lane, Pullen & Hudson, 2002; Snow et al., 1998). Students with strong phonemic awareness skills are typically skilled readers. Thus, the enhancement of phonemic awareness skills is linked with improved reading achievement (Sindelar et al., 2002).

As the student moves from understanding spoken language to understanding written language, the student acquires skill in *phonics* or the alphabetic principle—understanding the relationship between the letters of the written language (graphemes) and the individual sounds of the spoken language (phonemes) (Armbruster et al., 2001; Goodman, 1993). Students must acquire phonics, or decoding, skills in order to be successful readers (Armbruster et al., 2001; Fox & Hull, 2002; Snow et al., 1998).

To understand the words of the spoken and written language, the student must also develop a sufficient *vocabulary* base (Armbruster et al., 2001; Graves & Watts-Taffe, 2002). Oral vocabulary is the set of words individuals are able to recognize and use in spoken language, whereas reading vocabulary is the set of words individuals are able to recognize and use in written language (Armbruster et al., 2001). Vocabulary may be learned without formal instruction in early years, but is enhanced with a broader exposure to words when the student reaches the classroom setting. A student’s vocabulary base is strongly correlated with comprehension ability, as measured by standardized tests (Snow et al., 1998). In fact, the relationship between vocabulary knowledge and reading comprehension appears to be reciprocal, according to recent research (Baker, Gersten, & Grossen, 2002).

Additionally, the student’s ability to read the text accurately, quickly, and with expression (reading *fluency*) is important (Armbruster et al., 2001; Samuels, 2002). As described by Sindelar and colleagues (2002), fluent reading is “seemingly effortless, except that it requires extensive training to achieve” (p.719). The process of developing fluency begins with several successful attempts to phonologically decode a new printed word (Snow et al., 1998). Once the student is able to quickly and accurately identify words, experience helps to strengthen the
student's automaticity and fluency with word identification (Snow et al., 1998). Some degree of automaticity is needed for prompt comprehension of the printed text to avoid becoming fixed upon the correct pronunciation of isolated words (Sindelar et al., 2002; Snow et al., 1998). Reading fluency skill has been found to be a strong predictor of reading comprehension. Specifically, studies have indicated that the correlations between reading fluency and comprehension range between .70 and .90 (Baker et al., 2002).

Lastly, reading comprehension is the meaning derived from printed text. The ultimate purpose of learning to read is to comprehend, or understand, what is written (Armbruster et al., 2001). Students skilled in comprehension not only understand the literal meaning of the text, but are also able to draw inferences and use comprehension-monitoring and repair strategies when that understanding diminishes (Baker et al., 2002; Snow et al., 1998). Students may be able to retain the overall passage meaning only if a limited number of word identification problems occur while reading the passage.

Researchers recommend that reading instruction in the average Kindergarten through third grade classrooms address all five of these components of reading. As previously mentioned, it is important that by fourth grade students are able to read skillfully, with sufficient comprehension and fluency, to manage new text content and text types (Strickland, 2002). The primary reason for this is that when students reach fourth grade, the focus of instruction shifts from “learning to read” to “reading to learn”.

The National Research Council identified third grade as a key marker for academic success (Snow et al., 1998). The probability that the child will succeed in school and graduate high school is significantly reduced if a child does not learn to read with adequate skill by third grade (Snow et al., 1998). Thus, later occupational and economic outlook can be severely limited (Snow et al., 1998). This can have dramatic effects on a student’s skill in navigating an increasingly technological and literate society (Snow et al., 1998). Hiebert (2002) notes that according to the U.S. Department of Education, federal and state literacy initiatives, such as the No Child Left Behind Act (NCLB Act) of 2001, often specify third grade as the critical grade by which students should be reading grade-appropriate material.
Assessments of Reading Achievement

Traditional assessments of reading achievement are summative in nature in that they are used as assessments of reading skill at “key points in time after students are expected to learn and demonstrate the skills tested” (p.113) (Shinn, Shinn, Hamilton, & Clarke, 2002). Typically used reading assessments include published standardized reading tests, informal teacher-constructed tests, or observation (Madelaine & Wheldall, 1999). Generally, most standardized achievement tests, such as high-stakes testing, are summative assessments of reading achievement. Thus, it becomes important to distinguish between reading achievement (as measured by summative assessments) and reading growth (as measured by formative assessments, to be discussed below).

Typically summative measures are administered so infrequently and contain so few items of each of the reading achievement domains that reading growth is not feasibly measured. In other words, these assessment tools are not sensitive to small gains in reading skill (Shinn et al., 2002). For the same reasons, summative evaluations are not particularly effective in providing teachers with information to ameliorate specific student skill deficits and identify skill strengths given that they are administered too late in the curriculum or academic year (Shinn et al., 2002). By this time, students have often already established poor reading habits with little exposure to possibly effective interventions for remediating reading difficulties. For example, the FCAT reading assessment is not administered until the spring of the third grade, by which time students’ academic success is likely to have been already determined (Snow et al., 1998). Although summative evaluations are important in assessing student outcomes in aggregate clusters, individual students are likely to benefit more from formative evaluations of reading achievement and growth, which later impact scores on the summative evaluations, such as the FCAT.

According to Shinn and colleagues (Shinn, 1995; Shinn et al., 2002), formative evaluations are narrower in assessment focus, occur progressively throughout the student’s skill development, and compliment summative evaluations in providing a complete assessment of the student and the educational framework. Formative evaluations provide a more in-depth and ongoing assessment of a student’s skills to help teachers and parents assess student progress with current instructional methods. Using formative assessments, teachers make data-based decisions
regarding the effectiveness or need for modification of current instructional practices. For example, Stecker and Fuchs (2000) conducted a study comparing matched students, half of whom received instructional modification according to formative assessment data while the matched partners received the pre-determined interventions despite assessment data. Study results indicated that those students for whom modifications were data-based performed significantly better on a subsequent global achievement test.

By initiating formative assessments in the beginning of the academic year, teachers are able to obtain baseline data regarding student achievement. Additionally, teachers are able to obtain, at a glance, information regarding the relative standing of each student in the class with respect to reading skill. Thus, the teacher is able to note those students who need immediate assistance, those who may need additional assistance, and those who may need more challenging material. The identification of these students increases the opportunity for teachers to modify instruction and improve the effectiveness of available instructional time (Shinn & Bamonto, 1998; Shinn et al., 2002).

Reading instruction is most effective if student materials fall within their instructional reading level (Shinn et al., 2002). This is the level at which the reading material is the most difficult material that a student can manage with teacher assistance (Shinn et al., 2002; Walker, 2000). Therefore, the passage is not so difficult that the student is unable to decode or comprehend the text (frustration level), nor so undemanding that the student is able to read and understand without any difficulty (independent level) (Shinn et al., 2002; Walker, 2000).

In the research literature, procedures to determine a student’s instructional reading level are available. For example, Fuchs and Deno (1982) offer a procedure to identify instructional level according to the student’s rate of oral reading. Alternatively, Walker (2000) proposes establishing the instructional level separately for oral and silent reading material using oral reading performance in combination with comprehension checks via teacher-developed comprehension questions. This method provides the teacher with additional clues to determine if the student is experiencing difficulty with print processing (reading) or meaning processing (comprehending).

Good and Kaminski (2002b) have also developed an assessment battery to assist teacher’s in identifying students at-risk for reading difficulty given their present level of skill.
This battery, the Dynamic Indicators of Basic Early Literacy Skills (or DIBELS; Good & Kaminski), is currently available in the sixth edition. DIBELS evaluates students’ skills in the areas of initial word sounds, letter naming fluency, phoneme segmentation fluency, nonsense word fluency, oral reading fluency, and retell fluency to establish reading difficulty risk levels for students in beginning preschool through the end of third grade. The DIBELS assessment system will be discussed in greater detail later.

Shinn and colleagues (2002) argue that despite available research, the placement of students within the appropriate instructional level of the curriculum has been largely ignored. Instead, students are placed according to their grade level, not their achievement level. This may be due, in part, to educators’ efforts to ensure that students meet the “standards”, rather than progress appropriately. However, appropriate placement is critical to student progress (Shin et al., 2002).

High-Stakes Testing

Rise of High-Stakes Tests

As mentioned previously, federal and state governments have created a system of accountability in education and in the assessment of student achievement in recent decades (Hiebert, 2002). According to Shinn and colleagues (2002), governmental involvement has resulted primarily from a “national response to the problems of literacy and other academic concerns” (p. 114). To address those problems and concerns the federal and state governments have proposed the establishment of high academic standards for students and schools, using achievement tests to ensure those standards are being met.

Public Law 105-78 was passed in 1997 whereby Congress asked the National Research Council to review existing state and commercial tests of achievement with the goal of eventually evaluating the feasibility of linking these assessments to one another and to the federally funded National Assessment of Educational Progress (Feuer et al., 1999). In their report to the U.S. Congress and President Clinton’s administration, the Committee on the Equivalency and Linkage of Educational Tests from the National Research Council reported that states and local school
districts have attempted to revise their curricular goals, design customized tests to assess those goals, and to implement test-based accountability systems to realign classroom instruction with those goals. However, Feuer and colleagues noted that without specific guidelines or requirements, states began adopting content-based assessments that differed substantially from one another in content and format.

Currently, state educational testing systems function to communicate student performance results to parents, educators, policy makers, and the public in general (Feuer et al., 1999). These results may be aggregated and reported differently among the different states, depending upon the desired interpretation reference. For example, if test scores are provided in a norm-referenced mode, the results are given in percentile ranks, stanines, normal curve equivalents, and/or grade equivalents to provide the consumer with an indication of how the individual student’s performance compares with the performance of similar students (i.e., same age or grade) in the norming sample. Alternatively, when criterion-referenced, or standards-based results are provided, the information obtained is the individual student’s standing in relation to state academic skill standards and whether students are meeting the criteria or standards designated. Therefore, it is important to understand how state standards are set when one interprets the test results (e.g., aggregation method, norm vs. criterion-referenced).

The “No Child Left Behind” initiative. As mandated in the “No Child Left Behind” (NCLB) initiative, every child is expected to be a skilled reader by the end of the third grade (U.S. Department of Education, 2002). With the implementation of the NCLB Act in 2001, states were required to institute appropriate accountability systems. Although the individual states retain flexibility in choices regarding curricula and assessment plans due to the overall decentralized system of federal governance of educational programs (Feuer, et al., 1999), those schools that consistently do not meet “adequate yearly progress” would be subject to corrective action and restructuring measures. Those that do meet or exceed the state objectives would be eligible for State Academic Achievement Awards (USDOE, 2002). The No Child Left Behind Act was reauthorized by the federal government in 2002 requiring annual assessment to measure schools’ yearly progress (FASP, 2003; U.S. Department of Education, 2002).

Accountability. Assessments of reading are now not only important for classroom teachers and students, but also for administrators, schools, and school districts within a state
Accountability has become the running theme for schools and instructional personnel. Guthrie (2002) notes that the recent attention placed upon accountability emerged from an effort to improve schools and education for American students. Standards-based reforms were intended to improve the nation’s educational framework as well as student and school progress by spurring the reform of general state curricula and materials, encouraging professional development of teachers, and equalizing resources for the schools (Darling-Hammond, 2002). Furthermore, it was believed that the use of these assessments would help bolster the public’s confidence in the ability of the local schools to educate students (Heubert & Hauser, 1999). The consequences and repercussions of acceptable and unacceptable student progress have become so great that assessment of student progress has been coined as “high-stakes” testing.

High-Stakes Testing

Significance. In specific terms, a test is described as “high stakes” when important decisions about students, educators, or schools rest upon the results of that testing (Guthrie, 2002; Heubert & Hauser, 1999). Repercussions of test scores can be centered upon the individual student, the classroom teacher, the school administrator, and the school district. At the student level, common repercussions of high-stakes testing include promotion and retention decisions, tracking, and graduation (Guthrie, 2002; Heubert & Hauser, 1999). Decisions regarding promotion of students to the next grade affect all students and their families. If students are unable to meet the pre-determined criterion-referenced goal for each section of the test, their promotion to the next grade may be hindered. Some schools are able to counterbalance some of the impact of poor test scores with informal classroom measures, grades, and teacher judgments (Guthrie, 2002). Thus, the assessment scores may be used in conjunction with other assessments, paired with other performance tasks, and with portfolio assessments (Feuer et al., 1999).

Nonetheless, Guthrie (2002) reports that one critical and common consequence of poor performance on high-stakes tests is retention. Retention is considered by many school districts to be one of several acceptable academic interventions to improve student performance. In fact, Guthrie (2002) reports that in the 1990s, almost 18% of students aged 6 to 8 were retained. By
the age of 17, approximately 30% of this country’s student population had been retained at least once in their academic career (Guthrie, 2002).

Recent research has shown that retention without additional instructional support and remediation of skill deficits is unlikely to benefit a student and may actually harm a student’s academic progress and likelihood of remaining in school until graduation (Heubert & Hauser, 1999). In a meta-analysis of 20 research studies published in professional publications between the years 1990 through 1999, Shane Jimerson (2001) found that grade retention did not have the desired impact on later academic achievement and emotional adjustment in opposition to the goals of the retention. He further explained that retention itself does not “address the multiple factors influencing the students’ poor achievement or adjustment that resulted in the decision to retain the student” (p. 432). Furthermore, students who are retained, especially minority and low SES students, are significantly more likely to withdraw or “drop-out” of school (FASP, 2006a; Snow et al., 1998). In Florida the financial burden of ineffective grade retention is staggering. Concluding the 2002-2003 academic year, 208,296 students were retained in Pre-Kindergarten through twelfth grade costing the state of Florida over $1 billion (FASP, 2006b). In 2003-2004, 27,713 third graders were retained with a cost of $153 million for educating the students an additional year (FASP, 2006b).

Another consequence of high-stakes test results for individual students arises when schools use those results of student performance to assign students to specific academic programs or “tracks” (Guthrie, 2002; Heubert & Hauser, 1999). In middle school and high school, assignment to academic tracks often equates to the assignment to vocational training or college-preparatory programs. According to Guthrie (2002), once assigned, students’ tracks are rarely changed and can have a significant impact on that student’s future (e.g., future enrollment in post-secondary education and later occupational opportunities). In an attempt to identify districts or states that continue to use high-stakes tests in this manner, many websites were found to oppose the use of these tests in that manner. However, at the time of this review, a specific locale whereby high-stakes tests were used for “tracking” purposes could not be located. It is hoped that this is due to a popular trend against the practice.

In addition to the aforementioned uses of high-stakes testing, the results of these tests are also often used to evaluate student instruction in the schools (Guthrie, 2002). In fact, 43% of the
states in this country reported using high-stakes testing to improve instruction (Guthrie, 2002). As mandated by federal law, if a school’s population consistently performs poorly on high-stakes tests, that school may have to surrender control of educational decision-making and instruction to state departments of education who can then privatize the schools. State departments of education may also decide to implement curricular modifications and subject the personnel to evaluations and approval (Guthrie, 2002). Results of high-stakes tests are also often used in re-allocating school or district budgets (Heubert & Hauser, 1999). Guthrie (2002) asserts that this type of consequence of high-stakes test results result in high levels of publicity and high visibility of successful and unsuccessful schools.

Given the widespread uses of high-stakes test results and the far-reaching consequences of poor test scores, it is not surprising that students, parents, teachers, administrators, and researchers are keenly interested in the effects of using these accountability measures. In their report to Congress and President Clinton’s administration, Heubert and Hauser (1999) recognized that “test use may have negative consequences for individual students even while serving important social or educational policy purposes” (p. 2). In this report, the need for a balance between individual and collective costs and benefits for the various uses of high-stakes tests is highlighted.

Although researchers have noted potentially positive characteristics of testing programs (e.g., means of fostering growth, improving curriculum, providing additional data regarding student strengths and weaknesses, feedback to teachers regarding effectiveness of their teaching methods, and identifying schools in need of support and interventions), educators, parents and researchers are wary of the potential negative characteristics of using these testing programs in high-stakes decision making. First, Richard Allington (2002) warns the public regarding the use of legislative mandates to establish grade-level achievement standards for promotion to the next grade. He argues that the inherent error in doing so lies in the establishment of the “grade-level” standard. According to Allington, grade-level standards are psychometrically-based upon the average achievement of students in a given grade. Thus, “with grade level defined as the average performance, getting everyone on or above grade level means getting everyone on or above average—an immediate mathematical impossibility” (p. 261). Moreover, a problem emerges when requiring students to meet grade level standards for promotion decisions. According to
Allington, this requires the immediate improvement of 50% of the students when taking the next state standards-based exam.

Allington (2002) recommended shifting from the requirement that students meet an absolute standard to a requirement that students demonstrate skill improvement, relative to their standing in previous test administrations. He cited research that demonstrates that reading achievement among fourth-graders has improved consistently since 1990 in the U.S. and that when using recent international assessments of reading proficiency, in aggregate, U.S. fourth-graders ranked among the highest students in the world. Nonetheless, according to Allington (2002), the specific standards and the interpretation of adequate reading proficiency (i.e., whether a student is “on grade-level”) will influence how reading difficulties are perceived.

Guthrie (2002) notes another criticism of using high-stakes tests—teachers and students are generally unhappy with the use of test results. Many teachers argue that they and their students have associated negative feelings with the test, especially if results are poor. For students, Guthrie cites research that low-performing students often feel negatively toward the tests and themselves, which increases their level of resentment and anxiety and subsequently decreases their motivation to do well.

Furthermore, teachers report spending a lot of instructional time on test preparation and administration, sometimes reaching 100 instructional hours. Due to this intensive test preparation, it is argued that the curriculum is often narrowed to accommodate the specific standards of the test, thus eliminating other curricular activities that may foster reading skill, creativity, imagination and critical thinking, especially in the reading curriculum.

On the one hand, it is important that the test content matches the instructional content for the valid use of high-stakes test results in decision-making. This is referred to as “instructional validity” (Heubert & Hauser, 1999). Also, students need to be given an opportunity to experience practice with different types of test items. Students should also be provided with explicit test-taking skill instruction to ensure that low scores are not a result of a lack of “test-wiseness” on the part of the students.

Although, test results become invalidated when teachers “teach-to-the-test” whereby objectives are taught so narrowly that the items on the assessment no longer represent the domain at large (Heubert & Hauser, 1999). By doing so, Heubert and Hauser contend that the
scores are artificially raised “without actually improving the broader set of academic skills that the test is intended to measure”. However, it is impossible, given time and other classroom constraints, for a test to include items that adequately represent all of the processes and concepts of a large subject area such as reading (Heubert & Hauser, 1999; Shinn et al., 2002).

Shinn and colleagues (2002) address another criticism of high-stakes tests. They argue that high-stakes tests are incapable of informing teaching, contrary to one of the primary arguments used to promote the use of high-stakes tests. According to Shinn and colleagues, assessment that informs teaching is assessment that assists in the identification of at-risk students, helps teachers plan instructional content and format, and complements ongoing formative evaluations. Shinn and colleagues are concerned that high-stakes tests have yet to show “a significant and positive impact on the instructional practices of teachers, and consequently, the problem of illiteracy” (p. 116). According to them, a national focus on “accountability” rather than student growth will ensure that the disadvantages of using high-stakes tests will outweigh the advantages (namely, the potential for student achievement). But as noted by many, evaluations of student progress and school achievement are often two very different and sometimes opposing goals (North Carolina School Psychology Association, 2001).

Two potential risks of using state assessments for students with disabilities and foreign language learners are evident in many school districts. School accountability ratings often exclude the scores of students with disabilities or foreign language learners (FASP, 2003). Therefore, school administrators may find themselves encouraging the assessment and placement of students into special education classrooms when they demonstrate low achievement in the regular education classroom. This “push” is understandable given the rewards and punishments given to schools based upon high-stakes test results; however, it is unethical and illegal given federal mandates for placing students in the “least restrictive environment” (IDEA, 1990, 1997). Furthermore, placement in special education has its own set of negative consequences for the individual students, including the restrictions placed on students with special diplomas (certificates of completion of special education) in post-secondary school (FDOE, n.d.b.; Schwartz & Schwartz, 2003). These restrictions on special diplomas can affect later occupational opportunities (e.g., military service).
Nonetheless, Braden (1997) warns, “Schools that are committed to a single, uniform standard of success for all children may find themselves doing some unexpected (and unwanted) things to deliver on that commitment” (p. 245). Coffman (as cited in Braden, 1997) argues that holding students to uniform standards encourages a narrowing of educational experiences for most students, doom many students to failure, and limit the development of many talented students. Thus, Braden (1997) argues that the following two values are incompatible: 1) All children can and must achieve a common, high standard, and 2) Individual differences in student abilities and performance should be accommodated.

In contrast to the aforementioned criticisms of high-stakes tests, Feuer and colleagues (1999) argue that when a test is used in a “low-stakes” manner (i.e., no serious consequences may result from the scores), teachers and students may feel less motivated to dedicate time to the specific sample of content that commonly appears on those tests. Some argue that high-stakes testing conditions often motivate teachers and students alike to perform their best (Shinn et al., 2002). When the test is considered high-stakes, teachers and students are compelled to focus upon the specific types of items commonly occurring on the test. That said, administrators tend to favor high-stakes testing, believing that results can be used directly to modify instructional content and practices, aid in planning, and evaluate students and programs.

Remediating problems of high-stakes testing. Given the criticisms of high-stakes testing programs, it is still an admirable goal to use statewide tests to identify and assist struggling schools, and perhaps struggling students. Former President of the Wyoming School Psychology Association, Charles Smith (2002), questioned how one could criticize high standards without endorsing mediocrity.

In their investigation into the problems with their own state-adopted high stakes test, the North Carolina School Psychology Association (2001) publicized areas for improvement in use of the test. Later collaborating with 13 other interested groups, the Coalition for Fair Testing was founded (Breckheimer, 2003). Together the Coalition and some North Carolina legislators were successful in lobbying for a legislative bill supporting fair testing. The bill addressed and attempted to ameliorate many of the problems found at that time in the North Carolina high-stakes test at that time. For example, student promotion would be determined not only by test scores but also by classwork, grades, and teacher evaluation of student progress. Additionally,
the bill promoted the early identification of students that may be at-risk for academic difficulty, the implementation of academic interventions for students performing below grade level at any time, limitations on time spent in test preparation, a retention appeals process, and an external evaluation of the testing program. Achieving some of these goals and continuing to work on others, the Coalition was able to improve the use of high-stakes tests in North Carolina. Given the entrenchment of high-stakes testing in federal and state law, it is imperative that educators overcome the limitations of high-stakes tests and use these assessments in the most ethical manner.

**Best practices and high-stakes testing.** Heubert and Hauser (1999) presented a “best practices” approach in using state-adopted content tests as high-stakes tests as approved by the national Committee on Appropriate Test Use of the Board on Testing and Assessment. In a collaborative effort, this committee outlined guidelines that should be respected when using tests for high-stakes decisions. First, schools that use students’ scores on these measures for student decisions, classroom instruction, curricular modifications, or school accountability should be able to demonstrate the appropriate alignment of the curriculum, instruction and the test used. Second, educators, administrators, and policy-makers are reminded that these tests are not perfect and that no single test score should be used as a “definitive measure of a student’s knowledge”. Given the above, important educational decisions should not be made using a single test score, but should incorporate other informative measures of the student’s knowledge and progress. If it is decided that a student should be retained, remedial and instructional support should be added to the student’s academic program (Heubert and Hauser, 1999). Heubert and Hauser recommend that students be given the opportunity to demonstrate their skill mastery with multiple administrations of practice tests that familiarize the student with the respective testing format and the use of effective test-taking strategies.

Furthermore, while tests can assist educators and administrators in spotting differences among students that could be addressed by the school system, it is important to recognize that lower scores are common among certain groups such as racial minorities and low-income families (Heubert and Hauser, 1999). Research has documented that African-American children, Hispanic children, and children in urban school settings are at a greater risk of having difficulties learning to read than the middle-class, European-American suburban children (Snow et al.,
1998). Heubert and Hauser argue (1999) that these differences mirror the persistent inequalities present in the schools and should not be viewed as permanent truths about the respective groups of students. In fact, recent research points to poverty as one of the strongest predictors of low academic achievement and low reading and mathematics scores on Florida’s FCAT (FASP, 2003; Park, Turnbull & Turnbull, 2002).

The American Psychological Association, the American Educational Research Association, and the National Council on Measurement in Education collaborated to produce the Standards for Educational and Psychological Testing, or Standards, (FASP, 2003). In addition to those guidelines detailed by the Committee of Appropriate Test Use listed above, the Standards add the following number of principles to promote fairness. First, it is important that when authorities mandate the use of high-stakes tests that the intended use of those results are clearly detailed. When that test will be used for multiple purposes, such as school accountability and student promotion, the technical validity of the test (including standard error of measurement statistics) for each of those purposes should also be made available. School and state administrators should be able to verify that the content of the assessment is that which the students are given sufficient opportunity to learn. Additionally, when using test scores for decisions regarding students’ academic programs and progress, administrators and educators should be able to provide empirical support for the relationship between test scores, instructional programs, and other intended student outcomes. Lastly, if tests are to be used as determinants of student promotion or graduation, students should be given multiple opportunities to express content mastery through equivalent methods, if necessary (FASP, 2003).

