Retention and special education referral practices before and after the implementation of high-stakes testing

Sue L. Buslinger-Clifford
Florida International University

DOI: 10.25148/etd.FI14051899
Follow this and additional works at: https://digitalcommons.fiu.edu/etd
Part of the Special Education and Teaching Commons

Recommended Citation
Buslinger-Clifford, Sue L., "Retention and special education referral practices before and after the implementation of high-stakes testing" (2004). FIU Electronic Theses and Dissertations. 1821.
https://digitalcommons.fiu.edu/etd/1821

This work is brought to you for free and open access by the University Graduate School at FIU Digital Commons. It has been accepted for inclusion in FIU Electronic Theses and Dissertations by an authorized administrator of FIU Digital Commons. For more information, please contact dcc@fiu.edu.
FLORIDA INTERNATIONAL UNIVERSITY

Miami, Florida

RETENTION AND SPECIAL EDUCATION REFERRAL PRACTICES BEFORE AND AFTER THE IMPLEMENTATION OF HIGH-STAKES TESTING

A dissertation submitted in partial fulfillment of the requirements for the degree of

DOCTOR OF EDUCATION

in

EXCEPTIONAL STUDENT EDUCATION

by

Sue L. Buslinger-Clifford

2004
To: Dean Linda Blanton  
College of Education  

This dissertation, written by Sue L. Buslinger-Clifford, and entitled Retention and Special Education Referral Practices Before and After the Implementation of High-Stakes Testing, having been approved in respect to style and intellectual content, is referred to you for judgment.

We have read this dissertation and recommend that it be approved.

__________________________  
Judith Cohen  

__________________________  
Barry Greenberg  

__________________________  
Wendy Silverman  

__________________________  
Philip Lazarus, Major Professor  

Date of Defense: June 2, 2004  

The dissertation of Sue L. Buslinger-Clifford is approved.

__________________________  
Dean Linda Blanton  
College of Education  

__________________________  
Dean Douglas Wartzok  
University Graduate School  

Florida International University. 2004
© Copyright 2004 by Sue L. Buslinger-Clifford

All rights reserved.
DEDICATION

I dedicate this effort to my family; Henry, my husband, Alexander, Katherine and Garrett, my children; who provided encouragement, hope and love...you are everything to me. And to Jake, you were always there.
ABSTRACT OF THE DISSERTATION

RETENTION AND SPECIAL EDUCATION REFERRAL PRACTICES BEFORE AND AFTER THE IMPLEMENTATION OF HIGH-STAKES TESTING

by

Sue L. Buslinger-Clifford

Florida International University, 2004

Miami, Florida

Professor Philip Lazarus, Major Professor

The rewards and sanctions associated with high-stakes testing may induce educators to participate in practices that will ensure the elimination of the scores of low-achieving students from the testing pool. Two ways in which scores may be eliminated is through retention or referral to special education.

This study examined the use of these practices at 179 elementary schools in Miami-Dade County Public Schools, the 4th largest school district in the country. Between- and within-subjects designs were analyzed using repeated measures analysis of variance to compare retention and referral to special education practices over a five-year period of time, two years prior to and two years after the implementation of Florida’s high-stakes test, the Florida Comprehensive Assessment Test, FCAT.

Significant main effects for referral and retention over time were demonstrated. The use of retention steadily increased over the first three years, with its usage maintained during the fourth year. While the use of referral actually decreased from the first to second years, a significant change occurred after the implementation of the FCAT.
Examination of the use of these practices according to student and school characteristics revealed significant differences. Increases in the use of referral across time was significant for Black, non-Hispanic and Hispanic students, all limited English proficiency population categories, medium and low socioeconomic status category schools, all grade levels, and for schools with accountability grades of A, C, D and F with the most striking absolute increase occurring for F schools. Increases in the use of retention across time were significant for all ethnic groups, limited English proficiency categories, and socioeconomic status categories, for grades kindergarten through four and by gender. Significant increases occurred for schools with accountability performance grades of C, D and F; however the most dramatic increase occurred for the F schools. A direct relationship between performance category grade of school and their use of retention was demonstrated. The results suggest that schools changed their use of referral and retention in response to the implementation of the FCAT.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>CHAPTER</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>I. INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>The Present Study</td>
<td>3</td>
</tr>
<tr>
<td>Summary of the Chapter</td>
<td>6</td>
</tr>
<tr>
<td>Organization of the Study</td>
<td>7</td>
</tr>
<tr>
<td>II. LITERATURE REVIEW</td>
<td>8</td>
</tr>
<tr>
<td>High-stakes Assessment</td>
<td>9</td>
</tr>
<tr>
<td>Retention</td>
<td>32</td>
</tr>
<tr>
<td>Special Education</td>
<td>42</td>
</tr>
<tr>
<td>Summary of the Chapter</td>
<td>57</td>
</tr>
<tr>
<td>III. METHODS</td>
<td>59</td>
</tr>
<tr>
<td>Sample Data</td>
<td>60</td>
</tr>
<tr>
<td>Procedures</td>
<td>60</td>
</tr>
<tr>
<td>Limitations of the Study</td>
<td>63</td>
</tr>
<tr>
<td>Independent Variables</td>
<td>64</td>
</tr>
<tr>
<td>Dependent Variables</td>
<td>66</td>
</tr>
<tr>
<td>Threats to Internal Validity</td>
<td>66</td>
</tr>
<tr>
<td>Data Design and Analyses</td>
<td>67</td>
</tr>
<tr>
<td>Summary of the Chapter</td>
<td>69</td>
</tr>
<tr>
<td>IV. RESULTS</td>
<td>70</td>
</tr>
<tr>
<td>Referral to Special Education Practices</td>
<td>70</td>
</tr>
<tr>
<td>Retention Practices</td>
<td>71</td>
</tr>
<tr>
<td>Referral and Retention Practices for Identified Groups</td>
<td>73</td>
</tr>
<tr>
<td>Summary of the Chapter</td>
<td>90</td>
</tr>
<tr>
<td>V. DISCUSSION</td>
<td>92</td>
</tr>
<tr>
<td>Referral and Retention Practices: Review of the Research Questions</td>
<td>93</td>
</tr>
<tr>
<td>Conclusions</td>
<td>106</td>
</tr>
<tr>
<td>Policy Implications and Future Research Considerations</td>
<td>108</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>111</td>
</tr>
<tr>
<td>VITA</td>
<td>123</td>
</tr>
<tr>
<td>TABLE</td>
<td>PAGE</td>
</tr>
<tr>
<td>-------</td>
<td>------</td>
</tr>
<tr>
<td>1. Minimum FCAT Performance Criteria for C, D and F</td>
<td>15</td>
</tr>
<tr>
<td>2. Higher FCAT Performance Criteria for A and B</td>
<td>15</td>
</tr>
<tr>
<td>3. FCAT Mean Scale Score Performance by Grade and Subject</td>
<td>17</td>
</tr>
<tr>
<td>4. Percentages of Students at Each Performance Achievement Level</td>
<td>18</td>
</tr>
<tr>
<td>5. Elementary School Demographics in M-DCPS</td>
<td>59</td>
</tr>
<tr>
<td>6. Descriptive Statistics–Referral and Retention Percentages Prior to and After High-Stakes Implementation</td>
<td>71</td>
</tr>
<tr>
<td>7. Total Number of Referrals for Elementary Schools-Prior to and After High-Stakes Implementation</td>
<td>72</td>
</tr>
<tr>
<td>8. Total Number of Retentions for Elementary Schools-Prior to and After High-Stakes Implementation</td>
<td>72</td>
</tr>
<tr>
<td>9. Repeated Measures ANOVA for Individual Ethnic Groups-Referral Summary</td>
<td>75</td>
</tr>
<tr>
<td>10. Referral Population Means for Ethnic Groups Prior to and After High-Stakes Implementation</td>
<td>76</td>
</tr>
<tr>
<td>11. Repeated Measures ANOVA for Individual Ethnic Groups-Retention Summary</td>
<td>76</td>
</tr>
<tr>
<td>12. Retention Population Means for Ethnic Groups Prior to and After High-Stakes Implementation</td>
<td>77</td>
</tr>
<tr>
<td>13. Repeated Measures ANOVA for Individual Grade Levels-Referral Summary</td>
<td>78</td>
</tr>
<tr>
<td>14. Referral Population Means for Grade Levels Prior to and After High-Stakes Implementation</td>
<td>78</td>
</tr>
<tr>
<td>15. Repeated Measures ANOVA for Individual Grade Levels-Retention Summary</td>
<td>79</td>
</tr>
<tr>
<td>16. Retention Population Means for Grade Levels Prior to and After High-Stakes Implementation</td>
<td>80</td>
</tr>
<tr>
<td>17. Repeated Measures ANOVA for Individual Genders-Retention Summary</td>
<td>80</td>
</tr>
</tbody>
</table>
CHAPTER I
INTRODUCTION

In the not too distant past, Americans began to feel that the public education system was failing their children. Students were graduating with a high school diploma, but without the basic skills necessary to compete in the domestic workplace and the global economy (National Commission on Excellence in Education, 1983). Federal legislation mandated that states adopt challenging standards in content and performance. With the reauthorization of the Elementary and Secondary Education Act, (the No Child Left Behind Act of 2001), increased accountability measures for schools were required. Mastery of content in reading and mathematics was to be measured through yearly standardized assessment for children in grades three through eight, as well as one year in high school. Large scale assessment was seen as a way of ensuring school improvement through the use of public accountability (Stiggins, 1999). Teachers would teach to the content of the standards and objectives of the test in order to promote student mastery (Popham, 1993).

While the intent of standardized assessment is to push educators to improve student learning outcomes, a number of critical decisions are made from individual and school results, thus elevating these assessments to the status of "high-stakes" testing. For schools, achievement performance levels help to determine which will receive rewards and recognition and which will receive assistance and/or administrative punitive action. For individual students, their performances on these tests determine academic tracks, promotion or retention, and whether they will graduate from high school or not (Popham, 1993). These consequences may not be in the best interest of individual students and run
contrary to the standards set by the National Research Council regarding the use of large
scale assessments (Heubert & Hauser, 1999).

By connecting high-stakes assessment to rewards and sanctions, legislatures have been able to assure that public policy is able to drive the implementation of educational objectives and outcomes (Madaus & Kellaghan, 1993). If educators are to be judged by students' mastery of test content as evidenced by their performances on high-stakes assessment, they will attempt to excel in that judgment (Popham, 1993) through a number of available options. First, they may change the actual achievement and learning of their students. However, little evidence exists to support that the requirement of high-stakes assessment alone has much impact on student achievement outside of an increase in scores on the particular assessment being used (Mehrens, 1998). Second, educators may also teach to the test (Popham, 2001). Third, states may set cut-off scores for acceptable performance lower than initially set in order that more students pass (Heubert, 2001). A fourth option is to eliminate the scores of low-achieving students through retention or referral to special education (Allington & McGill-Franzen, 1992; Heubert, 2001). These practices may potentially prove harmful or provide questionable benefit to individual students (Kavale & Forness, 1999; Rafter & Carey, 1995).

Intended and unintended effects are expected upon the implementation of high-stakes testing. The intended effects are, of course, better schools with greater student achievement. Critics (Allington & McGill-Franzen, 1992; Heubert, 2001) fear that unintended effects for low-scoring students include higher incidence of retention and referral to special education. Excluding their scores raises the overall score for a school, thereby limiting punitive action. It is imperative to determine if retention and referral to
special education practices have changed in response to the implementation of high-stakes testing. And as these practices have historically been used disproportionately with students who are poor, male and/or of minority/ethnic background, it is particularly important to evaluate the usage of these practices on members of these groups.

Retention in grade is the single strongest predictor of dropping out of school (Shepard & Smith, 1989) and correlates negatively with a student's personal adjustment, self-concept and peer acceptance (Jimerson, 2001; Rafoth & Carey, 1995). Referring a child for special education evaluation almost certainly guarantees that a psychoeducational evaluation will occur, resulting in special education placement at a rate of 70 to 74% (Ysseldyke, Vanderwood, & Shriner, 1997). In many districts this means the student will be placed out of the general education classroom and into special education (Kirk, Gallagher, & Anastasiow, 2003), the efficacy of which has long been questioned (Kavale & Forness, 1999).

The Present Study

As the popularity of the accountability movement grows and the legal requirements are implemented, it is important to evaluate retention and special education referral practices. Although educators believe that retention and referral to special education are effective interventions for students demonstrating academic and/or behavioral difficulties (Byrnes, 1989; Podell & Soodak, 1993; Soodak & Podell, 1993; Tomchin & Impara, 1992), the long-term outcomes for children who have been retained are uniformly negative (Holmes, 1989; Jimerson, 2001), while the outcomes for children placed in special education are variable (Carlberg & Kavale, 1980; Kavale & Forness, 1999; Wang & Baker, 1985-1986). This is particularly troubling, as retentions (Jacob,
2002; Mc-Gill-Franzen & Allington, 1993) and placements in special education (Jacob, 2002; Figlio and Getzer, 2002) have been observed to increase as the demands of accountability have risen. To utilize these interventions in an effort to eliminate the scores of low-achieving students demonstrates at the very least, educational ignorance and naiveté, and at the worst, educational malpractice.

Through the evaluation of these practices, educators may be enabled to make better educational decisions that will benefit children. For example, it is important to determine how retention practices and special education referrals change as schools respond to the demands of high-stakes testing. Of particular importance is how decisions related to retention and special education referrals impact students from lower socioeconomic communities and those from culturally diverse backgrounds. Research has supported that students with the aforementioned backgrounds are more likely to be retained and referred for special education (Anderson, 1997; Brantlinger, 1997; Figueroa & Artiles, 1999; Harry & Anderson, 1995; McNamara, 1998). This information has important consequences related to the use of financial resources, the implementation of legislative action and the development of social policy.

It was assumed that in the present study, a relationship would exist between the usage of retention and referral to special education, and the implementation of high-stakes testing. In an effort to eliminate the scores of low-achieving students from the testing pool, a significant increase would be noted after the implementation of the Florida Assessment Comprehensive Test (FCAT) and the subsequent awarding of rewards and sanctions associated with this test. In an effort to avoid punishments and sanctions, an increase in the usage of these practices would be most pronounced the years after
mandatory high-stakes testing and by those with the most to lose, low-performing schools.

**Problem Statement**

While the intent of standardized high-stakes assessment is to push educators to improve student learning outcomes, a number of critical decisions are made from individual and school results. Achievement performance levels help to determine which schools receive rewards and recognition and which receive assistance and/or administrative punitive action. In an effort to guarantee that their schools will receive rewards and recognition, educators may attempt to eliminate the scores of low achieving students through retention or referral to special education (Heubert, 2001). A large-scale inquiry in an urban area was proposed to evaluate these practices prior to and after the implementation of Florida's high-stakes test, the Florida Comprehensive Assessment Test.

**General Purpose of the Study**

The primary purpose of this research was to examine retention and special education referral practices as they occurred prior to and after the implementation of high-stakes testing. The secondary purpose was to examine if these practices were more likely to be used for particular groups of students (based on minority/ethnic status, gender, and enrollment at low, medium and high achieving schools).

**Research Questions and Hypotheses**

To conduct this research, three major research questions and their related hypotheses were proposed:
Question #1: What were the referral rates to the special education referral process prior to and after the implementation of high-stakes testing in a large urban school system?

Hypothesis #1: The use of referral to the special education referral process will increase after the implementation of high-stakes testing.

Question #2: What were the retention rates prior to and after the implementation of high-stakes testing in a large urban school system?

Hypothesis #2: The use of retention will increase after the implementation of high-stakes testing.

Question #3: If retention and special education referral rates have changed after the implementation of high-stakes testing, are the changes disproportionate for any identified group of students in a large urban school setting?

Hypothesis #3: As a result of the implementation of high-stakes testing in a large urban school setting, a disproportionate representation of particular identified groups for retention and referral to special education will be evident.

Summary of the Chapter

The current study was intended to examine the practices of retention and referral to special education prior to and after the implementation of Florida's high-stakes assessment, the Florida Comprehensive Assessment Test. Federal legislation has mandated that children in grades three through eight, as well as one year in high school, be assessed on a yearly basis in the areas of reading and mathematics. Large scale assessment has been seen as the mechanism to ensure school improvement. In the state of Florida, accountability for student learning outcomes is largely assessed by student performances on the FCAT. Performances on this test determine rewards and sanctions at the district, school and individual student levels. In an effort to guarantee that their
schools receive rewards and recognition, scores of low achieving students may be eliminated through retention or referral to special education.

Organization of the Study

The remaining sections of the research study are divided into the following chapters.

Chapter II, “Literature Review” reviews and discusses the related literature on high-stakes testing, the Florida Comprehensive Assessment Test, and important issues concerning retention and special education.

Chapter III, “Method”, describes the research design, description of the participants, and procedures for data collection and analysis. Definitions of key terms and variables are also provided in this chapter.

Chapter IV, “Results”, presents the findings and evaluation of the data.

Chapter V, “Discussion”, is based upon the data presented in Chapter IV. Findings are summarized and discussed. Limitations of the study and recommendations for further research are also discussed.
The accountability movement was initiated with the publication of A Nation at Risk (National Commission of Excellence in Education, 1983). The commission argued that the American educational system was mediocre and had lost sight of its basic purpose, the setting of "high expectations and disciplined effort needed to attain them" (National Commission of Excellence in Education, 1983, A Nation at Risk section, ¶ 4). Their assertion was supported, in part, by a decrease in scores on standardized measures of achievement (Airasian, 1988). They challenged school systems to set higher standards for teachers and for learners. One of the recommendations included the standardized assessment of students through state and local systems. Assessments were to take place at major decision points in order to certify students’ skills and to provide feedback regarding the need for remedial or advanced instruction (National Commission of Excellence in Education, 1983). If students were taught the content related to higher standards, then a test could be administered to determine if the students met the learning expectations of those standards (Barksdale-Ladd & Thomas, 2000).

The development of national goals for the education of children was begun during the administration of President George H.W. Bush. During President William Clinton's first term, these goals were passed by the Congress in Goals 2000: Educate America Act (1994). Goal 3 of this act states children in grades 4, 8 and 12 will demonstrate competency of challenging academic standards. The content of various states' educational standards were also being developed during this same period of time (Barksdale-Ladd & Thomas, 2000). "Accountability" for student learning outcomes
became the rallying cry of the nation's governors and legislative bodies (Popham, 1999).

While education had historically been under local control, the politics of educating children had become a federal and state matter. Learning standards and the way to assess the attainment of those standards was no longer just a local concern. (Dorn, 1998).

With the reauthorization of the Elementary and Secondary Education Act, the "No Child Left Behind Act" of 2001, Congress mandated increased accountability for student learning outcomes. Higher standards were to be required with the intention of closing the achievement gap between rich and poor children. Students were to be assessed on reading and mathematics in grades three through eight and in one year of high school by the year 2005-06 (The No Child Left Behind Act, 2001). While federal legislation continued to define and formalize this process, many states already had in place measures for assessment of student learning outcomes.

High-stakes Assessment

Curriculum standards were established by all of the states during the 1990's (Barksdale-Ladd & Thomas, 2000). By the middle of that decade, the majority of states had measures in place that assessed these standards; 43 states used the assessments for instructional improvement, while 38 of the states used their assessments for program evaluations (Roeber, Bond & Braskamp, 1997). However, a number of critical decisions began to be made from the results of these measures, quickly leading to their becoming high-stakes in nature. Decisions for individual children include whether a child will graduate, be promoted or retained, or if they will be referred to special classes (Airasian, 1993). Inferences regarding current teachers' skills could also be made through the performances of the children in their classes on high-stakes assessments (Neill, 1999).
Rewards could be given to those whose students excelled and assistance given to those whose students did not. For school systems, entire instructional programs could be judged based upon the performances of the children under their jurisdiction (Dorn, 1998). Intricate systems of rewards and sanctions were established with the expressed purpose of providing a means for motivating changes in teaching and learning practices. Without such rewards and sanctions, it was thought change would not occur (Kohn, 2000b).

**Development of Florida's Accountability Measure: Florida Comprehensive Assessment Test**

In 1993, development of the Sunshine State Standards began. These standards defined the *content of the curriculum* in the public schools of the state of Florida and were adopted by the State Board of Education in 1996 (Florida Department of Education, 1999). During this same time, the Florida Commission on Education Reform and Accountability recommended that student learning be assessed. By requiring students to be assessed, it was believed that educational expectations would be raised and students would learn the necessary skills to be competitive in the job market. Their recommendations were adopted by the State Board of Education in 1995. A statewide assessment was developed to assess the content standards of reading and mathematics as articulated in the Sunshine State Standards, as well as the first four standards of Goal 3 of Blueprint 2000 (Florida Department of Education, 2001). The Sunshine State Standards defined statewide curriculum content, while the statewide assessment instrument, the Florida Comprehensive Assessment Test (FCAT), evaluated student achievement of this curriculum.
In June and July of 1996, the Content Advisory Committee of the Florida Department of Education was formed. Its members consisted of teachers, citizens, business leaders, administrators, and curriculum specialists. The committee determined which benchmarks of the Sunshine State Standards would be assessed and provided guidelines for test question development (Florida Department of Education, 2001). Test questions would not only include the traditional multiple-choice format but performance task questions requiring short and long written responses. Questions were to demand a student to "analyze, synthesize, and evaluate the information presented to apply strategies or procedures they have learned" (Florida Department of Education, 1999, p. 3). A separate committee, the Technical Advisory Committee, was formed to assist with the technical aspects of test development. They provided expertise regarding test procedures and analysis.