In the Code of Fair Testing Practices in Education, the Joint Committee on Testing Practices offered a further addition to the above principles. This addition is similar to Allington’s (2002) caution regarding the use of grade-level standards, described previously. The committee determined that authorities should adhere to reasonable procedures for setting passing scores on the high-stakes tests. The procedures used should be made available to the consumer and evidence supporting the appropriateness of the scores should be offered.

The above guidelines for the appropriate and ethical use of tests are supported by state professional agencies such as the FASP (2003) and the North Carolina School Psychology Association (2001). These associations serve as state professional organizations for school
psychologists who are in the position and profession of improving the educational needs and mental health of students. It is the responsibility of the educators and the administrators to ensure that high-stakes tests are used in the appropriate manner according to the best practices and recommendations outlined above.

**High-Stakes Assessments of Reading**

In their review of many state-adopted assessments of reading and the available research in the area, Feuer and colleagues (1999) have identified four dimensions commonly emphasized in high-stakes reading tests: word recognition, passage comprehension, vocabulary, and reading inquiry. *Word recognition* is defined as the complex cognitive, linguistic, and neurological process of learning to recognize words in print. *Passage comprehension* refers to the individual’s ability to understand the main idea of a passage using summarization skills, background knowledge, self-monitoring procedures, knowledge of word meanings, and ability to identify causal relationships (Feuer et al., 1999; Shinn et al., 2002). In high-stakes assessment of reading, passage comprehension is commonly measured via free recall, multiple-choice, and short-answer items across many text genres, including narrative text, expository text, and poetry. The *vocabulary* dimension essentially assesses one’s word knowledge via identification of synonyms, antonyms, or definitions. The last dimension emphasized in common reading tests is *reading inquiry*, or the use of cognitive strategies to judge relevance, locate information, and build a collective network of knowledge from a variety of text passages. Although related to passage comprehension, the skills involved in reading inquiry are deemed distinctive from comprehension. Given the broad domain of these reading skills, tests that measure reading ability often differ according to the purpose of the assessment, the age of the student assessed, and the test publisher/consumer’s belief about the reading process.

Success on these high-stakes reading tests is often determined not only by reading ability (accounting for approximately 40% of the difference between students), but also by error (15%), format knowledge (10%), motivation (15%), and content knowledge (20%) (Guthrie, 2002). Other correlates of success on these reading tests include teacher portfolio assessments of progress, general intelligence level, background knowledge of text content, and test-taking
strategies (Guthrie, 2002). Guthrie (2002) points to knowledge of these correlates as useful keys in constructing an approach to preparing students for high-stakes reading tests.

As mentioned previously, the Florida Comprehensive Assessment Test (FCAT) is Florida’s high-stakes assessment of student achievement. The background and general information regarding the FCAT is detailed in the following section. The technical adequacy of the FCAT is reviewed and recent FCAT data is provided.

**Florida Comprehensive Assessment Test (FCAT)**

*Background*

The Florida Comprehensive Assessment Tests, or FCAT, is currently used in Florida as an annual high-stakes measure of student achievement. The FCAT was developed as an integral part of educational reform in the state of Florida in an effort to raise standards for education, improve student education, and increase school accountability (Florida Department of Education [FDOE], 2001b, 2004a). The FDOE’s Assessment and Accounting Briefing Book (2004a) describes how the assessment program that now encompasses the FCAT was developed in 1972 with the first string of measures assessing students’ minimum competency skills under the “minimum competency testing program”. In 1976, the Florida Legislature approved an “accountability act” whereby students in third, fifth, eighth, and eleventh grades were assessed to ensure mastery of basic skills. By 1977, Florida enacted the nation’s first high school competency (graduation) test. Although challenged in courts, Florida’s use of the high school competency test as a graduation test was upheld.

In 1995, the Florida Commission on Education Reform and Accountability convened to devise an assessment program to measure student learning, raise educational standards, and help students “compete for jobs in the global marketplace” (FDOE, 2001b). Adopting the recommendations of the Commission, the State Board of Education approved the new “Comprehensive Assessment Design” in 1995 that detailed the development of statewide assessments to measure student performance in four broad areas—reading, mathematics, writing, and creative/critical thinking. According to the Comprehensive Assessment Design, standards
for educational content to guide instruction would need to be specified. Thus, the *Sunshine State Standards* (the curriculum frameworks for Florida for language arts, mathematics, science, social studies, health and physical education) were developed and subsequently adopted by the State Board of Education in order to provide an “equitable system of student assessment and school accountability” (FDOE, 2001b, 2002c, 2003c).

Thus, the FCAT was developed from the Comprehensive Assessment Design criteria to assess educational mastery in the areas defined by the *Sunshine State Standards* (FDOE, 2001b). According to the *FCAT Briefing Book* (2001b), the “FCAT measures the content specified within the strands, standards, and benchmarks of the *Sunshine State Standards* and does so in the context of real-world applications.” (p. 18). The “strand” is the category of knowledge being tested on the FCAT reading test (i.e., Reading and Literature). “Standards” are the general statements of expected student achievement within a strand and are the same for all grade levels. Narrowed further, the “benchmarks” are the specific objective statements of expected student achievement under each standard. The resulting FCAT-Sunshine State Standards (FCAT-SSS) was developed as a criterion-referenced measure (FDOE, 2001b, 2003c). FCAT items were developed for the state of Florida through contract with a private company, CTB/McGraw Hill. It was initially planned to be administered to four grade levels to assess achievement in elementary, middle, and high school, and was field-tested in 1997 as such. Subsequently, in 1998, the FCAT reading and mathematics sections were administered to students in grades 4, 5, 8, and 10. Although not used for accountability purposes at that time, the results were released to educators and parents.

In 1999, the FCAT was re-administered to students in grades 4, 5, 8, and 10 and “school accountability for student performance” began with the release of the 1999 results (FDOE, 2001b). Schools were assigned accountability grades according to student performance on the FCAT that year. Governor Jeb Bush’s A+ Plan of 1999, authorized by the Florida Legislature by amending Section 229.57 of the Florida Statutes, extended the FCAT testing to grades 3 through 10. The extended portions of the FCAT were then developed by the Harcourt Educational Measurement Company and included a norm-referenced component to the FCAT (FCAT-NRT), allowing the comparison of student scores to national norms. This became possible because the FCAT includes two subtests from the Stanford 9 test—Reading Comprehension and
In February and March 2001, students in grades 3 through 10 were assessed on the FCAT (FDOE, 2003b). In August of the same year, the State Board of Education designated the scores that would be considered passing, with passing scores becoming one of the requirements for high school graduation. Scaled scores range from 100 to 500 and as of February 1, 2002, scores of 300 and above on the reading and mathematics portions of the FCAT were established as “passing” scores for high school graduation. Otherwise the scores were reported in terms of achievement levels or developmental scores. Currently, the FCAT-SSS reading and mathematics portions are administered to students in grades 3 through 10 (FDOE, 2004a).

Based upon the aggregate student scores on the FCAT, each Florida school is assigned a “grade”. The grading system of schools is the state’s accountability and publicity measure to promote school improvement (Jefferson, 2003). School grades range from “A” (making excellent progress) to “F” (failing to make adequate progress) as detailed in Rule 6A-1.09981 of Florida Statutes, Section 229.57 (FDOE, 2001b). For example, for a school to earn an “A” grade, it must demonstrate that 95% of the standard curriculum students were assessed and that 90% of those students earned scores of Level 3 (minimum passing score) or above on the FCAT reading test (FDOE, 2001b). The complete school grading protocol is available in the *FCAT Briefing Book* (FDOE, 2001b; 2004a).

According to FDOE (2001b), “school grades are directly linked to accountability sanctions and rewards” (p. 11). Schools that receive “A” grades or that improve by a letter grade from one year to the next are eligible for monetary incentives under the “Florida School Recognition Program”. Lower performing schools are subject to potential external control by FDOE and/or restricted access to additional financial bonuses (FDOE, 2001b). Parents of students that attend schools receiving a failing grade for two out of four consecutive years are given options for “Opportunity Scholarships”, or voucher, whereby the students are permitted to attend a higher performing public or private school at the public expense (FDOE, 2001b). FDOE also notes that no statistical tests were used in developing the criteria for the school grades (e.g., cluster analysis). The impacts of this school grading system are currently being debated among educators, administrators, policy-makers, and interested parties throughout the state.
With respect to individual students, FCAT scores are commonly used to determine promotion eligibility, especially at the end of third grade, unless just cause can be demonstrated that a student should or should not be promoted (FDOE, 2001b). FDOE (2001b) states that grade promotion and retention decisions rest with the local school boards’ policies. Local school boards establish and abide by their own rules for promotion and retention as detailed in the respective pupil progression plan (FDOE 2001b, 2004a). FDOE policy states that if the test is used to determine promotion eligibility for a student, due process procedures must be provided that included the opportunity for the student to retake the test if a failing score is obtained in the initial administration (FDOE 2001b, 2004a). However, the educational plans of any student earning a Level 1 score (this type of score to be discussed further below) on the FCAT must be changed by the appropriate school personnel (FDOE, 2004a). Students with disabilities and students designated limited English proficiency (LEP) are assessed, generally with accommodations, if they seek a standard high school diploma (as opposed to a special diploma) (FDOE 2001b, 2004a). Past tenth grade, students are permitted six additional opportunities to pass the FCAT and graduate if needed, retaking only the section(s) needed (Reading and/or Math) (FDOE, 2004a).

*FCAT Scores*

Scores on the FCAT-SSS are reported in terms of scaled scores (range 100-500) and achievement levels. Students can score between 1 (lowest achievement score) and 5 (highest achievement score) (FDOE, 2001b, 2004a). However, the cutoff scores for each level vary according to grade and the mean and standard deviation of scores varies with each administration (FDOE, 2002c, 2006). Because it is difficult to determine student growth year-to-year using standard or scale scores, developmental scores are also provided (FDOE, n.d.a., 2004a). This is a “value-added assessment system” recently added in 2002 to the FCAT score reporting procedures. Developmental scores range from 0 to 3000, which allows the tracking of an individual student’s achievement progress and growth over time (FDOE, n.d.a., 2004a). As student achievement improves (as measured by FCAT scores), the developmental scores rise. Educators and parents are able to monitor academic progress each year in reading and mathematics with the use of the developmental scores (FDOE, n.d.a., 2004a). Lastly, there does
not appear to be any discussion to the use of developmental score growth and the determination of “adequate yearly progress” of Florida schools or the determination of adequate student growth for individual students if their absolute scores remain below designated **Sunshine State Standards**.

**General Information**

In 2001, the FDOE (2001b) reported the following costs for the administration of the FCAT. When tallying all grades and tests, the cost is approximately $4.64 per student, which is less than .25% (one-fourth of one percent) of the state’s educational budget for grades K-12. The total cost estimated to develop, administer, score and report all components of the FCAT are approximately $27 million per year (FDOE, n.d.b). With respect to investments of time, the FDOE (2001b) reported that the administration of the FCAT takes approximately eight hours over four days, which FDOE equates to less than one percent of the instructional time during the academic year (not including test preparation hours that are commonly included in teacher planning). In 2004, the published costs of administration appeared to have increased. FDOE (2004a) reported that the cost per student equaled $16.57 per student, or less than .33% (one-third of one percent of the state’s educational budget for grades K-12.

Additionally, some questions were raised by the general public regarding possible score adjustments for low socioeconomic status (SES) of schools or students. However, the FDOE asserted in the *FCAT Briefing Book* (2001b) and in the *Assessment and Accountability Briefing Book* (2004a) that it is the responsibility of all schools to teach all students regardless of SES. It is assumed that “all students are capable of making learning progress, and all schools are held to equally challenging performance standards” (FDOE, 2004a, p. 10). As such, the FCAT scoring criteria did not include double standards for these students and all students would be held to the same “challenging performance standards” (FDOE, 2001b).
Technical Adequacy

Reports of technical adequacy of the FCAT were limited in accessible FDOE published reports (e.g., FDOE, 2001b, 2004a). However, the following validity and reliability estimates were reported. Regarding reliability, FDOE reports that test-retest reliability is considered a “seldom calculated” index of reliability for measures such as the FCAT (FDOE, 2001b). Thus, the parallel forms reliability was reported instead (above .90 for grades 4, 5, 8, and 10) (FDOE, 2001b). The exact correlation statistics were not reported and the reliability estimates for the remaining grades was not published. In 2004, FDOE made some of these statistics more readily available in the Assessment and Accountability Briefing Book (2004a).

Regarding validity, the FCAT Briefing Book (FDOE, 2001b) reports that “there are various types of test score validity, but not all of them apply to the FCAT” (p. 9). However, according to the FCAT Briefing Book, the FCAT possesses “strong” content and concurrent validity. In the field of measurement and test construction, content validity is defined by Smith and Glass (1987) as “the overlap of the curriculum and objectives of the instructional program with the content of the test items and the level of their difficulty” (p. 107). The FCAT reportedly has content validity because it was developed by the Department with the assistance of commercial testing companies and was “validated” by committees of practicing teachers and curriculum specialists.

Regarding the concurrent validity of the FCAT, scores from the criterion-referenced component (FCAT-SSS) and the norm-referenced component (FCAT-NRT, or SAT-9) were correlated for grades 4, 5, 8, and 10 (FDOE, 2001b). The correlations between the two measures ranged from .70 to .81. Therefore, FDOE reports that the FCAT has adequate concurrent validity. Data regarding other indices of reliability and validity such as split-half reliability, test-retest reliability, construct validity, discriminant validity, and face validity were not readily available—the interested reader was referred to technical reports on the FCAT but was not directed in where to find this information. This researcher found it difficult to locate the same. This is a critical omission for the general public considering the use of the FCAT for high-stakes decision-making in the state of Florida. Nonetheless, the FCAT and the associated goals of the state assessment program are not likely to be abandoned in the near future. Thus, it remains important for educators to continue to seek student and school growth each year on the FCAT.
According to the Reading Test Item and Performance Task Specifications (FDOE, 2001a) reading achievement forms “the basis for assessing the benchmarks as identified in the Sunshine State Standards” and that “the purpose of the FCAT Reading test is to measure students’ achievement in constructing meaning from a wide variety of texts” (p. 3). These texts are either informational (subject-matter centered) or literary (fiction, nonfiction, poetry, or drama) texts. Scores on these types of passages form the basis for the two reading subscores on the FCAT—Constructs Meaning from Informational Text, and Constructs Meaning from Literature. In third grade, the proportion of informational text passages and literary text passages is 40% and 60%, respectively.

The Specifications (FDOE, 2001a) state that the reading material selected for use on the FCAT should be retrieved from sources such as basal reading series, popular children’s literature, and magazines with recognizable key concepts embodying all the elements of quality writing. Some passages may be created for the purpose of the FCAT as long as they possess the same qualities—cohesive, logically arranged, and stylistically consistent. Included passages are intended to be grade-level appropriate with respect to interest, appeal, concepts, and timeliness. The passages are intended to be understood without background knowledge preparation. Multicultural passages are considered desirable inclusions as long as the questions asked would not present a bias toward any group of students. It is specifically noted that passages should be free from bias or stereotyping (e.g., only male doctors), and offensive or emotionally charged content. Therefore, passage topics are restricted and listed in the appendix of the Specifications.

The Specifications also addresses the level of vocabulary and passage difficulty permitted on the FCAT Reading subtest (FDOE, 2001a). Regarding vocabulary, it is specified that only grade-level appropriate passages and vocabulary are appropriate for inclusion. However, the passage may be modified to revise a word or phrase that is above grade level. Grade-level appropriateness is determined by cross-referencing the vocabulary with accepted, published word lists. Nonetheless, more difficult words may still be included if the reviewers determine that the meaning of the word is clear from the surrounding text. Passage length guidelines are also provided. For example third grade passages should range from 100 to 700 words per passage with an average of 350 words.
The cognitive level of appropriate test items and performance tasks is also outlined in the Specifications (FDOE, 2001a). To evaluate the cognitive level, the B. S. Bloom taxonomy of levels of thinking were utilized to construct a two-level classification system. Level I cognitive factors include knowledge, comprehension, and application (only inasmuch that the application skill required in the test situation is similar to the learning situation). Level II cognitive factors include application (in novel situations), analysis, synthesis, and evaluation. For third grade it is expected that 60% of the test items will be from Level I factors and 40% will be from Level II factors. All items are multiple-choice, as performance tasks are not used in third grade testing. It is not specified whether inter-rater reliability estimates were employed in assigning items as Level I or Level II.

The content to be tested in the FCAT Reading test is that which is specified within the strands, standards, and benchmarks of the Sunshine State Standards (FDOE, 2001b, 2003c). Test items reflect the assessment of benchmarks. A benchmark clarification statement is provided with each benchmark describing how students should be able to demonstrate mastery of that benchmark. Students in third grade are administered 50 to 55 multiple choice questions for the FCAT-SSS and 50 multiple-choice questions for the FCAT-NRT.

The FDOE (2001b) provided the following information regarding specific FCAT item development. According to the FCAT Briefing Book, the respective test development contractors employed professional item writers to draft the initial FCAT test items. FDOE staff, in collaboration with committees of Florida classroom teachers and curriculum supervisors, reviewed and revised the test items. The items were deemed acceptable for each grade level if they corresponded with the Sunshine State Standards Grade Level Expectations (FDOE, 2003c) using the Reading Test Item and Performance Task Specifications as a guideline (FDOE, 2001a). The Specifications provided item-writers with guidelines for the “cognitive levels, style, format, and scope for all FCAT test items and performance tasks” (p. 2) (FDOE, 2001a). Prior to approval, each test item was reviewed for community sensitivity and possible bias by a variety of groups and organizations (FDOE, 2001b). The selection and specifics of these groups were not provided in the Specifications.
FCAT Reading Data

In the FDOE’s report *Lessons Learned* (2002c), FCAT data from the years of 1998, 1999, and 2000 for students in grades 4, 8, and 10 were analyzed to determine the extent to which students are learning and demonstrating the content of the *Standards* and to assist educators in improving the teaching and learning of the *Standards*. This report contains an analysis of the data from standard curriculum students only (not ESE or LEP) since their scores provide the basis of the statistical calibration and equating procedures performed annually. From the analysis of this data, the following observations and implications for instruction were reported.

First, overall FCAT Reading achievement (i.e., reading scores) rose slightly across the three years in aggregate, but the annual scores did not differ substantially (e.g., mean FCAT Reading scores for fourth graders were 294, 296, and 305 for years 1998, 1999, and 2000, respectively). Although the means remained similar, the proportion of fourth grade students that earned a Level 3 or above increased from 50% in 1998 to 59% in 2000.

In this analysis, items were clustered by relevant *Standards* benchmarks. These clusters include: words and phrases in context; main idea; plot and purpose; comparisons and cause and effect; and reference and research. Based upon the analysis of specific item clusters, *Lessons Learned* offered the following suggestions for improving instructional practices and increasing student achievement (FDOE, 2002b, 2002c). Teachers should teach the full scope of the benchmark incorporating skills from basic decoding and comprehension to higher-level cognitive reading skills. Teachers are advised to emphasize reading skills in all content areas and to highlight vocabulary acquisition. Students should be taught text comprehension strategies and skills in determining important text content, especially with informational or expository text. Teachers are advised to incorporate a wide variety of reading materials in the classrooms incorporating different lengths, genres, cultures, ability levels, and interests. Students should be provided the opportunity to demonstrate knowledge not only with multiple-choice test items, but also short and extended response items in all content areas. Students should also be taught to determine author’s intent, validity of information, faulty arguments, or potential author bias. Lastly, teachers are encouraged to utilize a full range of reading assessments to measure student progress and guide instruction. Given that reading fluency has a strong impact on reading
comprehension, it is deemed important to assess students’ general reading ability with fluency measures and then address the more complex cognitive activity of comprehension.

Data regarding FCAT Reading scores for each academic year are published on the FDOE website (FDOE, 2003b). Given that third grade is the first year students participate in the FCAT and that third grade students’ reading skills should be moderately developed to ensure future academic success (discussed in greater detail below) the FCAT Reading scores for third graders from the 2003 administration of the FCAT were explored (FDOE, 2003b). According to the FDOE (2003b), 188,107 students were administered the FCAT Reading test in 2003. The mean scale score (range 100-500) was 298. Thirty-eight percent of these students performed below the Level 3 achievement level (the minimum acceptable level of performance). Specifically, 23% performed at Level 1 and 15% performed at Level 2. The remainder of the students achieved adequate passing scores: 33% at Level 3, 25% at Level 4, and 5% at Level 5. These published proportions equal 101%, presumably due to rounding error. Statistics for each county are also provided on the website. Consequently, in 2003, approximately two out of five students (38%) in the state did not pass the minimum achievement level designated by FDOE. In 2004, 35% of Florida’s students failed the Reading FCAT-SSS and in 2005, 33% (or one-third) of the students failed the Reading FCAT-SSS (FDOE, 2004b, 2005a).

According to FDOE statistics (2003a) on FCAT-NRT Reading scores, Florida’s third grade students’ reading achievement increased in relative standing with the national average from 2000 to 2003. Specifically, Florida’s third graders scored at the 49th percentile in 2000, 56th percentile in 2001, 57th percentile in 2002, and 61st percentile in 2003. A similar trend was noted for fourth, fifth, sixth, and seventh graders on the FCAT-NRT Reading test. In the following sections, specific reading skills believed to significantly impact student performance on the FCAT Reading test are discussed.
Oral Reading Fluency

Background

Researchers Deno and Fuchs sought to develop a standardized reading assessment that related to other important measures of reading proficiency, could be implemented by teachers accurately, could effectively monitor student growth, could be used in the busy classroom setting, and could be used to enhance instructional programs to promote student progress (Fuchs, 1995). They found that reading connected text (versus a list of words) aloud was a critical indicator of general reading skill (Fuchs, 1995).

ORF is a measure of a student’s reading fluency when reading aloud determined by the student’s speed and accuracy (Shinn, 1989a). Shinn (1989a) detailed the standardized administration of the ORF procedure. To determine a student’s ORF, the student is assessed individually using two copies of a passage (numbered for the examiner to ease scoring, and un-numbered for the student). The student is instructed to read aloud from the beginning of the passage for one minute. When the assessment begins, the examiner tracks the number and type of oral reading errors. When one minute elapses (exactly), the teacher calculates the number of words read correctly and incorrectly and graphs the resulting scores. To increase technical adequacy of the assessment tool, it is recommended that three passages are administered and that the median score is used. Marston (1989) notes that this type of assessment has the advantage of giving the examiner the opportunity to observe the student’s mastery of the qualitative features of reading as well.

Advantages

First and foremost, researchers have found that when teachers use ORF data to designate student achievement goals, monitor the effects of the instructional program, and adjust interventions when little progress is made, student achievement improves (Shinn, 1995; Shinn et al., 2002). Although student achievement is most important, the use of ORF procedures has other advantages as well. L. S. Fuchs, D. Fuchs, and Hamlett (1990) outlined how ORF data can be used to guide instruction development. ORF data helps the classroom teacher evaluate the “adequacy of student progress” or the need to modify existing instruction or to contrast the
effectiveness of different reading interventions. Next, ORF procedures are ideal formative assessments for classroom teachers given their feasibility (Shinn et al., 2002). This assessment approach requires limited time investments, is easy to administer, score, and graph (especially with recently appearing computer software) and is an accurate measure of student skill (Shinn et al., 2002). Thirdly, ORF assessment has a high “testing-teaching” overlap and the ability to measure small changes in progress over short periods of time, a lack of which is a criticism of many standardized tests.

Additionally, the ORF procedure uses a response format that closely mirrors the skill assessed, a primary component of its content validity (Marston, 1989). Also, assessing a student’s ORF highlights that student’s specific strengths and weaknesses that can be addressed in future instruction. Thus, this assessment fits nicely into a problem-solving model or approach to improving student achievement (Knutson & Shinn, 1991; Shinn, 1995; Shinn & Bamonto, 1998; Shinn, Collins & Gallagher, 1998). Furthermore, because administration per child is limited, frequent administration and progress monitoring is possible. Thus, the ability of teachers to obtain repeated measurements of student skill allows student growth to be assessed. The capability for repeated measurement is important because it increases the reliability in observation of student skill, allows educators to identify trends in student progress, and allows quicker reaction when students begin to have difficulty (Marston & Tindal, 1995).

**Technical Adequacy**

Research has shown that ORF assessment of reading skill embodies strong technical adequacy (reliability, validity, and norms), thus increasing its usefulness in decision-making regarding individual student progress (Deno, 1985, 1989; Fuchs, Fuchs, & Maxwell, 1988; Good & Jefferson, 1998; Marston, 1989). Marston (1989) summarizes the research supporting the technical adequacy of ORF assessment. With respect to ORF scores, reliability measures were consistently high. Test-retest reliability estimates ranged from .82 to .97, with most of the estimates above .90. The test-retest intervals across studies ranged from 1 to 10 weeks. Using parallel forms to assess reliability, again the estimates across studies were strong (.84 to .96, with most above .90). Inter-rater reliability was also measured for ORF procedures and found to be .99 (Tindal, Marston, & Deno, 1983 as cited in Marston, 1989).
Studies of ORF validity were also found to be adequate to strong across the literature (Fuchs et al., 1988; Marston, 1989). Marston (1989) reviewed studies of criterion validity, which were determined by correlating generally accepted, published norm-referenced tests of reading achievement, such as the Stanford Diagnostic Reading Test and the Woodcock Reading Mastery Test (which measures different aspects of reading at different age levels) with ORF scores. Across studies, the correlation coefficients ranged between .63 and .91, with most above .80. Fuchs and colleagues (1988) compared the validity of ORF to three other informal measures of reading comprehension: question answering tests (most commonly used whereby students answer reading comprehension questions from the basal reading series), recall procedures (students read a passage and retell the passage in their own words), and cloze technique (students supply the missing words in a passage whereby every $n^{th}$ word is deleted). With respect to criterion validity, the oral passage reading test was correlated significantly higher with the Reading Comprehension subtest of the Stanford Achievement Test (.89) than the other three informal reading measures (.74, .66 to .82, and .66 to .79, respectively). These results support earlier findings by Deno (1985) in which ORF was strongly correlated with the Literal and Inferential subtests of the Stanford Achievement Tests (.78 and .80, respectively). To further clarify the significant relationship between oral reading fluency and reading comprehension, Kranzler and colleagues (Kranzler, Brownell, & Miller, 1998) utilized simultaneous multiple regression analyses and found that this relationship cannot be explained by one’s general cognitive ability, processing speed, and efficiency.

Regarding the criterion validity studies of ORF procedures, Marston (1989) reported that criterion validity studies assessing correlations between ORF and four basal reading series’ criterion-referenced mastery tests supported the use of ORF with correlations ranged from .57 to .86. This provided more support for the use of ORF assessments as a valid predictor of global reading proficiency (Marston, 1989). Conducting a more recent review of the criterion validity of ORF, Good and Jefferson (1998) used data that was disaggregated by grade, as opposed to data that was collapsed across grades commonly found in studies reviewed by Marston. Analyzing the data in this manner, the criterion validity for ORF for each of the grades is clearer. The coefficients in the studies reviewed by Good and Jefferson ranged from .62 to .73—lower than
studies used in Marston’s review, but still within the acceptable range of .60 to .80 as to support the criterion validity of ORF use across grade levels, if desired.

Fuchs and colleagues (1988) found that although many educators do not consider oral reading fluency to be a measure of comprehension, the correlation was, in fact, stronger with the Reading Comprehension subtest than the Word Study subtest on the Stanford Achievement Test. As cited in Good and Jefferson (1998), Shinn and colleagues found strong correlations between ORF and the constructs of Decoding (.89 to .90) and Comprehension (.74 to .75), stronger than correlations found with other measures such as the Stanford Diagnostic Reading Test (approximately .61 and .74, respectively). Shinn, Good, Knutson, and Tilly (1992) were able to demonstrate that ORF was indeed a strong measure of reading skill, including reading comprehension.

ORF assessment procedures have also been shown to have adequate to strong concurrent and discriminant validity with other measures of reading skill and reading comprehension. In the study conducted by Fuchs and colleagues (1988), intercorrelations among the four informal reading measures (oral passage reading, question answering tests, recall procedures, and cloze procedures) were moderate to strong. Regarding discriminant validity, ORF scores were able to adequately differentiate students who were learning disabled/Chapter I students and general education students (Marston, 1989) and to differentiate students of various reading proficiency levels (Fewster & MacMillan, 2002). In fact, ORF scores were found to predict the LD classification of students as effectively as aptitude-achievement discrepancy measures (Marston, 1989). Also, in a study by Wilson, Schendel and Ulman (1992) ORF was found to predict student need for special reading instruction as effectively as the Iowa Tests of Basics Skills reading scores and teacher’s ratings as measured by the Teacher Rating of Academic Performance, Revised scale. Lastly, longitudinal studies of reading growth further supported the use of ORF measures to assess small changes in reading growth. The DIBELS assessment system is discussed next, which includes a measure of ORF in addition to other emergent literacy skills.
DIBELS Assessment

Background

DIBELS was developed by Roland Good and Ruth Kaminski as a series of brief measures to assess acquisition of early literacy skills that were empirically predictive of later reading achievement (FCRR, n.d.; Good & Kaminski, 2002b; Kaminski & Good, 1996, 1998; Simmons, Kame’enui, Good, Harn, Cole, & Braun, 2002). According to the “Official DIBELS Home Page” on the University of Oregon’s DIBELS website, DIBELS is described in the following manner (http://dibels.uoregon.edu/index.php):

The Dynamic Indicators of Basic Early Literacy Skills (DIBELS) are a set of standardized, individually administered measures of early literacy development. They are designed to be…used to regularly monitor the development of pre-reading and early reading skills.

The measures were developed upon the essential early literacy domains discussed in both the National Reading Panel (2000) and National Research Council (1998) reports to assess student development of phonological awareness, alphabetic understanding, and automaticity and fluency with the code. Each measure has been thoroughly researched and demonstrated to be reliable and valid indicators of early literacy development and predictive of later reading proficiency to aid in the early identification of students who are not progressing as expected. When used as recommended, the results can be used to evaluate individual student development as well as provide grade-level feedback toward validated instructional objectives.

According to Kaminski and Good (1998), DIBELS is used to prevent severe reading difficulty those problems, in that it effectively identifies early reading skill deficits. The initial forms of DIBELS targeted early literacy skills of those students in preschool, kindergarten, and first grade In later editions, this was expanded to later grades and is currently offered for administration to students through sixth grade.

Kaminski and Good (1998) note that young children can demonstrate wide variability in early literacy skills. This is important in that DIBELS performance for these students may be
less predictive of reading achievement when they reach later elementary grades. However, as described by Hintze and colleagues, DIBELS is a desirable assessment method given its overall utility in making instructional decisions, its ease of administration, and its suitability within a problem-solving framework (Hintze et al., 2004).

The sixth edition of DIBELS contains five subtests: Letter Naming Fluency, Initial Sound Fluency, Phoneme Segmentation Fluency, Nonsense Word Fluency, and Oral Reading Fluency. Students are administered select subtests according to their grade level to provide a measure of grade-level skills that promote reading achievement (Good & Kaminski, 2002a). Letter Naming Fluency, Initial Sound Fluency, and Phoneme Segmentation Fluency are those subtests that continue the original DIBELS intent of Kaminski and Good (1998) to assess emergent literacy skills. These researchers note that the skills assessed by these three subtests are “transitory, enabling skills that facilitate the acquisition of reading” (Kaminski and Good, 1998).

Each subtest is described in detail in the Florida Center for Reading Research (FCRR) DIBELS Training Manual (2004) and discussed here. Letter Naming Fluency is a one-minute timed task assessing a student’s fluency (speed and accuracy) in naming upper and lower case letters. This subtest is administered to students in kindergarten and the beginning of first grade to identify those students that may be struggling with early phonological awareness skills. Initial Sound Fluency is administered to students in kindergarten to evaluate student skill in identifying and orally producing the initial sound(s) of orally presented words. Low scores on this subtest also indicate student difficulty with phonological awareness. Students in kindergarten and first grade are also administered the subtest Phoneme Segmentation Fluency, which is a one-minute timed task evaluating student skill in segmenting three and four-phoneme words into individual phoneme units. This skill is also related to later reading achievement and may be enhanced with additional phonological awareness instruction. Next, Nonsense Word Fluency is a decoding measure administered to students in kindergarten, first and second grades. The student is required to orally read a series of vowel-consonant or consonant-vowel-consonant nonsense words. This subtest assesses the student’s ability to read novel words by blending letters, which requires knowledge of letter-sound correspondences. Oral Reading Fluency is a one-minute timed task whereby first, second, and third grade students are asked to read aloud from grade-
As discussed previously, oral reading fluency has been shown to be an important determinant of reading comprehension and overall reading achievement. Lastly, the recent sixth edition also contains a final subtest, \textit{Word Use Fluency}, which is currently being researched but is not regularly administered at this time as part of the DIBELS assessment battery as used by FCRR (FCRR, 2004; Good & Kaminski, 2002a).

\textit{Technical Adequacy}

Good and Kaminski (2002) provide detailed descriptions, administration instructions and reliability and validity estimates for each of the DIBELS subtests. However, for the purposes of the present paper, the technical adequacy of two of these subtests will be discussed—NWF and ORF.

According to the researchers, NWF has been found to possess a one-month alternate-form reliability for first grade of .83 (Good & Kaminski, 2002). The DIBELS NWF was also correlated with the Woodcock-Johnson Psycho-Educational Battery, Revised Readiness Cluster score to determine concurrent criterion validity. The resulting estimates for first grade administrations were low (.36 in January and .59 in February). First grade January NWF scores appeared to possess strong predictive validity for end of first grade ORF scores (.82). The predictive validity was weaker for end of second grade ORF scores (.60) and the Woodcock-Johnson Psycho-Educational Battery Total Reading Cluster score (.66). Remaining estimates of reliability and validity were not reported, a limitation in the research at this time. Furthermore, the estimates reported herein focus on first and second grades, whereas the interest of the present paper is on third grade reading achievement. The authors did not intend for the NWF subtest to be administered to third grade students; therefore, there is currently no data to examine the reliability, validity, and predictive utility for this grade level.

The DIBELS ORF subtest was originally designed to be administered to students in mid-first grade to end of third grade (Good & Kaminski, 2002). Recently, this subtest was been expanded to include fourth through sixth grade students. However, the reliability and validity estimates for these grades are not yet available at this time. The ORF subtest is based upon the research and development of CBM in which the technical adequacy of this subtest is well established. The stimulus page presented to the student is the only difference between this
subtest and CBM. Otherwise, scoring and interpretation procedures remain the same. Good and Kaminski cite the published test-retest and alternate-form reliability as well as the criterion-related validity estimates of other researchers for the ORF subtest. A detailed description of those estimates is provided in the related section on ORF (above).

Uses

DIBELS has become an increasingly popular assessment method of reading skill within the last decade. At the time of this manuscript, the measures are offered free of charge for download on the DIBELS website (http://dibels.uoregon.edu/index.php), if used for educational purposes. In fact, FCRR (2004) uses the DIBELS assessment system as part of the agency goals. FCRR was established in 2002 by current Governor Bush, “[and] has as its mission the discovery, creation, cataloging and application of cutting-edge reading research in Florida classrooms” (FCRR, 2004, p. 1). In close connection with the “Just Read, Florida!” Office of the FDOE and the Reading First Professional Development at the University of Central Florida, FCRR seeks to provide Florida’s low-performing schools the needed structure and guidance in completing the objectives proposed by NCLB (FCRR, 2004). The DIBELS assessment system is used due to the battery’s utility in monitoring student achievement and identifying students at-risk for reading difficulty (FCRR, 2004). Also, data obtained from regular DIBELS assessments has been used by the agency to predict later FCAT performance as described in detail below (Buck & Torgesen, 2003).

DIBELS and High Stakes-Testing

Student performance on DIBELS ORF measures has been used to accurately predict performance on the Colorado State Assessment Program test, Florida’s FCAT test, and the Arizona Instrument to Monitor Standards (Buck & Torgesen, 2003; Shaw & Shaw, 2002; Wilson, 2005). It is not surprising given the documented relationship between ORF and reading comprehension that DIBELS data would prove useful in determining student outcomes on these high-stakes measures. Hiebert (2002) argues that reading rate is likely a factor in students’ performance on the FCAT. To do well on the FCAT, Hiebert asserts that students need a strong
rate of reading to support their ability to identify themes and supporting details (components of comprehension) over the eight-page text commonly found on this Florida high-stakes measure.

Buck and Torgesen (2003) studied the relationship between ORF and subsequent performance on the FCAT with 1,102 third grade students. Specifically, they studied the relationship between ORF and the reading comprehension portion of the FCAT-Sunshine State Standards (FCAT-SSS, a criterion-referenced measure), the FCAT norm-referenced test (FCAT-NRT), and the math section of the FCAT-SSS. Buck and Torgesen reported significant positive correlations between oral reading fluency and the Reading FCAT-SSS scores ($r = .70$, $p < .001$), math FCAT-SSS scores ($r = .53$, $p < .001$), and reading scores of the FCAT-NRT ($r = .74$, $p < .001$).

In their research, Buck and Torgesen (2003) also compared the ORF scores to the cutpoints proposed by Roland Good and colleagues (Good, Simmons, Kame’enui, Kaminski, & Wallin, 2002) to determine the accuracy of those cutpoints in predicting FCAT success. According to Good et al. (2002), those students who score below 80 correct words per minute on third grade ORF measures are considered at high risk for performing below grade level on reading comprehension measures. Those who score in the range of 80-109 on ORF are considered to be at moderate risk, and those who score above 110 are considered to be at low risk. In the study by Buck and Torgesen, 81% of students who were considered high risk for achieving below grade level on reading comprehension measures did, in fact, score below level 3 on the reading section of the FCAT-SSS (a score indicating below grade level). Additionally, 91% of the students considered low risk performed satisfactorily on the reading section of the FCAT-SSS. Buck and Torgesen report that, as expected, “students performing in the midrange on the ORF measures were…almost equally likely to attain either an adequate or inadequate score on the FCAT-SSS” (p. 3).

Next, Buck and Torgesen (2003) used the same data to compare the relationship of ORF and reading FCAT-SSS performance across ethnic groups of students. They found that the relationship of ORF and FCAT-SSS reading performance was roughly similar across ethnic groups (Caucasian, African-American, and Hispanic). However, ORF scores above 110 were somewhat less predictive of FCAT-SSS reading success for minority groups. In contrast, ORF scores below 80 were more predictive of failure on the FCAT-SSS reading portion for minority groups.
students. The predictive relationship was not significantly different, however, for the Caucasian and African-American groups contrast and the Caucasian and Hispanic groups contrast. It should be noted that the proportion of minority students was small as compared to that of Caucasian students (17% versus 83%), a limitation that would potentially limit confidence in the results.

Lastly, Buck and Torgesen (2003) compared the results between students on free or reduced lunch (a measure of low SES) to those not on free or reduced lunch to determine whether the correlation relationship differed between those students. The resulting correlations between ORF scores and FCAT-SSS Reading were .70 and .69, for students on free or reduced lunch and those that were not, respectively. Thus, low SES status did not statistically alter the predictive relationship ($\chi^2 = 0.877, p = .35$). The lack of significance among these group contrasts gives further evidence “that ORF scores significantly predict Reading FCAT-SSS scores” (p.7). More research in this area is needed.

Currently, Good and colleagues (2002) prescribe the administration of the ORF portion of DIBELS for third grade students. However, it is advised that second graders are administered both the ORF subtest as well as the Nonsense Word Fluency (NWF) subtests. Thus, there is no current research regarding the utility of using student NWF scores to predict FCAT-SSS reading success. Given that many third graders are performing below grade level as reported by FDOE, it may be useful to determine whether scores from NWF may be useful in improving the prediction of scores on the reading portion of the FCAT.

Summary

With the NCLB initiative, it was expected that every child would be able to read by the end of the third grade (USDOE, 2002). Thus, third grade became the critical grade for assessing student reading skill and state educational testing programs emerged to communicate student performance results to parents, educators, policy makers, and the public in general (Feuer et al., 1999; Hiebert, 2002).
The FCAT is currently used in Florida as an annual high-stakes measure of student achievement. Students’ scores on the FCAT impact individual students and schools at large. Reading achievement is a major focus on the FCAT due to its importance in students’ success in all aspects of education. It is important that by fourth grade students are able to read skillfully, with sufficient comprehension and fluency, to manage new text content and text types (Strickland, 2002). Given the importance placed on FCAT scores, educators seek accurate and efficient methods to predict student FCAT achievement in advance of the year-end testing.

In the domain of reading, decoding and fluency are significant skill requisites in reading achievement and success on the FCAT (Marston, 1989; Armbruster et al., 2001). Accurately decoding and fluently reading connected text (versus a list of words) aloud is a “critical indicator” of general reading skill (Fuchs, 1995). Oral reading fluency (ORF) is a measure of reading fluency as defined by oral reading speed and accuracy (Shinn, 1989a). Research has consistently demonstrated that ORF is a valid and reliable predictor of overall reading achievement and success on the Reading sections of the FCAT (Buck & Torgesen, 2003; Shinn, et al., 1992). To do well on the FCAT, students need a strong rate of reading to support their ability to identify themes and supporting details (Hiebert, 2002).

To date, however, no research appears to have examined word decoding skill and its predictive value with the FCAT. Additionally, there remains a paucity of research regarding the impact of other potentially important predictor variables, such as individual student characteristics, on FCAT scores. For example, it is unknown what effects a student’s attendance rate, retention status, and level of participation in class affect FCAT achievement. If, in fact, these variables have a significant impact on FCAT scores, educators may be better able to identify and assist those students.

The strength of relationship between ORF and FCAT performance notwithstanding, it is curious whether student characteristics or performance in other areas (e.g., reading instruction, NWF) may help to predict ORF performance. This may help to elucidate the variables that may help to explain or are significantly related to ORF success.

Lastly, Good and colleagues (2002) have provided educators with benchmark guidelines for the identification of students at low, moderate, and high risk for reading failure using oral reading fluency data. Buck and Torgesen (2003) have validated the use of those cutpoints in
identifying students in Florida at risk for failing the FCAT Reading test. It is important that this research is replicated to provide further confirmation of the utility of those benchmark guidelines for educators. It remains unknown whether cutpoints may be available for third grade NWF scores to classify students at risk, assuming that NWF is predictive of FCAT scores initially.
CHAPTER 3
METHODOLOGY

Research Questions

Stated as a relationship, the author explored the following research questions:

• What is the relationship between ORF (oral reading fluency) DIBELS scores, NWF (nonsense word fluency / decoding) DIBELS scores, student background and performance variables (gender, SES, class participation, attendance rate, prior grade retentions, Reading grades, and Accelerated Reader scores) and Reading FCAT-SSS scores?

• What is the relationship between student background and performance variables (gender, SES, class participation, attendance rate, prior grade retentions, Reading grades, and Accelerated Reader scores, NWF) and student ORF performance?

• What is the relationship between ORF risk level classification (and NWF scores if predictive of FCAT performance) and FCAT failure?

The author focused on the null hypotheses to avoid prematurely excluding any potential effects. This study was a correlational research design exploring the strength of the relationships detailed above.

Participants

Participants originally included 186 third grade students in a large school district in Florida. Eligible participants were those enrolled in the regular education curriculum and required to take the FCAT in Spring 2005. This researcher obtained the appropriate approval
from the university’s Institutional Review Board (IRB) to conduct the study (see Appendix A). Consistent with IRB requirements, informed consent documentation was drafted for prospective participants. As reported in the sampling procedure described below, only those students for whom their parents consented and whom the students themselves also assented were eligible for participation in the study.

**Sampling**

Participants in this study were recruited in the following manner. The school system was partitioned into five geographical regions. Four elementary schools from each region were selected so that the variability that exists among students in the different geographical locations was maximized. Also, the researcher selected both high and low-performing schools with respect to students’ scores on the FCAT to ensure that a range of FCAT scores were included for score variability. Twenty schools in total were approached to participate in the present study. The researcher detailed the study to the principals of these schools and 12 principals subsequently consented for their schools to participate in this study. Participating principals provided their verbal consent for their school to participate. Due to district regulations, the researcher was not permitted to reveal information that could potentially alert readers to the specific district, schools, or participants involved in this study. Thus, information regarding school demographics for each participating school was compiled but not provided here due to publication restrictions.

A description of the research goals and procedure was then given to every third grade teacher in each of the 12 schools. It was explained that for the purposes of treatment fidelity and confidentiality, the results of the present study would not be available until the study was concluded and only group-level results would be reported. The potential risks and benefits associated with participation in the study were summarized. Administrators and teachers were given the opportunity to seek clarification and decline participation, if desired. Teachers consenting to participate in the present study were asked to sign and informed consent form stating the same. The letter and teacher-directed informed consent is presented in Appendix A. None of the teachers declined participation in this study.
The sampling method for student recruitment used in this study was a criterion sample with the rules of inclusion described below. The accessible population of students was intended to approximate the target population of all third grade regular education students in the state of Florida. Conclusions drawn from the results of this study may be generalizable to those students that meet the same rules of inclusion (third grade regular education students) in Florida.

The researcher sought a sample size large enough to accommodate small to moderate effect sizes in the subsequent data analyses. According to Hair, Anderson, Tatham, and Black (1998), this would result in an effect size of approximately .35 (based upon the seminal work of Cohen). Desiring a confidence level of .95 and a statistical power of .80, the researcher worked to obtain at least 130 subjects using the sampling technique described herein (Hair et al., 1998). Due to the nature of the consent distribution, the researcher acknowledged the potential for a low return rate. Thus, a total of 384 consents were distributed in hopes of receiving at least 130 participants.

Thirty-two regular education third grade students were randomly selected from each of the 12 schools using a random numbers chart (Kirk, 1995) such that each student had an equal probability of being selected for inclusion. Consents were sent home with each selected student, distributed by the teacher (384 consents total). Each consent packet consisted of a letter of introduction to the research as well as an informed consent document stating the nature of the study, potential risks and benefits, and contact information of the principal investigator for additional information. The introduction letter and parental informed consent can be found in Appendix A.

Of the 384 informed consent packets distributed, 193 consent forms (or 50.3%) were received in return. Of those received, 186 parents/legal guardians consented to participation (48.4% of consent packets distributed, 96.4% of consent forms received in return), 6 parents/legal guardians declined participation (1.6% of consent packets distributed, 3.1% of consent forms received in return) and one consent packet was undeliverable because the student relocated (0.3% of consent packets distributed). Information regarding those families that did not return the consents was unavailable. Therefore, it was not possible to conduct an analysis of systematic error or bias that may have been present in the final sample.
Procedure

Data collection for the present study consisted of direct performance assessment of oral reading fluency and decoding skill (ORF and NWF) from the DIBELS assessment system (Good & Kaminski, 2002b), collection of desired demographic and student variable information from the school district’s database (gender, SES, attendance rate, and prior grade retentions), and a brief survey completed by the student’s classroom teacher rating that student’s level of participation, the student’s Accelerated Reader score, (if applicable), and grade letter evaluation in Reading for the third grading period. In cases where the questionnaire was not returned or it was not reported, school personnel were asked to provide the third quarter Reading grade for each student, if available.

Regarding the DIBELS assessment, each participant’s oral reading fluency and decoding skill was assessed using the ORF and NWF subtests, twice, prior to the FCAT administration in Spring 2005. The ORF and NWF protocols used for this study were obtained from the official DIBELS website (http://dibels.uoregon.edu/). These protocols are available via download and are published by Sopris West. They are provided to educators who register and are free of charge if used for educational purposes. A sample of each measure is provided in Appendix B, but the interested reader is referred to the DIBELS website for more information and complete protocols.

Each participant was administered the ORF and NWF subtests from the DIBELS as a set, resulting in a single interval administration. Two interval administrations of the ORF and NWF measures were conducted. The first administration occurred in December 2004 and the second administration occurred in February 2005; the FCAT was administered in March 2005. Volunteer school psychologists or school-based Reading Coaches administered the subtests, all of whom had received a minimum of six hours of formal in-service training from FCRR in the administration and scoring of the selected DIBELS assessment system subtests. Twenty percent (or 36) of the protocols from each administration interval were randomly selected for reliability checks by this author.

Additionally, the following information regarding each participant was obtained via district records: gender, SES (using free or reduced lunch as the observed variable), attendance
rate, and prior grade retentions. Each participant’s teacher was also provided with a brief questionnaire in which they were requested to evaluate the student’s general level of class participation in classroom instructional reading activities using a five-point Likert scale assessment (see Appendix C). Teachers were also requested to report the student’s third quarter reading grade and Accelerated Reader score, if used in their classroom. The questionnaires were to be returned to the principal investigator using an enclosed self-addressed stamped envelope. Lastly, each participant’s score on the Reading FCAT-SSS was to be obtained via district records.

**Instruments**

Instruments used in the present study include the FCAT test, the ORF and NWF subtests from the DIBELS assessment system and a brief survey administered to each participant’s teacher (Appendices B and C, respectively). The reliability and validity estimates for the selected DIBELS estimates were stated previously as provided by the researchers (Good & Kaminski, 2002b). This author developed the one-item class participation survey. Other than the FCAT (as described below), the collection of the remaining variables did not require the use of any specific instruments.