Test questions were written by the private contractor for grades 4, 5, 8 and 10 and reviewed by the various stakeholders from July to October, 1996. Field testing of more than 1200 questions occurred in March, 1997 and subsequent test question quality review was conducted (Florida Department of Education, 2001). From this pool of questions, the first FCAT was developed for children in grades 4, 5, 8 and 10.

The Florida Board of Education reports that the FCAT is reliable and valid. The FCAT Briefing Book (2001) reports reliability coefficients above .90. Concurrent validity of the FCAT with the SAT-9 is reported to be between .71 and .81. Studies conducted by the department suggest no indication of item bias. The standard error of measurement is approximately 20 at the center of the distribution. Towards the ends of the distribution, it is reportedly higher (Florida Department of Education. 2002c).
Administration of the FCAT

The FCAT was administered to obtain baseline data in February, 1998. Although schools were made aware of their results, it was not until November, 1998, that achievement performance levels were determined by a committee of business leaders, educators and private citizens and approved by the State Board of Education. Five levels were to be reported, from a high of 5 to a low of 1 (Florida Department of Education, 1999).

The second administration of the FCAT occurred in February, 1999. School accountability was determined for the first time using these results. Schools were assigned performance grades based largely on how their students performed on this test (Florida Department of Education, 2001).

Testing was expanded to include field testing of reading and mathematics questions for all students in grades 3 through 10 in February/March, 2000. While the format of the original FCAT design was maintained for students in grade 4, 5, 8 and 10, multiple-choice format reading tests were given to children in grades 3, 5, 6, 7 and 9 and mathematics multiple-choice format tests were given to children in grades 3, 4, 6, 7 and 9. A national norm-referenced test (SAT-9) was also included for all children to allow for comparison to a national standardization group. Scores on the national-norm-referenced tests were reported for all students as well as the FCAT scores of the previously assessed grade levels. School performance grades were assigned utilizing the FCAT score results for grades 4, 5, 8 and 10. It was not until 2001 that the additional grade levels were reported and included in the school performance grading system. It was also this year that
the writing test, Florida Writes, was included as part of the FCAT (Florida Department of Education, 2001).

The FCAT was administrated to all students in the specified grades with a few exceptions. Students who were considered Limited English Proficient (LEP) and had been in the program for two years or less, or who a Limited English Proficient committee decided to exempt from testing, were exempt; as well as exceptional student education students whose cognitive ability or behavior prevented them from participating, even if accommodations were provided. An Individualized Educational Plan would provide documentation for exclusion of special education students and alternate assessments were to be developed (Florida Department of Education, 2002b). Other limited English proficient and special education students were required to take the FCAT but their scores were not reported in the Standard Curriculum reporting groups (Florida Department of Education, 1999). Therefore, their scores were not utilized when determining school grades.

School Improvement through Accountability

The key focus of school improvement in education in the State of Florida was accountability. The FCAT was the tool to measure the performance of students' mastery of the Sunshine State Standards in reading, writing and mathematics (Florida Department of Education, 2001). As there is a direct link between the curriculum standards and the content of the FCAT, the FCAT was seen as a way to ensure improved learning outcomes. For students to perform well on the FCAT, it would be necessary for teachers to teach the curriculum of the Sunshine State Standards. The data from the FCAT would be used to report progress of individuals, schools and districts. Schools were to be
evaluated on aggregate student performance data which would be linked to accountability rewards and sanctions. Initially, this was to occur in grades 4, 5, 8 and 10. However, with the legislative establishment of the state governor's A+ plan in 1999, achievement levels of grades 3 through 10 were to be reported by 2001.

Student performance on the FCAT ranges from a high of 5 to a low of 1. Each performance level is associated with a range of scores. The range of scores associated with each performance level was established in November, 1998 by the State Board of Education. The associated scores remained constant over the 1999, 2000, and 2001 administrations. The ranges of scores for minimal level performance at each performance level were to be increased beginning with the 2002 administration (Florida Department of Education, 1999).

In determining rewards and sanctions, data from the most recent two years were to be considered. School grades were assigned to individual schools based on FCAT performance and other criteria of the Florida School Recognition Program Section 231.2905. F.S. For the years 2000 and 2001, the minimum FCAT performance criteria established for grades C, D and F. and the higher FCAT performance criteria established for grades A and B. are presented in Tables 1 and 2 (Florida Department of Education, 2001). In determining what FCAT performance criteria are met, percentages of students performing at different achievement performance levels were calculated for reading, mathematics and writing.

Additional criteria were considered before the final determination of school grades. For example, if a school received a grade of A, not only would they have met the higher performance criteria established for A and B, but they also would have tested at
least 95% of eligible students and either improved or maintained the reading scores of their lowest performing students. In addition, these schools would have maintained their mathematics and writing performance levels as well as have met other established criteria regarding dropout, suspension and absenteeism rates. For schools assigned an F, they would have scored below the minimum criteria in reading, mathematics and writing and tested less than 90% of their students (Florida Department of Education, 2001).

Table 1

*Minimum FCAT Performance Criteria for C, D and F*

<table>
<thead>
<tr>
<th>Instructional Level</th>
<th>Reading</th>
<th>Mathematics</th>
<th>Writing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elementary</td>
<td>60% score</td>
<td>60% score</td>
<td>50% score</td>
</tr>
<tr>
<td></td>
<td>Level 2 &amp; Above</td>
<td>Level 2 &amp; Above</td>
<td>Level 3 &amp; Above</td>
</tr>
<tr>
<td>Middle</td>
<td>60% score</td>
<td>60% score</td>
<td>67% score</td>
</tr>
<tr>
<td></td>
<td>Level 2 &amp; Above</td>
<td>Level 2 &amp; Above</td>
<td>Level 3 &amp; Above</td>
</tr>
<tr>
<td>High</td>
<td>60% score</td>
<td>60% score</td>
<td>75% score</td>
</tr>
<tr>
<td></td>
<td>Level 2 &amp; Above</td>
<td>Level 2 &amp; Above</td>
<td>Level 3 &amp; Above</td>
</tr>
</tbody>
</table>

Table 2

*Higher FCAT Performance Criteria for A and B*

<table>
<thead>
<tr>
<th>Instructional Level</th>
<th>Reading</th>
<th>Mathematics</th>
<th>Writing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elementary</td>
<td>50% score</td>
<td>50% score</td>
<td>67% score</td>
</tr>
<tr>
<td></td>
<td>Level 3 &amp; Above</td>
<td>Level 3 &amp; Above</td>
<td>Level 3 &amp; Above</td>
</tr>
<tr>
<td>Middle</td>
<td>50% score</td>
<td>50% score</td>
<td>75% score</td>
</tr>
<tr>
<td></td>
<td>Level 3 &amp; Above</td>
<td>Level 3 &amp; Above</td>
<td>Level 3 &amp; Above</td>
</tr>
<tr>
<td>High</td>
<td>50% score</td>
<td>50% score</td>
<td>80% score</td>
</tr>
<tr>
<td></td>
<td>Level 3 &amp; Above</td>
<td>Level 3 &amp; Above</td>
<td>Level 3 &amp; Above</td>
</tr>
</tbody>
</table>
Schools graded D and F receive district assistance. Intervention plans are devised with a copy going to the Department of Education for the grade F schools. They may receive grant award preferences as well as technical assistance priority. Action will be taken if a school receives an F for two consecutive years, goes back to an F within a five year period, or receives an F after three years. There is also a potential for action if a school receives a grade of D after three years (Florida Department of Education, 2001).

Children at schools graded F for two years may apply for Opportunity Scholarships so they may attend a different school with higher performance levels (Florida Department of Education, 2001). Action by law, or through the withholding of lottery dollars, may be taken against the school. Conversely, schools that receive a grade of A may be eligible for monetary incentives. Schools that improve at least one letter grade may also be eligible for incentives (Florida Department of Education, 2001).

Performance on the FCAT

Starting with the 1999 administration, students’ performance scores were determined and the percentage of students achieving at the various levels was reported by schools, districts and the state. The proportion of students at each performance level was to be compared from year to year to determine growth (Florida Department of Education, 1999). The ranges of scores associated with the various performance levels remained constant through December, 2001. They were increased starting in January, 2002 (Florida Department of Education, 2001). Performances at level 5 indicate students are able to demonstrate competency of the most challenging of the Sunshine State Standards. Performances at level 1 suggest the opposite; students have little success in demonstrating competency (Florida Department of Education, 1999).
When comparing performances of children across the state's districts over time, it is necessary to compare performance on the parts of the FCAT that have been tested since 1998; reading in grades 4, 8, and 10; mathematics in grades 5, 8 and 10.

Table 3

*FCAT Mean Scale Score Performance by Grade and Subject*

<table>
<thead>
<tr>
<th>Grade/Subject</th>
<th>1998</th>
<th>1999</th>
<th>2000</th>
<th>2001</th>
<th>2002</th>
</tr>
</thead>
<tbody>
<tr>
<td>4/Reading</td>
<td>294</td>
<td>288</td>
<td>293</td>
<td>298</td>
<td>299</td>
</tr>
<tr>
<td>5/Math</td>
<td>300</td>
<td>303</td>
<td>314</td>
<td>314</td>
<td>318</td>
</tr>
<tr>
<td>8/Reading</td>
<td>298</td>
<td>295</td>
<td>290</td>
<td>295</td>
<td>295</td>
</tr>
<tr>
<td>8/Math</td>
<td>299</td>
<td>296</td>
<td>303</td>
<td>308</td>
<td>305</td>
</tr>
<tr>
<td>10/Reading</td>
<td>299</td>
<td>302</td>
<td>298</td>
<td>304</td>
<td>303</td>
</tr>
<tr>
<td>10/Math</td>
<td>302</td>
<td>308</td>
<td>311</td>
<td>323</td>
<td>319</td>
</tr>
</tbody>
</table>

Comparisons may be made by looking at changes in the average mean scores (Table 3) of students as well as the percentages (Table 4) of students scoring at each of the five achievement levels. In 1997-98 when baseline information was obtained, scores were reported for Standard Curriculum students. For subsequent years, data from other students was also included, accounting for some of the decrease in performances at some of the grade levels after the baseline year. However over time, slight increases in average mean score performances generally have been noted in reading at grade levels 4 and 10 with a slight decrease in performance for grade 8. In mathematics, stronger increases in performances are evident in grades 5 and 10 with a more modest increase in performance for grade 8. The same observation may be made regarding the percentages of students at
the five achievement levels. Generally over time, the percentages of students scoring at
the lowest achievement levels 1 and 2 have decreased, while the percentages of students
at the highest achievement levels 4 and 5 have increased in both reading and mathematics
(Florida Department of Education, 2001; Florida Department of Education, 2002a).

Table 4

Percentages of Students at Each Performance Achievement Level

<table>
<thead>
<tr>
<th>Subject</th>
<th>Grade</th>
<th>Year</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reading</td>
<td>Grade 4</td>
<td>1999</td>
<td>36</td>
<td>17</td>
<td>29</td>
<td>17</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2000</td>
<td>33</td>
<td>16</td>
<td>29</td>
<td>19</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2001</td>
<td>31</td>
<td>16</td>
<td>28</td>
<td>18</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2002</td>
<td>30</td>
<td>15</td>
<td>28</td>
<td>21</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>Grade 8</td>
<td>1999</td>
<td>28</td>
<td>28</td>
<td>31</td>
<td>12</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2000</td>
<td>32</td>
<td>29</td>
<td>27</td>
<td>11</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2001</td>
<td>30</td>
<td>27</td>
<td>26</td>
<td>13</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2002</td>
<td>29</td>
<td>26</td>
<td>28</td>
<td>14</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Grade 10</td>
<td>1999</td>
<td>33</td>
<td>37</td>
<td>19</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2000</td>
<td>35</td>
<td>36</td>
<td>19</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2001</td>
<td>33</td>
<td>32</td>
<td>19</td>
<td>7</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2002</td>
<td>32</td>
<td>33</td>
<td>21</td>
<td>8</td>
<td>7</td>
</tr>
<tr>
<td>Math</td>
<td>Grade 4</td>
<td>1999</td>
<td>33</td>
<td>32</td>
<td>21</td>
<td>12</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2000</td>
<td>26</td>
<td>29</td>
<td>24</td>
<td>17</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2001</td>
<td>27</td>
<td>25</td>
<td>22</td>
<td>20</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2002</td>
<td>25</td>
<td>27</td>
<td>23</td>
<td>19</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>Grade 8</td>
<td>1999</td>
<td>33</td>
<td>23</td>
<td>27</td>
<td>11</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2000</td>
<td>29</td>
<td>20</td>
<td>28</td>
<td>13</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2001</td>
<td>24</td>
<td>21</td>
<td>31</td>
<td>14</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2002</td>
<td>25</td>
<td>22</td>
<td>31</td>
<td>14</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>Grade 10</td>
<td>1999</td>
<td>27</td>
<td>26</td>
<td>24</td>
<td>19</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2000</td>
<td>26</td>
<td>23</td>
<td>23</td>
<td>22</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2001</td>
<td>19</td>
<td>20</td>
<td>24</td>
<td>27</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2002</td>
<td>19</td>
<td>21</td>
<td>25</td>
<td>27</td>
<td>8</td>
</tr>
</tbody>
</table>
The same observation may be made when evaluating overall school performances. Seventy-six schools were classified as low performing, receiving the grade of “F” after the 1998-99 FCAT administration. This number fell to four after the 1999-2000 administration and to zero after the 2000-2001 administration. The accountability performance bar was raised after this time and the number of low performing schools jumped to 64 in 2001-2002. Results for the 2002-2003 FCAT administration indicate only 35 schools received a failing grade (Florida Department of Education, 2003).

High-Stakes Accountability and the Public

The establishment of curriculum standards and the demonstration of skills competencies by students (as assessed through accountability measures) appear to have been positively received by the public (Popham, 1987). The public’s opinion regarding the efficacy of public schooling has steadily declined over the past 30 years and state mandated standards with accompanying high-stakes assessment has been seen as a way to fix a crippled system (Airasian, 1988; Madaus & Kellaghan, 1993). The thought is, if meaningful curricula standards are developed, then teachers will teach the content of these standards (Airasian, 1988). The mastery of these standards would be reflected in the performances of children on state accountability measures. Increases in performances would, therefore, indicate an increase in student achievement. "State-mandated, high-stakes testing programs have gained widespread popular support in the quest for heightened educational standards because they symbolize order and control, desired educational outcomes, and traditional moral values" (Airasian, p 301). Those who
worked hard and could demonstrate mastery of curriculum standards would be rewarded, while those who could not would receive sanctions.

The results of standardized assessment are powerful and are perceived by the public, at least, as being objective and legitimate indicators of students' mastery of state standards (Airasian, 1988; Dorn, 1998). Accountability test results have become the primary way in which the effectiveness of entire school systems, individual schools, teachers and students have been judged (Dorn, 1998). Because of the popularity of the accountability movement, other ways of judging schools and students have been lost. The public's judgments regarding the performances of schools have been reduced down to just one: performance on the state's accountability measure.

Even though accountability measures have been elevated to a prominent level of importance in the mind of the public, rarely is the public made aware of the shortcomings of such assessments. Only a narrow range of abilities can be assessed (Airasian, 1988) by any given accountability instrument. Rarely is the public aware of which curriculum objectives are being measured. Inferences regarding generalized functioning are made from performances on tests assessing limited content and scope (Canner, 1992). As high-stakes assessments are administered only once a year, they cannot provide appropriate feedback adequate to guide everyday classroom instruction. However, as long as performances on the state accountability measures are acceptable, few questions appear to be asked about classroom instructional practices (Dorn, 1998).

While high-stakes tests may be appropriate for informing the public about relative performances of children, they are not valid for many of their uses, such as making decisions regarding promotion and graduation (Smith & Fey, 2000). According to the
American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education (1999), tests must be valid for their intended purpose and should not be used to make academic placement decisions; such as retention or low-ability tracking; of questionable efficacy. Never should a single measure be used to support a critical decision, but rather decisions should be made after a reflection upon the data provided through multiple sources of information. These issues, however, do not appear to be a concern for the public at large. The results of 70 polls and surveys on test related questions, indicated the general public, parents, students, and even some teachers, wanted more high-stakes assessment (Phelps, 1998).

*High-Stakes Accountability and the Classroom*

Assessment has long been viewed as a means of gathering data to inform decision-making regarding student learning outcomes. According to the International Reading Association (1999), "...the primary purpose of assessment is to help students by providing information about how instruction can be improved" (p. 258). In order to accomplish this, assessment needs to be frequent so that changes in classroom instructional practices may be made to accommodate the needs of individual students. As high-stakes assessments are generally given only once a year, they do not appear to be able to provide data easily for improvement of everyday instructional practices (Dorn, 1998).

Although the expressed purpose of "accountability" testing is to ensure greater student learning outcomes, in actuality it is a way to ensure that states have control over the content of the individual classroom curriculum (Airasian, 1993). The higher the stakes associated with the results of said assessments, the greater the chances that
assessment content will be reflected in the curriculum of every day instruction; thereby creating measurement-driven instruction (Popham, 1998). As the curriculum aligns with the content of the assessment, a narrowing of the curriculum occurs. Activities in the classroom tend to be concentrated on reading, mathematics and writing--areas assessed by accountability measures. The instruction of other subjects such as science, social studies, fine arts and physical education suffers (International Reading Association, 1999; Kohn, 2000b; Madaus, 1998). In studies conducted in various high-stakes testing states, investigators documented that most teachers spend the majority of their time teaching the basics that are assessed by high-stakes testing (Gordon & Reese, 1997; Jones et al., 1999; Smith, 1991b). No new learning occurs for weeks prior to actual testing (Gordon & Reese, 1997). Teaching to the test is even more likely to be engaged in by teachers at schools with large populations of minority and poor children (Madaus, West, Harmon, Lomax, & Viator, 1992; McNeil & Valenzuela, 2000; Orfield & Wald, 2000) as these schools demonstrate a greater need to elevate test scores.

Questionable teaching practices may be utilized by those that feel "under pressure" from administrators and the public. As the pressure to perform increases, teachers are more likely to devote greater time to test preparation and teaching to the test (Johnston, 1998; Smith, 1991a). In interviews conducted with teachers (Barksdale-Ladd & Thomas, 2000), informants indicated they believed that high-stakes assessment was in response to a public who found public school teachers ineffectual. However, many of the informants did not believe that high-stakes testing had improved teaching practices. Responses regarding daily classroom activities lend support to this belief. In response to pressure exerted by administrators, teachers indicated they were primarily engaged in
four different activities: teaching test content information, test format practice, test-taking skills, and daily test preparation (Barksdale-Ladd & Thomas, 2000). A survey of Texas teachers (Gordon & Reese, 1997) indicated that classroom teaching primarily consisted of familiarizing students with test items, not having them attain mastery of underlying learning concepts. Instruction consists of giving multiple examples, looking for cues, repeated demonstrations and rote practice with item variations. Few teachers are willing to be innovative (Gordon & Reese, 1997; Meier, 1997). Instead of addressing the issues that may be responsible for low achievement -- poverty, class size, lack of resources (Florida Association of School Psychologists, 2002)-- teachers at low achieving schools feel that the only control they have is controlling of test preparation (Smith, 1991a).

The publishing of test scores causes further difficulties for teachers. Teachers of low-achievers experience increased levels of alienation and shame and a decrease in self-esteem (Smith, 1991a). Teachers of students with both high-and low-achievement experience anxiety, particularly as the stakes of accountability measures increase (Kohn, 2000a, Mehrens, 1998). Students report experiencing resentment, anger and anxiety (Gordon & Reese, 1997). They are less confident in their abilities and there is a negative impact on their love of learning (Jones et al., 1999).

When performance is either rewarded or sanctioned, it is not surprising that some individuals resort to unethical practices. Incidences of cheating have been documented in some classrooms (Yang, 2000; Zlatos, 1996) and teaching to the test has become teaching the test (Canner, 1992; Smith, 1991b; Yang, 2000). In other classrooms, certain groups of students have been targeted for extra instruction (Heubert & Hauser, 1999) so that they may perform well on the tests. While extra instruction seems a desirable outcome, it is
the intention behind the instruction that is problematic--singling out only those students whose scores have the best chance to effect overall school score results as a result of additional attention. Which students becomes the focus is dependent upon which statistic is used to assess school performance. If the mean score is used, efforts will be concentrated on those students at the extreme of the distribution. If, however, the percent passing score is used, the focus will be on those whose scores are close to the cut-off passing score (Rafferty & Treff, 1994). Still in other classrooms, teachers turn against the very students that need the most assistance, the "undesirables" whose performances represent the low end of the high-stakes assessment bell curve (Kohn, 2000a).

Experienced teachers are leaving urban settings, which have greater concentrations of poor and minority students, for suburban ones (Lee, 1998). The neediest of students are being left with the highest numbers of new, unprepared teachers (Hilliard, 2000), many of which are teaching out-of-field. Other teachers are simply abandoning the profession altogether (Kohn, 2000b). And, when students with disabilities are to be included in the testing pool as mandated by the Individual with Disabilities Educational Act of 1997, they may be encouraged to stay home or be sent on a field trip on the day of testing (Kantrowitz & Springen, 1997).

**Accountability Outcomes**

Unethical and questionable practices occurring in the classroom lead to serious concerns regarding the validity of high-stakes test results (Kohn, 2000a; Mehrens, 1998). Inferences are being drawn by the public that cannot be supported.

When any new test is developed, failure rates are usually the highest right after the test's introduction. For example, on graduation tests, rates of failure gradually decline
in the years following their introduction (Linn, 2000). Actual changes in student learning may not be occurring, as these same improvements are not seen on the National Assessment of Educational Progress (NAEP) (Heubert, 2001). Changes in teaching to the content domains of the test may account for these results. Scores increase as a result of teaching to the test and familiarity with test content and format (Linn, 2000; Mehrens, 1998).

This trend is noted for various measures of high-stakes assessment. Increases in test performances do not appear to generalize to other measures of achievement, a fact suggesting no actual increase in student learning and mastery (Mehrens, 1998). For students in Texas, scores on the state's high-stakes test, the Texas Assessment of Academic Skills (TAAS), have increased over time and the achievement gap between whites and minorities has narrowed (Hoffman, Assaf & Paris, 2001). In 1994, only 55% of the students had passed the TAAS. Whereas by 1997, that proportion of students had risen to 74%. However, after careful evaluation (Haney, 2000), a number of problems were found. Overall comparisons of performances to that of performances on the NAEP indicated that TAAS results may not be valid. Increases in performances on the TAAS may have been due to teaching to the test (Haney, 2000) and intensive test preparation (McNeil & Valenzuela, 2000), with an inordinate amount of time devoted to these practices at schools with large minority populations. Also, contributing to this phenomena was the exemption or exclusion of students from the testing pool, particularly minority students, through retention in grade, dropping out of school, or referral to special education program placements (Haney, 2000). When comparing the performances
of minority students to that of majority students, it was noted they obtained lower passing scores (Haney, 2000).

The structure of the NAEP strongly influenced the development of Kentucky’s high-stakes test, the Kentucky Instructional Results Information System (KIRIS), making comparisons particularly valuable. Even though the two tests were similar in their make-up, score gains on the two tests were very different between the years of 1992 and 1996 (Koretz & Barron, 1998). When comparing performances in reading for the first two years of this time period for the fourth grade, KIRIS scores increased by three-fourths of a standard deviation. Virtually, no change in the NAEP occurred. In mathematics, it was possible to compare performances for both fourth and eight grades from 1992 to 1996. For the fourth grade, score gains on the KIRIS were .61 standard deviation and on the NAEP, .17 standard deviation. For the eighth grade, score gains on the KIRIS were .56 standard deviation and on the NAEP, .13 standard deviation. Results did not generalize and the authors concluded that even the rise in scores on the NAEP was similar to that experienced by other states during that same period of time.

Similar results have been obtained by Smith, Fey, Heinecke, Miller-Kahn, and Noble (as cited in Smith & Fey, 2000) when examining the performance outcomes for students in the Chicago Public Schools. In Chicago, student achievement is assessed on the Iowa Test of Basic Skills (ITBS). While these scores have continued to rise, the same increases in performance have not been noted on the NAEP. Factors such as test preparation and the exclusion of students through transfer or dropping-out were not considered by the Chicago Public School system when analyzing their results on the ITBS.
The generalization of results was examined by comparing test results under low and high-stakes conditions (Koretz, Linn, Dunbar & Shepard, 1991). From 1980 to 1986, the state of Kentucky assessed achievement under low-stakes conditions. In 1987, a new test was introduced under high-stakes conditions. In 1990, the low-stakes test was reintroduced and comparisons were made between the results of the low-stakes and high-stakes assessments. Students on the high-stakes assessment scored six months higher in reading and mathematics and four months higher in vocabulary than they did on the low stakes assessment, suggesting a lack of generalization of achievement outcomes. To further evaluate the question of generalization, a parallel form, in content and difficulty, of the high-stakes test was developed. When comparing the performances of students on both of these forms, it was found that students scored seven months higher in mathematics and three months higher in reading on the original high-stakes form. The authors concluded that taught-to-the-test results do not generalize from one assessment to another.

In addition to comparison to the NAEP, achievement gains have also not been evidenced on other measures of achievement; SAT, ACT or Advanced Placement tests (Amrein & Berliner, 2002). In comparing the results from 18 high-stakes testing states to that of the average performance of the nation, very little benefit has been noted. After implementation of high-stakes graduation examinations, in the 18 states, 56% experienced a decrease in SAT scores, 67% experienced a decrease in ACT scores and 57% had less students passing AP examinations with passing scores of 3 or higher. Participation rates on the SAT and AP class placements were also lower when compared to the nation. On the ACT, a slight increase was noted in participation rate.
The performances of these 18 high-stakes states were also compared to that of the nation’s performance on the NAEP. For the years 1992-2000, 50% of the states had increases in the Composite Mathematics score for the fourth grade, with 50% of the states having decreases or no change in scores. Composite Mathematics scores for eighth grade showed that only 36% of the states posted increases. For reading in the fourth grade, only 46% of the states had increases in scores as compared to the nation. The only “encouraging” data came from analysis of cohort achievement data as students progressed from the fourth to the eighth grades. While 62% of the states evidenced a decrease in mathematics achievement as compared to the nation in the years 1996 to 2000, 69% of the states posted increases in reading from 1994 to 1998. This artifact may have been more a result of the exclusion of high numbers of students from testing rather than actual gains in achievement (Armein & Berliner, 2002).

"Unintended" Consequences of High-Stakes Assessment

The choice of statistics used to evaluate the achievement of students can often have an effect upon the delivery of services to children (Rafferty & Treff, 1994). Accountability data is often expressed in one of two ways. When reported as the percentage of students passing, instruction and resources are often focused on those students whose scores fall close to the cut-off for passing. However, when data is reported as mean scores, the focus is targeted at those students achieving at the extreme ends. By targeting those students on the low extreme, the overall mean score may be increased. Intensive instruction targeting academic deficits is one way to accomplish this. However, this may also be accomplished by utilizing questionable testing practices, such
as test preparation and teaching to the test, as well as through the elimination of potentially problematic low achieving students' test scores (Rafferty & Treff, 1994).

Incentives, through the awarding of rewards for appropriate performances, often provide teachers and administrators a reason to exclude the scores of low-achieving students. At the high school level, students have been actively encouraged to leave school (International Reading Association, 1999). States with high-stakes assessment report a decrease in graduates, with an increase in drop outs (Jacob, 2001).

Students may be referred for placement to exceptional student education so that their scores are no longer eligible for inclusion in assessment results (International Reading Association. 1999) or tests are taken with accommodations and/or reported separately from the scores of standard curriculum students (McGill-Franzen & Allington, 1993). Increases in placements in special education have been observed as the demands of accountability have increased, most notably in the kindergarten through second grade levels, the years prior to mandatory third grade testing. Other students have been retained in grade in order to give them an additional year of instruction before taking the high-stakes test (Mc-Gill-Franzen & Allington, 1993). Their subsequent performances are then compared to same-grade cohorts instead of same-age cohorts, thereby contaminating accountability results.

Achievement, retention and special education trends were examined in the Chicago Public Schools system prior to and after the implementation of Illinois’ accountability program (Jacob, 2002). Large achievement gains were noted, particularly for previously low-achieving schools, on the Iowa Test of Basic Skills. These gains were greater than what would have been predicted in light of previous achievement patterns.
However, achievement gains on the ITBS did not generalize to other state and district mandated measures of achievement. A sharp increase in the numbers of students placed in special education occurred after the introduction of the accountability measure, with the largest increase being seen in the low-performing schools. Retention rates began to increase in the year prior to the implementation of the accountability policy and peaked the first year of the policy. In comparison to the rates of 1993-1995, the retention rates of 1997 increased by 33% in the first grade, 100% in second grade, and 150-200% in the fourth, fifth, and seventh grades. Somewhat of a decline in retention rates was experienced after this peak year. This was perhaps a result of teachers and parents not wanting to retain low-achieving students for more than one year, thereby, diluting the pool of potential students that could be retained. That is, many of those low-achieving students had already been retained.

Similar special education placement patterns were found in an investigation of six anonymous Florida school systems by Figlio and Getzer (2002). This study spanned the years 1991-92 to 1998-99 and compared special education placement rates prior to and after the introduction of an accountability measure to identify poor performing schools, the Florida Writes. While their study analyzes data prior to the introduction of accountability school grades and the full implementation of the FCAT, data from this study indicates that low-achieving students and those from low SES schools are more likely to be classified, or reclassified, to a test-exempt special education category. When controlling for rate increase trends into special education over time, overall 7.3% of the students were classified in exempted special education categories at the beginning of the study. This number rose to 10.8% by 1998-99. Grade-to-grade trends indicated this
practice was most prevalent transitioning into the fourth grade, the first year of accountability assessment. The lower a child’s previous year’s test scores were (as assessed on other measures such as the SAT-9), the greater the likelihood of being classified (or reclassified) into a test-exempt category. The authors suggested that this practice would not change even when school systems are required to be accountable for the progress of all children, even those classified into subgroups such as special education. Schools will be required to demonstrate either mean minimal performance levels or an increase of 10% annually for defined subgroups. By placing what the authors call “ringers” into special education, schools will be able to demonstrate adequate yearly progress of these groups. Also, testing accommodations will be available to all students classified as special education, thus providing further incentives to place children into special education (Figlio & Getzer, 2002).

High-Stakes Testing Summary

Curriculum standards were established by all of the states during the 1990's (Barksdale-Ladd & Thomas, 2000). By the middle of that decade, the majority of states had measures in place that assessed these standards; 43 states used the assessments for instructional improvement, while 38 of the states used their assessments for program evaluations (Roeber, Bond & Braskamp, 1997). However, a number of critical decisions began to be made from the results of these measures, quickly leading to their becoming high-stakes in nature. Decisions for individual children include whether a child will graduate, be promoted or retained, or if they will be referred to special classes (Airasian, 1993). Inferences regarding current teachers’ skills could also be made through the performances of the children in their classes on high-stakes assessments (Neill, 1999).
Rewards could be given to those whose students excelled and assistance given to those whose students did not. For school systems, entire instructional programs could be judged based upon the performances of the children under their jurisdiction (Dorn, 1998). Intricate systems of rewards and sanctions were established with the expressed purpose of providing a means for motivating changes in teaching and learning practices. Without such rewards and sanctions, it was thought change would not occur (Kohn, 2000b).

Although achievement gains have been noted on high-stakes assessments, actual changes in student learning may not be occurring, as these same improvements are not seen on the National Assessment of Educational Progress (NAEP) (Heubert, 2001). Changes in teaching to the content domains of the test may account for these results. Scores increase as a result of teaching to the test and familiarity with test content and format (Linn, 2000; Mehrens, 1998). Incentives, through the awarding of rewards for appropriate performances, often provide teachers and administrators a reason to exclude the scores of low-achieving students. If the test score is the goal of schooling, rather than the mastery of the state curriculum, then teachers, administrators and students will attempt to meet the expectations of that goal (Ravitch, 1995) through whatever means are available. If actions result in the unintended consequences of retention in grade or referral to special education, individual students may suffer unnecessary harm in the name of high-stakes accountability.

Retention

The use of retention as an educational intervention has increased over the past 25 years. Estimates of retention rates for American school children run from a low of 5 to 10% (Dawson, 1998) to a high of 15% (National Association of School Psychologists
These figures represent over 2.4 million children at an annual cost of over 14 billion dollars to provide for the extra year of schooling (Dawson, 1998). Approximately 30 to 50% of all students are retained prior to the ninth grade at least once (NASP, 2003).

**Characteristics Related to Retention**

Retention practices also appear to be more heavily utilized with students who are male, inner-city, poor and of minority status (Cosden & Zimmer, 1993; Dauber, Alexander, & Entwisle, 1993; Gottfredson, Fink & Graham, 1994; Jimerson, 1997; Jimerson, 1999; Meisels & Liaw, 1993; NASP, 2003). These children come from families with few economic and environmental resources as many of their parents have less than a high school education (Dauber et al., 1993).

Certainly academic failure and lack of school readiness skills are characteristics of many students who are retained. However, some students with equal academic failure and lack of school readiness skills may be promoted to the next grade in spite of these difficulties. Additional student and family characteristics emerge when comparing children who have been retained in kindergarten or retained in a transition classroom, with children who have been regularly promoted, and to those that have been promoted despite being recommended for retention (Ferguson, Jimerson, & Dalton, 2001). In kindergarten, the retained children were found to have parents that valued education less, were from lower SES strata, had mothers with lower levels of education, and had poorer personal-social skills when compared against students who had been promoted.

Academic and intellectual variables do not discriminate the participants when comparing
retained children to equally low-achieving promoted children (Ferguson et al., 2001; Jimerson et al., 1997).

**Academic and Behavioral Outcomes for Students Who Have Been Retained**

The efficacy of retention as an educational intervention has little support in the literature (NASP, 2003). Longitudinal studies have suggested that long-term academic achievement benefits do not exist when comparing the performances of students who have been retained to the performances of matched groups of low-achieving students who have been promoted (Jimerson et al., 1997; Mantzicopoulos & Morrison, 1992; Meisels & Law, 1993; Reynolds, 1992; Westbury, 1994). Also students who have been retained are more likely to dropout from school (Jimerson, Gabrielle, & Whipple, 2002; Roderick, 1994). Fifty percent of all dropouts have been retained at least one grade (Afolayan, 1991) and being retained two or more years, increases the chance of dropping out to over 90% (Dawson, 1998). In adulthood, individuals who were retained in school are less likely to be working full time or be working and attending post-secondary educational opportunities. They earn less per hour when compared to matched groups of low-achieving promoted individuals (Jimerson, 1999).

Positive short-term effects may be found for academic achievement and behavioral functioning when looking at the outcomes in the year following the decision to retain a student (Gottfredson et al., 1994; Jimerson et al., 1997; Mantzicopoulos & Morrison, 1992; Pianta, Tietbohl & Bennett, 1997). For the year following the decision to retain, students are generally repeating the same academic material and taking the same achievement assessments. It is not surprising that their performances would yield higher results the second time around. Positive effects on teachers’ ratings and self-perceptions
of competence have also been noted (Reynolds, 1992), perhaps as a result of teachers and retained students comparing themselves to students that are actually one year younger.

Meta-analysis has been utilized to measure the differences between groups of retained students with matched groups of low-achieving promoted students with respect to academic and behavioral outcomes. In his meta-analysis of the research conducted between 1929 and the mid to late 80’s, Holmes (1989) analyzed the results of 63 studies. Retention was associated with negative academic and behavioral outcomes in 54 of the studies. Benefits were reported in nine studies but these positive outcomes appeared to dissipate over time.

Studies conducted during the 1990’s were analyzed by Jimerson (2001) also utilizing meta-analysis. Of the twenty studies reviewed, 175 analyses of academic achievement outcomes were analyzed. Most analyses (84 outcomes) yielded no statistically significant differences. Eighty-two favored matched low-achieving promoted groups, while only nine favored retained groups of students. Of these nine outcomes, six of the nine involved outcomes in achievement only in the year following the year students had been retained. Overall, the retained students mean effect size was -.39. Retained students were lower in all areas: language arts, reading, math, composite scores and grade point average.

Social-emotional outcomes analyses yielded even more startling results. Of the 148 analyses conducted, 127 yielded no statistically significant differences between the two groups. Eight analyses favored the retained students, while 13 favored matched groups of low-achieving comparison students. The mean effect size for measures of socio-emotional and behavioral adjustment was -.22. Retained students were lower in all
subareas: social, emotional, behavioral, self-concept, and ratings of adjustment. The effect size for attendance was -.65 for retained students.

While authors of 4 of the 20 studies concluded that retention produced positive outcomes for students, these conclusions appeared to have been reached as a result of only considering the data revealed at the conclusion of the first year after students had been retained. Meta-analysis of this first year would support such a conclusion as analyses yielded an effect size of .09 in favor of the retained students. However, the mean effect size for longitudinal data was -.31 for the retained students.

Meta-analysis overwhelmingly demonstrates the negative outcomes associated with retention. However, there are individuals for which retention is determined to be a “successful” or “highly successful” intervention (Ferguson et al., 2001; Meisels & Liaw, 1993). More positive test score outcomes have been demonstrated for children whose mothers are more highly educated (Meisels & Liaw, 1993). Successful retentions have also been demonstrated for those students who were considered more ready for school, as evidenced by no significant delays on readiness measures, were of families from higher SES, had few personal-social functioning deficits and had parents who valued education (Ferguson et al., 1993). One might inquire why students who possess such characteristics would be retained.

RetentionPolicy

When analyzing retention rates, a number of patterns emerge. Retention appears to be most heavily utilized at the entrance to an instructional level; elementary, middle and senior high. The first year(s) of each of these instructional levels show the largest retention rates, with rates decreasing over subsequent years. When patterns vary from this
model, it is thought that outside influences, such as curriculum, instructional strategies and testing policies may be the cause for the variation (Morris, 1991).

Retention Patterns in Miami-Dade County Public Schools

Retention patterns in Miami-Dade County Public Schools were compared to that of twelve American states in the early 1980’s (Morris, 1991). The general pattern of greater numbers of retentions in the first year and lower numbers of retentions in the last year of an instructional level was found. Unexpected peaks in the data were noted at what was termed “promotional gates”, grade levels in which state assessment was given to assess school programs and individual student achievement. State assessment identified students who were in need of remediation. Many of their teachers felt they had not mastered curriculum content and therefore recommended that they be retained. Composite student achievement scores formed the basis for identification of deficient school programs. Even though schools were not sanctioned, their performances were a matter of public record. Administrators felt under pressure to report that their students and programs met state standards. There was a need to identify students with skill deficits skills early. Retention began to be a strategy utilized a year prior to the “promotional gate” assessment years. The data suggested

Introduce a diagnostic test to identify failing students so that they can be helped, and retention rates immediately increase in the test years. Use the same diagnostic test to identify weak school programs for improvement, and the increases in retention rates immediately tend to spread to other grades... Well meaning reforms can have unintended consequences. (Morris, 1991, p. 17).

Retaining low-achieving students can have great effects on composite test score data (McDaniel, 1986; Walker & Levine, 1988). Spending two years in one
grade allows greater time for content mastery. These students will not be tested with their cohorts but rather with a younger group of children as the retained students will be a full year, or more, older. The retained students’ test score performances will be calculated by comparing their performances to a grade-level normative sample that is at least a full year younger. Retaining low-achieving students will produce two testing outcomes. First, their low performances will be kept out of the composite score for their same-age cohorts. Secondly, their scores will be calculated with their same-grade cohorts. The outcomes for each group will be the same, higher average test scores. Without disaggregating test data according to promotion and retention status of the students, school administrators may be overconfident in their data, and obtaining a true picture regarding student achievement and progression may not be possible (Walker & Levine, 1988).

Beliefs about Retention

Teachers and administrators appear to support the practice of retention for those children that do not meet grade level expectations (Byrnes, 1989; Tomchin & Impara, 1992). It is believed to be an acceptable intervention to be used for those children that are seen as immature or absent excessively. It is seen as being less harmful to younger children than to older children (Tomchin & Impara, 1992). Teachers believe that retaining a student prevents the child from experiencing possible future difficulties such as frustration, struggle, future retentions and peer pressure. These beliefs are largely based on anecdotal experiences (Smith, 1989). By giving children an extra year at grade level, teachers believe that they will have the opportunity to master curriculum content and become more mature (Tanner & Combs, 1993). Teachers usually have personal
knowledge concerning the academic achievement of children the year following a retention decision. As children do demonstrate higher levels of achievement during this year, teachers’ personal perceptions appear to be supported, thereby strengthening their convictions. An acknowledgement of the current research base regarding the long-term effects of retention appears to be missing.