Unpublished data regarding the psychometric properties for the 2005 administration of the Reading FCAT-SSS were obtained by contacting FDOE (2006). The mean Reading FCAT-SSS for third grade *standard curriculum students* (non-ESE students) on the 2005 administration was 317.22 (sd = 56.97). FDOE reported the standard error of measurement (SEM) as a function of Reading FCAT-SSS scores, assuming that the statistic would not be stable across ability levels. Thus, the SEM for third grade students during the 2005 administration of the Reading FCAT-SSS ranged from 13 points near the mean to 25 points as the Reading FCAT-SSS score approached 394. Reliability as measured by Cronbach’s alpha was strong at .893 for these students, an increase from the 2004 administration. The overall internal consistency index for FCAT Grade 3 Reading was 0.61.
Consideration was given to potential administration error of the DIBELS instruments used in this study. It was not feasible given personnel resources to monitor each DIBELS administration for integrity. However, the researcher reviewed the protocols and selected 36 protocols randomly from NWF #1, NWF #2, ORF #1, and ORF #2 using a random numbers table (Kirk, 1995). Only those subtests administered by this researcher or another recruited examiner were available for reliability checks; protocols from administrations conducted by the Reading Coaches that align with district and state monitoring requirements were unavailable for examination. The results of the reliability checks are provided in the following chapter where appropriate.
CHAPTER 4
RESULTS

Sample

This study included 186 third graders in a large school district in Florida, five of whom were no longer enrolled in school at the conclusion of the study (2.7% attrition). None of the students’ parents rescinded consent and no student refused participation during the assent process. Therefore, the final sample included data from 181 participants for analysis.

The demographic composition of the sample is summarized in Table 1. Seventy-two (39.8%) participants were male, 109 (60.2%) were female. Ethnic distribution was as follows: 34.8% Caucasian, 59.1% African-American, 2.2% Hispanic, 1.1% Asian, 2.8% Other (including mixed ethnicities). See Table 1 for comparison with district and state statistics. Participant socio-economic status (or SES) was characterized in this study by whether the student received free or reduced-lunch benefits as approved by the district according to family financial need. Of the 120 students that applied for free or reduced-lunch price benefits, 98 students (54.1% of entire sample, 81.7% of those applying) received free lunch, 18 (9.9% of entire sample, 15.0% of those applying) received reduced-lunch, and 4 were ineligible (2.2% of entire sample, 3.3% of those applying). Thus, 116 students received some level of financial support in the form of lunch benefits. Sixty-one of the participants did not submit applications for free or reduced-lunch benefits. Including the four students that applied for lunch benefits and were ineligible, 65 total students in the sample received no reduced or free-lunch benefits from the district. Due to the limited number of students that received reduced-lunch benefits versus free lunch benefits, this level of the SES predictor variable was collapsed with the larger group of students that received free-lunch benefits, creating a combined group. Therefore, SES is comprised of two levels, regular and economically disadvantaged in this study (Table 1).
Table 1

Descriptive Statistics for Gender, Ethnicity, and SES for Third Grade, Fall 2004

<table>
<thead>
<tr>
<th></th>
<th>Sample</th>
<th>District</th>
<th>State</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n</td>
<td>%</td>
<td>n</td>
</tr>
<tr>
<td>Male</td>
<td>72</td>
<td>39.8</td>
<td>5,220</td>
</tr>
<tr>
<td>Female</td>
<td>109</td>
<td>60.2</td>
<td>4,845</td>
</tr>
<tr>
<td>Caucasian</td>
<td>63</td>
<td>34.8</td>
<td>4,690</td>
</tr>
<tr>
<td>African-American</td>
<td>107</td>
<td>59.1</td>
<td>4,304</td>
</tr>
<tr>
<td>Hispanic</td>
<td>4</td>
<td>2.2</td>
<td>468</td>
</tr>
<tr>
<td>Asian</td>
<td>2</td>
<td>1.1</td>
<td>266</td>
</tr>
<tr>
<td>Other</td>
<td>5</td>
<td>2.8</td>
<td>337</td>
</tr>
<tr>
<td>Economically</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Disadvantaged</td>
<td>116</td>
<td>64.1</td>
<td>—</td>
</tr>
</tbody>
</table>

Next, descriptive data was obtained from district records regarding each participant’s number of grade retentions prior to this academic year of 2004-2005 and current attendance rate. With respect to grade retention, 50 (27.6%) students in this study were retained in the same grade at least once in their academic history due to low academic performance. The frequency of grade retentions for this sample is listed in Table 2.

Of those students retained, 12 (24.0%) were retained in Kindergarten, 14 (28.0%) were retained in first grade, 13 (26.0%) were retained in second grade, and 24 (48.0%) were retained in third grade. It should be noted that the cumulative percentage of these figures was greater than 100% as some students were retained multiple times. Specifically, 38 students were retained once, 11 students were retained twice, and one student was retained three times prior to this academic year. An additional 20 students were retained at the conclusion of the 2004-2005 academic year in third grade (after the completion of the present study), 18 of whom failed the Reading FCAT-SSS (obtained a Level 1 or 2).
Table 2

*Number of Grade Retentions Prior to 2004-2005*

<table>
<thead>
<tr>
<th>Number of Retentions</th>
<th>f</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>131</td>
<td>72.4</td>
</tr>
<tr>
<td>1</td>
<td>38</td>
<td>21.0</td>
</tr>
<tr>
<td>2</td>
<td>11</td>
<td>6.1</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>0.6</td>
</tr>
</tbody>
</table>

District records were also accessed to obtain attendance data. Attendance rate was obtained by dividing the number of days the student was enrolled by the number of days the student was present for the academic year. The sample mean attendance rate was .953 (sd = .047, range = .68 to 1.00).

Data on students’ class participation in Reading activities and Accelerated Reader performance were obtained via questionnaire distributed to teachers. Questionnaires were distributed twice in an effort to obtain the largest response possible. A total of 120 questionnaires were returned (66.3% return rate). Of those questionnaires returned, all teachers responded to the participation item. As previously discussed, teachers rated each student’s participation on a scale of zero (none) to four (constant) in reading instruction and activities including: attending to lesson and task, asking questions, and attempting to complete assignments (see Appendix C). As evidenced in Table 3, participation ratings were positively skewed towards greater student participation in this sample.

The majority of teachers did not provide data regarding student Accelerated Reader achievement; some teachers left that item blank whereas many indicated that they did not use the Accelerated Reader program in their classroom. Ultimately, the Accelerated Reader data was eliminated from the study for this reason. A detailed description of missing data, including Accelerated Reader score data, is offered in a subsequent section for the reader.
Lastly, reading skill (observed performance) data were collected for each participant. NWF (decoding skill) and ORF (oral reading fluency) scores were obtained twice prior to FCAT administration via direct performance assessment, in December 2004 and February 2005. Each student then participated in statewide FCAT testing in March 2005. Descriptive statistics for NWF, ORF, and Reading FCAT-SSS are displayed in Table 4.

Table 3
*Descriptive Statistics for Participation Ratings*

<table>
<thead>
<tr>
<th>Participation Rating</th>
<th>f</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>11</td>
<td>6.1</td>
</tr>
<tr>
<td>2</td>
<td>26</td>
<td>14.4</td>
</tr>
<tr>
<td>3</td>
<td>41</td>
<td>22.7</td>
</tr>
<tr>
<td>4</td>
<td>42</td>
<td>23.2</td>
</tr>
</tbody>
</table>

Table 4
*Descriptive Statistics for NWF, ORF and FCAT Scores*

<table>
<thead>
<tr>
<th></th>
<th>n</th>
<th>M</th>
<th>SD</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>NWF #1</td>
<td>163</td>
<td>80.86</td>
<td>40.76</td>
<td>6-232</td>
</tr>
<tr>
<td>NWF #2</td>
<td>153</td>
<td>89.36</td>
<td>44.44</td>
<td>12-237</td>
</tr>
<tr>
<td>Median ORF #1</td>
<td>158</td>
<td>102.44</td>
<td>33.70</td>
<td>0-180</td>
</tr>
<tr>
<td>Median ORF #2</td>
<td>156</td>
<td>103.15</td>
<td>27.27</td>
<td>35-169</td>
</tr>
<tr>
<td>SSS Reading Scale Score</td>
<td>181</td>
<td>307.24</td>
<td>52.26</td>
<td>134-446</td>
</tr>
</tbody>
</table>
Data Analysis Overview

Sequence of Analyses

In order to explore the identified research questions, the author employed statistical procedures described below using SPSS 11.0 (2001). First, preliminary analyses were conducted in the form of a missing data analysis (for the data set) and case analyses (for each procedure). These analyses were conducted to explore systematic errors in the data set resulting from missing data or errant unrepresentative results due to observations with excessive influence.

Regarding the first research question, multiple regression analysis was used to identify significant predictors of FCAT Reading performance (Reading FCAT-SSS) among the predictor variables (gender, SES, retention, attendance, class participation, Reading grades, NWF, and ORF). For this research question, two analyses were conducted resulting in two models. The first model was constructed to tease out the effects of the predictor variables that have not been researched as thoroughly (more on this below). The second model utilized stepwise regression to determine the best subset of predictors to maximize the prediction of Reading FCAT-SSS.

Next, research has shown ORF to be a strong predictor of FCAT Reading success. For the second research question, the researcher was interested in determining which, if any, of the predictor variables in this study (gender, SES, retention, attendance, class participation, Reading grades, and NWF) could be used to reliably predict ORF scores. Multiple regression analysis was used to determine the best subset of predictors of ORF scores from the available predictor variables.

Lastly, the third research question was posited to determine the relationship between ORF risk level (high risk, moderate risk, low risk, and above average) and Reading FCAT-SSS failure. The researcher used frequency analysis and chi-square analysis to analyze the proportion of students at each benchmark level that did not achieve the minimal passing score on the FCAT test. These analyses were employed to survey the predictive utility of ORF cutpoints used by researchers and state educators for identifying students at risk for reading difficulty. Given that NWF was not found to be a significant predictor of Reading FCAT-SSS performance, adding no significant unique variance above that of ORF to the prediction, the analysis was not re-conducted with the published NWF cutpoints as originally planned. The predictive utility of
NWF cutpoints for predicting FCAT success for third graders was deemed to be without practical importance.

*Control of Family-wise Error Rate*

In an effort to control family-wise error rate for the family of hypothesis testing herein, the commonly used procedures of protected testing and simultaneous inference were utilized. As such, each of the multiple regression analyses contained an initial global test of the null hypothesis. This global test for multiple regression was the overall test of $R^2$. If the overall null hypothesis was rejected then follow-up significance testing of the coefficients was conducted.

In research practice, researchers typically examine the circumstances of their study to determine which simultaneous inference procedure is appropriate for the data analysis. The Bonferroni Correction is an often-used simultaneous inference procedure to protect against family-wise error, (Hair et al., 1998; Tate, 1998). This procedure allows the researcher to compute an adjusted family-wise error rate by dividing the desired alpha (.05 in this case) by the number of statistical tests to be utilized in the study. One can be too liberal and identify statistical significance when, in fact, there are no differences among the greater population. On the contrary, one can be over-conservative, requiring too great of a critical value and, thus, obscuring very real effects in the population. The latter is especially true when the research design is more complex incorporating more than one hypothesis.

For the purposes of this study, simultaneous inference procedures were employed to set the desired family-wise error rate to .05, resulting in a confidence level of .95. In an effort to balance the effects of excessive liberalism and conservatism, the researcher chose to employ Bonferroni Correction and to divide the test-wise alpha by the number of family of tests (hypotheses), rather than each individual comparison (R. Tate, personal communication, October 8, 2004). Thus, statistical significance for this study was determined to be .05 divided by 3 (for three hypothesis tests), or .0167.
Preliminary Analyses

The data set was examined to determine the effects of instrument administration reliability, missing subjects, and missing data. Case analyses and assessment of validity of assumptions were then addressed with each primary analysis.

Instrument Administration Reliability

The researcher conducted a reliability check on the NWF and ORF protocols. The results of the reliability checks are as follows: NWF #1 = .72; ORF #1 = 1.00; NWF #2 = .69; ORF #2 = 1.00. No errors were found in the scoring of any of the ORF protocols selected. However, several errors in the scoring of the NWF protocols were identified. Common errors in scoring included: addition errors, neglect of maximum correct number of phonemes per line (NWF only), and omission of completion time, if under one minute. Due to the measurement error evidenced by the reliability check, this researcher re-calculated each NWF protocol. The NWF protocols that did not note completion time constituted 8.8% (or 16 of 181) of the total number of NWF protocols. These protocols were deemed spoiled data and were eliminated from further analysis.

Analysis of Missing Subjects and Data

Data for this study were obtained via a variety of data collection methods including archival review of district database records, direct performance assessment, and questionnaires to teachers. As previously mentioned, 384 informed consent packets were distributed with a 50.3% return rate. Information regarding those families that did not return the consents was unobtainable. Thus, it was not possible to conduct an analysis of systematic error that may be present due to the omission of those families that did not respond to the request. However, the resulting sample was compared to district and state demographics for comparison (Table 1).

Comparing the sample composition to district and state statistics revealed the following trends. This sample reflected a statistically significant greater representation of females than males whereas district-wide enrollment was almost equal with respect to gender (sample-district contrast \( z = -3.26 \); sample-state contrast \( z = -3.23 \)). Additionally, although the number of
African-Americans in this sample was more comparable to district statistics than statewide statistics, the proportion of African-American students in this sample was significantly greater ($z = 4.43$). Sample and district proportions Hispanic students were comparable ($z = -1.54$), but the sample proportion of Hispanic students was significantly less than statewide prevalence ($z = -6.39$). Lastly, this sample also included greater representation of those students that are economically disadvantaged as indicated by whether or not they receive free or reduced lunch prices. Due to lack of published data on the specific number of students, tests of significance between sample proportions and district and state proportions could not be conducted.

In the present study, the researcher attempted to select schools that were both high and low-performing schools on FCAT during the previous academic year to ensure a varied distribution. This selection was guided by the use of schools that were classified as low performing schools on FCAT receiving grants from the Reading First program. This had an additional benefit in that DIBELS data was already routinely collected on these students at the same time frame as the interval assessments for this study. Students in these district schools are predominantly economically disadvantaged (92.1%) and African-American (84.5%), but relatively comparable with respect to gender (51.2% males) (FCRR, 2006). Despite attempts to select enough students from high-performing schools to diversify the study’s sample, the final sample reflected a greater number of economically disadvantaged students than anticipated. These differences in sample composition with respect to district and state composition may result in a sample bias towards students with these characteristics and may have potential effects on the generalizability of the results herein.

Five participants for whom signed consents were received were no longer enrolled at the conclusion of data collection resulting in 2.7% attrition rate. Therefore, district data regarding attendance rate and FCAT scores were available for 97.3% of the participants. These students were eliminated from the study for a total of 181 participants in the final sample. In this final sample, data obtained from district records (gender, SES, retentions, attendance, FCAT scores) did not have any missing information.

Conversely, missing data occurred among the performance assessment variables (NWF #1, NWF #2, ORF #1, and ORF #2) as well as variables reported by questionnaire (participation ratings, third quarter Reading grades, and Accelerated Reader scores). The proportion of missing
data for each variable of interest is provided in Table 5. Reasons for missing data on these measures include absenteeism on the day of assessment, assessment administration error, questionnaire return rate, and extensive omission of Accelerated Reader scores.

Table 5

*Presence of Missing Data*

<table>
<thead>
<tr>
<th></th>
<th>f</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>0</td>
<td>0.0</td>
</tr>
<tr>
<td>SES</td>
<td>0</td>
<td>0.0</td>
</tr>
<tr>
<td>Retentions</td>
<td>0</td>
<td>0.0</td>
</tr>
<tr>
<td>Attendance Rate</td>
<td>0</td>
<td>0.0</td>
</tr>
<tr>
<td>Reading Grade</td>
<td>51</td>
<td>28.2</td>
</tr>
<tr>
<td>Participation</td>
<td>61</td>
<td>33.7</td>
</tr>
<tr>
<td>Accelerated Reader Score</td>
<td>150</td>
<td>82.9</td>
</tr>
<tr>
<td>NWF #1</td>
<td>18</td>
<td>9.9</td>
</tr>
<tr>
<td>NWF #2</td>
<td>28</td>
<td>15.5</td>
</tr>
<tr>
<td>ORF #1</td>
<td>23</td>
<td>12.7</td>
</tr>
<tr>
<td>ORF #2</td>
<td>25</td>
<td>13.8</td>
</tr>
<tr>
<td>Reading FCAT-SSS</td>
<td>0</td>
<td>0.0</td>
</tr>
</tbody>
</table>

*Note.* Accelerated Reader scores were later excluded due to the excessive amount of missing data for this variable.

Regarding the missing data for the DIBELS scores, 16 of the 181 participants’ NWF #2 scores were deemed invalid by the principal investigator; the examiner who administered that measure to those students appeared to allow the student to read the entire protocol (a valid method if recording total time elapsed) but failed to record the total elapsed time. To explain further, during the administration of NWF, standard administration procedures required that each student be assessed for precisely one minute or allowed to read the entire page noting the total
elapsed time. It was the examiner’s prerogative to use either method, as both methods are sanctioned for standardization purposes. If given the whole page, the correct phonemes per minute were calculated using a formula to adjust to the extended time period. The examiner at one school did not note time or indicate any stopping points during NWF #2 administration and the final scores did not accurately reflect the variable of interest. Without a distinct mark on the protocol where the student stopped or a total elapsed time, the protocol could not be correctly rescored. Thus, these scores were deemed spoiled and were excluded from further analysis. Additionally, another examiner at a different school did not obtain ORF scores on either assessment interval for the participants at that school (19 students). Therefore, this data was also unavailable for analysis. These events reflect as a limitation to this study with potential effects on sample bias (to be addressed later). The number of students for which scores were available on each measure was reported in Table 4. The number of students missing complete performance data for the assessments were as follows: 36 (19.9%) for the first interval assessment, 48 (26.5%) for the second interval assessment, and 62 (34.3%) for both intervals.

Lastly, 120 questionnaires were received from teachers in which they were asked to rate the participants’ class participation during Reading instruction and to report the students’ third quarter Reading grades and Accelerated Reader scores if the program was used in their classroom. As discussed previously, 120 returned questionnaires reflects a 66.3% return rate. Unfortunately, Accelerated Reader scores were reported for only 31 students (25.8% of returned questionnaires, 17.1% of total sample). Of the questionnaires returned, most teachers reported that they do not use the Accelerated Reader program in their classroom. This resulted in an insufficient number of Accelerated Reader data points to include in final data analysis and this variable was eliminated from further analysis. The potential for additional problems resulting from missing data for the remaining variables is considered next.

To determine whether there exists any systematic patterns in the missing data, each variable was coded for “missingness” and the correlations were computed between this missingness indicator and the remainder of the variables included in the study (Tate, 1998). If these correlations were small, this suggested that using a listwise deletion of subjects for which there is any missing data would not affect the representativeness of the sample to the population.
The bivariate Pearson product-moment correlations for the missingness indicators and the original variables were provided for the reader in Table 6.

The missingness indicator for NWF #1 was significantly correlated with the number of grade retentions and attendance rate (r\textsubscript{ret} = -.17, p = .021, n = 181; r\textsubscript{att} = .23, p = .002, n = 181). The missingness indicator for NWF #2 was significantly correlated with third quarter Reading grade (r\textsubscript{RG} = .25, p = .004, n = 130). The missingness indicator for ORF #2 was significantly correlated with ORF #1 and attendance rate (r\textsubscript{O1} = .20, p = .013, n = 158; r\textsubscript{att} = .33, p < .001, n = 181). Next, the missingness indicator for participation was significantly correlated with NWF #1, NWF #2, SES, number of grade retentions, attendance rate, and Reading FCAT-SSS (r\textsubscript{N1} = .22, p = .005, n = 163; r\textsubscript{N2} = .18, p = .027, n = 153; r\textsubscript{SES} = -.19, p = .009, n = 181; r\textsubscript{ret} = -.15, p = .048, n = 181; r\textsubscript{att} = .15, p = .041, n = 181; r\textsubscript{SSS} = .23, p = .002, n = 181). Lastly, the missingness indicator for third quarter Reading grade was significantly correlated with SES, number of grade retentions, attendance rate, NWF #1, NWF #2, ORF #1, and Reading FCAT-SSS (r\textsubscript{SES} = -.19, p = .012, n = 181; r\textsubscript{ret} = -.16, p = .027, n = 181; r\textsubscript{att} = .16, p = .029, n = 181; r\textsubscript{N1} = .26, p = .001, n = 163; r\textsubscript{N2} = .22, p = .006, n = 153; r\textsubscript{O1} = .16, p = .047, n = 158; r\textsubscript{SSS} = .23, p = .002, n = 181).

Intuitively, the significant correlations between missingness indicators for NWF and attendance rate as well as ORF and attendance rate is not unexpected given that those students with higher absenteeism were not as likely to be in attendance on the days of direct performance assessment. The significant correlation between the missingness indicator for ORF #1 and ORF #2 suggests that those for whom data is not available for the first ORF administration likewise were not available for the second ORF administration. This is highly likely given that ORF scores at one of the 12 schools were not collected for either interval administration by error of the examiner. Due to the extent of missing data for the variables of participation and third quarter Reading grade, it is not surprising that there are large correlations between the respective missingness indicators and several other study variables.
<table>
<thead>
<tr>
<th>Miss. Indic.</th>
<th>Gen</th>
<th>SES</th>
<th>Ret</th>
<th>Att</th>
<th>Par</th>
<th>RG</th>
<th>N1</th>
<th>O1</th>
<th>N2</th>
<th>O2</th>
<th>SSS</th>
</tr>
</thead>
<tbody>
<tr>
<td>xGen</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>xSES</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>xRet</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>xAtt</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>xPar</td>
<td>.04</td>
<td>**-.19</td>
<td>-.15</td>
<td>.15</td>
<td>—</td>
<td>-.01</td>
<td>**.22</td>
<td>.13</td>
<td>*.18</td>
<td>.14</td>
<td>**.23</td>
</tr>
<tr>
<td>xRG</td>
<td>.01</td>
<td>*-.19</td>
<td>*-.16</td>
<td>.16</td>
<td>.09</td>
<td>—</td>
<td>**.26</td>
<td>.16</td>
<td>**.22</td>
<td>.15</td>
<td>**.23</td>
</tr>
<tr>
<td>xN1</td>
<td>-.15</td>
<td>-.10</td>
<td>*-.17</td>
<td>**.23</td>
<td>.11</td>
<td>.12</td>
<td>—</td>
<td>.02</td>
<td>.04</td>
<td>.03</td>
<td>.05</td>
</tr>
<tr>
<td>xO1</td>
<td>.01</td>
<td>-.08</td>
<td>.03</td>
<td>.13</td>
<td>-.01</td>
<td>.09</td>
<td>.05</td>
<td>—</td>
<td>.08</td>
<td>.00</td>
<td>-.02</td>
</tr>
<tr>
<td>xN2</td>
<td>.10</td>
<td>-.10</td>
<td>.04</td>
<td>.02</td>
<td>.06</td>
<td>**.25</td>
<td>-.13</td>
<td>.03</td>
<td>—</td>
<td>-.04</td>
<td>.03</td>
</tr>
<tr>
<td>xO2</td>
<td>-.00</td>
<td>-.10</td>
<td>.07</td>
<td>**.33</td>
<td>-.01</td>
<td>.08</td>
<td>.10</td>
<td>*2.20</td>
<td>.13</td>
<td>—</td>
<td>.06</td>
</tr>
<tr>
<td>xSSS</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>

Note. Gen = gender; SES = socio-economic status as represented by receipt of free/reduced lunch benefits (range 0-1, 0 = no benefits, 1 = benefits); Ret = number of grade retentions; Att = attendance rate; Par = participation rating (range 0-4); RG = 3rd quarter Reading grade (range 0-4); N# represents decoding ability as measured by scoring correctly verbalized phonemes (phonemes per minute) and assessment interval; O# represents oral reading fluency as measured by scoring correctly verbalized standard English words in context (words per minute) and assessment interval; SSS = FCAT Sunshine State Standards Reading Scale Score.

*p < .05, **p < .01, — indicates no missing data or correlation not computed.
However, the large correlation between the missingness index for NWF #1 and number of grade retentions may indicate a potential bias in NWF #1 statistics with respect to number of grade retentions. The relationship between the missingness indicator for NWF #1 and number of grade retentions may be due to chance or may, in fact, represent a real difference. One possible explanation is that for those students who are retained, especially more than once, events may have prevented their participation in the administration of NWF. However, this is unlikely given that a similar correlation would be expected for the missingness indicator for ORF #1 since the two subtests are administered sequentially to the student in the same session. Additionally, there is no strong correlation between NWF#1 and number of grade retentions that may help to explain this event.

Due to the large correlations noted above, the use of the listwise deletion strategy for subsequent multiple regression analyses is not advisable (Tate, 1998). This compensation strategy results in a dramatically reduced sample size having a subsequent impact on the statistical power to detect true population effects (Tate, 1998). Secondly, it is possible that the extent of missing data for the participation variable (33.7%) resulted in an under-representation of certain unidentified subgroups of the sample for whom the teacher did not return the survey, potentially causing a systematic bias in the study results concerning that variable. Therefore, any observed effect of this variable is viewed with some caution.

Another type of deletion strategy, pairwise deletion, is also an option when the extent of missing data is problematic for one or more variables. Tate (1998) describes the detrimental effects of utilizing this strategy for data management, explaining that this strategy computes pairwise correlations for the subsample having values for both variables. Thus, the resulting correlations may not represent the full sample or the population of interest and can have “unacceptable properties for further analysis” (Tate, 1998, p. 47).

Any data management strategy to compensate for missing data is potentially problematic as a source of bias. Nonetheless, observation replacement is oftentimes desirable because it addresses the dilemma associated with missing data and ultimately producing a subsequent nonrepresentative subsample via deletion techniques. Additionally, this strategy eliminates difficulty associated with loss of statistical power due to sample size reduction. Considerations of various observation replacement strategies for use in the present study are discussed next.
First, *mean substitution* is an observation replacement strategy that substitutes the mean value of the target variable for any missing data within that variable (Tate, 1998). Tate cautions that this strategy, although simple, can result in distorted sample statistics (especially variances) that impacts the legitimacy of further data analysis (i.e., creating data that only fictitiously meets the statistical assumptions required for use of the chosen statistical technique). Other frequently used observation replacement strategies include *linear interpolation* and *linear trend at point* (both available in SPSS software). The former, linear interpolation, uses an interpolation equation calculated with the last valid value before the missing value and the first valid value after the missing value (SPSS, 2001). The latter, linear trend at point, replaces missing values with their predicted values by regressing the value series on an index variable scaled 1 to \( n \) (SPSS, 2001).

In the present study, it was advantageous to address the significant impact of missing data for the variables of participation and third quarter Reading grades using one of the above data management compensation strategies. Participation was considered first. To determine which observation replacement strategy more closely replicates the descriptive statistics of the original participation variable all three strategies were conducted. The results of these analyses are presented in Table 7 for comparison.

Given that the greatest concern for observation replacement lies with the distortion of the variances (Tate, 1998), the method of linear interpolation was determined to be the best method used to replace missing participation values; the variance of this method most closely approximated the variance of the original variable (Participation variance = .94 versus Participation \(_{\text{Linear Interpolation}}\) variance = .73). The descriptive statistics of the new variable with replaced values were very similar to the descriptive statistics of the original variable. Therefore, further references to the participation variable incorporated the replaced values from the linear interpolation method of observation replacement.
Table 7

Comparison of Observation Replacement Strategies of Missing Participation Values

<table>
<thead>
<tr>
<th></th>
<th>Particip</th>
<th>Mean Subst</th>
<th>Linear Interp</th>
<th>Linear Trend</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Valid</td>
<td>120</td>
<td>181</td>
<td>181</td>
<td>181</td>
</tr>
<tr>
<td>Missing</td>
<td>61</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>M</td>
<td>2.95</td>
<td>2.950</td>
<td>2.997</td>
<td>2.949</td>
</tr>
<tr>
<td>Std. Error of Mean</td>
<td>.088</td>
<td>.0586</td>
<td>.0637</td>
<td>.0586</td>
</tr>
<tr>
<td>SD</td>
<td>.969</td>
<td>.7878</td>
<td>.8568</td>
<td>.7880</td>
</tr>
<tr>
<td>Variance</td>
<td>.939</td>
<td>.6206</td>
<td>.7341</td>
<td>.6210</td>
</tr>
<tr>
<td>Skewness</td>
<td>-.519</td>
<td>-.635</td>
<td>-.623</td>
<td>-.631</td>
</tr>
<tr>
<td>Std. Error of Skewness</td>
<td>.221</td>
<td>.181</td>
<td>.181</td>
<td>.181</td>
</tr>
</tbody>
</table>

Note. Particip = original Participation variable; Mean Subst = missing values replaced with mean; Linear Interp = missing values replaced using linear interpolation equation; Linear Trend = missing values replaced using regression-based predicted values.

This researcher then chose to compensate for missing data among the third quarter Reading grades in the following manner. Third quarter Reading grades were available for 130, or 71.8% of the sample. Unfortunately, the district database did not retain quarterly grades once final grades were input. In cases where the third quarter grade was unavailable due to teacher or staff omission, the final reading grade for the academic year was used, obtained from district records. Using the final grade was preferable as a replacement method given that the use of a statistical invention of data should be used cautiously (Tate, 1998). This researcher chose to refrain from using a replacement procedure more than once. Also, the final reading grade is based upon the aggregate evaluation of that student during the academic year and was believed to be representative of that student’s performance. Therefore, further references to the Reading grade variable indicates this transformed variable for those who were missing third quarter Reading grades in the data set. The frequency and percentage rate of each letter grade in this sample are listed in Table 8.
Table 8

*Reading Grades for 2004-2005*

<table>
<thead>
<tr>
<th>Reading Grade</th>
<th>f</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>9</td>
<td>5.0</td>
</tr>
<tr>
<td>D</td>
<td>27</td>
<td>14.9</td>
</tr>
<tr>
<td>C</td>
<td>55</td>
<td>30.4</td>
</tr>
<tr>
<td>B</td>
<td>66</td>
<td>36.5</td>
</tr>
<tr>
<td>A</td>
<td>24</td>
<td>13.3</td>
</tr>
</tbody>
</table>

*Note.* Using replaced values for missing third quarter data.

**Primary Analyses**

*Hypothesis 1. FCAT Prediction*

As described above, the first research question was approached using two different multiple regression analysis techniques, each for a different purpose—improved understanding and prediction. In the first model, the hypothesis regarding the relationship between student background variables, performance variables, and FCAT Reading achievement is tested with forced-order entry multiple regression. This approach allowed the examination of the strength of the relationship between the predictor variables and the criterion variable, highlighting unique contributions of each predictor variable to the overall explained variance of the model in specific variable blocks (to be discussed further). It was believed that information obtained from this analysis could help identify other variables that also contribute to FCAT success (purpose of improved understanding). Regarding the second model, stepwise multiple regression was used to identify the best subset of predictors of FCAT Reading performance among the available predictor variables (gender, SES, retention, attendance, class participation, Reading grades, NWF, and ORF). This model reflected the research goal of prediction using the most parsimonious subset of predictor variables.

*Multicollinearity.* It should be noted that it is erroneous to conclude that predictor variables that are not retained in the final model derived from a stepwise procedure are, in fact, inconsequential (Hair et al., 1998; Tate, 1998). It is often the case that the entire set of predictor variables...
variables includes some variables that are highly correlated with one another—a phenomenon labeled *multicollinearity* (Tate, 1998). When this happens, the criterion for inclusion and deletion in the stepwise model precludes the entry of highly correlated variables since the subsequent variable does not contribute a significant amount of *unique* variance (Hair et al., 1998; Tate, 1998). Therefore, in the search for the most effective and parsimonious model for prediction, variables may be ignored regardless of theoretical value (Tate, 1998). For this reason, both model approaches were used in this study and presented to fulfill both goals—a desire for understanding additional variables that contribute to one’s success on FCAT Reading (Model 1) as well as an efficient method for early identification for at-risk students (Model 2).

To examine the extent of multicollinearity among the predictor variables, the intercorrelations were computed. This correlation matrix is provided in Table 9. Correlations greater than .90 reflect substantial collinearity and correlations greater than .80 reflect relatively high levels of collinearity (Hair et al., 1998). In this particular set of predictor variables, no intercorrelations were greater than .90. The two administrations of each DIBELS subtest were relatively highly correlated (ORF #1 and ORF #2, \( r = .88 \); NWF #1 and NWF #2, \( r = .77 \)).

Given the strong potential for problems associated with high levels of multicollinearity, the severe reduction in sample size that occur with listwise deletion of variables missing any score, and the lack of a third interval administration which would allow a growth curve model approach, this researcher chose to utilize a single set of DIBELS scores. Alternatively, the use of a change score analysis (e.g., the difference score between ORF #1 and ORF #2) could have been an interesting approach in the analysis of the effects of NWF and ORF. In that type of analysis, the change (hopefully growth) in skill between the two interval administrations would become a separate predictor variable. However, only 118 of 181 participants (65.2%) had all four interval administration scores available for analysis. Therefore, this researcher chose to utilize the scores from a single interval administration.
Table 9

*Descriptive Statistics and Intercorrelations Among Variables*

<table>
<thead>
<tr>
<th></th>
<th>Gen</th>
<th>SES</th>
<th>Ret</th>
<th>Att</th>
<th>Par</th>
<th>RG</th>
<th>N1</th>
<th>O1</th>
<th>N2</th>
<th>O2</th>
<th>SSS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gen</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>SES</td>
<td>-.00</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Ret</td>
<td>.11</td>
<td>.27</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Att</td>
<td>.01</td>
<td>-.06</td>
<td>-.00</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Par</td>
<td>-.13</td>
<td>-.23</td>
<td>-.02</td>
<td>.01</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>RG</td>
<td>-.03</td>
<td>-.34</td>
<td>-.26</td>
<td>.10</td>
<td>.44</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>N1</td>
<td>.08</td>
<td>-.29</td>
<td>-.17</td>
<td>.01</td>
<td>.24</td>
<td>.40</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>O1</td>
<td>-.03</td>
<td>-.30</td>
<td>-.24</td>
<td>.05</td>
<td>.27</td>
<td>.52</td>
<td>.63</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>N2</td>
<td>.04</td>
<td>-.29</td>
<td>-.21</td>
<td>.07</td>
<td>.14</td>
<td>.37</td>
<td>.77</td>
<td>.63</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>O2</td>
<td>-.03</td>
<td>-.24</td>
<td>-.27</td>
<td>.04</td>
<td>.28</td>
<td>.51</td>
<td>.61</td>
<td>.88</td>
<td>.60</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>SSS</td>
<td>-.07</td>
<td>-.42</td>
<td>-.32</td>
<td>.03</td>
<td>.29</td>
<td>.64</td>
<td>.48</td>
<td>.70</td>
<td>.45</td>
<td>.68</td>
<td>—</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>M</th>
<th>SD</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>.40</td>
<td>.49</td>
<td>.35</td>
<td>.64</td>
<td>.95</td>
<td>.30</td>
<td>2.38</td>
<td>80.86</td>
<td>102.44</td>
<td>89.36</td>
<td>103.15</td>
</tr>
<tr>
<td></td>
<td>.64</td>
<td>.48</td>
<td>.95</td>
<td>.62</td>
<td>.62</td>
<td>.30</td>
<td>2.38</td>
<td>80.86</td>
<td>102.44</td>
<td>89.36</td>
<td>103.15</td>
</tr>
<tr>
<td></td>
<td>3.00</td>
<td>4.06</td>
<td>3.00</td>
<td>1.05</td>
<td>1.05</td>
<td>4.06</td>
<td>4.06</td>
<td>80.86</td>
<td>102.44</td>
<td>89.36</td>
<td>103.15</td>
</tr>
<tr>
<td></td>
<td>2.38</td>
<td>1.05</td>
<td>2.38</td>
<td>1.05</td>
<td>1.05</td>
<td>1.05</td>
<td>1.05</td>
<td>80.86</td>
<td>102.44</td>
<td>89.36</td>
<td>103.15</td>
</tr>
<tr>
<td></td>
<td>80.86</td>
<td>40.76</td>
<td>80.86</td>
<td>40.76</td>
<td>40.76</td>
<td>40.76</td>
<td>40.76</td>
<td>80.86</td>
<td>102.44</td>
<td>89.36</td>
<td>103.15</td>
</tr>
<tr>
<td></td>
<td>102.44</td>
<td>33.70</td>
<td>102.44</td>
<td>33.70</td>
<td>33.70</td>
<td>33.70</td>
<td>33.70</td>
<td>80.86</td>
<td>102.44</td>
<td>89.36</td>
<td>103.15</td>
</tr>
<tr>
<td></td>
<td>89.36</td>
<td>44.44</td>
<td>89.36</td>
<td>44.44</td>
<td>44.44</td>
<td>44.44</td>
<td>44.44</td>
<td>80.86</td>
<td>102.44</td>
<td>89.36</td>
<td>103.15</td>
</tr>
<tr>
<td></td>
<td>103.15</td>
<td>27.27</td>
<td>103.15</td>
<td>27.27</td>
<td>27.27</td>
<td>27.27</td>
<td>27.27</td>
<td>80.86</td>
<td>102.44</td>
<td>89.36</td>
<td>103.15</td>
</tr>
<tr>
<td></td>
<td>307.24</td>
<td>52.26</td>
<td>307.24</td>
<td>52.26</td>
<td>52.26</td>
<td>52.26</td>
<td>52.26</td>
<td>80.86</td>
<td>102.44</td>
<td>89.36</td>
<td>103.15</td>
</tr>
</tbody>
</table>

*Note.* Correlation is the Pearson product-moment correlation, expressed with two-tailed significance levels. Variables are coded as follows: 1 = gender; 2 = SES; 3 = number of grade retentions; 4 = attendance rate; 5 = participation rating; 6 = 3rd quarter Reading grade; 7 = NWF #1; 8 = ORF #1; 9 = NWF #2; 10 = ORF #2; 11 = Reading FCAT-SSS. Means and standard deviations obtained from correlation computation, which excluded cases pairwise.
In deciding between the two administrations, the ultimate goal of the research study was reconsidered—to provide educators with an efficient method of identifying students at risk for failing the Reading FCAT-SSS as early in the year as possible. The interval administration from December 2004 was the earliest data available. Therefore, this researcher utilized the ORF and NWF scores from the first administration, ORF #1 and NWF #1, in the remainder of the data analyses.

**Model 1**

*Analysis of outliers.* A case analysis was first conducted to evaluate the presence of potential outlier observations and observations that exerted excessive influence on the regression results. Outliers, defined as “observations with excessively large residuals”, are those fit very poorly by the resulting regression model (Tate, 1998, pp. 48-49). Existing outliers may or may not exert excessive influence on the results and both types of case analyses were conducted.

The presence of potential outliers was explored using the studentized residuals as a case index. As a rule of thumb, any studentized residual greater than 3.0 (absolute value) indicates a possible outlier (Tate, 1998). Given that rule of thumb, this data set included one outlier with a studentized residual of approximately –3.3. To further determine whether this observation is performing like an outlier, the residuals plot was inspected for separation from the remainder residual points (Tate, 1998). This statistically identified outlier did not appear to have a clear sense of separation from the rest of the observations via visual inspection of the residuals plot. Unfortunately, an in-depth study to explore why the Reading FCAT-SSS score for this student was fit so poorly by the model was not possible via interview or further record review. In fact, an in-depth study was not possible for any of the observations in any of the remaining analyses as well. Therefore, in all other respects this observation appeared to be legitimate.

To explore the potential for this outlier observation to be exerting excessive influence, an examination of the case indices that reflect the impact of individual observations on the regression coefficients, the “delta betas”, was conducted. This analysis indicated that at least one observation appeared to be exerting excessive influence on the estimated coefficients of gender, number of retentions, attendance rate, participation, and NWF. Subsequently, a sensitivity study was conducted by re-analyzing the data twice, once with the outlier and once without the outlier.
The resulting change in $R^2$ between the final models was unimpressive (change in $R^2 = .009$), indicating that the outlier did not exert undue influence on the model $R^2$. Thus, the observation was retained and the reported results reflect the inclusion of all participants for this hypothesis.

*Regression assumptions.* A scatter plot of the model residuals versus predicted outcomes was visually inspected to determine the presence of any violations of the regression assumptions (Tate, 1998). This graphical representation did not suggest any violations of the correct fit, constant variance, or normality assumptions required for the legitimacy of the regression results. In addition, there did not appear to be any conditions that might result in a violation of the independence assumption, especially given the exclusion of the second DIBELS scores. The reliabilities of the independent variables were strong enough to support the robustness of the multiple regression procedure to any minor violation of the assumption that all independent variables are known exactly.

*Forced-order entry structure.* Forced-order entry was utilized to better understand the individual and additive effects of each predictor variable. Blocks were created as follows. The first block incorporated fixed attributes—gender and SES. The second block was labeled “academic history” and was comprised of number of grade retentions and attendance rate. Block three consisted of the student’s participation in Reading instruction. Block four was defined by the Reading grade. The fifth block consisted of decoding skill as measured by NWF and the sixth block consisted of oral reading fluency as measured by ORF. Previous research noted in the literature review as well as personal theory drove the distinction between block configurations. Most importantly, due to the strong prediction power of ORF and FCAT performance established in the literature, ORF was entered last.

*Regression results.* The fixed attributes (gender and SES) block resulted in a model $R^2$ of 0.192, reflecting the overall strength of relationship between FCAT Reading success and the attributes variables ($F = 16.92$, df = 2, 142, $p < .001$). The adjusted $R^2$, which compensates for the positive bias in $R^2$, was 0.181, reflecting a relatively modest overall strength of relationship. The standard error of the estimate was 47.45.

The addition of the second block, academic history, resulted in the following model specifics. The model $R^2$ for the variables of gender, SES, number of retentions and attendance rate was 0.257, indicating a statistically significant relationship at the .0167 level between
Reading FCAT-SSS and these variables \( (F = 12.09, \text{df} = 4, 140, p < .001) \). The addition resulted in a change in \( R^2 \) of 0.064 and adjusted \( R^2 \) of 0.235. The standard error of the estimate was 45.85.

Next, the predictor variable of participation was added to the model for block three. The model \( R^2 \) for this block was statistically significant at 0.295 \( (F = 11.62, \text{df} = 5, 139, p < .001) \). By adding participation into the model at this point, the change in \( R^2 \) was 0.038 and the adjusted \( R^2 \) was 0.269. The standard error of the estimate was 44.82.

The fourth block included the addition of the Reading grade, as a measure of the teacher’s evaluation of the student’s achievement in Reading instruction. The addition of this predictor variable resulted in a model \( R^2 \) of 0.494 (change in \( R^2 \) was 0.199). This indicated a statistically significant relationship \( (F = 22.42, \text{df} = 6, 138, p < .001) \) with a corresponding adjusted \( R^2 \) of 0.472. The standard error of the estimate was 38.12.

NWF, a measure of student decoding ability, was added to create the fifth block. The model \( R^2 \) of 0.540 for this block was statistically significant \( (F = 22.95, \text{df} = 7, 137, p < .001) \). The change in \( R^2 \) with the addition of this block was 0.046 and the adjusted \( R^2 \) was 0.516. The standard error of the estimate was 36.47.

Lastly, ORF was added as the sixth and final block, resulting in a model with all predictor variables included. The new model \( R^2 \) with the inclusion of ORF was 0.633 (change in \( R^2 \) of 0.094) and was statistically significant \( (F = 29.35, \text{df} = 8, 136, p < .001) \). The new adjusted \( R^2 \) for the final model was 0.612 and the standard error of the estimate was 32.67.

At each block, the proportion of variability of the Reading FCAT-SSS scores explained by the predictor variables included in that block increased. The greatest increase occurred with the addition of the Reading grade \( (R^2 \text{ change} = 0.199) \). The effects of the individual predictor variables on FCAT Reading achievement are summarized by looking at the individual pairwise correlations between each predictor variable and the dependent variable as was provided in Table 9.

Examining the unstandardized regression coefficients at each block entry provides an alternate perspective. This perspective includes a test of the coefficients. The results of this analysis are provided in Table 10 with each value appropriately interpreted as the effect of that variable, controlling for the remaining variables (Tate, 1998).
Table 10

Regression Results Summary for Hypothesis 1, Model 1—Final Model

<table>
<thead>
<tr>
<th>Effect</th>
<th>Estimate</th>
<th>95% Confidence Interval</th>
<th>Standardized Coefficient</th>
<th>ΔR²</th>
</tr>
</thead>
<tbody>
<tr>
<td>Block 1</td>
<td></td>
<td></td>
<td></td>
<td>0.192*</td>
</tr>
<tr>
<td>Gender</td>
<td>-1.716</td>
<td>-13.406, 9.975</td>
<td>-0.016</td>
<td></td>
</tr>
<tr>
<td>SES</td>
<td>-15.923*</td>
<td>-28.171, -3.675</td>
<td>-0.149</td>
<td></td>
</tr>
<tr>
<td>Block 2</td>
<td></td>
<td></td>
<td></td>
<td>0.064*</td>
</tr>
<tr>
<td>Retentions</td>
<td>-7.345</td>
<td>-17.274, 2.583</td>
<td>-0.084</td>
<td></td>
</tr>
<tr>
<td>Attendance rate</td>
<td>2.799</td>
<td>-117.999, 123.596</td>
<td>0.002</td>
<td></td>
</tr>
<tr>
<td>Block 3</td>
<td></td>
<td></td>
<td></td>
<td>0.038*</td>
</tr>
<tr>
<td>Participation</td>
<td>-3.228</td>
<td>-10.310, 3.854</td>
<td>-0.055</td>
<td></td>
</tr>
<tr>
<td>Block 4</td>
<td></td>
<td></td>
<td></td>
<td>0.199*</td>
</tr>
<tr>
<td>Reading grade</td>
<td>18.251*</td>
<td>11.741, 24.760</td>
<td>0.377</td>
<td></td>
</tr>
<tr>
<td>Block 5</td>
<td></td>
<td></td>
<td></td>
<td>0.046*</td>
</tr>
<tr>
<td>NWF</td>
<td>0.030</td>
<td>-0.142, 0.203</td>
<td>0.024</td>
<td></td>
</tr>
<tr>
<td>Block 6</td>
<td></td>
<td></td>
<td></td>
<td>0.094*</td>
</tr>
<tr>
<td>ORF</td>
<td>.671*</td>
<td>0.445, 0.896</td>
<td>0.429</td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>212.144</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. The point estimates for each predictor variable are based on entry of Block 6 with all predictor variables included. The ΔR² result is the increase in R² due to adding each block incrementally, given the other predictor variables.

* p < .0167

In summary, the first block reflected the significant effect of SES on Reading FCAT-SSS scores (BSES = -46.34, t = -5.74, p < .001). The second block revealed an added significant effect of number of retentions (Bret = -23.24, t = -3.48, p = .001) and retained the significant effect of SES (BSES = -39.31, t = -4.87, p < .001). In the third block, the effect estimate of participation was statistically significant (Bpart = 11.97, t = 2.74, p = .007); the effects of SES and number of...
retentions remained statistically significant in this block of the model ($B_{\text{SES}} = -33.25, t = -4.06, p < .001; B_{\text{ret}} = -24.58, t = -3.75, p < .001$). Addition of Reading grade in the fourth block resulted in the loss of number of retentions and participation from statistical significance. However, the significance of the effect of SES was retained and the effect of Reading grade was added ($B_{\text{SES}} = -22.32, t = -3.13, p = .002; B_{\text{grd}} = 26.45, t = 7.36, p < .001$). The fifth block included the addition of NWF, which was statistically significant in addition to SES and Reading grade ($B_{\text{NWF}} = 0.31, t = 3.70, p < .001; B_{\text{SES}} = -18.52, t = -2.69, p = .008; B_{\text{grd}} = 22.97, t = 6.45, p < .001$). Lastly, the effect of ORF was evident via its addition in the final block. This final block now reflects all predictor variables. In this final block, NWF no longer contributed a significant effect. However, SES and Reading grade remained significant ($B_{\text{SES}} = -15.92, t = -2.57, p = .011; B_{\text{grd}} = 18.25, t = 5.55, p < .001$) and ORF produced a significant individual effect ($B_{\text{ORF}} = 0.67, t = 5.89, p < .001$). It is important to again note that significance at this level of examination means that the variable is producing a significant effect on Reading FCAT-SSS achievement, holding all other predictor variables constant.

Model 1, therefore, contributes to understanding of the effects of these additional predictor variables on FCAT Reading achievement. Of particular interest, the teachers’ Reading grade-letter evaluations appear to be highly predictive of FCAT reading performance, an effect not noted in current FCAT prediction literature. Additionally, NWF does appear to be a significant predictor of Reading FCAT-SSS scores in isolation from ORF; however, when ORF is added to the model, NWF no longer produces a significant individual effect. This model also highlights the deleterious effect of student economic disadvantage on FCAT Reading success. On the other hand, it is of ultimate interest by practitioners to be able to efficiently predict student FCAT achievement as early in the year as possible, using the minimum number of success indicators. Thus, Model 2 is presented.

Model 2

Model 2 was a statistical variation of Model 1. Whereas in the first model each predictor variable was entered in blocks to determine individual effects, in Model 2 all of the predictor variables were simultaneously considered for entry using the stepwise regression analysis technique. The stepwise regression technique produced a regression variate that included only
the subset of predictor variables that best explained the variability in Reading FCAT-SSS scores in the most parsimonious prediction equation (Hair et al., 1998; Tate, 1998).