It is not surprising then that parents do not have an accurate knowledge base regarding retention (Elam, 1990). In the 22nd annual Gallop poll when parents were asked about their attitudes regarding retention, 54% of the respondents felt that children who had failed achievement tests but were socially promoted would be more likely to eventually drop out of school. Only 32% of respondents indicated children who failed achievement tests and were retained would be more likely to drop out of school.

Retention Policies of Miami-Dade County Public Schools

Retention and promotion policies are contained in the M-DCPS’ Pupil Progression Plan (M-DCPS, 1996; 1997b; 1998b; 1999b; 2000b). This plan is updated each year to comply with revisions to state law and regulations as well as to any changes in district policies. Policy changes are approved by the school board prior to the beginning of the academic year. Annual revisions in place during the years of this study were reviewed.

Many parts of the plan as it relates to retention remained consistent from 1996 through the summer of 2001. Retention could occur twice during the elementary years, once in grades K – 2 and once in grades 3 – 5. For kindergarten students, approval by the building principal was required. Teachers and administrators were reminded of their
responsibility in providing academic interventions in an effort to prevent retentions. They were also obligated to consider the negative consequences of retention.

Various activities associated with the retention process also remained consistent across time. These activities included a review of the student’s educational progress, as well as consideration and initiation of additional activities that would assist in retention decisions and improve the performance of the student. For those students that had been retained, they were to be promoted if after the second week of attendance of the year following retention, they had not received intervention strategies. Additionally, if students did not appear to respond to a change in interventions or programs, they could be considered for a referral to Special Education.

When determining if a student should be retained, teachers and administrators had to determine if the student could demonstrate grade level proficiency. For the 1996-97 school year (M-DCPS, 1996), students in grades 1, 2, 3 and 4 were required to attend summer school if they could not perform at grade level in reading comprehension and mathematics applications as evidenced on district approved standards, student performance and/or teacher recommendation. Determination of possible promotion was to be conducted after the summer session. The policy for grade 5 was the same but only the area of reading comprehension was considered.

The terms remediation and retention were further clarified in the 1997-98 (M-DCPS, 1997b) document in response to state legislation and regulation. Students in grades 1, 2, and 3 were to be given intensive reading intervention if they demonstrated “substantially deficit” skills in reading on locally determined assessments or by teacher recommendation. If by the end of second or third grade a student who had received a year
of intensive remediation could not demonstrate proficiency, the student had to be
retained. Also students scoring at deficit levels on the statewide assessment in reading
and writing given at the fourth grade were to be retained. In all cases, students could be
promoted for good cause.

Although data were not evaluated for the 1998-99 year, retention policies would
certainly affect use of this practice during the period of this study. A student was to be
retained if they were designated as having a reading deficiency and that deficiency was
not remediated by the end of the following year. A student could be promoted for good
cause (M-DCPS, 1998b).

Policies for the 1999-00 (M-DCPS, 1999b) school year required students at
selected grade levels to demonstrate specified performance levels in reading,
mathematics and writing. Student performance was determined on District assessments
and indicators. Students determined to have “substantially deficit” skills in reading were
to receive one year of intensive reading instruction. If by the end of the grades 2, 3, or 4
following the year of intensive intervention, their skills were not remediated, students
were to be retained. A student could be promoted for good cause.

During 2000-01 (M-DCPS, 2000b), school retention policy was focused on fourth
grade. Reading deficiencies were to be targeted at all elementary grade levels. Students,
whose reading skills were determined to be deficient in grades 1, 2, or 3, were to receive
intensive remedial interventions. If their performance on the fourth grade FCAT
continued to yield deficit results, students had to be retained unless good cause was
determined. Specifically, students must have been identified through various assessment
procedures during the first, second and third grades, received actual intensive remediation
for one year and scored at a deficit level on the fourth grade FCAT Reading test before retention could occur.

Retention Summary

The use of retention as an educational intervention continues despite the overwhelming negative academic and behavioral outcomes for the majority of retained students. As school districts are under increasing pressure to raise student achievement levels as evidenced through the results of high stakes testing, retention has been seen as a way to raise test scores. While scores can be raised by an average of 20 points just through the practice of retention (Slavin, 1991), it is the individual student that will pay the ultimate price. According to Smith and Shepard (1989)

Retention has high cost and virtually no value, save the public relations advantages for the schools. Those children who are retained or otherwise failed by public schools are thereby deprived of rightful learning opportunities and, more important, opportunities to succeed in life beyond school (p. 235).

Special Education

The educational needs of children with exceptionalities were inconsistently addressed across the United States during the first half of the twentieth century. As parents sought equal treatment for their children with disabilities through the courts, it became clear the issue would need to be addressed at the federal level. In 1958, the U.S. Congress passed the Expansion of Teaching in the Education of the Mentally Retarded Children Act which supported research and personnel development in the areas of deafness and mental retardation. Further funding for research and training was authorized in 1963; however, inconsistencies across state programs continued to exist. In order to assist states in providing court-mandated programs and uniformity of services, Congress
passed PL 94-142, the Education for All Handicapped Children Act, in 1975. This measure assured a free and appropriate education to all children with disabilities. The six key principles of the measure included zero reject (no students may be excluded), nondiscriminatory evaluation, individualized education program, least restrictive environment, due process and parental participation. Through its reauthorization, this law has been amended to include additional handicapping conditions as well as to include services to infants and toddlers. In 1990, the act became known as the “Individuals with Disabilities Education Act” or IDEA (Yell, 1998).

Through referral, evaluation and placement procedures, children may be identified as meeting the eligibility criteria of one of fourteen categorical disabilities. For the 2000-01 academic school year, 5,775,722 students aged 6 to 21 were served in special education programs (U.S. Department of Education, 2002). The majority of students, approximately 90% (Kirk et al., 2003), were served in “high incidence” or “mild handicapped” programs for children with learning disabilities, speech and language disabilities, mental retardation and emotional handicaps (U.S. Department of Education, 2002). Sixty-two percent of the students were classified as white, 19.8% as black, 14.5% as Hispanic, 1.9% as Asian/Pacific Islander and 1.5% as American Indian/Alaska Native.

The numbers of students served in special education programs continually grows from year to year. In 1991-92, 10.6% of students were classified as disabled. By 2000-01, the percentage of students classified as disabled had risen to 12.3% (Green & Forster, 2002). Slight growth has been noted in all of the disabling categories over the past 10 years; however, significant growth has been documented in the categories of learning disabilities and other health impaired (Kirk et al., 2003). Growth in all categories since
the implementation of PL 94-192 may be due to better identification procedures and the funding incentives provided for the education of these students once they have been identified (Algozzine, Christenson, & Ysseldyke, 1982).

The inclusion of children with attention-deficit hyperactivity disorder accounts for much of the increase in the other health impaired category. However, the increase in the learning disabilities category is not so easily explained. Some of the increase has been due to the change in the definition of mental retardation so that children previously served under that category are now served under learning disabilities. Another explanation is simply that low achieving children have been misclassified as learning disabled due to various factors external to themselves (Green & Forster, 2002).

Referral to Placement Practices

In order to ensure students have access to an appropriate education, it is necessary to accurately identify and document their learning needs. Before a student is found eligible for services under IDEA, it is necessary to evaluate that student to determine if a disability exists, and whether because of that disability, the student demonstrates an educational need for the provision of special education and related services (Yell, 1998).

A student’s difficulties may first be evidenced in the classroom. Interventions may be initiated by the classroom teacher in an attempt to ameliorate or remediate any skills deficits of a student prior to a referral (Yell, 1998). When these efforts are thought to be ineffective, it is usually the classroom teacher that seeks a more formalized remedy by referring a child to the special education identification process (Lloyd, 1991). The two most common reasons for referring a student to this process are academic and behavioral
However, not all students are equally referred to this process. Boys appear to be referred at higher rates than girls, at approximately a ratio of 2 to 1 (Christenson et al., 1983; Flynn & Rahbar, 1994; Lloyd, 1991). Males and females behave differently in the classroom and the more externalizing behavior of boys, hyperactivity and/or disruptiveness, may influence the teacher’s referral decision (Lloyd, 1991; Smart, Sanson, & Prior, 1996; Ysseldyke et al., 1983) as they will not tolerate disruptive or destructive behavior (Campbell, Dobson, & Bost, 1985; Schumm & Vaughn, 1992). Teacher’s perceptions and beliefs that the etiology of learning disabilities is primarily genetically male sex-based may also help to account for this discrepancy (Vogel & Nass, 1993).

Teachers most often attribute the academic difficulties of children to causes other than their own instructional practices deficits (Christenson et al., 1983). While teachers are more than willing to accept credit for a student’s academic success, they rarely will accept the blame. Very often the “blame” is assigned to within-student characteristics or other outside circumstances, such as difficulties at home. Teachers who believe that teaching cannot overcome these external circumstances tend to refer and believe that special education for these children is an appropriate placement (Podell & Soodak, 1993; Soodak & Podell, 1993).

Once having referred a student, the eventual outcomes can be reliably predicted. In their review of the referral-evaluation-placement practice, Ysseldyke et al. (1997) found that 90 to 92% of all students who are referred are evaluated. Of those students
who are evaluated, 70 to 74% are found eligible for special education services. Some variability was noted across districts by community characteristics and by geographic regions. For the years 1990-91 to 1992-93, rates were highest in suburban districts and in the Western region. Rates were the lowest in urban districts and in the Southern region of the country. Results of this study replicated an earlier study conducted in 1982 (Algozzine et al., 1982) demonstrating the consistency of the referral-test-place model.

Once a child has been referred for an evaluation, the assessment that is conducted may seek to confirm the initial reason for referral. The school psychologist will tend to focus the assessment to obtain “referral consistent information” (O’Reilly, Northercraft, & Sabers, 1989) that will justify placement in special education. If a teacher refers a child because of an academic or behavioral difficulty, an assessment is conducted that will seek to find out what is wrong with the student (Ysseldyke & Algozzine, 1982). At times this will mean that assessment is conducted until a disability is found, with as many as twenty different diagnostic tests being administered (Mehan, 1986).

This “confirmatory bias” (O’Reilly et al., 1989) also extends to eligibility and placement decisions. Eligibility decisions tend to confirm initial reasons for referral, sometimes in spite of the actual data presented at eligibility and staffing committee meetings. Even when test data indicates a child functions in the average range, eligibility committees may find the child to be handicapped (Algozzine & Ysseldyke, 1981) basing their decisions more on what they perceive the child to need instead upon the psychometric data generated through the assessment process (Gresham, MacMillan, & Bocian, 1998).
Little correlation exists between the definitions of various handicapping conditions and the data that is presented from an assessment in support of a child receiving services under that category. Various studies across time have demonstrated the consistency of this finding. Shepard, Smith, and Vojir (1983) looked at the records of 800 students who had been determined to be learning disabled. Only 30% of the students actually met the ability/achievement criteria of the learning disabilities category. Thirty-percent appeared to have difficulties learning due to low achievement, second language acquisition or environmental concerns. Data analyzed from the Connecticut Longitudinal Study by Shaywitz, Shaywitz, Fletcher, and Escobar (1990) indicated that only 45% of the students identified as being learning disabled met the research criteria for that categorization. The learning profiles of students with learning disabilities and those with low achievement are similar, making differentiation between the two, at the very least, difficult (Algozzine, Ysseldyke, & McCue, 1995; Lyon et al., 2002). It is not surprising then that when later analyzing data from the Connecticut Longitudinal Study (Shaywitz et al., 1999), that the academic achievement of reading disabled and non-disabled low achievers in reading was virtually indistinguishable. The learning curve for both groups was similar and found to be consistently below that of non-reading impaired students from the age of six to eighteen.

Misidentification was also found by MacMillan, Gresham, Siperstein, and Bocian (1996). Of 43 children with subaverage performance, 35 placement decisions were made. Six students were determined to be mentally retarded, 19 as learning disabled, 4 as speech and language impaired and 6 as ineligible for services. Factors such as teacher beliefs and perceptions, teacher referral, assessment focus and confirmatory bias, and
eligibility team perceptions regarding what is in the best interest of the child all seem to bias assessment and eligibility decisions. Child specific characteristics such as attractiveness, low socioeconomic status and academic difficulty may also influence eventual eligibility outcomes (Ysseldyke & Algozzine, 1982).

**Overrepresentation of Minorities in Special Education**

Bias in the labeling of minority and low-status children as disabled has been argued by Dunn (1968) and Mercer (1973). Concerns regarding the disproportionate representation of minority and low-income students continue (Brantlinger, 1997; Figueroa & Artiles, 1999; Harry & Anderson, 1995). African-American children have been consistently overrepresented in all disabling categories, particularly in the “high incidence” categories of mental retardation, speech-impaired, and emotionally disturbed (Harry & Anderson, 1995) and Hispanic students are overrepresented in programs for learning disabilities and speech and language impairments (Harry, 1992).

The practice of labeling minorities as mentally deficient is not new with the advent of special education. This practice of identification of minorities as deficient was seen at the beginning of the 20th century with the arrival of large populations of immigrants (Figueroa & Artiles, 1999). It has been suggested (Prasse & Reschly, 1986) that this practice was extended to minority students as a way to segregate African-American students after court-mandated integration of schools.

Court cases have questioned the validity of utilizing measures of intelligence to support the placement of minority children in special education. *Diana v. State Board of Education* (1970) challenged the classification of Spanish-speaking students as mentally retarded based on an I.Q. test that had been administered in English. *Larry P. v. Riles*
(1979) argued that African-American children were classified as mentally retarded based upon IQ tests that were biased. Although each of these cases established difficulties in the assessment process, they also established that overrepresentation existed and was a problem. This was, especially disconcerting, as educable mentally retarded class placements were seen as “dead end” placements (Larry P. v Riles, 1979).

Students who are determined to be disabled, particularly as mentally or emotionally handicapped, are often educated outside of the general education classroom (Harry & Anderson, 1995) thereby losing out on the academic and social advantages of that environment. Labeling a child as disabled also carries with it considerable social stigma and lack of opportunity. Teachers maintain lower expectations for learning success creating a “confidence meltdown” in the child, until the “problem” for which the child was initially referred, begins to “appear endemic” to the child (Harry & Anderson, 1995, p. 611). These difficulties along with the perceived ineffectiveness of special educational placement further confound the problem of overrepresentation of minorities.

The Effectiveness of Special Education

In light of the problems associated with identification and eligibility determination of students into special education, particularly those students classified in one of the higher incidence categories, it is crucial to prove the efficacy of special education. However, results have been “ambiguous and contradictory” (Semmel, Gerber, & MacMillan, 1995). Methodological difficulties may be responsible for this perception (Hocutt, 1996, Kavale & Forness, 1999). Sample sizes are usually small, random assignment is not possible as the education each child is dictated by their Individualized Education Plan (IEP), and comparison groups are rarely comparable (Hocutt, 1996). In
order to address the difficulties associated with sample size, meta-analysis has been utilized. The research regarding special education effectiveness has looked at intervention usage, educational placements and outcomes by disability category.

Historically what has made special education “special” has been the utilization of specialized techniques and instruction (Kavale & Forness, 1999). These interventions are generally not utilized as part of the general education curriculum, but rather are administered or supervised by special education teachers. A meta-analysis of some of these interventions (perceptual motor training, diet modification, modality instruction, social skills training for students who were classified emotionally/behaviorally disabled and for those classified as learning disabled, psychotropic medication, psycholinguistic training and stimulant medication) by Kavale and Forness (1999) yielded somewhat discouraging results, ranging from an effect size of .08 for perceptual-motor training to an effect size of .58 for stimulant medication. Many of the standard deviations associated with the effect sizes were large indicating that for many children the intervention has a negative effect. As these children already demonstrate great academic or social deficits, it would seem necessary to provide interventions that could provide more than limited or simply modest gains. Even in light of much of this research, the authors conclude that the classroom practitioner continues to believe in the efficacy of such specialized intervention and continues to use techniques with which they are familiar and comfortable.

A summary of the research regarding other instructional techniques utilized by general education, but ones that have also been utilized by special education, yielded decidedly different results (Kavale & Forness, 1999). Kavale and Forness examined the
efficacy of reading comprehension interventions, mnemonics, direct instruction, peer tutoring, computer assisted instruction, formative evaluation, behavior modification, and early intervention. They found effect sizes ranging from a low of .53 for peer tutoring to a high of 1.62 for mnemonic instruction. The majority of interventions yielded large effect sizes with standard deviations indicating limited variability. Kavale and Forness concluded that when the “education” in special education is emphasized over the “special” of special education, greater outcomes are achieved for the students.

Only modest effects have been demonstrated for outcomes in different educational settings (Kavale & Forness, 1999). When utilizing meta-analysis to compare achievement and social/personality outcomes, a slight negative effect size was noted for special versus general class placement (Carlberg & Kavale, 1980). Achievement variables of mainstreamed, non-mainstreamed and non-handicapped students (Wang & Baker, 1985-1986) analyzed with meta-analysis indicated that mainstreamed students outperformed those that were educated in self-contained classrooms. However, mainstreamed students did not do as well their non-disabled peers.

Outcome differences have also been noted depending upon disability category. Meta-analysis (Carlberg & Kavale, 1980) demonstrated that negative effect size existed for students labeled as educable mentally retarded and slow learner. Small positive effects were noted for students labeled as learning disabled or emotionally/behaviorally disabled. The authors concluded that the average student with a learning disability was better off in a special class. However, a different conclusion was reached by Wang and Baker (1985-1986). In their meta-analysis, they concluded that special class placement
for children with hearing impairments and mental retardation was successful, but this was not so for students with learning disabilities.

Factors inherent to special education may account for successful outcomes for some students. Teachers of special education are highly trained with 59%, as opposed to 49% of general education teachers, having earned a master’s degree or higher (U.S. Department of Education, 2002). They are more likely to use a wider variety of teaching approaches, monitor students’ progress and behavior more often, and have a wider variety of techniques to manage disruptive and inattentive students (Hocutt, 1996). Class sizes are smaller than that found in general education. Reducing class size alone has been found to produce moderate achievement effect sizes (Glass & Smith, 1979).

Outcomes for Students in Special Education

For the 1999-2000 academic years, the graduation rate of students with disabilities with a standard diploma was 56.2% (U.S. Department of Education, 2002). Steady improvement in this rate has been documented since 1995-96. The rate of graduation varied across the disability categories. Rates were lowest for students with mental retardation (39.5%) while the highest rate of graduation (73.4%) was for students with visual impairment. Graduation rates for students with learning disability, hearing impairments, other health impairments, speech and language impairments, traumatic brain injury, and orthopedic impairments were at 60% or more. Students with multiple disabilities, autism, deaf-blindness, and emotional disturbance were below the graduation mean for all students with disabilities.

Conversely, the dropout rate showed a decline over the same period of time, from 34.1% in 1995-6 to 29.4% in 1999-2000 (U.S. Department of Education, 2002). This rate
also varied across the disability categories. Emotional disturbance had the highest dropout rate of 51.4%. Four categories; students with deaf-blindness, autism, visual impairments and hearing impairments; had dropout rates below 15%. Dropout rates have improved most over time for the categories of learning disabilities, speech and language impairments, orthopedic impairments, hearing impairments and emotional disturbance.

Graduation rates for students with exceptionalities were highest for students classified as White (62.5%) or Asian/Pacific Islander (56.3%). Graduation rates were the lowest for students classified as Black (39.7%). Students classified as Black dropped out at the highest rate (37.0%) while students classified as Asian/Pacific Islander dropped out at the lowest rate (19.3%).

Employment outcomes are also poor for students with exceptionalities (Hanley-Maxwell, 1999). Low rates of competitive employment and high rates of underemployment are reported. Enrollment rates in postsecondary training options are lower than for non-disabled students. However, a positive correlation was found between time spent in the general education classroom and post-school outcomes.

Referral to Special Education Policies of Miami-Dade County Public Schools

Students who demonstrate academic and behavioral difficulties are referred to the CST process. The procedures developed for this process are routinely reviewed and revised as new information develops in the field. During the 1990-91 school year, the concept of Collaborative Consultation was implemented for use in all of the schools. For any student demonstrating academic and behavioral difficulties, the CST would convene and develop interventions for use in the classroom. Team members were to determine
who was going to administer the interventions, how the implementation was to be monitored and when the team would meet again to discuss the student.

A review of the CST process was initiated during the 1999-00 school year (J.L. Jackson, personal communication, November 29, 1999). The process was found to be flawed at many levels. Academic and behavioral difficulties were only generically described (e.g. reading problems, math problems, and behavior problems). As a result, the interventions that were developed were non-specific and often unrelated to the difficulties the child was experiencing (e.g. change the child’s seat in the classroom). Intervention monitoring was not occurring but often left up to teacher self-report. When committees would reconvene to review the progress of a student, it was found that academic and behavioral difficulties were not remediated and the student would be referred for an evaluation.