**Analysis of outliers.** To begin, a case analysis was conducted once again to identify the presence of any potential outliers in this new regression model. One observation, in fact, the same observation as in Model 1, reflected a studentized residual greater than 3.0 (absolute value). This observation had a studentized residual of –3.6. The residuals plot was inspected for separation from the remainder residual points to explore whether it appeared as an outlier. This observation did not have a clear sense of separation from the rest of the observations via visual inspection of the residuals plot. As before, this observation appeared to be legitimate.

To explore the potential for this observation to be exerting excessive influence, the delta betas were examined. The exploration of the delta betas indicated that no observations appeared to be exerting excessive influence on the estimated coefficients. Therefore, a sensitivity study was not needed. Thus, the observation was retained and the results reflected the inclusion of all participants for this hypothesis.

**Regression assumptions.** A scatter plot of the model residuals versus predicted outcomes was visually inspected to determine the presence of any violations of the regression assumptions. This graphical representation did not suggest any violations of the correct fit, constant variance, or normality assumptions required for the legitimacy of the regression results. In addition, there did not appear to be any conditions that might result in a violation of the independence assumption. As stated previously, the reliabilities of the independent variables were strong enough to support the robustness of the multiple regression procedure to any minor violation of the assumption that all independent variables are known exactly.

**Regression results.** The stepwise procedure was then employed as planned. This researcher incorporated the commonly used criteria for entering and removing variables from the model (p-value of 0.05 to enter, p-value 0.10 to remove) (Tate, 1998). The final stepwise regression model resulted from the following statistical steps.

**Step 1:** The predictor variable with the largest bivariate correlation and associated $R^2$ was ORF ($r = 0.69$, $R^2 = 0.474$, $F = 128.90$, df = 1, 143, $p < .001$). The associated $p$-value of less than 0.001 satisfied the “$p$ to enter” criterion of 0.05. Therefore, ORF was entered as the first variable in the model.
Step 2: Of the remaining predictor variables (gender, SES, retentions, attendance rate, participation, Reading grade, and NWF), Reading grade was associated with the next largest increase in $R^2$, given ORF. This was revealed by the largest $t$-value of 6.76. Based on the associated $p$-value of less than 0.001, Reading grade was entered into the model. Once Reading grade was entered, the effect of ORF was re-evaluated. The $p$-value of 0.001 compared to the “$p$ to remove” of 0.10 resulted in the decision to retain ORF in the model.

Step 3: Considering the remaining predictor variables, SES then provided the next largest increase in $R^2$ (as reflected by the largest $t$-value of -2.80). The associated $p$-value of 0.006 compared to 0.05 results in the decision to enter SES into the model. Once entered, the unique effects of the two previously entered variables (ORF and Reading grade) were re-examined. The associated values of less than 0.001 for each resulted in both variables being retained.

Step 4: The variable of retentions was next to be considered for entry. However, with the associated $p$-value of –0.094, this variable did not meet the criteria for “$p$ to enter” of 0.05. Thus, the variable retentions was not retained in the final model. The final model, therefore, was comprised of ORF, Reading grade, and SES (the same variables found to be significant in the final block of Model 1). The results of this final model are summarized in Table 11.

The model $R^2$ was 0.623 and was statistically significant at the 0.0167 level ($F = 77.69$, df = 3, 141, $p < .001$). The adjusted $R^2$, compensating for the positive bias in $R^2$, was 0.615. This indicated a strong overall strength of relationship. The standard error of estimate was 32.53. Interestingly, correcting for attenuation that occurs due to imperfect measures resulted in an attenuated $R$ ($R_{\text{att}}$) of 0.80 (McNemar, 1969). Interpretations of the individual effects of each predictor variable follow.

The positive effect of ORF performance on FCAT Reading achievement, significant at the 0.0167 level, reflects an estimated increase of 0.71 Reading FCAT-SSS score points for every score increase in ORF, controlling for Reading grade and SES. A common rule of thumb for determining practical importance of a statistically significant interval predictor variable is to compare the standardized regression coefficient to 0.1 (Tate, 1998). The corresponding threshold for a raw score coefficient is determined by dividing the standard deviation of the dependent variable by the standard deviation of the predictor variable in question and then multiplying by 0.1.
### Table 11

*Stepwise Regression Results Summary for Hypothesis 1, Model 2—Complete Model*

<table>
<thead>
<tr>
<th>Variables</th>
<th>Effect Estimate</th>
<th>95% Confidence Interval</th>
<th>Standardized Coefficient</th>
<th>$\Delta R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Step 1</strong></td>
<td></td>
<td></td>
<td></td>
<td>0.474</td>
</tr>
<tr>
<td>Constant</td>
<td>197.385</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ORF</td>
<td>1.076</td>
<td>0.889, 1.264</td>
<td>0.689</td>
<td></td>
</tr>
<tr>
<td><strong>Step 2</strong></td>
<td></td>
<td></td>
<td></td>
<td>0.128</td>
</tr>
<tr>
<td>Constant</td>
<td>181.920</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ORF</td>
<td>0.759</td>
<td>0.571, 0.947</td>
<td>0.485</td>
<td></td>
</tr>
<tr>
<td>Reading grade</td>
<td>19.908</td>
<td>14.087, 25.728</td>
<td>0.412</td>
<td></td>
</tr>
<tr>
<td><strong>Step 3</strong></td>
<td></td>
<td></td>
<td></td>
<td>0.021</td>
</tr>
<tr>
<td>Constant</td>
<td>201.703</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ORF</td>
<td>0.713</td>
<td>0.527, 0.900</td>
<td>0.456</td>
<td></td>
</tr>
<tr>
<td>Reading grade</td>
<td>17.877</td>
<td>12.013, 23.741</td>
<td>0.370</td>
<td></td>
</tr>
<tr>
<td>SES</td>
<td>-16.823</td>
<td>-28.714, -4.933</td>
<td>-0.157</td>
<td></td>
</tr>
</tbody>
</table>

For ORF, the calculation is as follows: $(52.43 / 33.54)(0.1) = 0.16$. It should be noted here that the standard deviation of the Reading FCAT-SSS is slightly different in this regression analysis (52.43) than that reported in the descriptive statistics for the whole sample (52.26). This is common when one executes multiple analyses that incorporate listwise deletion of cases with any missing data. All points in the 95% confidence interval for ORF in Table 11 were greater in magnitude than this threshold (0.16), lending confidence that the population effect of ORF on Reading FCAT-SSS is not only statistically significant, but also of practical importance. The unique contribution of ORF to the model $R^2$ ($\Delta R^2$ in the table) was 0.474.

As defined by its inclusion in the final stepwise multiple regression model, the effect of Reading grade on Reading FCAT-SSS was also statistically significant. Using the point estimates provided in Table 11, the effect of Reading grade is interpreted as follows. There was an estimated increase of 17.88 points in Reading FCAT-SSS associated with each letter grade
improvement in Reading, controlling for ORF and SES. The practical importance of the effect of Reading grade was likewise considered. The raw score coefficient threshold was computed to be \((52.43 / 1.08) (0.1) = 4.85\). All points in the confidence interval for Reading grade were greater in magnitude than the calculated threshold of practical importance. Thus, the effect of Reading grade was considered to be not only statistically significant, but also of practical importance. The unique contribution of Reading grade to the model \(R^2\) was 0.128.

Lastly, SES was also found to be statistically significant. The predictor variable of SES is actually a contrast of those receiving free or reduced-lunch benefits (economically disadvantaged) and those not receiving benefits (not economically disadvantaged). The results of this analysis indicated that economically disadvantaged students scored an estimated 16.82 points lower than those without disadvantage on Reading FCAT-SSS, controlling for ORF and Reading grade. The threshold of practical importance for a contrast is calculated differently than above. The contrast threshold is approximately 0.3 times the standard deviation of the dependent variable (Reading FCAT-SSS scores in this hypothesis) (Tate, 1998). In the present analysis, the calculation of the threshold for practical importance is as follows: \((0.3) (52.43) = 15.73\). Comparing this threshold to the confidence interval for SES, suggested that the effect of SES was inconclusive with respect to practical importance. The unique contribution of SES to the model \(R^2\) was 0.021.

Given that Model 2 more closely reflected the ultimate goal of the present study, the results of this model were considered further. The final regression equation can be utilized as a prediction equation for decisions about individuals if there exists an adequate level of precision of prediction for the sample (Tate, 1998). The prediction equation for the final model was determined by the coefficients displayed in Table 11. Certainly, a perfect prediction of FCAT Reading success was not achieved; therefore the next step is to estimate the range of predicted values that may be expected when using the prediction equation. To determine whether the prediction equation provided a sufficient precision of prediction of Reading FCAT-SSS achievement, the following analysis was conducted.

The half-width of the prediction interval is frequently used to represent the precision of prediction (Tate, 1998). Tate suggests that the researcher decides the level of precision that is desirable for prediction. For this study, the desired precision of the prediction interval width was
set at 30% of the range of Reading FCAT-SSS. The half-widths can be calculated in either raw or standardized score variations. In this study, the half-width of the 95% prediction interval was equal to \( \pm (1.96)(32.531) \), or \( \pm 63.76 \) units in raw score form and \( \pm (1.96)(32.531 / 52.432) \), or \( \pm 1.22 \) standard deviation units in standardized form. Additionally, the ratio of the width of the prediction interval to the approximate range of Reading FCAT-SSS scores was 0.62, a value larger than a target range of near 0.3 or less (Tate, 1998). Therefore, the precision of future predictions of individual outcomes was not acceptable using the prediction equation derived herein. Prediction accuracy is explored using a different perspective in Hypothesis 3.

**Hypothesis 2. ORF Prediction**

Both in the literature base and in the present study, ORF was found to be a statistically significant and practically important predictor of Reading FCAT-SSS. This researcher was interested to see which, if any, of the available predictor variables were significant predictors of ORF performance. The remaining set of predictor variables (gender, SES, retentions, attendance rate, class participation, Reading grades, and NWF performance) was used to predict ORF performance. The significance level was set to 0.0167, as discussed previously, and the simultaneous form of multiple regression (without forced-order entry) was employed to determine the relationships therein.

**Analysis of outliers.** Exploring the potential for outliers, the studentized residuals were examined. The greatest studentized residual was –2.836, below the 3.0 (absolute value) cutoff used in this study to identify outliers. A follow-up analysis of the delta betas and a sensitivity study were, therefore, not necessary.

**Regression assumptions.** The plot of the model residuals versus the predicted outcomes was visually inspected to ascertain whether there existed any violations of the regression assumptions. This inspection revealed that there did not appear to be any violations of the assumptions of correct fit, constant variance, or normality. As previously mentioned, there was also no evidence to suggest that the assumption of independence of predictor variables was violated.

**Regression results.** The resulting model \( R^2 \) of 0.492 suggested a moderately strong relationship between the predictor variables (gender, SES, retentions, attendance rate, class
participation, Reading grades, and NWF performance) and ORF. This $R^2$ was statistically significant at the 0.0167 level ($F = 18.94$, df = 7, 137, $p < .001$). The adjusted $R^2$ of 0.466 provided a relatively unbiased estimate of the model $R^2$ and similarly reflected a moderately strong relationship. The standard error of the estimate was 24.51.

The effects of the individual predictor variables on ORF were presented for the reader in summary in Table 12. Specifically, the effects of gender, SES, number of retentions, attendance rate, and participation were not statistically significant at the 0.0167 level. However, Reading grade and NWF were significant predictors of ORF performance.

Table 12

<table>
<thead>
<tr>
<th>Effect</th>
<th>Estimate</th>
<th>SE of the Estimate</th>
<th>95% Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>91.015</td>
<td>45.065</td>
<td></td>
</tr>
<tr>
<td>Gender</td>
<td>-6.193</td>
<td>4.403</td>
<td>-14.900, 2.515</td>
</tr>
<tr>
<td>SES</td>
<td>-3.872</td>
<td>4.635</td>
<td>-13.037, 5.293</td>
</tr>
<tr>
<td>Retentions</td>
<td>-5.846</td>
<td>3.734</td>
<td>-13.229, 1.537</td>
</tr>
<tr>
<td>Attendance rate</td>
<td>-33.890</td>
<td>45.738</td>
<td>-124.333, 56.553</td>
</tr>
<tr>
<td>Participation</td>
<td>0.010</td>
<td>2.687</td>
<td>-5.215, 5.411</td>
</tr>
<tr>
<td>Reading grade</td>
<td>7.034*</td>
<td>2.395</td>
<td>2.298, 11.771</td>
</tr>
<tr>
<td>NWF</td>
<td>.409*</td>
<td>.055</td>
<td>.300, .519</td>
</tr>
</tbody>
</table>

* $p < .0167$

The effect of Reading grade was statistically significant, indicating a 7.03 score increase in ORF associated with each Reading grade letter improvement, controlling for the other predictor variables. Exploring practical significance, the raw score coefficient threshold for Reading grade was $(33.54 / 1.08)(0.1) = 3.11$. Many, but not all, points in the confidence
interval for Reading grade were greater in magnitude than this threshold. Therefore, the assessment of the practical significance of the effect of Reading grade on ORF is inconclusive. However, so few points overlap that the effect of Reading grade on ORF is likely of practical importance.

The effect of NWF on ORF performance was also statistically significant at the 0.0167 level. This effect indicated that there was a 0.41 increase in ORF score associated with a one point increase in NWF score, controlling for the other predictor variables. The raw score threshold for NWF was \((33.54 / 41.98)(0.1) = 0.08\). All points in the associated confidence interval for NWF were greater in magnitude than this threshold. It is therefore concluded that the population effect of NWF on ORF is not only statistically significant, but also of practical importance.

The present hypothesis was posited to improve understanding of the relationships of the variables considered. It was not the goal to use the derived variate as a measure to predict ORF given that ORF itself is a very efficient and effective assessment tool. Thus, the precision of predictions of individual outcomes was not explored.

**Hypothesis 3. Accuracy of FCAT Pass or Fail prediction**

FCAT Reading achievement is reported as a scaled score, which is then transformed into a Level score designated by FDOE. The FCAT Level score ranges from Level 1 through Level 5, according to the respective Reading FCAT-SSS score. The Reading FCAT-SSS score ranges for each level during the 2005 administrations are as follows: Level 1 = 100 to 258; Level 2 = 259 to 283; Level 3 = 284 to 331; Level 4 = 332 to 393; and Level 5 = 394 to 500 (FDOE, 2005c). Passing scores on the FCAT are those that are Level 3 or above. Failing scores are those that fall within Level 1 or 2. The number of students and percentage of the sample scoring at each Level are reported in Table 13. Fifty-five (30.4%) of participants in this study failed the Reading FCAT-SSS as compared to 33% of students statewide in 2005.
Table 13

*Frequency and Proportion of Reading FCAT-Level Scores*

<table>
<thead>
<tr>
<th>FCAT Level</th>
<th>Sample n</th>
<th>Sample %</th>
<th>State %</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>29</td>
<td>16.0</td>
<td>20</td>
</tr>
<tr>
<td>2</td>
<td>26</td>
<td>14.4</td>
<td>13</td>
</tr>
<tr>
<td>3</td>
<td>65</td>
<td>35.9</td>
<td>33</td>
</tr>
<tr>
<td>4</td>
<td>55</td>
<td>30.4</td>
<td>28</td>
</tr>
<tr>
<td>5</td>
<td>6</td>
<td>3.3</td>
<td>6</td>
</tr>
</tbody>
</table>

The final hypothesis of this study reflected an interest in determining the accuracy of pass or fail predictions of FCAT Reading achievement using ORF cutpoints found in the literature base. This study examined two sets of cutpoints proposed by researchers (see Appendix D). First, Good and colleagues (2002) used ORF scores to classify students into three risk categories (low, moderate, and high risk) for achieving below grade level performance on a measure of broad reading achievement. Alternatively, FCRR published a DIBELS (ORF) Risk Levels Chart (2004) to assist educators in determining risk for scoring below grade level in classroom reading-based activities. It is important to note that neither of these risk models were derived with the intent of predicting FCAT success or failure. Nonetheless, the models were intended to provide educators with a useful tool to identify students in need of additional reading instruction and intervention. These cutpoints are powerful graphical tools for teachers to use at a glance to determine the need for additional instruction or intervention for students in their classroom, based upon ORF performance, throughout the year.

This researcher sought to explore the accuracy of these benchmark resources. Using ORF scores from the first administration, three months prior to FCAT, students were classified into risk groups congruent with cutpoints proposed by each group of researchers in turn. The proportion of students at each risk level that failed the FCAT was then compared to both sets of cutpoints to determine the accuracy of those benchmark guidelines ($n = 158$). A summary of the proportion of students that failed the Reading FCAT-SSS for each risk model is reported in Table 14. Detailed tables for each model then follow.
Table 14

Comparison of Proportion of Risk Group Failures on Reading FCAT-SSS

<table>
<thead>
<tr>
<th></th>
<th>High Risk</th>
<th>Moderate Risk</th>
<th>Low Risk</th>
<th>Above Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Good et al.</td>
<td>73.2 %</td>
<td>30.0 %</td>
<td>7.5 %</td>
<td>—</td>
</tr>
<tr>
<td>FCRR</td>
<td>86.7 %</td>
<td>55.0 %</td>
<td>28.6 %</td>
<td>7.4 %</td>
</tr>
</tbody>
</table>

In summary, those students that were classified at high risk on the Good et al. (2002) and the FCRR (2004) risk models, the reported frequency of students that failed the Reading FCAT-SSS represent “true positives” (73.2% ad 86.7%, respectively). Those students that were classified at low risk or above average (FCRR model only), the proportion of students that failed represent “false negatives”.

Good et al., cutpoints. First, the prediction accuracy of cutpoints proposed by Good and colleagues (2002) was explored. Good et al. theorized that students with ORF scores at or above 110 correct words per minute (cwpm) were at low risk for below-grade level performance on a broad measure of reading comprehension. Students considered at moderate risk were those that scored between 80 and 109 cwpm. High-risk students were characterized as those that scored below 80 cwpm on the ORF measure. Good and colleagues did not establish or define an “above average” group as in the FCRR (2004) model. The number and percentage of students that scored at each Level of the Reading FCAT-SSS for each risk category proposed Good et al. is reported in Table 15.

Thirty students that were classified by Good et al., cutpoints as high-risk performers failed the Reading FCAT-SSS (73.2% of the high-risk group). The moderate risk group reflected those students who theoretically may be at some risk for low achievement. Among this sample, 15 students classified as moderate risk (30% of all moderate-risk students) did, in fact, fail the Reading portion. Only 5 students classified as low-risk performers (7.5% of all low-risk students) failed the Reading portion. From a slightly different perspective, the greatest number of students that failed the Reading portion were, in fact, classified in the high-risk group by these cutpoints (60.0%). Thirty percent of students that failed the Reading portion were classified as
moderate risk and 10% of those that failed were classified as low risk. A chi-square analysis was conducted to determine if the probability of students passing or failing the FCAT occurred at chance. This analysis revealed that the proportion of students at each level was significantly different than what would be expected given chance ($\chi^2 = 50.86$, df = 2, $p < .001$).

### Table 15
*Good et al. (2002) Risk Group Failures per Reading FCAT Level Score*

<table>
<thead>
<tr>
<th>Reading FCAT Level</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>High Risk (0 – 79 cwpm)</td>
<td>21</td>
<td>9</td>
<td>9</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Mod. Risk (80 – 109 cwpm)</td>
<td>4</td>
<td>11</td>
<td>22</td>
<td>12</td>
<td>1</td>
</tr>
<tr>
<td>Low Risk (110+ cwpm)</td>
<td>1</td>
<td>4</td>
<td>25</td>
<td>32</td>
<td>5</td>
</tr>
</tbody>
</table>

**FCRR cutpoints.** The FCRR DIBELS Risk Levels (2004) chart provides similar DIBELS-based subtest cutpoints. The FCRR DIBELS Risk Levels Chart is provided for educators to assist in score interpretation for student performance on DIBELS measures, including ORF. This chart classified students into four risk categories based upon ORF scores. The risk categories included: high risk, moderate risk, low risk, and above average. High-risk performance was defined as “seriously below grade level and in need of substantial intervention” (FCRR, 2004). Moderate risk indicated performance “moderately below grade level and in need of additional intervention”. Students in the low-risk category were those that were predicted to
be performing at grade level whereas students in the above average category were those that were predicted to be performing at or above the 60\textsuperscript{th} percentile on those measures. This researcher compared the FCRR cutpoints to sample statistics to determine accuracy of FCAT failure predictions in this study.

ORF scores obtained in the second academic quarter in this data set were compared to the FCRR DIBELS Risk Levels Chart for the second quartile 2004-2005, congruent with the ORF #1 administration period (FCRR, 2004). The number and percentage of students that scored at each Level of the Reading FCAT-SSS for each risk category proposed FCRR is reported in Table 16.

Table 16

\textit{FCRR (2004) Risk Group Failures per Reading FCAT Level Score}

<table>
<thead>
<tr>
<th>Reading FCAT Level</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>High Risk (0 – 61 cwpm)</td>
<td>11</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Mod. Risk (62 – 87 cwpm)</td>
<td>12</td>
<td>10</td>
<td>13</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>Low Risk (88 – 105 cwpm)</td>
<td>2</td>
<td>8</td>
<td>16</td>
<td>9</td>
<td>0</td>
</tr>
<tr>
<td>Above Average (106 + cwpm)</td>
<td>1</td>
<td>4</td>
<td>25</td>
<td>32</td>
<td>6</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Reading FCAT Level</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>High Risk (0 – 61 cwpm)</td>
<td>73.3 %</td>
<td>13.3 %</td>
<td>13.3 %</td>
<td>0 %</td>
<td>0 %</td>
</tr>
<tr>
<td>Mod. Risk (62 – 87 cwpm)</td>
<td>30.0 %</td>
<td>25.0 %</td>
<td>32.5 %</td>
<td>12.5 %</td>
<td>0 %</td>
</tr>
<tr>
<td>Low Risk (88 – 105 cwpm)</td>
<td>5.7 %</td>
<td>22.9 %</td>
<td>45.7 %</td>
<td>25.7 %</td>
<td>0 %</td>
</tr>
<tr>
<td>Above Average (106 + cwpm)</td>
<td>1.5 %</td>
<td>5.9 %</td>
<td>36.8 %</td>
<td>47.1 %</td>
<td>8.8 %</td>
</tr>
</tbody>
</table>
Of those students in the high-risk group, 86.6% ultimately failed the Reading FCAT-SSS (26.0% of the total number of participants that failed). FCAT success was more variable among the moderate risk ORF achievement group with 55.0% failing (44.0% of the total number of participants that failed). Alarmingly, of participants classified as low-risk based upon ORF performance in this study, 28.6% (almost 1/3) failed the Reading FCAT-SSS (20.0% of the total number of failures in this sample). Likewise, the above average group also contained participants that failed the Reading FCAT-SSS (7.4% of the group failed accounting for 10.0% of the total number of failures). Chi-square analysis confirmed that these proportions were not likely due to chance ($\chi^2 = 49.78$, df = 3, $p < .001$).

Comparison of Models. A table comparison of the accuracy of the failure predictions for each model was presented in Table 14. In summary, the FCRR (2004) cutpoints for high-risk students more accurately predicted failure on the Reading FCAT-SSS than the Good et al. cutpoints (2002) (86.7% FCRR high-risk students failed versus 73.2% of Good et al. high-risk students). Of the moderate risk groups, students were almost equally likely to fail or pass the Reading FCAT-SSS using the FCRR cutpoints (55.0%) whereas only 30.0% of the moderate risk students in the Good and colleagues’ cutpoints failed. The low risk group for the Good et al. cutpoints more closely resembled the above average group for the FCRR cutpoints (7.5% Good et al. low-risk students failed versus 7.4% FCRR above average students failed). Lastly, FCRR also incorporated a low-risk group that more closely approximated the moderate risk group of Good et al.; 28.6% of the students in FCRR’s low-risk group failed the Reading FCAT-SSS while 30.0% of the Good et al. moderate-risk students failed. The implications of this comparison will be discussed further in the following chapter.
CHAPTER 5
DISCUSSION

Summary

This research study examined a variety of student demographic, background, and performance variables, exploring their relationship with student reading achievement on a comprehensive statewide achievement test, the FCAT. Due to the widespread impact of student performance on this measure, educators, parents, and the public have grown increasingly interested in predicting performance in an effort to identify those students that may be at risk for failure. Ultimately, students that are identified as at-risk for failure on the FCAT may benefit from appropriately timed and administered supplementary instruction and intervention. If initiated early enough in the academic year, this intervention may help to ameliorate academic deficits and improve FCAT performance.

Research on the predictability of FCAT achievement is still in its infancy. Previous research has shown ORF to be a strong predictor of general reading achievement and a limited number of studies have shown ORF to be a relatively strong predictor of Reading FCAT-SSS scores (e.g., Buck & Torgesen, 2003). Current research efforts are needed to replicate earlier studies as well as to improve the overall prediction model. Therefore, in the present study additional student-based variables in the form of demographic, historical, and performance variables were added to explore the utility of adding these predictor variables to the prediction equation. Specifically, student demographic (gender, SES), academic history (number of retentions, attendance rate), and performance variables (participation, Reading grade, decoding skill via NWF, and oral reading fluency via ORF) were incorporated into the present study to explore the impact of each on the Reading FCAT-SSS prediction model. NWF and ORF were both administered three months prior to
Next, the predictor variables were then used to identify their relationships with ORF to determine which, if any, helped to explain the variance in this highly effective predictor variable. Lastly, models for reading risk in the literature base were compared to ORF scores in this data set to determine the overall utility of a risk model in identifying students at risk for FCAT failure.