As a result of this review, the School Support Team (SST) was developed (M-DCPS, 2002). Changes to the process included specific identification of academic and behavioral deficits. Through a review of pre-referral diagnostic screening, a child’s problems were to be specifically identified (e.g. consonant blends, regrouping during multi-digit addition, and out-of-seat behavior to gain attention). Interventions were to be developed to target the specific problems. Monitoring no longer would be through self-report but rather by a person not responsible for delivering of the interventions. Upon reconvening of the SST, a child’s progress was to be reviewed. If a child was making progress, the SST process was to be continued and interventions were to be continued or changed as needed. If the child was not making progress, the committee needed to confirm that interventions had been delivered as designed, and decide if the child would
remain in the SST process and/or be referred for a psychoeducational evaluation. The SST process actively encourages classroom teachers to seek assistance from their colleagues when a student in their classroom has academic and/or behavioral difficulties. For those students that have, for whatever reasons, not been granted the opportunity to learn, early identification of difficulties and specific, targeted intensive interventions would increase chances for remediation within the general education classroom and reduce reliance on special education placement. Special education would remain a viable and necessary option for those students that did not respond to appropriately developed and implemented interventions. Although teachers at SST schools continued to refer children for significant academic and behavioral difficulties to the SST process, there was the possibility they might utilize the process for students with less significant difficulties as they understand referral to the process does not necessarily mean placement in special education. At schools where this does occur, initial referrals to the process would increase, and in fact, would be promoted. However, the SST process was not implemented until the 2000-01 school year. Only three of the 179 schools involved in this study were part of the pilot project during the first year.

When evaluating the percentages of students enrolled in special education at the elementary level over the years of the study, a significant pattern develops. The percentages of students enrolled prior to FCAT is considerably less than the percentages of students enrolled after FCAT: 1996-97 – 8.5%; 1997-98 – 8.5%; 1999-00 – 10.6%; 2000-01 - 10.7% (M-DCPS, 1997a; 1998a; 2000a; 2001).
Summary

The research suggests numerous problems with the referral-test-place process. Bias may exist in each step of this process leading to inappropriate identification of students for special education services. This bias may also be responsible for the significant overrepresentation of minority students in particular disability categories. “Special” teaching interventions have demonstrated limited efficacy, while data presented on effectiveness by placement or category has been contradictory.

The learning profiles of students who are learning disabled and those that are low achievers are similar, further confounding this problem. Rarely are the general education instructional programs evaluated to determine if the learning deficiency exists in the instructional program rather than the child. It has been suggested (Heller, Holtzman, & Messick, 1982) that the general education program’s appropriateness and the teacher’s effectiveness be evaluated before any child is referred for special education consideration.

Effective pre-referral interventions have also been suggested as a way to decrease inappropriate referrals. Effective pre-referrals are rarely used (Flugum & Reschly, 1994) and the quality of the ones that are used is poor. However, successful pre-referral interventions programs have been shown to decrease the numbers of students referred, increase student performance outcomes, increase teacher’s skills and improve teachers’ attitudes regarding students. By improving the instruction and management skills of the general education teacher, a reduction in the numbers of students determined to have academic and behavioral disorders may be noted (Donovan & Cross, 2002).
Summary of the Chapter

With the reauthorization of the Elementary and Secondary Education Act, (the No Child Left Behind Act of 2001), increased accountability measures for schools were required. Large scale assessment was seen as a way of ensuring school improvement through the use of public accountability (Stiggins, 1999). While the intent of standardized assessment is to push educators to improve student learning outcomes, a number of critical decisions are made from individual and school results, thus elevating these assessments to the status of "high-stakes" testing. For schools, achievement performance levels help to determine which will receive rewards and recognition and which will receive assistance and/or administrative punitive action. For individual students, their performances on these tests determine academic tracks, promotion or retention, and whether they will graduate from high school or not (Popham, 1993). These consequences may not be in the best interest of individual students and run contrary to the standards set by the National Research Council regarding the use of large scale assessments (Heubert & Hauser, 1999).

The establishment of curriculum standards and the demonstration of skills competencies by students (as assessed through accountability measures) appear to have been positively received by the public (Popham, 1987). However, actual changes in student learning may not be occurring. Even though test score gains have been demonstrated on various state accountability measures, these same gains are not noted on other measures of achievement (Heubert, 2001). Scores on accountability measures may be increasing as a result of teaching to the test, familiarity with test content and format.
(Linn, 2000; Mehrens, 1998) and the exclusion of low-performing students through retention or special education placement.

Increases in placements in special education have been observed as the demands of accountability have increased, most notably in the kindergarten through second grade levels, the years prior to mandatory third grade testing. Other students have been retained in grade in order to give them an additional year of instruction before taking the high-stakes tests (Mc-Gill-Franzen & Allington, 1993). While schools may reap the benefits of “demonstrated” test score increases, individual students who have been retained or placed inappropriately in special education may bear the burden of these practices. The long-term outcomes for children who have been retained are uniformly negative, while the outcomes for children placed in special education are variable.
CHAPTER III

METHOD

The effects of high-stakes testing on retention and referral to special education practices were studied in a large, urban school district, Miami-Dade County Public Schools (M-DCPS). The district is located in southeast Florida and is the fourth largest school district in the country. Permission to conduct this study was obtained from the Office of Evaluation and Research of the M-DCPS. The Florida International University Institutional Review Board determined the study was exempt.

Table 5

<table>
<thead>
<tr>
<th>Academic Year</th>
<th>White Non-Hispanic (%)</th>
<th>Black Non-Hispanic (%)</th>
<th>Hispanic (%)</th>
<th>Other (%)</th>
<th>Total # of students in district</th>
<th>Total minority students</th>
</tr>
</thead>
<tbody>
<tr>
<td>1996-97</td>
<td>13.38</td>
<td>32.85</td>
<td>52.17</td>
<td>1.59</td>
<td>162,793</td>
<td>141,004</td>
</tr>
<tr>
<td>1997-98</td>
<td>12.67</td>
<td>32.48</td>
<td>52.98</td>
<td>1.86</td>
<td>163,969</td>
<td>143,187</td>
</tr>
<tr>
<td>1998-99</td>
<td>12.03</td>
<td>31.89</td>
<td>54.03</td>
<td>2.05</td>
<td>165,571</td>
<td>145,650</td>
</tr>
<tr>
<td>1999-00</td>
<td>11.29</td>
<td>30.90</td>
<td>55.58</td>
<td>2.23</td>
<td>167,008</td>
<td>148,146</td>
</tr>
<tr>
<td>2000-01</td>
<td>10.67</td>
<td>29.93</td>
<td>57.07</td>
<td>2.33</td>
<td>167,721</td>
<td>149,826</td>
</tr>
</tbody>
</table>

The student population of M-DCPS is diverse, the majority of which are identified as members of minority ethnic or racial groups. Changes in racial and ethnic group (see Table 5) memberships are noted for the elementary student population over the identified years of this study (M-DCPS, 1997a; 1998a; 1999a; 2000a; 2001). The numbers of White, non-Hispanic and Black non-Hispanic students have steadily decreased while the numbers of Hispanic and other students have increased over time.
Population totals reflect the number of students registered in kindergarten through grade five at all elementary schools.

Sample Data

The sample data represented the records of children in kindergarten through grade five who had either been referred to a Child Study Team (CST) for intervention and assistance (a discussion of the CST may be found in the Special Education section of the Literature Review) or who had been retained, prior to and after the implementation of the state's high-stakes test, the Florida Comprehensive Assessment Test (FCAT). All elementary schools were considered for inclusion in the sample. However, schools with less than a kindergarten through grade five configuration and those that were not in operation for the entire number of years under study were dropped from the sample. Schools that included a sixth grade for part of the study were included, but data from that grade level was not included in the sample as the majority of schools did not contain this grade level. The final sample included data from 179 elementary schools.

Procedures

Retention and referral data were collected for 179 schools from school district records for the years prior to and after the first state-wide administration of the FCAT. "Prior to" years were defined as 1996-7 and 1997-98, while "after" years were defined as 1999-00 and 2000-01. Data were not evaluated for the first year of state-wide administration, 1998-99, as schools could not accurately predict how their students would perform on the brand-new assessment and the subsequent effect their performances would have on their assigned school performance grades. Therefore, it was hypothesized that administrators and teachers might not be engaged in questionable practices, such as
referral or retention in order to ensure favorable outcomes for their school’s performance grade during the first state-wide administration.

Retention data were obtained from M-DCPS' surveys of retained students. This report is generated in the fall (Fall Student Survey) of each year and reflects the number of students who were not promoted during the previous academic year. Information is reported for each school by grade level by ethnicity, and grade level by gender. Data were obtained for each school in the sample under the years of study.

Referral to the Child Study Team process data were generated from M-DCPS’ Division of Psychological Services district records. Referral data were reported for each school by grade level by ethnicity, and by grade level by gender. Data were obtained for each school in the sample under the years of the study. The data represented the records of students considered to have deficit skills or disabilities and needing further assistance and/or intervention. All referral records were considered, as distinctions are not made between different types of referrals. Children are referred based upon difficulties noted with learning and/or behaving, not based on particular suspected disabilities, such as mental retardation or learning disability. Possible outcomes for students referred to the Child Study Team process include developing interventions and/or psychoeducational evaluation.

Additional data were obtained on school retention and referral records. Absolute referral and retention totals for each ethnicity and gender were converted to percentages for each grade level of each school. Class enrollment figures by ethnicity for each school were obtained from yearly district reports titled Miami-Dade County Public Schools District & School Profiles (M-DCPS, 1997a; 1998a; 1999a; 2000a; 2001). Profiles
provide a “snapshot” of a school’s population at one point in time during the school year. Retention and referral records are collected at different points during the school year. It is possible for retention and referral records to reflect an occurrence for a particular student who would not have been counted within the school’s population profile. For example, for a particular school, their population profile may have indicate that no students in the third grade were classified in the ethnic category of Other. However, after the population profile data were obtained, a child classified in the ethnic category of Other transfers into the school and is subsequently either referred or retained for the year. In order to maintain a conservative approach, absolute retention and referral data were dropped during such occurrences. Eight out of 9,868 retention records and 31 out of 47,069 referral records were not considered in the final ethnicity analysis. All records were considered in the final gender analysis, and therefore, were used when considering overall population trends for limited English proficiency, school socioeconomic status, grade level and school performance grades classifications.

The profiles also provided information regarding each school’s participation in the free and reduced lunch program and percent of limited English proficient population. Data from the 1998-99 school year was considered as this was the first year of state-wide administration of the FCAT.

Percent of participation in the free and reduced lunch program was used to assign a socioeconomic status for each school. In order to prevent extreme unequal assignment to categories, a frequency analysis was conducted. Schools with less than 50% of their students on free or reduced lunch \( (n = 43) \) were considered high socioeconomic status; schools with 50 to 75% \( (n = 43) \) were considered medium socioeconomic status; and
schools with greater than 75% ($n = 93$) were considered low socioeconomic status. Actual percentages ranged from 8.2 to 99.0%. Although realistically, schools with up to 50% of their students qualifying for free or reduced lunch would not be considered to be from a high socioeconomic status, Miami and many of its surrounding areas are considered extremely poor. Therefore, the use of the terms “high”, “medium” and “low” takes on a different quality when used to categorize the populations of schools in M-DCPS.

Likewise, three categories were developed regarding limited English proficiency populations of individual schools. Schools with less than 11.5% limited English proficient students ($n = 60$) were considered to have a low percentage of LEP students, schools with 11.5-23.5% limited English proficient students ($n = 59$) were considered to have a medium percentage of LEP students and schools with greater than 23.5% limited English proficient students ($n = 60$) were considered to have a high percentage of LEP students. Actual percentages for schools ranged from 0 to 58.7%.

School accountability grades were obtained for each school from M-DCPS accountability reports. The grade assigned after the 1998-99 administration of the FCAT was utilized as this was the first year that schools were ranked according to the standards associated with FCAT performances. Four schools received the grade of “A”, 10 received the grade of “B”, 62 received the grade of “C”, 84 received the grade of “D” and 19 received the grade of “F”.

Limitations of the Study

Due to the static nature of the data obtained, limitations are evident. The data associated with retention practices at the group, school and individual student
characteristics levels were obtained after the completion of the school year and represent the final retention outcomes for particular schools and individuals. These data may, however, somewhat underrepresent intended outcomes as students who were recommended for retention may have been withdrawn from the public school system and placed in the private sector in an effort to avoid a retention. These occurrences would not be represented in the data collected. The data associated with referral practices presents with a limitation for the interpretation of school characteristic data. Data classified by student characteristics such as gender, ethnicity, and grade level maintained its integrity across the years of this study as these characteristics do not change. However, due to the high mobility rates of many of the schools in M-DCPS, the static nature of the referral data is more problematic. Referral data were generated at the end of the academic year. The possibility exists that a student may have been referred while in attendance at one school, transferred in the middle of the year to another school, and be reflected in the data of the new school.

Independent Variables

Socio-cultural characteristics served as a means of grouping data. In addition to these characteristics, data were also grouped by grade level and school performance grade. If schools were participating in practices that eliminated the scores of low achieving students, it was anticipated that the use of these practices might be more pronounced at certain grade levels or particular school achievement levels.

Gender. Data grouped for male and female students.
Ethnicity. Ethnic membership is reported by parent(s) when a student initially enters school. The categories are White non-Hispanic, Black non-Hispanic, Hispanic and Asian/Indian Multiracial (other).

Grade level. Data only from elementary schools was considered. The usual configuration of the majority of elementary schools in Miami-Dade County Public Schools is Pre-kindergarten through grade five. Pre-kindergarten students were eliminated from this study as the majority of these students represent children with identified disabilities. Of those schools with a grade six, these records were not considered in order to maintain a consistent comparison.

School socioeconomic status. The socioeconomic status of individual schools was determined by the percentage of students eligible for free or reduced lunch. For repeated measures analyses, schools were divided into three categories: high (<50%), medium (50-75%) and low (>75%) socioeconomic levels.

Limited English proficiency. Through parent report and language screening, students may be classified as having limited English proficiency. For repeated measures analyses, schools were divided into three categories according to the percentage of students classified as limited English proficient: high (>23.5%), medium (11.5-23.5%), or low (<11.5%).

School performance grade. School grades are assigned to individual schools based on FCAT performance and other criteria of the Florida School Recognition Program. Schools may be assigned the grades of "A", "B", "C", "D", or "F". The first performance grades were assigned after the February, 1999 administration and were
based largely on how students performed on the FCAT. Data were grouped by the school performance grade received after the 1999 administration of the FCAT.

Dependent Variables

Retention. Students who are not promoted to the next grade but rather must repeat the grade they have just finished the following year.

Referral to Child Study Team process. Students who demonstrate academic and behavioral difficulties in the general education classroom may be referred to the Child Study Team (CST) process for assistance and intervention. It is the first required step in the Special Education evaluation process. Prior to evaluation, a student’s progress may be discussed at one or more Child Study Team meetings.

Threats to Internal Validity

Data from all qualifying schools was used. Therefore, random assignment was not possible and a quasi-experimental condition existed. In this study, threats to internal validity must be considered as design equivalency cannot be assured (Gribbons & Herman, 1997). A rich description of factors that may influence outcomes is advised in order to minimize threats to internal validity. Any systematic changes in retention and referral to special education policies would certainly have an influence upon potential outcomes of this study.

Referral to Special Education Policies

Referral policies remained consistent from 1996 to the beginning of the 2000-01 academic year. While most schools continued to rely upon the Child Study Team (CST) process for referrals, three schools were part of the School Support Team pilot. CST’s revision, in the 2000-01 school year. SST was developed to provide specific diagnostic
and intervention assistance to students in the general education classroom. It was anticipated that teachers might utilize the process for children considered to have less significant difficulties. While teachers at the three schools continued to refer students with significant difficulties, it was not found that they generally requested the assistance of their colleagues for students experiencing difficulties that were considered less severe (J.L. Jackson, personal communication, November 29, 1999). Due to this, and the fact that only three schools were involved in the SST process during the last year of this study, minimal effects to the study outcomes should be noted.

Retention Policies

It is anticipated that retention policies over the time of this study would affect retention practice outcomes. Policies would suggest that retention rates would increase from the first year (1996-97) to the second year (1997-98), at particularly the second, third and fourth grade levels, as the requirement to retain is specified. Few changes should be noted between the second (1997-98) and third (1999-00) years of the study. A slight decrease at the fourth grade level might be evidenced as the requirement to retain based solely on assessment performance results was not maintained. Finally, during the last year of the study (2000-01), retention policies should affect a dramatic increase at the fourth grade level as retention practices were to be focused upon outcomes at this grade level.

Data Design and Analyses

Data analyses were conducted using repeated measures analysis of variance for the referral and retention dependent variables. This analysis is appropriate for a quasi-experimental condition when there is a concern regarding trends over time (Stevens,
The assumptions: independence in observations, multivariate normality and sphericity; in repeated measures analysis are maintained. For both the retention and the referral to special education dependent variables, independence in observations exists. Multivariate normality is assumed due to the large sample size. Sphericity is not violated as the within-subjects factor did not have more than two levels and the multivariate test, Wilks's $\Lambda$, was utilized (Green & Salkind, 2003).

In all designs, within-Ss variable of time existed across four time periods: "prior to" academic years of 1996-97 and 1997-98 and "after" academic years of 1999-00 and 2000-01. Test designs were created to assess practices on various grouping independent variables. The grouping variables of ethnicity, gender, and percent of student population classified as limited English proficient and on free/reduced lunch were chosen as retention and referral to special education historically has included an overrepresentation of students who are poor, male and/or of minority status. The grouping variables of grade level and school performance grades were chosen to assess their relationship to the use of retention and referral to special education practices.

Repeated measures ANOVA included a (6x4x4) design, the two between-Ss variables were grade level (K-5 inclusive) and ethnicity (White non-Hispanic, Black non-Hispanic, Hispanics and Other) and the within-Ss variable was time (Prior Year 1, Prior Year 2, After Year 3, After Year 4); a (6x2x4) design, the two between-Ss variables were grade level (K-5 inclusive) and gender (M,F) and the within-Ss variable was time (Prior Year 1, Prior Year 2, After Year 3, After Year 4); a (5x2x4) design, the between-Ss variables were assigned school performance grades (A,B,C,D,F) and gender (M,F) and the within-Ss variable was time (Prior Year 1, Prior Year 2, After Year 3, After Year 4); a
(3x2x4) design, the between-Ss variables were socioeconomic status as determined by percentage of students on free and reduced lunch (high, medium, and low) and gender (M,F) and within-Ss variable was time (Prior Year 1, Prior Year 2, After Year 3, After Year 4); and a (3x2x4) design, the between-Ss variables were percentage of students classified as limited English proficient (low, medium and high) and gender (M,F) and within-Ss variable was time (Prior Year 1, Prior Year 2, After Year 3, After Year 4). Means and standard deviations were determined through descriptive statistics. Main and interaction effects were analyzed for all significant findings of repeated measures analyses.

Summary of the Chapter

This study examined retention and referral to special education practices prior to and after the implementation of the state's high-stakes test, the Florida Comprehensive Assessment Test. Between- and within-subjects designs were analyzed using a repeated measures ANOVA to compare retention and referral to special education practices over a five-year period of time, two years prior to (1996-97 and 1997-98) and two years after (1999-00 and 2000-01) the 1998-99 implementation of the FCAT. Means and standard deviations were determined through descriptive statistics. Main and interaction effects were analyzed for all significant results from repeated measures analyses.
CHAPTER IV

RESULTS

This study was designed to examine retention and referral practices prior to and after the implementation of Florida’s high-stakes test, Florida Comprehensive Assessment Test. Repeated measures ANOVA, multivariate test, Wilks’s $\Lambda$, was utilized in conducting this examination. Significant interaction effects were further analyzed by individual repeated measures ANOVA. Means and standard deviations were determined through descriptive statistics. This chapter presents and discusses the findings of these analyses.

Referral to Special Education Practices

*Research Question 1: What were the referral rates to the special education referral process prior to and after the implementation of high-stakes testing in a large urban school system?*

Table 6 presents the aggregate sample data. Referral percentage means and standard deviations were calculated for the two years prior to and the two years after the 1998-99 FCAT administration. While referral percentage means decrease slightly from the first year to the second, a significant change from the second year to the third occurred. Referral percentage means rose from 7.03 to 8.51, an increase of 1.48 percentage points. Referrals are not only maintained at these higher rates for the fourth year but are increased by .36 percentage points.