First, this study explored the relationships among the predictor variables described above and Reading FCAT-SSS achievement. Secondly, given that oral reading fluency (ORF) has been shown to be an efficient and effective method for predicting student achievement on the FCAT, it was of interest to determine which, if any, of the predictor variables may help to predict ORF performance. Lastly, this researcher compared the ORF cutpoints proposed by Good and colleagues (2002) and FCRR (2004) to determine the prediction accuracy of those cutpoint models using scores from this sample.

Prediction of Reading FCAT-SSS achievement was conducted in two ways. In the first model, the predictor variables were added into the regression equation in a forced-order entry with designated blocks to determine the unique effects of the variables. ORF was entered last due to its known strong predictive power as evidenced by previous research. Block configuration was entered in the following sequence: demographics / fixed attributes (gender, SES), academic history (number of prior grade retentions, attendance rate), participation in Reading activities (participation), Reading grade (teacher’s evaluation of Reading performance via third quarter Reading grade), word decoding skill (NWF score), and oral reading fluency (ORF score).

The first analysis revealed that each block added significant, unique variance to the model. In other words, the proportion of variability of Reading FCAT-SSS explained by the model improved with the addition of each block of predictor variables. Specifically, within the first block (fixed attributes), gender was not found to be statistically significant whereas SES was significantly related to Reading FCAT-SSS achievement. By adding academic history variables to the model (number of prior grade retentions and attendance rate) it was found that retentions and SES were significant while gender and attendance rate were not. With the addition of the participation block, the effect of participation was significant and the significant effects of SES and retentions were retained. Gender and attendance rate remained not statistically significant. The significant effect of participation and retentions was then eliminated with the addition of
Reading grade to the model, leaving Reading grade and SES as significant effects. In the fifth block entry NWF was added and found to be statistically significant in addition to Reading grade and SES. Lastly, ORF was added to the model. With the addition of ORF, NWF was found to no longer produce a significant effect although Reading grade and SES remained significant. In summary, this model demonstrated the significant effects of ORF, Reading grade, and SES with respect to Reading FCAT-SSS achievement. Specifically, ORF scores and Reading grade had a significant positive effect on Reading FCAT-SSS; students with economic disadvantage (as defined by receipt of free or reduced lunch benefits) were found to perform significantly worse on Reading FCAT-SSS. The variables NWF, gender, attendance rate, participation, and number of retentions were not found to be statistically significant predictors of Reading FCAT-SSS achievement.

In particular, the addition of Reading grade to the first model produced the largest increase in explained variance (i.e., largest increase in $R^2$). Prior to the addition of Reading grade, the variables of participation and grade retentions were significantly related to Reading FCAT-SSS achievement. When Reading grade was added to the model, the statistical significance of these variables was eliminated. One explanation for this may be that both participation in Reading activities and the number of prior grade retentions have significant effects on FCAT success, however, the teacher’s evaluation of the student’s Reading ability may subsume or be somewhat based upon student participation and other attributes that were linked to prior grade retention as well (e.g., persistent learning difficulties, persistent lack of assistance at home, continued lack of appropriate instructional modality or materials for that student). In fact, the correlation between Reading grade and participation in this study was moderately positively correlated ($r = 0.44$).

The relationship between students’ Reading grades and Reading FCAT-SSS success had not previously been explored in the literature and was an unexpected finding in this study. However, previous research regarding accuracy of teachers’ predictions of reading skill indicates that teachers are good predictors of a variety of reading skills. Feinberg and Shapiro (2003) demonstrated a strong correlation between student oral reading fluency skill (as measured by curriculum-based measurement) and teacher’s prediction of oral reading fluency rate ($r = .70$). Additionally, Bates and Nettelbeck (2001) examined the accuracy of teacher judgments in
reading achievement among students with and without classroom behavioral problems. Those researchers hypothesized that the presence of classroom behavior problems would negatively impact teachers’ evaluations of reading ability. Bates and Nettelbeck found that teachers remained accurate judges of reading accuracy ($r = .77$) and reading comprehension ($r = .62$), despite the presence of classroom behavior problems. Although these students tended to perform more poorly on the reading measures, the teachers did not underestimate reading skill. The results of the present study add to the body of literature that identify teachers as good predictors and evaluators of reading skill and achievement. This suggests that teachers’ early evaluations of students may be regarded as good predictors of later FCAT failure; educators and parents of students who receive low Reading grades in the first quarter of the academic year should be concerned about the potential for poor student FCAT performance four months later.

The next largest increase in explained variance derived from adding ORF to the model. Interestingly, NWF did appear to be a significant contributor to the explained variance of the model prior to entering ORF; however, once ORF was entered, NWF was no longer a significant contributor. This suggests that although NWF explains a portion of the variability in Reading FCAT-SSS scores, it is not variance that is unique from the effect of ORF. Therefore, in choosing a screening tool for FCAT performance, educators would be well-advised to choose ORF over NWF if using only one of the two measures. In fact, FCRR does not provide NWF benchmarks for third grade students, and these results support that decision. As ORF continues to be supported as a strong predictor of FCAT success, educators and families would do well to explore how they may improve student ORF skill. Armbruster, Lehr and Osborn (2001) report that research has shown the use of frequent oral reading monitored by a teacher or parent to be an effective activity for improving reading fluency and overall reading achievement. This is a low-cost activity that teachers and families can use to improve ORF ability, increasing the likelihood of improved reading ability and FCAT success.

Lastly, student economic disadvantage contributed a significant amount of explained variance in this study. Specifically, economically disadvantaged students (those receiving free or reduced lunch prices) scored an estimated 15.92 points lower on the Reading FCAT-SSS than those students not economically disadvantaged, controlling for the other variables. This finding supports the importance of considering student economic disadvantage as a predictor of FCAT
Reading success—those students from lower income families tend to perform significantly poorer on the Reading FCAT-SSS.

Students from economically disadvantaged families may also present with concomitant experiential disadvantage for early educational experiences and access to intellectually stimulating environments. These students potentially arrive at school with multiple strikes against them and are expected to achieve similar to non-disadvantaged students on FCAT. Disadvantaged students are more likely to perform worse on the high-stakes test and then are subsequently more likely to be retained in third grade based upon FCAT failure. Given the many deleterious effects of grade retention on current and future student success discussed previously, disadvantaged students may potentially be caught in a spiraling reciprocal interaction between failure and retention. Schools may likewise be inadvertently set up for failing school grades when the student body includes a large number of disadvantaged students (e.g., inner-city schools). Therefore, the schools may face the removal of external resources and financial incentives that may have helped to initialize school improvement otherwise.

To compensate for the deleterious effects of economic disadvantage, educators may want to explore potential areas of difficulty for these students. For example, these families may benefit from increased access to age-appropriate reading materials for the student as well as other family members. Literacy workshops could be conducted with parents and adult family members to improve reading skill, comprehension, vocabulary, and to promote other techniques parents can use to develop a language rich environment for their children. These types of programs may result in a slow progression of skills for students and families over time.

Ultimately, SES, Reading grade, and ORF were the statistically significant variables given the set of the predictor variables in the first model. These findings were reproduced in Model 2, which utilized a stepwise regression procedure to determine the subset of variables that created the most effective parsimonious prediction model. ORF, SES, and Reading grade were again statistically significant and practically important in this model. The regression equation resulting from the second model yielded a large coefficient of determination ($R^2 = 0.623$), indicating that approximately 62% of the total variability of Reading FCAT-SSS scores could be explained using this model. Although this precision of prediction may not be statistically adequate to base major decisions regarding purchase of costly interventions, the results strongly
suggest that students can generally be identified as having difficulty based upon ORF, Reading grade, and SES. From a cost-benefit analysis perspective, incorrectly identify students as needing additional intervention based on these three predictors may be preferable to ignoring these variables and subsequently overlooking students who might benefit from such intervention. Factors in opposition to this idealized proposition would include limited district funds for intervention, limited teacher time and resources, limited student tolerance for intensive remedial instruction.

Consistently in the literature as well as in this study, ORF has been found to be a powerful predictor of reading skill. Less known are the determinants of ORF success, especially student-based variables such as those explored in the present study. In an effort to improve understanding about key variables impacting ORF performance, the same set of predictor variables was used to predict ORF.

Specifically, simultaneous multiple regression was conducted using gender, SES, number of prior grade retentions, attendance rate, participation, Reading grade, and NWF to predict ORF. The resulting regression equation indicated a moderately strong predictive relationship with a model $R^2$ of 0.492. The tests of the coefficients for Reading grade and NWF were found to be statistically significant. This significance was found to be of practical importance (i.e., have a sufficiently strong effect size) for NWF, but inconclusive with respect to practical importance for Reading grade.

This analysis highlighted the predictive value of NWF as a measure of decoding skill and Reading grade as an evaluation of student performance in reading instruction with respect to ORF performance. Both NWF and Reading grade added significant unique variance to the explanation of ORF. Once again, teachers were found to be good evaluators of reading ability as measured by ORF. Likewise, high NWF scores as a measure of decoding ability are indicative of high ORF scores as a measure of general oral reading fluency. Theoretically, discovering significant predictors, particularly if those predictors that are unfixed (unlike gender and SES) and easy targets for educator focus (such as decoding skill and oral reading fluency), might allow educators to indirectly support FCAT success.

Next, the accuracy of published recommended ORF cutpoints in determining student Reading FCAT-SSS achievement was investigated. Two cutpoint models were evaluated using
this sample, one model proposed by Good and colleagues (2002) and one proposed by FCRR (2004). These models and the associated ORF risk level cutpoints were applied to this dataset to determine the accuracy of predictions of Reading FCAT performance (refer to Appendix D). Good and colleagues defined a risk model with three classifications (low, moderate, and high) whereas FCRR defined a risk model with four classifications (above average, low, moderate, and high). The cutoffs for the risk categories were identical for the moderate and high groups between the two models. However, FCRR partitioned the low risk category of Good et al into low risk and above average). The predictive value of both cutpoint models was then compared.

Results of that comparison indicated the following. Students identified as high-risk via FCRR cutpoints during the second quarter of the academic year were more likely to later fail the Reading FCAT-SSS as compared to those students identified as high-risk via Good et al. cutpoints. Otherwise, the predictions for moderate risk and low risk students using the Good et al. cutpoints closely resembled the predictions for low risk and above average students in the FCRR risk categories. Nonetheless, data from this study reflected an unacceptably large proportion of students considered at low risk or above average that later failed the FCAT. In a previous section, the implications of overlooking students considered at low risk (false negatives) versus over-identifying students as needing additional intervention (false positives) were discussed concluding that over-identification of students for additional remediation was preferable than under-identifying needy students.

In an effort to contribute to the efficiency, effectiveness, and practicality of identifying at-risk students, this researcher advises the use of one risk assessment model. Therefore, this researcher argues that the cutpoints proposed by Good and colleagues (2002) are more useful for educators. This model provides a simpler risk assessment structure in that it contains only three risk categories. In contrast, FCRR’s (2004) risk assessment model contains four risk categories. The addition of a fourth category does not appear to be of greater help or accuracy of prediction. Also, the distinction between low risk and above average classification is somewhat confusing and it does not provide teachers with a clear direction or decision regarding need for additional instruction.

Lastly, this researcher originally proposed to explore the value of creating a risk classification scheme based upon NWF scores. This endeavor would only logically follow if
NWF scores were, in fact, predictive of FCAT performance in their own right given ORF administration. It would not be advantageous to remove ORF subtest use due to the strong predictive power it possesses; however, knowledge of a student’s NWF performance does not significantly add to the prediction of FCAT performance (based upon the results of the analysis of Hypothesis 1). Because NWF scores did not bear out to be a particularly effective predictor of FCAT performance above and beyond ORF scores, the addition of a NWF risk level chart would likely not prove beneficial. Reflecting back to the research goal, cutpoints for NWF performance were, therefore, not analyzed nor proposed.

**Limitations**

Inherent in any research study, limitations can be mild to severe and ultimately affect the results, conclusions, validity, and generalizability of the findings. Limitations in this study were rooted primarily in the lack of consenting participants, missing data, and subtest administration adherence. Other potential threats to internal validity (i.e., history, experimenter bias, demand characteristics, Hawthorne effect, mortality/attrition, quality of measures, selection bias, instrumentation, and statistical power) will also be discussed with respect to the presence of absence of problems associated with each in this study.

First, the overall representativeness of the sample to the target population was examined to explore the extent of population validity in this study (a potential threat to external validity). Regarding the consent process, the return rate of consent packets sent home to parents was consistent with average return rates for mailed or sent-home documentation. Nonetheless, any return rate less than 100% poses potential problems. Unfortunately, it was not possible to obtain additional information (i.e., demographics) regarding those families or students that did not return the consents. The collection of additional information regarding these families may have helped the researcher in evaluating sample-based systematic bias. For example, in this study there may be the potential for systematic bias towards families who are possibly more involved in their children’s education. However, it was heartening that a sufficient number of participants
from even the low-performing schools with transient populations did return the consents permitting the students’ participation in the present study.

To compensate inasmuch as possible, this researcher sought to evaluate the effect of missing subjects by comparing sample demographic data to district and state data (see Table 1). The greatest differences in demographic composition between sample data and district or statewide data were found in gender and ethnicity. The majority of participants in this study were female and of non-Caucasian ethnicity. Consequently, it can be concluded that the results of this study may not be universally generalizable to males or Caucasians in the state of Florida.

Missing data also posed difficulty for this study as described in the pertinent sections above. One predictor variable, Accelerated Reader scores, had to be eliminated from the study completely due to the small number of students that were enrolled in that program for which Accelerated Reader scores would be available. It was not known at the time of the research design that so few students and schools utilized the program. Nonetheless, this researcher does not believe that participation in the Accelerated Reader program would have had a direct, strong relationship with Reading FCAT-SSS success. This hypothesis would have to be explored in other research studies.

Also with respect to missing data, survey return rates were not as high as expected. These surveys included the participation ratings and the third quarter Reading grade for each identified student. In general, missing data can have profound effects on statistical power for the data analysis and validity of the research findings (Tate, 1998). This is especially true when using regression analyses in which listwise deletion excludes any participant from the analysis for whom there are any missing values on any of the predictor variables. In this study, listwise deletion of any subject for which there was missing data (data not retrievable from district records) would have resulted in a severely crippling decrease in sample size with a concomitant loss of statistical power. If listwise deletion of many participant values was permitted, it could have potentially obscured real population effects and/or create an under-representation of particular subgroups of the population, introducing bias into the results and interpretation (Tate, 1998). For that reason, missing data were addressed systematically.

Missing data was attended to in the following manner. Missingness indicators were used to determine the severity of the impact of missing data. Particularly concerning for data analysis
was missing data for Accelerated Reader scores (as discussed above), participation scores, and third quarter Reading grades. Observation replacement methods were utilized in an effort to ameliorate the negative impact of the missing data for these variables. Linear interpolation, a statistical replacement method, was used to estimate values for missing participation observations. In the case of missing third quarter Reading grades, the cumulative final Reading grade for the academic year was used. Although these replacement procedures are supported in the statistical literature or make intuitive sense, there is a possibility that the results may be somewhat skewed by false data (e.g., decreasing observation residuals resulting in decreased variance). Therefore, the presence of missing data is offered as a potential limitation to this study.

Subtest administration error also appeared to be problematic in the present study. For instance, a portion of data (8.8%) from the second DIBELS administration was unavailable for use as the data were spoiled. Specifically, the examiner failed to record the one-minute mark on the protocol or to provide the total time of completion. Thus, it was not possible to calculate a “correct per minute” score. As all of the examiners had received formal instruction on the administration of the DIBELS subtests in addition to having extensive experience administering many other standardized assessment measures, this type of error is not only surprising, but also unacceptable. Not only does this type of error impact internal validity of this study (i.e., quality of measures), but also the ramifications of this type of error extend beyond the present research interests. Vital decisions regarding students are made daily based upon the results of standardized assessments. It is hoped that more care is taken with those assessments when used for such significant purposes.

To combat the hazard of further administration error, the remainder of the protocols were checked for accuracy and corrected if needed. Reliability was improved by checking the scoring accuracy on each protocol. However, it must be surmised that the presence of scoring errors indicated the likelihood of other administration errors that could not be identified. For example, errors such as a lack of strict adherence to the standardized administration protocol (an internal validity threat via instrumentation administration) could artificially inflate or deflate student production and performance. Inasmuch as possible, this potential threat was limited by utilizing standardized instruments that were intended to be administered under standard conditions. In
other words, assent procedures and DIBELS measures were scripted and examiners were instructed to limit rapport-building to introductions only.

Nevertheless, it appeared that only the reliability statistics for the NWF subtest administration were affected. The NWF subtest is used less frequently than ORF and is reported by some testers as more difficult to score, given that exact pronunciation of individual phonemes is required for score credit. Due to differences in student and listener dialect as well as the examiner’s auditory processing, the examiner is more likely to inaccurately score student responses on this subtest as opposed to the ORF subtest. Unfortunately, it was not possible to determine the extent of these potential errors and subsequent interpretations from data analyses must take this into account. Summarized below are the potential impacts of other indicators of threats to internal validity.

Next, this researcher attempted to control the effects of history, environmental events that may impact data, in this study. When available, the researcher utilized students’ ORF scores collected during the same interval administration by district staff to meet district requirements. This helped to avoid retesting students with familiar material, potentially creating artificially inflated scores due to retesting. In these cases, it was not possible to monitor the test administration to each participant. Nonetheless, there is no indication of or control over systematic contextual influences that may have been present during testing. In summary, history may be one potential threat to internal validity in the present study.

Other potential threats to internal validity in this study and the prospective impact of each follow. Experimenter bias may have been a concern for 8.0% of the observations for NWF #1 and ORF #1 because the researcher assessed these students personally. Demand characteristics may also have impacted the students’ DIBELS scores given that participants are instructed to do their best reading. On the other hand, this is minimized by not disclosing to the student their scores from any administration. Attrition was also of limited concern given the low attrition rate (2.7%), but present nonetheless.

The quality of the measures was generally very good with respect to the reliability and validity estimates of each, except the participation survey. The one-item participation survey was developed by the researcher and not previously studied for statistical properties. As discussed in detail above, measures administered by volunteer school psychologists were scored
surprisingly inaccurately, despite FCRR training and re-orientation by the researcher. As previously mentioned, these instruments were re-scored as needed.

The remaining potential threats to internal validity were not likely to have been problematic in this study—selection bias, Hawthorne effect, regression to the mean, and statistical power. First, students were randomly selected according to specific criteria or rules of inclusion such that all participating students were third grade regular education students. The researcher also attempted to incorporate schools that were both low and high performing with respect to the previous year’s FCAT achievement. Problems associated with the Hawthorne effect were not likely a powerful threat in the present study given that the examiner observed students only during direct assessment and there was no control group. Next, specific groups (i.e., only high or low performers) were not used in this study so the effects of regression to the mean was not a likely confound in this study. Lastly, sufficient power was obtained for the present study due to the final number of participants maintained in the study, thus limiting the risk of Type II errors.

Future Directions

This research focused upon students within the regular education third grade classrooms. First, it may be of interest to simply replicate the present study with a larger, more diverse sample with respect to gender and ethnicity. This would increase the generalizability of the results and conclusions. Furthermore, cross validation procedures could then be utilized with greater confidence to estimate how well the stepwise multiple regression equation developed from this sample would predict in other random samples from the same population (Tate, 1998). Also, one might explore whether there exists any difference in the predictive validity of the stepwise regression equation across levels of economic disadvantage. This study did not explore this, in part due to sample size.

Another possible direction for future research may include a research study that incorporates a larger sample size such that the prediction model for specific subgroups could then be compared. For example, it is plausible that the significant predictor variables for
students with Specific Learning Disabilities (SLD) may differ than those found to be significant in the present study with regular education students. Notably, Reading grade was found to be a significant predictor of Reading FCAT-SSS success for regular education third graders. On the contrary, grades of students with SLD may be based upon special standards that are below the reading proficiency standards for regular education students. The options for grading students in ESE programs are generally outlined in the pupil progression plan for the respective school district. For example, one district in Florida outlines four course option plans for ESE students whereby students are expected to meet regular or special standards depending upon the course option elected for the student on the individualized ESE education plan (i.e., IEP) (Duval County Public Schools, 2005). Students that are enrolled in one of the two options that allow for special standards receive subject grades based on those standards specified on the IEP. Grades based on special standards are denoted with an asterisk (*) on the student’s report card. Therefore, students with SLD that opt for a special diploma track later in their academic career have more leniency as determined by the IEP on the level of performance required for adequate progress. In such cases, a Reading grade of an ‘A’ may reflect a different level of student Reading achievement for a regular education student as opposed to an ESE student enrolled in a course option with special standards. Research incorporating this group should be very explicit in the description of these students, including eligibility information and possibly utilizing a measure of reading ability so that the groups can be compared via a norm-referenced test.

Finally, the present study was focused on Reading FCAT-SSS achievement. Previous research by Buck and Torgesen (2003) examined the correlation between ORF and Math FCAT-SSS achievement as well and found a significant positive correlation between the two ($r = .54$, $p < .001$). Similarly, the predictor variables in the present study could be applied to predict math achievement also amongst third graders. Similarly, DIBELS procedures could be administered to students in later grades to determine the predictive quality of these measures for FCAT and other measures of academic achievement for those students.
Conclusion

Despite the limitations described above it is believed that the goals of this research study were achieved. That is, this study provided evidence to support the inclusion of other very relevant and powerful predictors of FCAT Reading achievement, specifically Reading grade and SES. The overall model $R^2$ for the stepwise multiple regression including the predictors ORF, Reading grade, and SES was 0.623, reflecting a strong theory in which these predictors explained approximately 62% of the variance in Reading FCAT-SSS performance. In this study, as in previous research, ORF continues to be a simple and effective predictor; however, administrators and parents would be remiss to overlook teachers’ evaluations of their students’ success in Reading activities as measured by quarterly grades. Referring back to the original research goal (i.e., to provide educators with an efficient and effective method to identify students at risk for FCAT failure), the findings in this study support that goal. For example, teachers have no difficulty retrieving students’ Reading grades. The benefit of doing so outweighs the small investment of time and energy.

In fact, Reading grades are such a powerful predictor of success that teachers may use a low first or second quarter Reading grade to generate an alert to themselves, the student, and the student’s family that this student is at risk for FCAT failure and may benefit from added instruction and support both in the classroom and at home. A more formal analysis or validation of this alert procedure could be conducted by calculating the frequency and proportion (with a corresponding test of statistical significance) of students that failed the Reading FCAT-SSS by Reading grade, similar to the analysis in Hypothesis 3 of this study. For example, a post-hoc analysis of the same in this study revealed that approximately 4% of the students who earned a third quarter Reading grade of an A failed the Reading FCAT-SSS, approximately 8% of those who earned a B failed, 44% who earned a C failed, 67% who earned a D failed, and 78% who earned an F failed. Although these findings should be replicated with other studies and with either first or second quarter grades, the above initially supports teachers providing alerts for students earning a C or below in Reading.

The results of this study also indicate that when controlling for ORF ability, students that are economically disadvantaged are at a significantly greater risk for lower Reading FCAT-SSS
achievement than students that are not economically disadvantaged. This finding indicates the risk of double penalty for these students in that financial disadvantage impacts FCAT performance, which impacts academic failure and limited future academic success. Furthermore, schools with a large number of economically disadvantaged students are more prone to lower school grades as assigned by the Florida Department of Education due to low FCAT scores among these students. The effects of policy regarding the uses of the FCAT scores for student promotion and school grades are exacerbated among the economically disadvantaged student population.

Finally, ORF cutpoints appear to be of continued assistance in identifying students at risk for FCAT failure. These cutpoints serve as a powerful and efficient tool in the early identification of students with academic difficulty. The results of this study support a three-category risk assessment model (low, moderate, and high risk) for teachers. It is recommended that this type of resource (risk category model) be used in any instructional setting that already utilizes DIBELS or other ORF measures.
APPENDIX A.1

HUMAN SUBJECTS COMMITTEE APPLICATION

FLORIDA STATE UNIVERSITY  Application No.: 04.774

Human Subjects Application

to the INSTITUTIONAL REVIEW BOARD

for RESEARCH INVOLVING HUMAN SUBJECTS

The Federal Government and University policy require that the use of human subjects in research be monitored by the Institutional Review Board (IRB). The following information must be provided when humans are used in research studies, whether internally funded, extramurally funded or unfunded. Research in which humans are used may not be performed in the absence of IRB approval.