*Hypothesis 1: The use of referral to the special education referral process will increase after the implementation of high-stakes testing.*
Table 6

<table>
<thead>
<tr>
<th>Academic Year</th>
<th>Referral M</th>
<th>Referral SD</th>
<th>Retention M</th>
<th>Retention SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prior to</td>
<td>7.17</td>
<td>6.05</td>
<td>1.01</td>
<td>2.20</td>
</tr>
<tr>
<td>1996-97</td>
<td>7.03</td>
<td>6.06</td>
<td>1.47</td>
<td>2.76</td>
</tr>
<tr>
<td>After</td>
<td>8.51</td>
<td>6.99</td>
<td>2.11</td>
<td>3.41</td>
</tr>
<tr>
<td>1999-00</td>
<td>8.87</td>
<td>7.84</td>
<td>2.09</td>
<td>3.57</td>
</tr>
<tr>
<td>2000-01</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The corresponding null hypothesis that referral to special education percentage means prior to and after the implementation of high-stakes testing are equal was tested utilizing a repeated measures ANOVA. The results for the ANOVA indicated a significant time effect, Wilks’s $\lambda = .93$, $F(3,2145) = 52.6$, $p < .01$, multivariate $\eta^2 = .07$. The null hypothesis is rejected. Over time, referral percentages increased from a mean of 7.17 in 1996-97 to a mean of 8.87 in 2000-01 with the most notable increase occurring between the 1997-98 and 1999-00 years. From the second to the third year, absolute referrals rose from 10,524 to 12,705 (Table 7).

Retention Practices

Research Question 2: What were the retention rates prior to and after the implementation of high-stakes testing in a large urban school system?

Table 6 presents the aggregate sample data. Retention percentage means and standard deviations were calculated for the two years prior to and the two years after the 1998-99 FCAT administration. Retention percentages increased steadily over the first three years, with the greatest increase occurring between the second and third years, .64
percentage points. Retention percentage means are maintained from the third to fourth year.

Table 7

**Total Number of Referrals for Elementary Schools—Prior to and After High-Stakes Implementation**

<table>
<thead>
<tr>
<th>High-Stakes Implementation</th>
<th>Year</th>
<th>Total Referred</th>
<th>Total Population</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prior to</td>
<td>1996-97</td>
<td>10,826</td>
<td>150,720</td>
</tr>
<tr>
<td></td>
<td>1997-98</td>
<td>10,524</td>
<td>150,401</td>
</tr>
<tr>
<td>After</td>
<td>1999-00</td>
<td>12,705</td>
<td>151,047</td>
</tr>
<tr>
<td></td>
<td>2000-01</td>
<td>13,014</td>
<td>150,188</td>
</tr>
</tbody>
</table>

Table 8

**Total Number of Retentions for Elementary Schools—Prior to and After High-Stakes Implementation**

<table>
<thead>
<tr>
<th>High-Stakes Implementation</th>
<th>Year</th>
<th>Total Retained</th>
<th>Total Population</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prior to</td>
<td>1996-97</td>
<td>1,539</td>
<td>150,720</td>
</tr>
<tr>
<td></td>
<td>1997-98</td>
<td>2,142</td>
<td>150,401</td>
</tr>
<tr>
<td>After</td>
<td>1999-00</td>
<td>3,092</td>
<td>151,047</td>
</tr>
<tr>
<td></td>
<td>2000-01</td>
<td>3,095</td>
<td>150,188</td>
</tr>
</tbody>
</table>

**Hypothesis 2:** The use of retention will increase after the implementation of high-stakes testing.

The corresponding null hypothesis that retention percentage means prior to and after the implementation of high-stakes testing are equal was tested utilizing repeated measures ANOVA. The results for the ANOVA indicated a significant time effect, Wilks's $\Lambda = .88$, $F(3,2145) = 93.8$, $p < .01$, multivariate $\eta^2 = .12$. The null hypothesis is
rejected. Over time, retention percentages increased from a mean of 1.01 in 1996-97 to a mean of 2.09 in 2000-01 with the most notable increase occurring between the 1997-98 and 1999-00 years. From the second to the third year, absolute retentions rose from 2142 to 3092 (Table 8).

Referral and Retention Practices for Identified Groups

Research Question 3: If retention and special education referral rates have changed in response to high-stakes testing, are the changes disproportionate for any identified group of students in a large urban school setting?

Through repeated measures ANOVA analyses, significant effects for changes in retention and referral practices over time, two years prior to and two years after the implementation of the FCAT, were demonstrated using the Wilks’s lambda (Λ) multivariate criteria. Designs were created to assess these practices for particular identified groups and were created according to gender, ethnicity, grade level, and for schools according to school performance grade, percent of population determined to be limited English proficient and percent of population qualifying for the free and/or reduced lunch program. Hypothesis 3 was tested for these designs for both retention and referral practices dependent variables. Time main effects for referral and retention dependent variables were significant, \( p < .01 \), and have been discussed in the presentation of results for research questions 1 and 2. Numerous interactions were found to be significant in support of Hypothesis 3.

Following significant interaction effects, repeated measures ANOVA were conducted individually on the identified groups. In order to control for Type I error across multiple ANOVAs, Bonferroni-like controls were used (Stevens, 1996).
Hypothesis 3: As a result of the implementation of high-stakes testing in a large urban school setting, a disproportionate representation of particular identified groups for retention and referral to special education will be evident.

**Ethnicity**

*Referral to special education dependent variable.* The interactions of Grade x Time, Ethnicity x Time, and Grade x Ethnicity x Time were tested. The Time main effect was significant, $\Lambda = .991$, $F(3,4270) = 12.9$, $p < .01$. The interaction of Grade x Ethnicity x Time was nonsignificant, $\Lambda = .989$, $F(45,12685) = 1.02$, $p = .43$. The interaction of Grade x Time was significant, $\Lambda = .994$, $F(15,11787) = 1.78$, $p < .05$, as well as the interaction of Ethnicity x Time. $\Lambda = .996$, $F(9,10392) = 2.00$, $p < .05$.

Outcomes of individual repeated measures ANOVA for each of the four ethnicity groups are presented in Table 9. Each measure was tested at the .0125 level utilizing a Bonferroni-like control ($\alpha = .05/4 = .0125$). Significant changes occurred for Black, non-Hispanic, and Hispanic students, but not for White, non-Hispanic, or Other students across the years of this study. The significant interaction of Grade x Time is discussed in the Gender and Grade Level section.

Population means are presented in Table 10. For the two groups of significance, Black, non-Hispanic and Hispanic students, the most noticeable percentage mean increases occur from the second to the third year. For Black, non-Hispanic students, means increased by .12 from the first to second year and 1.85 from the second to the third year. Although referral percentage means decrease from the third to the fourth year, the percentage mean of the fourth year is considerably greater than pre-FCAT percentage means. For Hispanic students, means decreased by .44 from the first to second year, and
increased by 1.52 for the second to the third year and .73 for the third to fourth year.

Overall, across the years of this study, the percentage of students, from least likely to most likely, to be referred followed this trend: Other, White, non-Hispanic, Black, non-Hispanic, and Hispanic.

Table 9

Repeated Measures ANOVA for Individual Ethnic Groups-Referral Summary

<table>
<thead>
<tr>
<th>Ethnicity</th>
<th>df</th>
<th>F</th>
<th>$\eta^2$</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>White, non-Hispanic</td>
<td>3</td>
<td>2.03</td>
<td>.006</td>
<td>.108</td>
</tr>
<tr>
<td>Black, non-Hispanic</td>
<td>3</td>
<td>8.25</td>
<td>.023</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Hispanic</td>
<td>3</td>
<td>11.88</td>
<td>.032</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Other</td>
<td>3</td>
<td>.25</td>
<td>.001</td>
<td>.859</td>
</tr>
</tbody>
</table>

Retention dependent variable. The interactions of Grade x Time, Ethnicity x Time, and Grade x Ethnicity x Time were tested. The Time main effect was significant, $\Lambda = .982, F(3,4270) = 26.46, p < .01$. The interaction of Grade x Ethnicity x Time was nonsignificant, $\Lambda = .99, F(45,12685) = 1.00, p = .47$. The interaction of Grade x Time was significant, $\Lambda = .991, F(15,11787) = 2.67, p < .01$, as well as the interaction of Ethnicity x Time, $\Lambda = .995, F(9,10392) = 2.31, p < .05$.

Outcomes of individual repeated measures ANOVA for the four ethnicity groups are presented in Table 11. Each measure was tested at the .0125 level utilizing a Bonferroni-like control ($\alpha = .05/4 = .0125$). Significant changes occurred for all ethnic groups with the exception of the Other ethnic group. The significant interaction of Grade x Time is discussed in the Gender and Grade Level section.
Table 10

*Referral Population Means for Ethnic Groups Prior to and After High-Stakes Implementation*

<table>
<thead>
<tr>
<th>Ethnicity</th>
<th>1996-97</th>
<th>1997-98</th>
<th>1999-00</th>
<th>2000-01</th>
</tr>
</thead>
<tbody>
<tr>
<td>White, non-Hispanic</td>
<td>6.19</td>
<td>5.54</td>
<td>6.05</td>
<td>6.86</td>
</tr>
<tr>
<td>Black, non-Hispanic</td>
<td>6.83</td>
<td>6.95</td>
<td>8.80</td>
<td>8.45</td>
</tr>
<tr>
<td>Hispanic</td>
<td>7.48</td>
<td>7.04</td>
<td>8.56</td>
<td>9.29</td>
</tr>
<tr>
<td>Other</td>
<td>3.50</td>
<td>3.24</td>
<td>3.63</td>
<td>3.69</td>
</tr>
</tbody>
</table>

Table 11

*Repeated Measures ANOVA for Individual Ethnic Groups-Retention Summary*

<table>
<thead>
<tr>
<th>Ethnicity</th>
<th>df</th>
<th>F</th>
<th>η²</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>White, non-Hispanic</td>
<td>3</td>
<td>5.43</td>
<td>.015</td>
<td>.001</td>
</tr>
<tr>
<td>Black, non-Hispanic</td>
<td>3</td>
<td>14.15</td>
<td>.038</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Hispanic</td>
<td>3</td>
<td>23.59</td>
<td>.062</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Other</td>
<td>3</td>
<td>3.62</td>
<td>.010</td>
<td>.013</td>
</tr>
</tbody>
</table>

Population means are presented in Table 12. An increase in means across the four years is noted for the White, non-Hispanic students, with the largest mean increase occurring from the third to fourth year. For the Black, non-Hispanic and Hispanic students, steady increases occur from the first to second to third years. A decrease occurs from the third to fourth years for both of these groups; however, percentage means are considerably higher than pre-FCAT levels. Identical to the finding for the referral dependent variable and the ethnicity group factor, the percentage of students retained in the fourth year, from least likely to most likely, followed the same trend: Other; White, non-Hispanic; Black, non-Hispanic and Hispanic students.
Table 12

Retention Population Means for Ethnic Groups Prior to and After High-Stakes Implementation

<table>
<thead>
<tr>
<th>Ethnicity</th>
<th>1996-97</th>
<th>1997-98</th>
<th>1999-00</th>
<th>2000-01</th>
</tr>
</thead>
<tbody>
<tr>
<td>White, non-Hispanic</td>
<td>.65</td>
<td>.86</td>
<td>1.10</td>
<td>1.50</td>
</tr>
<tr>
<td>Black, non-Hispanic</td>
<td>1.05</td>
<td>1.39</td>
<td>2.20</td>
<td>2.03</td>
</tr>
<tr>
<td>Hispanic</td>
<td>1.02</td>
<td>1.47</td>
<td>2.07</td>
<td>2.04</td>
</tr>
<tr>
<td>Other</td>
<td>.65</td>
<td>.38</td>
<td>.77</td>
<td>1.09</td>
</tr>
</tbody>
</table>

Gender and Grade Level

Referral to special education dependent variable. The interactions of Grade x Time, Gender x Time, and Grade x Gender x Time were tested. The Time main effect was significant, $\Lambda = .931, F(3,2134) = 52.74, p < .01$. The interaction of Grade x Gender x Time was nonsignificant, $\Lambda = .994, F(15,5891) = .863, p = .61$ as well as the interaction of Gender x Time, $\Lambda = .999, F(3,2134) = .88, p = .45$. The interaction of Grade x Time was significant, $\Lambda = .985, F(15,5891) = 2.19, p < .01$.

Following the significant finding for the interaction of Grade x Time, repeated measures ANOVA were conducted individually on each of the six grade level groups. These outcomes are presented in Table 13. Each measure was tested at the .008 level utilizing a Bonferroni-like control ($a = .05/6 = .008$). Significant changes occurred across the years of this study for all grade levels.

Population means are presented in Table 14. With the exception of kindergarten which demonstrates a steady increase in the percentages of students referred, the most significant increase in referral percentage means occurs from the second to the third years for all the grade levels. The most dramatic increase occurs at the third grade level. Whereas, a decrease of .62 was noted from the first to the second years, an increase of
2.57 occurred between the second and third year. Other large increases were noted for the second and fourth grade levels but none quite as dramatic as at the third grade level.

Table 13

Repeted Measures ANOVA for Individual Grade Levels-Referral Summary

<table>
<thead>
<tr>
<th>Grade Level</th>
<th>df</th>
<th>F</th>
<th>η²</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kindergarten</td>
<td>3</td>
<td>7.95</td>
<td>.063</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>First Grade</td>
<td>3</td>
<td>7.15</td>
<td>.057</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Second Grade</td>
<td>3</td>
<td>9.99</td>
<td>.078</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Third Grade</td>
<td>3</td>
<td>17.14</td>
<td>.127</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Fourth Grade</td>
<td>3</td>
<td>9.51</td>
<td>.074</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Fifth Grade</td>
<td>3</td>
<td>12.26</td>
<td>.094</td>
<td>&lt;.001</td>
</tr>
</tbody>
</table>

Table 14

Referral Population Means for Grade Levels Prior to and After High-Stakes Implementation

<table>
<thead>
<tr>
<th>Grade Level</th>
<th>1996-97</th>
<th>1997-98</th>
<th>1999-00</th>
<th>2000-01</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kindergarten</td>
<td>5.33</td>
<td>5.71</td>
<td>6.16</td>
<td>7.11</td>
</tr>
<tr>
<td>First Grade</td>
<td>9.03</td>
<td>9.46</td>
<td>10.48</td>
<td>11.12</td>
</tr>
<tr>
<td>Second Grade</td>
<td>9.38</td>
<td>9.19</td>
<td>10.96</td>
<td>11.23</td>
</tr>
<tr>
<td>Third Grade</td>
<td>8.63</td>
<td>8.01</td>
<td>10.58</td>
<td>10.84</td>
</tr>
<tr>
<td>Fourth Grade</td>
<td>6.71</td>
<td>6.74</td>
<td>8.44</td>
<td>8.17</td>
</tr>
<tr>
<td>Fifth Grade</td>
<td>3.96</td>
<td>3.05</td>
<td>4.44</td>
<td>4.77</td>
</tr>
</tbody>
</table>

Retention dependent variable. The interactions of Grade x Time, Gender x Time, and Grade x Gender x Time were tested. The Time main effect was significant, Λ = .879, $F(3,2134) = 97.61, p < .01$. The interaction of Grade x Gender x Time was nonsignificant, Λ = .995, $F(15.5891) = .776, p = .71$. The interaction of Grade x Time was significant, Λ = .957, $F(15.5891) = 6.26, p < .01$ as well as the interaction of Gender x Time, Λ = .996, $F(3.2134) = 2.65, p < .05$. 

78
Following the significant finding for the interaction of Grade x Time, repeated measures ANOVA were conducted individually on each of the six grade level groups. These outcomes are presented in Table 15. Each measure was tested at the .008 level utilizing a Bonferroni-like control ($\alpha = .05/6 = .008$). Significant changes occurred for all grade levels except grade five.

Table 15

Repeated Measures ANOVA for Individual Grade Levels-Retention Summary

<table>
<thead>
<tr>
<th>Grade Level</th>
<th>df</th>
<th>F</th>
<th>$\eta^2$</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kindergarten</td>
<td>3</td>
<td>19.07</td>
<td>.139</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>First Grade</td>
<td>3</td>
<td>27.84</td>
<td>.190</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Second Grade</td>
<td>3</td>
<td>19.54</td>
<td>.142</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Third Grade</td>
<td>3</td>
<td>18.27</td>
<td>.134</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Fourth Grade</td>
<td>3</td>
<td>15.20</td>
<td>.114</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Fifth Grade</td>
<td>3</td>
<td>2.06</td>
<td>.017</td>
<td>.105</td>
</tr>
</tbody>
</table>

Population means are presented in Table 16. For grades kindergarten through four, retention percentage means increase over the first three years, with the largest increases occurring between the second and third years. While the first grade posts the largest changes in retention percentages, the third grade posts the next largest change, surpassing the increase noted for the second grade. Between the third and fourth years, only the kindergarten increase is of note. However, for the other grades, the higher third year retention percentage means are basically maintained from the third to fourth year.

Following the significant finding for the interaction of Gender x Time, repeated measures ANOVA were conducted individually on each of the gender groups. These outcomes are presented in Table 17. Each measure was tested at the .025 level utilizing a
Bonferroni-like control ($\alpha = .05/2 = .025$). Significant changes occurred for both males and females.

**Table 16**

*Retention Population Means for Grade Levels Prior to and After High-Stakes Implementation*

<table>
<thead>
<tr>
<th>Grade Level</th>
<th>1996-97</th>
<th>1997-98</th>
<th>1999-00</th>
<th>2000-01</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kindergarten</td>
<td>.86</td>
<td>1.22</td>
<td>1.74</td>
<td>2.19</td>
</tr>
<tr>
<td>First Grade</td>
<td>1.63</td>
<td>2.43</td>
<td>3.60</td>
<td>3.66</td>
</tr>
<tr>
<td>Second Grade</td>
<td>1.31</td>
<td>1.94</td>
<td>2.71</td>
<td>2.58</td>
</tr>
<tr>
<td>Third Grade</td>
<td>1.24</td>
<td>1.74</td>
<td>2.68</td>
<td>2.36</td>
</tr>
<tr>
<td>Fourth Grade</td>
<td>.61</td>
<td>1.11</td>
<td>1.39</td>
<td>1.20</td>
</tr>
<tr>
<td>Fifth Grade</td>
<td>.39</td>
<td>.38</td>
<td>.53</td>
<td>.54</td>
</tr>
</tbody>
</table>

**Table 17**

*Repeated Measures ANOVA for Individual Genders-Retention Summary*

<table>
<thead>
<tr>
<th>Gender</th>
<th>df</th>
<th>$F$</th>
<th>$\eta^2$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Males</td>
<td>3</td>
<td>44.33</td>
<td>.110</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Females</td>
<td>3</td>
<td>60.33</td>
<td>.145</td>
<td>&lt;.001</td>
</tr>
</tbody>
</table>

Population means are presented in Table 18. While population means of males and females tend to parallel one another across the four years, the percentage increase for females from the first to second year is double that of males. Percentage increases from the first to second, and the second to the third, are similar for both genders.

**Gender and School Performance Grade**

*Referral to special education dependent variable.* The interactions of School Performance Grade $x$ Time, Gender $x$ Time, and School Performance Grade $x$ Gender $x$ Time were tested. The Time main effect was significant. $\Lambda = .964$, $F(3,2136) = 26.49$, $p$
<.01. The interaction of School Performance Grade x Gender x Time was nonsignificant, 
\( \Lambda = .998, F(12,5651) = .392, p = .967 \) as well as the interaction of Gender x Time, \( \Lambda = 1.0, F(3,2136) = .20, p = .896 \). The interaction of School Performance Grade x Time was significant, \( \Lambda = .944, F(12,5651) = 10.32, p < .01 \).

Table 18

Retention Population Means for Gender Prior to and After High-Stakes Implementation

<table>
<thead>
<tr>
<th>Gender</th>
<th>1996-97</th>
<th>1997-98</th>
<th>1999-00</th>
<th>2000-01</th>
</tr>
</thead>
<tbody>
<tr>
<td>Males</td>
<td>1.33</td>
<td>1.80</td>
<td>2.61</td>
<td>2.55</td>
</tr>
<tr>
<td>Females</td>
<td>.69</td>
<td>1.14</td>
<td>1.61</td>
<td>1.63</td>
</tr>
</tbody>
</table>

Following the significant finding for the interaction of School Performance Grade x Time, repeated measures ANOVA were conducted individually on each of the five school performance grade categories. The results are presented in Table 19. Each measure was tested at the .01 level utilizing a Bonferroni-like control (\( \alpha = .05/5 = .01 \)). Significant changes occurred across the years of this study for schools graded A, C, D and F.

Table 19

Repeated Measures ANOVA for Individual School Performance Grade Categories-Referral Summary

<table>
<thead>
<tr>
<th>School Performance Grade</th>
<th>df</th>
<th>F</th>
<th>( \eta^2 )</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>3</td>
<td>5.12</td>
<td>.254</td>
<td>.004</td>
</tr>
<tr>
<td>B</td>
<td>3</td>
<td>1.09</td>
<td>.027</td>
<td>.356</td>
</tr>
<tr>
<td>C</td>
<td>3</td>
<td>6.60</td>
<td>.026</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>D</td>
<td>3</td>
<td>24.02</td>
<td>.067</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>F</td>
<td>3</td>
<td>37.42</td>
<td>.333</td>
<td>&lt;.001</td>
</tr>
</tbody>
</table>
Population means are presented in Table 20. The largest increases for all of the groups, including school performance grade B, occurred between the second and third years even though percentage means from the first to second year suggest a decrease in this practice for at least schools in three of the five categories. What is most striking is the 4.32 increase in the second to the third year for the school performance grade of F. Even with this substantial increase, a 1.21 increase is posted for the third to the fourth year for this category.