PLEASE COMPLETE AND SUBMIT PAGES 1 AND 2 plus YOUR ANSWERS TO THEQUESTIONS (on page 3) IN TYPEWRITTEN FORM TO: HUMAN SUBJECTS COMMITTEE, Mail Code 2763, or

2035 E. Paul Dirac Drive, Box 15
100 Sliger Bldg., Innovation Park
Tallahassee, FL 32310

Researcher: Angela I. Canto  Date: 10/25/04

Project Title: “Predicting Third Grade Students’ FCAT Reading Achievement Using Measures of Oral Reading Fluency, Decoding, and Student Characteristics

Project Period (starting/ending dates): 11/1/04—11/1/05

Position in University (faculty, etc.) If student, please indicate FSU Faculty Advisor:
Doctoral student with FSU Faculty Advisor—Briley Proctor, Ph.D.

Department: Department of Educational Psychology and Learning Systems

Telephone: (904) 477-0697   E-Mail Address: aitcanto@hotmail.com
(850) 644-3742   proctor@coe.fsu.edu
(where you can be reached in case of a problem with your application)

Mailing Address (where your approval will be mailed): 4456 Delano Court
Jacksonville, FL 32257
Project is (please check one):  __X__ dissertation  _____ teaching  ____ thesis (equivalency)____ other

Project is:  ____X____ unfunded  _________ funded (if funded, please complete the following):  N/A

Funding Agency (actual/potential):  1. ___________________________
                                          2. ___________________________

Contract/Grant No. (if applicable):  ________________________________________

FOR EVALUATION OF YOUR PROJECT, PLEASE CHECK THE FOLLOWING WHICH APPLY:

☐ Mentally or Physically Challenged Subjects  ☐ Subjects studied at FSU
☐ Children or Minor Subjects (under 18 years old)  ☐ Subjects studied at non-FSU location(s)
☐ Prisoners, Parolees or Incarcerated Subjects  ☐ Students as Subjects
☐ Filming, Video or Audio Recording of Subjects  ☐ Employees as Subjects
☐ Questionnaires or Survey(s) to be administered  ☐ Pregnant Subjects
☐ Review of Data Banks, Archives or Medical Records  ☐ Fetal, placental or surgical pathology tissue(s)
☐ Subjects’ major language is not English  ☐ Involves Blood Samples
☐ Involves Deception (if yes, fully describe at Question No. 7)  ☐ Subjects to be paid
☐ Exclusion of Women or Children Subjects (must explain why they are being excluded)

Survey Techniques: Check applicable category if the only involvement of human subjects will be in one or more of the following categories:

___X___ Research on normal educational practices in commonly accepted educational settings

___X___ Research involving educational tests (cognitive, diagnostic, aptitude, achievement)

___X___ Research involving survey or interview procedures (if checked, please see below)

___X___ Research involving the collection or study of existing data, documents, records, specimens

If research involves use of survey or interview procedures to be performed, indicate:

1. Responses will be recorded in such a manner that human subjects cannot be identified, by persons other than the researcher, either directly or through identifiers linked to the subjects.

   _X_ yes  ___ no

2. Would subject’s responses, if they became known outside the research, reasonably place the subject at risk of criminal or civil liability or be damaging to the subject’s financial standing or employability.

   ___ yes _X_ no

3. The research deals with sensitive aspects of the subject’s own behavior, such as illegal conduct, drug use, sexual behavior, or use of alcohol.

   ___ yes _X_ no

100
1. **Give a complete description of your research procedures as they relate to the use of human subjects.**

**Project Significance:** The Florida Comprehensive Assessment Test (FCAT) is Florida’s response to President George W. Bush’s No Child Left Behind (NCLB) Act mandates. It was developed as an integral part of educational reform in the state of Florida in an effort to raise standards for education, improve student education, and increase school accountability (Florida Department of Education [FDOE], 2001b). Students’ scores on the FCAT impact individual students and schools at large. Given the importance placed on FCAT scores, educators seek accurate and efficient methods to predict student FCAT scores.

In the domain of reading, research has demonstrated that decoding and fluency are significant skill requisites in reading achievement. Limited research is currently available to demonstrate the relationship between these skills and success on the FCAT. This proposed study seeks to examine the relationship between oral reading fluency and decoding skill, as measured by brief subtests from the Dynamic Indicators of Basic Early Literacy Skills (DIBELS) assessment system (Good & Kaminski, 2002), and third grade student achievement on the Florida Comprehensive Assessment Test (FCAT) in reading. Student demographic variables (gender and SES) and individual characteristics (absenteeism, number prior retentions, level of participation in class, Accelerated Reader scores, and reading grades) are included to determine whether these variables exert influence on the prediction of FCAT achievement. Also, researchers have developed benchmark guidelines based on a student’s performance on the DIBELS subtests to assist educators in identifying those students who may be at risk for reading difficulty. The
present study attempts to validate those benchmarks for third grade ORF scores and explore the utility of establishing benchmarks for third grade NWF scores.

Procedure Details: Participants will include 160 third grade students in the Duval County Public Schools in Jacksonville, Florida. Eligible participants will be those that are enrolled in the regular education curriculum and required to take the FCAT in Spring 2005. The Duval County Public School system is partitioned into five geographical regions. Four elementary schools from each region will be selected for inclusion to maximize the variability inherent among students in the different geographical locations. Eight third grade regular education students will be randomly selected from each participating school ensuring that each student will have an equal probability of being selected for inclusion.

Each student’s legal guardian will have the opportunity to consent for participation. Students’ parents or legal guardians will be asked to sign informed consent forms to participate in the research study. Parents will be given the opportunity to seek clarification and decline participation, if desired. In the event a randomly selected student or guardian declines participation, another student will be randomly selected from that school. This process will continue until 160 students have consented and those 160 children have given their assent.

This researcher will meet with the school administrators and teachers identified in each of the selected school sites to discuss the present study. It will be explained that for the purposes of treatment fidelity and confidentiality, the results of the present study will not be made available until the study is concluded. The potential risks and benefits associated with participation in the study will also be discussed. Administrators and teachers will be given the opportunity to seek clarification and decline participation, if desired. Consenting teachers will be asked to sign informed consent forms to participate in the research study.

The desired data will be obtained in the following manner. Each participant’s oral reading fluency and decoding skill will be assessed using two subtests from the DIBELS assessment system (Good & Kaminski, 2002), Oral Reading Fluency (ORF) and Nonsense Word Fluency (NWF), twice (in equal intervals) prior to the FCAT administration in Spring 2005 (see appendix for example). Thus, the direct performance assessment will result in two interval administrations and four total subtest administrations. Twenty percent (or 32) of the protocols from each administration interval will be randomly selected for reliability checks by this researcher. The results of the reliability checks will be reported and discussed. Additionally, the following demographic information of each participant will be obtained via school records: absenteeism, ethnicity, SES (using free or reduced lunch as the observed variable), and the number of prior grade retentions. Each participant’s teacher will also be requested to evaluate the student’s general level of class participation via a brief survey (see appendix) using a five-point Likert scale assessment each time the participant is administered the DIBELS subtests. Lastly, each participant’s scores on the FCAT will be obtained via school records. At the conclusion of data retrieval, the principal investigator will destroy the case code record, which is the only record of participating subjects and identifying information.

The data will be analyzed to identify the best subset of predictors of FCAT Reading performance among the available predictor variables. The results will be written for publication in professional journals and/or presentation at professional conferences.

2. HAVE THE RISKS INVOLVED BEEN MINIMIZED AND ARE THEY REASONABLE IN RELATION TO ANTICIPATED BENEFITS OF THE RESEARCH, IF ANY, TO THE SUBJECTS AND THE IMPORTANCE OF THE KNOWLEDGE THAT MAY REASONABLY BE EXPECTED TO RESULT?

The principal investigator has attempted to outline a procedure above that minimizes any risk to the confidentiality of the participants, using appropriate subject coding procedures. Each participant will be assigned a subject number beginning with #001. The assigned number is not connected in any manner to
identifying information from the student. A master list of participants and subject numbers will be maintained solely by the principal investigator in a locked file cabinet until the conclusion of data collection and analysis. At that time, the master list will be destroyed and no other documents with identifying participant information will be available. As evidenced by the project significance above, results from this study may be instrumental in assisting teachers, students, and the public at large in identifying those students who may be at risk for failing FCAT with its corresponding negative outcomes.

WHAT PROVISIONS HAVE BEEN MADE TO INSURE THAT APPROPRIATE FACILITIES AND PROFESSIONAL ATTENTION NECESSARY FOR THE HEALTH AND SAFETY OF THE SUBJECTS ARE AVAILABLE AND WILL BE UTILIZED?

The principal investigator and recruited examiners (school psychologists) received formal training from the Florida Center for Reading Research in the proposed research procedures and policies, including the strictest confidentiality standards.

3. DESCRIBE PROCEDURES TO BE USED TO OBTAIN INFORMED CONSENT.

Each student’s legal guardian will have the opportunity to consent for participation. Students’ parents or legal guardians will be asked to sign informed consent forms, sent home from the student’s participating school, to participate in the research for the upcoming academic year. The associated risks and benefits of the study will be outlined in detail in the consent forms. Parents will have the opportunity to ask questions prior to consent. There are no consequences to the student, teacher, or family for consenting or denying consent in this study. Parents will be given contact information for the principal investigator if questions arise or if withdrawal from the study is desired.

(A) WHO WILL BE OBTAINING INFORMED CONSENT? Principal Investigator

(B) WHEN WILL THE SUBJECTS BE ASKED TO PARTICIPATE AND SIGN THE CONSENT FORM? Prior to any data collection

(C) IN USING CHILDREN, HOW WILL THEIR ASSENT BE OBTAINED? See appendix for assent script to be used by each examiner. Each student will be given the opportunity to decline participation at any time.

4. DESCRIBE HOW POTENTIAL SUBJECTS FOR THE RESEARCH PROJECT WILL BE RECRUITED.

Participants will include 160 third grade students in the Duval County Public Schools in Jacksonville, Florida. Eligible participants will be those that are enrolled in the regular education curriculum and required to take the FCAT in Spring 2005. The Duval County Public School system is partitioned into five geographical regions. Four elementary schools from each region will be selected for inclusion to maximize the variability inherent among students in the different geographical locations. Eight third grade regular education students will be randomly selected from each participating school ensuring that each student will have an equal probability of being selected for inclusion. Informed consent will be obtained from each student’s legal guardian. In the event a randomly selected student or guardian declines participation, another third grade regular education student will be randomly selected from that school. Assent from the child will be obtained prior to data collection.

5. WILL CONFIDENTIALITY OF ALL SUBJECTS BE MAINTAINED? HOW WILL THIS BE ACCOMPLISHED? PLEASE ALSO SPECIFY WHAT WILL BE DONE WITH ALL AUDIO AND/OR VISUAL RECORDINGS, IF APPLICABLE, PICTURES AND PERSONAL DOCUMENTATION OF SUBJECTS BOTH DURING AND AFTER COMPLETION OF THE RESEARCH.

Participant confidentiality will be maintained for this study, to the extent allowed by law. This will be accomplished as described above. Identifying information of subjects will be destroyed
immediately following the data collection. No cumulative records will be removed from the schools.

6. **IS THE RESEARCH AREA CONTROVERSIAL AND IS THERE A POSSIBILITY YOUR PROJECT WILL GENERATE PUBLIC CONCERN? IF SO, PLEASE EXPLAIN.**

It is not anticipated that any public concern will be generated or that the research area of FCAT prediction is controversial. In fact, this type of research is limited and public understanding of these relationships is limited. Any information obtained may be of use to the Florida citizenry.

7. **DESCRIBE THE PROCEDURE TO BE USED FOR SUBJECT DEBRIEFING AT THE END OF THE PROJECT. IF YOU DO NOT INTEND TO PROVIDE DEBRIEFING, PLEASE EXPLAIN.**

Group-level results obtained from this study may be presented to participating schools and teachers (omitting identifiable student information). Parents, students, and school personnel will be given the opportunity to ask questions at any time during or after the study by contacting the Principal Investigator via the provided contact information.
Dear Parent or Legal Guardian of ____________________:

We are all aware of current state testing efforts (FCAT) to ensure that Florida’s students are achieving the minimum standards in reading and math. We also know that students’ scores on this important test are used to determine students’ promotion eligibility and school accountability. It is in the interest of all students that school personnel are better able to predict those students that may be struggling academically and may be at risk for failing the FCAT. If at-risk students are identified early, teachers may be able to implement interventions that will help to improve reading skill. To do this, researchers conduct studies.

Your child has been randomly selected among all third grade regular education students in Duval County to have the opportunity to participate in a research study. This study seeks to find a way to identify those students who may be at risk for FCAT failure and may need additional instructional assistance in advance of FCAT testing. Your child’s teacher, as well as district and school administration, have approved this research study. However, your child’s participation in the present study will only be possible if you consent to participation and sign the attached informed consent. More information regarding the study is provided below.

Project Title: “Predicting Third Grade Students’ FCAT Reading Achievement”

Principal Investigator: Angela I. Canto
Doctoral Candidate, Florida State University
School Psychology Intern, Duval County Public Schools
(904) 390-2474

Procedure Details:
If you agree for your student to participate in this study, your student will be given a brief reading test twice prior to FCAT testing in Spring 2005. This test involves approximately 10 minutes of direct student contact by myself or a certified Duval County School Psychologist in your child’s school. Each participating student will read aloud for approximately four minutes to the examiner and then be returned to class. No other direct student contact is required.

In addition to the brief reading test described above, general information regarding student demographics and characteristics will be obtained from the student’s cumulative records and his/her teacher. I will be the only person that will maintain access to any identifying student information—no identifying information will be written on any forms that may be seen by others. Your child will be assigned a “student number” so that his/her name and any other identifying information can be kept confidential to the extent allowed by law.

Students and parents will be given the opportunity to ask any questions and/or to withdraw from the study at any time by contacting me. Benefits of participation include direct individual contact with students in effort to evaluate reading ability. Risks of participation include 10 minutes of removal from class instruction twice this academic year (20 minutes total). Efforts will be made to avoid missing important academic instruction.

Directions: If you agree for your student to participate in this important study, please read and sign the attached informed consent form and return to your student’s teacher the following day. Thank you for your time.

Sincerely,
Angela I. Canto
I freely and voluntarily and without element of force or coercion, consent to be a participant in the research project entitled “Predicting Third Grade Students’ FCAT Reading Achievement.”

This research is being conducted by Angela I. Canto who is a Doctoral Candidate at Florida State University and a School Psychology Intern in the Duval County Public Schools. I understand the purpose of the research project is to better understand how students’ achievement on the FCAT Reading test can be predicted. I understand that if I allow my child to participate in the project, demographic information will be collected and reading performance will be assessed. The total time commitment would be approximately 20 minutes during the academic year. Ms. Canto will answer any questions I may have or she will refer me to a knowledgeable source.

I understand that my child’s participation is totally voluntary and either my child or I may stop participation at anytime. All information obtained regarding my student will be kept confidential and identified by a subject code number. My child’s name will not appear on any of the results. No individual responses will be reported to the public—only group findings will be reported. My child’s results will be kept confidential to the fullest extent allowed by law.

I understand there is a possibility of a minimal level of risk involved if I allow my child to participate in this study. My child may briefly miss instructional time (approximately 10 minutes). However, every attempt will be made to assess my child during non-academic classroom time. I understand there are benefits for participating in this research project. I will be providing educators and researchers with valuable insight into how to better identify students struggling with reading who may be at risk for failing the FCAT Reading test. This knowledge can assist them in providing additional instructional time and resources to struggling students in the future.

I understand that this consent may be withdrawn at any time without prejudice, penalty or loss of benefits. I have been given the right to ask any questions concerning the study. Questions, if any, have been answered to my satisfaction.

I understand that I may contact Ms. Canto at (904) 390-2474, Dr. Briley Proctor at (850) 644-3742, or the Florida State University Human Subjects Committee at (850) 644-8633 for answers to questions about this project or my rights. Group results will be sent to me upon my request.

I have read and understand this consent form.

________________________________________
Please print Child’s Name

________________________________________
Please print Parent’s Name

________________________________________
Parent’s Signature

________________________________________
Date
Hi, my name is ____________. I’d like to work with you for about 10 minutes today on short reading activities. I’d like to get your permission also. There are no rewards or punishments and you aren’t graded on your reading but I’d like you to do your best work. May we begin?

If student agrees, begin with the first subtest using the provided standardized procedure.

If student declines, return student to classroom and contact Angel Canto at (904) 477-0697 to receive the name of another randomly selected student in your school.

I freely and voluntarily and without element of force or coercion, consent to participate in the research project entitled “Predicting Third Grade Students’ FCAT Reading Achievement.”

This research is being conducted by Angela I. Canto who is a Doctoral Candidate at Florida State University and a School Psychology Intern in the Duval County Public Schools. I understand the purpose of the research project is to better understand how students’ achievement on the FCAT Reading test can be predicted. I understand that I will be asked to complete a brief, one-item questionnaire regarding student class participation. Also, the selected student(s) (whose parents have given consent) will be taken out of the classroom twice for approximately 10 minutes to assess reading skills. Ms. Canto will answer any questions I may have or she will refer me to a knowledgeable source.

All information obtained regarding my survey response and the participating student will be kept confidential to the extent allowed by law. No individual results will be reported to me—only group findings will be reported at the conclusion of the project at my request.

I understand there is a possibility of a minimal level of risk involved if I participate in this study. The completion of the survey will require approximately 3 minutes of my time. I also understand there are potential benefits for participating in this research project. I will be providing educators and researchers with valuable insight into how to better identify students struggling with reading who may be at risk for failing the FCAT Reading test.

I understand that my participation is totally voluntary and I may stop participation at anytime. I understand that this consent may be withdrawn at any time without prejudice, penalty or loss of benefits. I have been given the right to ask any questions concerning the study. Questions, if any, have been answered to my satisfaction.

I understand that I may contact Ms. Canto at (904) 390-2474, Dr. Briley Proctor at (850) 644-3742, or the Florida State University Human Subjects Committee at (850) 644-8633 for answers to questions about this project or my rights. Group results will be sent to me upon my request.

I have read and understand this consent form.

_______________________________________
Please PRINT your name

_______________________________________
Teacher’s SIGNATURE

_______________________________________
Date
APPENDIX A.2
HUMAN SUBJECTS COMMITTEE APPROVAL LETTER

[Letter content]

Florida State
UNIVERSITY

Office of the Vice President for Research
humanresearch@fsu.edu
Telephone: 848-3311
Fax: 848-3371

APPROVAL MEMORANDUM

To:
Angela Ones
4900 Olene Court
Jacksonville, FL 32257

From:
John T. Emerson, CSE

The Human Subjects Committee has reviewed your proposal for research and recommends approval of the proposed protocol for the use of human subjects in the project. The Human Subjects Committee was convened in accordance with the rules established by the Institutional Review Board (IRB) and the standards set by the Federal Regulations governing the protection of human subjects. The project has been determined to be exempt from further review and approval.

The protocol has been reviewed by the Office of the Vice President for Research and has been determined to be in compliance with all relevant regulations and guidelines. The protocol has been reviewed by the Institutional Review Board (IRB) and has been determined to be in compliance with all relevant regulations and guidelines. The project has been determined to be exempt from further review and approval.

The protocol has been reviewed by the Office of the Vice President for Research and has been determined to be in compliance with all relevant regulations and guidelines. The protocol has been reviewed by the Institutional Review Board (IRB) and has been determined to be in compliance with all relevant regulations and guidelines. The project has been determined to be exempt from further review and approval.

This institution will ensure that the Office for Protection from Research Risks is notified in accordance with the regulations and guidelines of the IRB.

[Signature]
[Date]
The original measures for DIBELS measures are available at http://dibels.uoregon.edu/. The samples below are provided to the reader as an example only.

**NWF Administration**

The following administration instructions are used prior to allowing the student to read from the NWF probe (FCRR, 2004):

*Here are some more make-believe words* (point to the student probe). *Start here* (point to the first word) *and go across the page* (point across the page). *When I say ‘begin’, read the words the best you can.* *Point to each letter and tell me the sound or read the whole word.* *Read the words the best you can.* *Put your finger on the first word.* *Ready, begin.*

(p. 24)

**NWF Probe**

<table>
<thead>
<tr>
<th>rak</th>
<th>bim</th>
<th>tes</th>
<th>ked</th>
</tr>
</thead>
<tbody>
<tr>
<td>lab</td>
<td>fug</td>
<td>bof</td>
<td>yos</td>
</tr>
<tr>
<td>ig</td>
<td>han</td>
<td>lub</td>
<td>jad</td>
</tr>
</tbody>
</table>
APPENDIX B.2

SAMPLE OF ORF ADMINISTRATION MATERIALS

**ORF Administration**

The following administration instructions are used prior to allowing the student to read from the ORF probe (FCRR, 2004):

> When I say begin start reading aloud at the top of the page (point). Read across the page (point). 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. Ready, begin. (p. 30)

**ORF Probe**

Many monkeys have very long tails. They use their tails to help them grab on to trees and swing from branches. Small berries, fruits, and leaves from tees are their favorite foods. The tails are often used by the monkeys to hold on to the limb while reaching for food. Someday I’d like to see some monkeys in the wild.
Please answer the following 3 questions regarding this student:

On a scale of one (0) to five (4), please rate this student’s level of participation in classroom reading instruction and reading activities as compared to his/her peers this academic year.

Participation is defined by the following:
- Attends to lesson and task instructions
- Raises hand to volunteer answers or ask appropriate questions
- Attempts to complete class assignments
- Asks questions when unsure of what to do
- Completes homework

None  Occasional  Sometimes  Frequent  Constant
0     1           2           3          4

Please circle the student’s Reading Grade as of the most recent report card:

A      B      C      D      F

Please provide the student’s current Accelerated Reader score (if used in your classroom):

_______
APPENDIX D.1
REPRODUCED GOOD ET AL. (2002) ORF RISK LEVELS TABLE

The following is a reproduced table of the reading risk level based on the ORF cutpoints from Good et al. (2002):

Table 17
Reproduced Good et al. (2002) ORF Risk Level Classification

<table>
<thead>
<tr>
<th>Risk Classification</th>
<th>ORF Score (cwpm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>&lt; 80</td>
</tr>
<tr>
<td>Moderate</td>
<td>80 - 109</td>
</tr>
<tr>
<td>Low</td>
<td>&gt; 110</td>
</tr>
</tbody>
</table>

*Note. ORF score is measured in correct words per minute (cwpm).*
APPENDIX D.2

REPRODUCED FCRR (2004) ORF RISK LEVELS TABLE

The following is a reproduced table of the reading risk level based on the ORF cutpoints from FCRR (2004):

Table 18
Reproduced FCRR (2004) ORF Risk Level Classification for Third Grade, First Winter Quarter (2004-2005 School Year)

<table>
<thead>
<tr>
<th>Risk Classification</th>
<th>ORF Score (cwpm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>High Risk</td>
<td>&lt; 62</td>
</tr>
<tr>
<td>Moderate Risk</td>
<td>62 - 87</td>
</tr>
<tr>
<td>Low Risk</td>
<td>88 - 105</td>
</tr>
<tr>
<td>Above Average</td>
<td>&gt; 106</td>
</tr>
</tbody>
</table>

Note. ORF score is measured in correct words per minute (cwpm).
REFERENCES


115


Farstrup & S. J. Samuels (Eds.), What research has to say about reading instruction (3rd


Heubert, J. P. & Hauser, R. M. (Eds.). (1999). High-stakes testing for tracking, promotion, and

Samuels (Eds.), What research has to say about reading instruction (3rd ed.) (pp. 337-

the Dynamic Indicators of Basic Early Literacy Skills and the Comprehensive Test of
Phonological Processing. Unpublished manuscript, University of Massachusetts at
Amherst.

Individuals with Disabilities Education Act (IDEA) of 1990, Public Law 101-476, 104 Stat. 1142

Individuals with Disabilities Education Act (IDEA) of 1997, Public Law 105-17, 101 (1997).

from http://www.palmbeachpost.com/opinion/content/auto/epaper/editions/saturday/
opinion_e33bd0db83fac17310c0.html


model: Dynamic Indicators of Basic Early Literacy Skills. In In M. R. Shinn (Ed.),
Advanced applications of curriculum-based measurement (pp. 113-142). NY: The
Guilford Press.


underpinnings and integration into problem-solving assessment. Journal of School
Psychology, 29, 371-393.


BIOGRAPHICAL SKETCH

Angela Canto began her undergraduate education at the University of South Florida and completed her B.A. at the University of North Florida in 1997. After working as a case manager she returned to complete her doctoral degree at Florida State University in the Combined Doctoral Program in Counseling Psychology and School Psychology. She completed her pre-doctoral internship at the University of Florida – Shands Jacksonville and the Duval County Public Schools. She is currently employed in a post-doctoral position as a Psychology Resident with Trauma Psychological Services in the Department of Surgery at the University of Florida – Shands Jacksonville.