Table 20

**Referral Population Means for School Performance Grade Categories Prior to and After High-Stakes Implementation**

<table>
<thead>
<tr>
<th>School Performance Grade</th>
<th>1996-97</th>
<th>1997-98</th>
<th>1999-00</th>
<th>2000-01</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>4.31</td>
<td>4.73</td>
<td>5.69</td>
<td>6.55</td>
</tr>
<tr>
<td>B</td>
<td>6.14</td>
<td>5.68</td>
<td>6.76</td>
<td>6.10</td>
</tr>
<tr>
<td>C</td>
<td>7.53</td>
<td>7.49</td>
<td>8.25</td>
<td>8.41</td>
</tr>
<tr>
<td>D</td>
<td>7.40</td>
<td>6.96</td>
<td>8.40</td>
<td>8.82</td>
</tr>
<tr>
<td>F</td>
<td>6.16</td>
<td>7.04</td>
<td>11.36</td>
<td>12.57</td>
</tr>
</tbody>
</table>

Retention dependent variable. The interactions of School Performance Grade x Time, Gender x Time, and School Performance Grade x Gender x Time were tested. The Time main effect was significant, $\Lambda = .962, F(3,2136) = 27.76, p < .01$. The interaction of School Performance Grade x Gender x Time was nonsignificant, $\Lambda = .996, F(12,5651) = .788, p = .663$ as well as the interaction of Gender x Time. $\Lambda = .998, F(3,2136) = 1.23, p = .296$. The interaction of School Performance Grade x Time was significant, $\Lambda = .957, F(12,5651) = 7.80, p < .01$. 

82
Following the significant finding for the interaction of School Performance Grade x Time, repeated measures ANOVA were conducted individually on each of the five school performance grade categories. The results are presented in Table 21. Each measure was tested at the .01 level utilizing a Bonferroni-like control (\( \alpha = \frac{.05}{5} = .01 \)). Significant changes occurred for schools with grades of C, D and F.

Table 21

Repeated Measures ANOVA for Individual School Performance Grade Categories—Retention Summary

<table>
<thead>
<tr>
<th>School Performance Grade</th>
<th>df</th>
<th>F</th>
<th>( \eta^2 )</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>3</td>
<td>2.31</td>
<td>.134</td>
<td>.089</td>
</tr>
<tr>
<td>B</td>
<td>3</td>
<td>2.00</td>
<td>.049</td>
<td>.118</td>
</tr>
<tr>
<td>C</td>
<td>3</td>
<td>26.01</td>
<td>.095</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>D</td>
<td>3</td>
<td>43.99</td>
<td>.116</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>F</td>
<td>3</td>
<td>29.89</td>
<td>.285</td>
<td>&lt;.001</td>
</tr>
</tbody>
</table>

Population means are presented in Table 22. While the largest increases are again generally noted between the second and third years, there is great variation in patterns among the individual categories. Schools with the performance category of A and B decreased their use of retention from the first to the second year. Their largest increases were posted between the second and third year. While performance grade A schools decrease their use of retention from the third to the fourth year and performance grade B schools slightly increase their use, percentage means for both A and B schools during the fourth year are substantially higher than the years prior to school accountability grading. Steady increases are noted for performance grade C and D schools across time with the exception of the third to the fourth year for the C schools. Performance grade F schools...
post the largest increase from the second to the third year. Overall, their use of retention from the first to the fourth year demonstrates the most substantial rise, with an overall increase of 2.54 percentage points. While several of the other categories almost double their use of retention, schools with the performance grade of F increased their use of retention by more than fivefold over the years of the study. Also in the third and fourth year, there is a direct relationship between the performance category of the school and their use of retention. That is, the lower the grade of the school, the more likely the school will use retention as an intervention strategy. During the fourth year, performance grade F schools were more than 5 times as likely to retain their students than were performance grade A schools.

Table 22

*Retention Population Means for School Performance Grade Categories Prior to and After High-Stakes Implementation*

<table>
<thead>
<tr>
<th>School Performance Grade</th>
<th>1996-97</th>
<th>1997-98</th>
<th>1999-00</th>
<th>2000-01</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>.31</td>
<td>.20</td>
<td>.70</td>
<td>.59</td>
</tr>
<tr>
<td>B</td>
<td>.82</td>
<td>.69</td>
<td>1.01</td>
<td>1.17</td>
</tr>
<tr>
<td>C</td>
<td>1.07</td>
<td>1.80</td>
<td>2.10</td>
<td>1.98</td>
</tr>
<tr>
<td>D</td>
<td>1.11</td>
<td>1.48</td>
<td>2.11</td>
<td>2.12</td>
</tr>
<tr>
<td>F</td>
<td>.56</td>
<td>1.00</td>
<td>3.00</td>
<td>3.10</td>
</tr>
</tbody>
</table>

*Gender and School Socioeconomic Status*

*Referral to special education dependent variable.* The interactions of School Socioeconomic Status x Time, Gender x Time, and School Socioeconomic Status x Gender x Time were tested. The Time main effect was significant, \( \Lambda = .954, F(3,2140) = 34.67, p < .01 \). The interaction of School Socioeconomic Status x Gender x Time was
nonsignificant, $\Lambda = .999, F(6,4280) = .208, p = .975$ as well as the interaction of Gender x Time, $\Lambda = .999, F(3,2140) = .751, p = .522$. The interaction of School Socioeconomic Status x Time was significant, $\Lambda = .985, F(6,4280) = 5.29, p < .01$.

The outcomes for the individual repeated measures ANOVA for each of the three socioeconomic status categories is presented in Table 23. Each measure was tested at the .016 level utilizing a Bonferroni-like control ($\alpha = .05/3 = .016$). Significant changes occurred across the years of this study for schools classified as medium and low socioeconomic status.

Table 23

Repeated Measures ANOVA for Individual School Socioeconomic Status Categories-Referral Summary

<table>
<thead>
<tr>
<th>Socioeconomic Status Group</th>
<th>df</th>
<th>$F$</th>
<th>$\eta^2$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>3</td>
<td>3.11</td>
<td>.018</td>
<td>.026</td>
</tr>
<tr>
<td>Medium</td>
<td>3</td>
<td>9.97</td>
<td>.055</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Low</td>
<td>3</td>
<td>42.43</td>
<td>.103</td>
<td>&lt;.001</td>
</tr>
</tbody>
</table>

Population means are presented in Table 24. The largest increase for all the categories again occurs between the second and third years, .52 for high, 1.58 for medium and 1.89 for low socioeconomic status schools. This occurred even with the medium and low socioeconomic status categories demonstrating a decrease in referrals from the first to second year. The high and medium socioeconomic status categories maintained their use of referrals from the third to the fourth year, while the low socioeconomic status category continued to demonstrate an increase in their reliance on this practice for the same time period.
Table 24

*Referral Population Means for School Socioeconomic Status Categories Prior to and After High-Stakes Implementation*

<table>
<thead>
<tr>
<th>Socioeconomic Status Group</th>
<th>1996-97</th>
<th>1997-98</th>
<th>1999-00</th>
<th>2000-01</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>6.12</td>
<td>6.21</td>
<td>6.73</td>
<td>6.72</td>
</tr>
<tr>
<td>Medium</td>
<td>7.95</td>
<td>7.64</td>
<td>9.22</td>
<td>9.14</td>
</tr>
<tr>
<td>Low</td>
<td>7.31</td>
<td>7.12</td>
<td>9.01</td>
<td>9.74</td>
</tr>
</tbody>
</table>

*Retention dependent variable.* The interactions of School Socioeconomic Status x Time, Gender x Time, and School Socioeconomic Status x Gender x Time were tested.

The Time main effect was significant, $\Lambda = .910, F(3,2140) = 70.81, p < .01$. The interaction of School Socioeconomic Status x Gender x Time was nonsignificant, $\Lambda = .997, F(6,4280) = 1.04, p = .396$ as well as the interaction of Gender x Time, $\Lambda = .997, F(3,2140) = 1.98, p = .115$. The interaction of School Socioeconomic Status x Time was significant, $\Lambda = .986, F(6,4280) = 5.06, p < .01$.

The outcomes for individual repeated measures ANOVA for each of the three socioeconomic status categories are presented in Table 25. Each measure was tested at the .016 level utilizing a Bonferroni-like control ($\alpha = .05/3 = .016$). Significant changes occurred for each of the socioeconomic status categories.

Population means are presented in Table 26. The largest increases for the high and medium socioeconomic status categories occur from the first to the second year. However, the largest increase for the low socioeconomic schools occurs between the second and third year. While all schools increase their use of retention from the first year to the fourth, the largest increase occurred for the low socioeconomic status category.
Table 25

Repeated Measures ANOVA for Individual School Socioeconomic Status Categories—Retention Summary

<table>
<thead>
<tr>
<th>Socioeconomic Status Group</th>
<th>df</th>
<th>F</th>
<th>( \eta^2 )</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>3</td>
<td>13.90</td>
<td>.075</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Medium</td>
<td>3</td>
<td>23.48</td>
<td>.121</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Low</td>
<td>3</td>
<td>58.94</td>
<td>.137</td>
<td>&lt;.001</td>
</tr>
</tbody>
</table>

Gender and Percent of Limited English Proficiency Population

Referral to special education dependent variable. The interactions of Limited English Proficiency x Time, Gender x Time, and Limited English Proficiency x Gender x Time were tested. The Time main effect was significant, \( \Lambda = .931, F(3,2140) = 52.61, p < .01 \). The interaction of Limited English Proficiency x Gender x Time was nonsignificant, \( \Lambda = .999, F(6,4280) = .331, p = .921 \) as well as the interaction of Gender x Time, \( \Lambda = .999, F(3,2140) = .879, p = .451 \). The interaction of Limited English Proficiency x Time was significant, \( \Lambda = .981, F(6,4280) = 6.85, p < .01 \).

Table 26

Retention Population Means for School Socioeconomic Status Categories Prior to and After High-Stakes Implementation

<table>
<thead>
<tr>
<th>Socioeconomic Status Group</th>
<th>1996-97</th>
<th>1997-98</th>
<th>1999-00</th>
<th>2000-01</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>.85</td>
<td>1.18</td>
<td>1.43</td>
<td>1.61</td>
</tr>
<tr>
<td>Medium</td>
<td>1.09</td>
<td>1.72</td>
<td>2.29</td>
<td>1.92</td>
</tr>
<tr>
<td>Low</td>
<td>1.04</td>
<td>1.49</td>
<td>2.34</td>
<td>2.39</td>
</tr>
</tbody>
</table>

Following the significant interaction of Limited English Proficiency x Time, repeated measures ANOVA were conducted on each of the three Limited English...
Proficiency categories. These outcomes are presented in Table 27. Each measure was tested at the .016 level utilizing a Bonferroni-like control ($\alpha = .05/3 = .016$). Significant changes occurred across the years of this study for all of the Limited English Proficiency categories.

Population means are presented in Table 28. The largest increase for all the categories again occurs between the second and third years. 1.8 for low, 1.02 for medium and 1.61 for high Limited English Proficiency schools. This occurred even with the medium and high Limited English Proficiency categories demonstrating a decrease in referrals from the first to second year, suggesting that the increase noted from the second to the third years is well beyond that which would have been predicted from previous behavior. For the third to the fourth year, rates continued to be maintained or increased.

Table 27

Repeated Measures ANOVA for Individual Limited English Proficiency (LEP) Categories-Referral Summary

<table>
<thead>
<tr>
<th>LEP Group</th>
<th>df</th>
<th>$F$</th>
<th>$\eta^2$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>3</td>
<td>30.60</td>
<td>.113</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Medium</td>
<td>3</td>
<td>10.61</td>
<td>.043</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>High</td>
<td>3</td>
<td>27.11</td>
<td>.102</td>
<td>&lt;.001</td>
</tr>
</tbody>
</table>

Table 28

Referral Population Means of Limited English Proficiency (LEP) Categories Prior to and After High-Stakes Implementation

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>6.15</td>
<td>6.82</td>
<td>8.62</td>
<td>8.55</td>
</tr>
<tr>
<td>Medium</td>
<td>7.38</td>
<td>6.77</td>
<td>7.79</td>
<td>8.12</td>
</tr>
<tr>
<td>High</td>
<td>8.00</td>
<td>7.50</td>
<td>9.11</td>
<td>9.94</td>
</tr>
</tbody>
</table>
Retention dependent variable. The interactions of Limited English Proficiency x Time, Gender x Time, and Limited English Proficiency x Gender x Time were tested. The Time main effect was significant, $\Lambda = .884, F(3,2140) = 94.01, p < .01$. The interaction of Limited English Proficiency x Gender x Time was nonsignificant, $\Lambda = .998, F(6,4280) = .676, p = .669$ as well as the interaction of Gender x Time, $\Lambda = .996, F(3,2140) = 2.60, p = .051$. The interaction of Limited English Proficiency x Time was significant, $\Lambda = .993, F(6,4280) = 2.45, p < .05$.

The outcomes for individual repeated measures ANOVA for each of the three Limited English Proficiency categories are presented in Table 29. Each measure was tested at the .016 level utilizing a Bonferroni-like control ($\alpha = .05/3 = .016$). Significant changes occurred for each of the categories.

Table 29

<table>
<thead>
<tr>
<th>LEP Group</th>
<th>df</th>
<th>$F$</th>
<th>$\eta^2$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>3</td>
<td>35.02</td>
<td>.128</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Medium</td>
<td>3</td>
<td>27.35</td>
<td>.104</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>High</td>
<td>3</td>
<td>36.21</td>
<td>.132</td>
<td>&lt;.001</td>
</tr>
</tbody>
</table>

Population means are presented in Table 30. A steady increase in the use of retention is noted for all three classifications. While all the percentages for all three classifications are nearly identical for the 1996-97 year, the increase from the first to the third year for the high and low Limited English Proficiency classifications surpasses that of the medium Limited English Proficiency classification for the same period of time. For
all classifications, the reliance upon the use of retention is maintained from the third to
the fourth year.

Table 30

*Retention Population Means for Limited English Proficiency (LEP) Categories Prior to
and After High-Stakes Implementation*

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>1.02</td>
<td>1.59</td>
<td>2.25</td>
<td>2.19</td>
</tr>
<tr>
<td>Medium</td>
<td>1.02</td>
<td>1.18</td>
<td>1.88</td>
<td>1.89</td>
</tr>
<tr>
<td>High</td>
<td>0.98</td>
<td>1.63</td>
<td>2.19</td>
<td>2.17</td>
</tr>
</tbody>
</table>

Summary of the Chapter

This study was designed to examine retention and referral practices prior to and
after the implementation of Florida’s high-stakes test, Florida Comprehensive
Assessment Test. Repeated measures ANOVA and descriptive statistics were utilized in
conducting this investigation.

Significant main effects for referral and retention over time were demonstrated.
The use of retention steadily increases over the first three years, with the greatest increase
occurring between the second and third years. While not increasing in usage for the
fourth year, its usage continues to be maintained. While the use of referral actually
decreases from the first to second years, there is a significant change in practice from the
second to third year. The use of referral continues to increase in the fourth year.

Examination of the use of these practices according to student gender, ethnicity
and grade level and school socioeconomic status, percent of limited English proficient
population and school accountability performance grade revealed significant differences
according to student and school characteristics. Increases in the use of referral across time
was significant for Black, non-Hispanic and Hispanic students, all limited English proficiency population categories, and medium and low socioeconomic status category schools. While increases were noted for all grade levels, the most dramatic increase occurred for the third grade level. Significant increases occurred in schools with performance accountability grades of A, C, D and F with the most striking absolute increase occurring for F schools. Increases in the use of retention across time were significant for all ethnic groups and all limited English proficiency categories. While all socioeconomic categories showed significant changes, the largest increase was noted for low socioeconomic status schools. Significant changes occurred in grades kindergarten through four. The largest increase was noted for the first grade with the next largest increase occurring in the third grade. Differences were also found according to gender. While trends for males and females tended to parallel one another across the years of this study, increases for females exceeded that of males from the first to second year. Significant increases occurred for schools with accountability performance grades of C, D and F; however the most dramatic increase occurred for the F schools. A direct relationship between performance category grade of school and their use of retention exists. The lower the grade, the more likely the school will use retention as an intervention strategy. F schools were five times more likely to use retention than were A schools.
CHAPTER V
DISCUSSION

With the passing of the No Child Left Behind Act of 2001, increased accountability for student learning outcomes was required. Mastery of curriculum content was to be demonstrated in reading and mathematics by students in grades three through eight, as well as in one year in high school. States were to adopt achievement standards and determine mastery of these standards through accountability measures.

A number of decisions are made from the results of these standardized measures, thus elevating these assessments to the status of “high-stakes” testing. For schools, rewards and punishments are awarded based upon the achievement levels of their students. For individual students, performance on these tests can determine if they are promoted or retained, if they will meet eligibility criteria for graduation or the academic track in which they will be placed (Popham, 1993).

By connecting high-stakes assessment to rewards and sanctions, legislatures have been able to assure that public policy is able to drive the implementation of educational objectives and outcomes (Madaus & Kellaghan, 1993). If the consequences connected with poor achievement outcomes are substantial, educators may engage in any number of practices that might help ensure a positive outcome. Actual achievement and learning outcomes for students may be changed, they may teach to the test (Popham, 2001), or cut-off scores for acceptable performances may be set low (Heubert, 2001). Educators may also help inflate their student’s scores by cheating (Yang, 2000; Zlatos, 1996). Another option is for educators to eliminate the scores of low-achieving students from the testing pool through the practices of retention and referral to special education (Allington...
McGill-Franzen, 1992; Heubert, 2001). These practices may potentially prove harmful or provide questionable benefit to individual students (Kavale & Forness, 1999; Rafoth & Carey, 1995).

This study was designed to examine retention and referral practices prior to and after the implementation of Florida’s high-stakes test, the Florida Comprehensive Assessment Test (FCAT). Designs for repeated measures analysis of variance were created to determine if these practices changed over time, and if so, to determine if these practices were utilized differently for various groups.

Referral and Retention Practices: Review of the Research Questions

Three research questions were proposed regarding referral and retention practices prior to and after the implementation of the FCAT. The results suggest significant changes in the use of referral and retention after the introduction of this test, and that practices varied depending upon individual and group characteristics.

The first research question addressed referral rates prior to and after the implementation of FCAT. A significant main effect (p < .01) was demonstrated for referral rates over time. Population percentage means increased from 7.17 in 1996-97 to 8.87 in 2000-01, with the most notable increase occurring between the second, 1997-98 (7.03), and third, 1999-00, years (8.51). Absolute referrals rose from 10,524 to 12,705 during this time, an increase of 2,181 referrals. This rise in referrals is mirrored in the special education placement data for the elementary schools of M-DCPS; 1996-97 – 8.8%; 1997-98 – 8.5%; 1999-00 – 10.6%; 2000-01 – 10.7%, thus suggesting support for the strong relationship between referral and placement practices (Ysseldyke et al., 1997).
Although nationally, special education placement rates have been steadily rising, this explanation does not adequately account for the movement in referral rates in M-DCPS. In fact, if future behavior were to be predicted from prior behavior, the drop from the first to the second years, would suggest that referrals would have been relatively unchanged over the course of this study.

Policies and practices of the school system as they relate to referral to the special education evaluation process remained somewhat consistent over the years of this study. The only exception occurred in the fourth year of the study when the Child Study Team process began its revision to the School Support Team. However, this only involved three schools for this initial pilot year and would only have influenced behavior of these three schools during that year alone.

Retention policies may have had an effect upon the rise of referrals over the course of the study. For any child who had been retained, a referral to the process was an option. It would be expected that as retentions rose, so would referrals. This was not the case. While retentions steadily rise over the first three years of this study, the number of referrals did not demonstrate this steady increase from the first to second year.

Referrals to the special education evaluation process appear to be additionally influenced by other factors. The possibility exists that teachers became concerned for the achievement needs of individual students after they were provided with feedback of the performance of these students on the FCAT of 1998-99. The low scores of these students “sent up red flags” indicating that the needs of these students could not be fully addressed within general education; therefore the referral process was initiated. This may be a valid explanation for some referrals. However, if there is a concern for a particular student’s
achievement, that concern would most likely have existed before the student took the FCAT. Their academic difficulties would be evident in the day-to-day instructional activities and they would be referred because of their difficulties in the general education classroom. It would be expected that a consistent pattern of referrals or a steady increase would be noted if this explanation accounted for the rise in referrals.

Another possibility might be the need to eliminate low-achieving students from the FCAT testing pool. After the implementation of 1998-99 FCAT and the subsequent grading of schools, a noticeable increase in referrals and placements is noted. The small increase in referrals from the third to the fourth year may be an indication that most referrals of low-achieving students had already been made, that the referral system resources were overloaded or that teachers became more aware of the possible negative outcomes for individual students associated with special education placements. However, even though reliance on referral did not appreciably increase during this time, the practice was certainly maintained.

Other factors may also be responsible for the increase in referrals noted between the second and third years. And although these factors may not be evident within the scope of this study, their influences may have had a significant influence upon the outcomes.

As there was a change in referral practices, research question three asked if changes were disproportionate for any particular identified groups. Referral practices according to student ethnicity, gender, and grade level and by school performance accountability grade, socioeconomic status, and percent of population on free/reduced lunch were examined. Significant interaction effects were found for referral practices by
student ethnicity (p < .05), student grade level (p < .01), school performance accountability grade (p < .01), school socioeconomic status (p < .01), and school limited English proficiency population (p < .01) across time. When examining the mean scores for all of these categories, it is not surprising the largest increases occurred between the second and third years.

Referral to the special education evaluation process seems to have been used most often for Black, non-Hispanic and Hispanic students. The Black, non-Hispanic group experienced an increase of 23.7% in the mean referral percentage from 1996-97 to 2000-01, while the Hispanic students experienced an increase of 24.2%. This same increase did not occur for White, non-Hispanic (10.8%) or for Other students (5.4%).

Schools from the lowest socioeconomic category also experienced the greatest increase in referrals. Their rise of the mean percentage from 7.31 in 1996-97 to 9.74 in 2000-01 represents an increase of 33.2%. This was followed by the schools categorized in the medium socioeconomic category with a 15% increase. The smallest increase, 9.8%, was noted for the schools categorized in the high socioeconomic category.

Children who are poor or of minority status are often referred and placed into the special education system. This is reflected in the data for M-DCPS. It is not surprising that this school system would be following national trends. However, the use of referral by schools according to limited English proficiency is not as clear. Certainly, schools with a high percentage of limited English proficient students increased their use of referral by 24.2% over the years of the study. However, schools considered to have a low percentage of limited English proficient students demonstrated an increase of 39%. One possible explanation for this occurrence may be the ethnic make-up of the school. If a
school is low-SES and primarily Black, non-Hispanic, their use of referral may be more a reflection of their SES make-up than their LEP population percentage.

When considering student grade level, a steady increase at the kindergarten and first grade level is noted over time. This mirrors the national trend of increase for special education placements. While the greatest overall increase occurs for kindergarten (33.4%), the next largest increase is at the third grade (25.6%). Most of this increase for the third grade occurs from the second to third years, the years after the introduction of the FCAT and the awarding of school performance grades, with the mean percentage of 8.01 in 1997-98 rising to 10.58 in 1999-00. This significant rise would not have been predicted when considering that the change from 1996-97 to 1997-98 represented a decrease in referrals. And while, there is a two-year time period from 1997-98 to 1999-00, percentages would have been expected to continue to decrease or gradually increase during this time. Referring students at the third grade level would eliminate their performances from the upcoming fourth grade testing pool.

What is of equal interest is the use of referral to the special education evaluation process according to school performance accountability grade. Schools, who were awarded the grade of “F” after the 1998-99 FCAT administration, increased their use of referral from a mean percentage of 7.04 in 1997-98 to 11.36 in 1999-00, representing an increase of 61.4%. Even with this large increase in referrals from the second to third years, a 10.7% increase is noted from the third to fourth years. Over the years of this study, “F” schools increased their reliance on referral by 104%, only a small part of which occurred from the first to second years. The use of referral by “F” schools may represent a response to students who were genuinely in need of extra services. However,
“F” schools may also have been attempting to raise the grades of their schools by eliminating the scores of low-achieving students.

The next largest increases for the use of referral from the second to third years were for the “A” and “D” schools. While “A” schools have steadily increased referrals over the years of the study, the largest increases are noted between the second and third years. The same can be said for the “D” schools. It is hypothesized that “A” schools would find it in their best interest to maintain their status, while the “D” schools would attempt to keep from becoming “F” rated.

The second research question addressed retention practices prior to and after the implementation of FCAT. A significant main effect (p < .01) was demonstrated for retention rates over time. Percentages increased from a mean of 1.01 in 1996-97 to a mean of 2.09 in 2000-01 with the most notable increase occurring between the second, 1997-98 (1.47), and third, 1999-00, years (2.11). During the years of this study, absolute retentions rose from 1,539 to 3,095, an increase of 1,556 retentions. While there was a steady increase in retentions for the first three years, reliance on retention as an intervention was basically maintained from the third to fourth year.

Policies and practices of the school system as they relate to retention have changed over the course of this study and it was anticipated that retention policies may have an affect upon retention practice outcomes. Changes in policies would suggest an increase in retention from the first to second years, particularly at the second, third and fourth grade levels. Few changes should be noted between the second and third years but a slight decrease might be noted for the fourth grade. However, for the last year of this study, retention policies should dictate a dramatic increase at the fourth grade level.
When considering overall retention patterns, an increase is noted from the first to second years that could be explained by retention policies alone. However, the large increase from the second to the third year cannot be explained by retention policies as few changes were anticipated. Although there is a consistent progressive increase across time, this increase is not dictated by the retention policies of the district.

The increase in retention from the second to third years appears to be additionally influenced by other factors. Certainly factors outside of the scope of this study may have had an influence upon the outcomes. However, the possibility exists, as it did for the use of referrals, that teachers became concerned for the achievement needs of individual students after they were provided with feedback of the performance of these students on the FCAT of 1998-99. The low scores of these students may have indicated to teachers that children had not mastered the curriculum content of their grade level, and perhaps they should not be promoted onto the next grade. This may account for some of the retentions evidenced in the data. However, retention policy was specific, and FCAT performance by itself was not cause for retention.

Another possibility might be the need to eliminate low-achieving students from the FCAT testing pool. Just as in the case of referrals, after the implementation of 1998-99 FCAT and the subsequent grading of schools, an increase in retentions is noted.

As there was a significant change in retention practices, research question three asked if changes were disproportionate for any particular identified groups. Retention practices according to student ethnicity, gender, and grade level and by school performance accountability grade, socioeconomic status, and percent of population on free/reduced lunch were examined. Significant interaction effects were found for
retention practices by student ethnicity (p < .05), student gender (p < .05), student grade level (p < .01), school performance accountability grade (p < .01), school socioeconomic status (p < .01), and school limited English proficiency population (p < .05) across time.

When examining the mean scores for all of these categories, it is not surprising the largest increases occurred between the second and third years.

A steady increase in retention rates is noted for White, non-Hispanic students. An increase of 130.7% is demonstrated over the years of this study. The same pattern is not evidenced for the other ethnic groups. While a steady increase for Black, non-Hispanic (93.3%) and Hispanic (100%) students is noted over the first three years, a decrease from the third to fourth years is noted. For the other students, a 67.7% increase is noted over time with a decrease occurring between the first and second years. Even though the largest percentage increase occurs for White, non-Hispanic students, due to the small number of White, non-Hispanic students in the district population, minimal effect on overall retention numbers occurred. However, large Black, non-Hispanic and Hispanic student populations are represented in M-DCPS’ population and changes for these demographics should exert the most influence on overall retention outcomes. While overall rates rise from the first to third years, a minimal increase of only three absolute retentions occurs from the third to fourth years. This slight increase appears to be most heavily influenced by the decrease occurring for the Black, non-Hispanic and Hispanic groups.

When comparing retention rates for male and female students, a significant difference was found. While population mean trends for males and females tend to parallel one another across the four years, the percentage increase for females from the
first to second year is almost double that of males. Percentage increases from the second to third year, and from the third to fourth year, are similar for both genders. No explanation for the occurrence from the first to second year for the female students may be found within the boundaries of this study. However, males are retained at higher percentages than females overall.

Schools from the lowest socioeconomic category also experienced the greatest increase in retentions. Their increase of mean percentage from 1.04 in 1996-97 to 2.39 in 2000-01 represents an increase of 129.8%. This was followed by the schools categorized in the high socioeconomic category with an 89.4% increase. The smallest increase, 76.1%, was noted for the schools categorized in the medium socioeconomic category.

Poor and/or minority children are more likely to be retained in school. This is reflected in the data for M-DCPS. However, the data also reveals that schools categorized as high socioeconomic status have increased their use of retention at a greater rate than those schools categorized as medium socioeconomic status. And while substantial increases are noted in the second to third year, for two of the three categories; high and medium socioeconomic status; the largest increases are evidenced during the first to second year at these schools. The use of retention was firmly in place and in use prior to the introduction of the FCAT and subsequent school accountability grades. The smallest changes in retention usage occur between the third and fourth years for all three of the categories. The pool of low-achieving students may have been severely depleted through the generous usage of retention in prior years. However, percentage means are considerably higher than pre-FCAT levels, particularly for the high and low categories.
When analyzing the data according to percent of population considered being limited English proficient, the schools with a high percentage of students increased their use of retention by 121.4% over time. However, schools categorized as having a low percentage of limited English proficient students demonstrated an increase of 114.7% over this same time. As was hypothesized for the use of referrals, one possible explanation for this occurrence in the use retentions may be the ethnic make-up of the school. If a school is low-SES and primarily Black, non-Hispanic, their use of referral may be more a reflection of their SES make-up than their LEP population percentage.

The most notable increases for schools considered high LEP population and for schools considered low LEP population, occurred from the first to second years. While respectable gains continue for the second to third years, decreases in the usage of retention occurred from the third to fourth years. However, for the schools considered medium LEP population, the greatest increase occurred between the second and third years, the year prior to and the year after the implementation of FCAT and school accountability grades. Although the high and low LEP population schools utilized retention as an established practice prior to the accountability grading of schools, the medium LEP population schools use of retention may be in response to influences not previously experienced by these schools.

When analyzing retention data according to student grade level, it is necessary to consider the M-DCPS’ retention policies and procedures as well as information regarding retention trends (Morris, 1991). Typically, retentions are the highest at the beginning of an instructional grouping and decrease over the levels of that grouping. For the elementary instructional level, that would mean that retentions would be the highest at the
kindergarten and first grade levels and decrease from the second to the fifth grades. The retention mean percentage data for each year of this study follows this trend. Retention mean percentages are greatest at the first grade level and steadily decrease from the second to fifth grades. Kindergarten retentions are seen at a rate less than used in the first to third grades, but greater than that used in the fourth and fifth grades.

Although the mean retention percentages follow theoretical trends, it is necessary to ask if they are reflective of M-DCPS’ policies and procedures. As previously stated, these policies and procedures have changed over the course of this study. It was anticipated that through these changes alone, an effect on retention outcomes would be evidenced. Specifically, there would be an increase in retention from the first to second years, particularly at the second, third and fourth grade levels; few changes should be noted between the second and third years but a slight decrease might be noted for the fourth grade; and for the last year of the study, retention policies should dictate a dramatic increase at the fourth grade level.

Analysis of overall retention patterns did support the hypothesized increase in retentions from the first to second years. The largest increases should have been seen at the second, third and fourth grade levels. While the largest increase was seen at the fourth grade (82%), the next largest increase occurred at the first grade (49.1%), followed closely by the second grade (48.1%). Kindergarten continued to show heavy reliance upon retention (41.9%) in what would seem to be contrary to its usage for that instructional level as outlined in district policies and procedures. Third grade followed closely behind with a 40.3% increase.
It was hypothesized that little change would exist for the second to the third year based upon district policies and procedures. This was not the case. Large gains in retention mean percentages occurred for all grade levels, again a finding contrary to the district’s policies and procedures. The largest increase of 54% occurred at the third grade level, with mean percentages rising from 1.74 to 2.68. In fact, the mean percentage of 2.68 for the third grade in the third year approaches the mean percentage of 2.71 seen in the second grade. However, the increase from the second to third years for the second grade was considerably less, having gone from a mean percentage of 1.94 to that of 2.71. In absolute terms this represents 689 students at the third grade and 640 students at the second grade who were retained during the 1999-00 school year.

As previously stated, the increase in retention from the second to third years appears to be additionally influenced by other factors other than district policy and procedures. The “bump” experienced for the third grade may be a response to “preemptive” retention, allowing a student an additional year of instruction before taking FCAT in fourth grade. However, this phenomenon does not explain the significant increases noted for all grade levels when little changes were anticipated.

These large increases are not evidenced during the third to fourth years. While an increase at the fourth grade might have been anticipated based upon district policies and procedures, this did not occur. Most grades levels posted little change (1.7% for first and 1.9% for fifth) or decreases (13.7% for fourth, 12% for third and 4.8% for second) for the use of retention. The reasons for this are not clear. As retention usage was on the rise for first through the third years, the pool of available low-achieving students may have been decreased and district policies did not allow for a second retention at an instructional
grouping (K-2 and 3-5) to occur if a student had previously been retained in that

grouping. This would not have affected practices in kindergarten. For this level, reliance
upon retention continued to be demonstrated, again, running contrary to district policies
and procedures.

Just as with referrals, analysis according to school performance grades is of
particular interest. Schools, who were awarded the grade of “F” after the 1998-99 FCAT
administration, increased their use of retention from a mean percentage of .56 in 1997-98
to 3.10 in 1999-00, representing an increase of 454%. The most significant amount of
increase occurred from the second to third years. Although this time period represents
actually two years in time, the mean percentage increase from 1.0 to 3.0 exceeds that
which would have been anticipated when considering the percentage means from the first
to second years, .56 to 1.0. Just as with their use of referral, the use of retention by “F”
schools may represent a response to low-achieving students who were genuinely in need
of extra services. However, this concern seems to be attenuated after the implementation
of FCAT and the subsequent grading of schools.

The next largest increases for the use of retention over time were for the “A” and
“D” schools. While “D” schools steadily increased retention over the years of the study,
the largest increases are noted between the second and third years. Although representing
small numbers of students, the increase for the second to third years for the “A” schools
represents the largest percentage increase for any of the grade categories.

Steady increases are noted for “C” schools from the first to third years. Their most
substantial increase occurs from the first to second years. District policies and procedures
might account for this increase. For “B” schools, the largest increase is noted between the
second and third years when little changes in the use of retention were anticipated. The dramatic increase for all grade categories between the second and third years, when no particular changes were anticipated, suggests a response to influences other than district retention policies and procedures.

Conclusions

The purpose of this study was to examine referral and retention practices prior to and after the implementation of the FCAT and the subsequent awarding of school performance accountability grades. The results of this study indicate that reliance upon these practices significantly increased over the years of the study. In general, retention percentage means increased from the first to the third years, with means being maintained from the third to fourth years. The pattern for referral was somewhat different. A noticeable increase was evidenced for this practice from the second to the third year.

As indicated in the review of the literature, a number of decisions are made as a result of high-stakes testing. Rewards and punishments are awarded to schools based upon the achievement levels of their students on these tests. In order to avoid punishments and sanctions, educators may seek to eliminate the scores of low-achieving students through the use of retention and referral to special education. It is suggested that the increase in the use of both of these practices over the time of this study may have been at least partially in response to a perceived need to eliminate the scores of low-achieving students. By eliminating these scores, a school might be able to avoid the negative consequences associated with a low school performance grade or gain the rewards associated with a high school performance grade.
Support for this conclusion may be found in the analysis of practices according to school and individual characteristics. While children who are poor and/or of minority status continue to be retained or referred at high levels, further individual and school characteristic data indicate a surprising development and reaction. While increases in the use of referral were noted for all grade levels, the most dramatic increase occurred for the third grade. A large increase was also revealed for the use of retention at the third grade level. By holding students back or placing them in special education, these students’ scores would be eliminated from the upcoming fourth grade testing pool, the grade level from which school performance accountability grades were determined.

When examining the data according to school performance accountability grade, dramatic results were noted for those schools with the grade of F. Their heavy reliance upon the use of retention and referral practices may have been an attempt to avoid the extreme sanctions associated with being an “F” school for more than one year. Some caution is prudent when interpreting referral outcomes for these schools due to the limitations cited earlier in the Methods section. Results for school level characteristics may either under- or over-represent true occurrences. However, the increase in percentage means for these “F” schools is so dramatic and striking that even if the data over-represented the use of referral by these schools, their use of referral far outdistances the use by the other school performance accountability categories. Their use of retention, generated from data without the same associated limitations, follows that of referrals as well as trends reported for Chicago and other Florida schools (Figlio & Getzer, 2002; Jacob, 2002).
Policy Implications and Future Research Considerations

The practices of retention and referral to special education may potentially prove harmful or provide questionable benefit to individual students (Kavale & Forness, 1999; Rafoth & Carey, 1995). However, the public and educators alike do not appear to have full knowledge regarding these harmful or questionable benefits. The research regarding retention is uniformly negative; however, educators are willing to engage in this practice. The data regarding retention outcomes needs to be an integral part of inservice and continuing education training programs. Educators in turn need to provide this information to parents and legislators. Armed with this knowledge, the public and educators might not be so willing and accepting of retention, especially when it is practiced by a school in order to avoid sanctions and punishments associated with high-stakes testing.

The issues surrounding referral to special education are not as clear as the benefits vary. For those students who are suspected of having a disability, a referral to this process and the protection of rights it affords them is essential. However, for those students who have simply not been taught or whom educators are attempting to eliminate from the testing pool, the practice is simply unacceptable. By utilizing effective prereferral practices to ensure the development of appropriate general education interventions strategies and the monitoring of student’s response to interventions, the chances of inappropriate referrals should be decreased. For this to occur, all schools would need to be trained in effective prereferral methods and be provided continued support and training to ensure uniform implementation of the methodology. This is the process currently underway in M-DCPS with the School Support Team.
It is anticipated that the use of retention will continue as an intervention response to high-stakes testing, but the pattern associated with it, will change. The emphasis for accountability purposes and the awarding of school performance accountability grades has shifted to the third grade. Retention is also mandatory for those third grade students that do not pass the FCAT. Along with this shift, it is expected that retentions will increase for the second grade. Not only so that children will be given that extra year before taking the test, but also to eliminate the scores of low-achievers for at least an additional year. Further increases might also be noted in kindergarten and first grade in an effort to delay low achieving students from the testing pool as long as possible.

While schools must now test children in special education and demonstrate adequate yearly progress of these students, it is suggested that referral will continue to be used as intervention first choice. Schools may feel the pressure to place some students into special education that they might not have placed in the past. These children would essentially be the testing “ringers”. Their scores would help to boost the overall scores of the special education population for the school and therefore, the school would be able to demonstrate adequate yearly progress of that group. Testing accommodations, such as additional time, would be provided thereby providing “extraordinary conditions” for particularly “ordinary” students (Figlio & Getzer, 2002).

The use of retention and referral practices should be monitored closely by all school districts. Analysis of overall outcomes will reveal only so much. It is critical to evaluate by individual and school characteristics so as to determine new, emerging patterns of behavior.
Even if the use of retention and referral fell out of favor, new practices would be initiated. As long as the possibility for rewards or punishments and sanctions exist, schools and individuals will engage in behaviors that will help ensure particular outcomes. Policies need to change so that educators will choose practices with a sound research base that provide positive outcomes for all students.
REFERENCES


Larry P. v. Riles. 343 F. Supp. 1306 (D.C.N.D. Cal. 1972). *aff’d, 502 F.2d 963 (9th Cir. 1974), further proceedings, 495 F. Supp. 926 (D.C.N.D. Cal. 1979), aff’d, 502 F.2d 693 (9th Cir. 1984)*.


Yang, J. (2000). Before the board: High-stakes cheating; lowering the bar; court rejects vouchers; small schools are better; math problems. *American School Board Journal, 187*(2), 10-12, 14, 16-18.


VITA

SUE L. BUSLINGER-CLIFFORD

Born, Coral Gables, Florida

<table>
<thead>
<tr>
<th>Year</th>
<th>Degree/Credential</th>
<th>Institution</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>1974</td>
<td>B.S. Liberal Studies</td>
<td>Florida International University</td>
<td>Miami, Florida</td>
</tr>
<tr>
<td>1978</td>
<td>Ed.S. Counselor Education</td>
<td>University of Florida</td>
<td>Gainesville, Florida</td>
</tr>
<tr>
<td>1992</td>
<td>Ed.S. School Psychology</td>
<td>Florida International University</td>
<td>Miami, Florida</td>
</tr>
<tr>
<td>1999 – 2000</td>
<td>Adjunct Instructor</td>
<td>Florida International University</td>
<td>Miami, Florida</td>
</tr>
<tr>
<td>2000 – 2004</td>
<td>Graduate Assistantship</td>
<td>Florida International University</td>
<td>Miami, Florida</td>
</tr>
<tr>
<td>2003 – present</td>
<td>Instructional Supervisor</td>
<td>Division of Psychological Services</td>
<td>M-DCPS</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Miami, Florida</td>
<td></td>
</tr>
</tbody>
</table>