2014

Predictive Utility and Achievement Outcomes of Two Simultaneous District-Developed Interim Assessment Programs

Tavymae W. Chen
University of North Florida, tavymae@gmail.com

PREDICTIVE UTILITY AND ACHIEVEMENT OUTCOMES OF TWO SIMULTANEOUS DISTRICT-DEVELOPED INTERIM ASSESSMENT PROGRAMS

by

Tavymae Wells Chen

A dissertation submitted to the Doctoral Program Faculties

in partial satisfaction of the requirements for the degree of

Doctor of Education in Educational Leadership

UNIVERSITY OF NORTH FLORIDA

COLLEGE OF EDUCATION AND HUMAN SERVICES

Spring 2014

Certificate of Approval

The dissertation of Tavymae Wells Chen is approved: **Date**

Dr. Warren Hodge

Dr. Tracy Packiam Alloway

Dr. Terence Cavanaugh

Dr. Larry G. Daniel, Committee Chair

Accepted for the Department:

Dr. Jennifer Kane, Chair

Accepted for the College:

Dr. Larry G. Daniel, Dean

Accepted for the University:

Dr. Len Roberson
Dean of the Graduate School

Dedication

This work is dedicated to the memory of my brave and beautiful mother, Pamila Clemons Yoder, who taught me that taking a risk is its own reward.

I also dedicate this dissertation to those who have ever doubted themselves or whether an academic journey is possible or worthwhile.

It is, so get going.

Acknowledgements

My husband and best friend, Edem, has encouraged me throughout this process in every way possible. He has selflessly and tirelessly supported my doctoral journey emotionally, intellectually, and financially. I cannot imagine a better life partner and I hope to make him proud.

I wish to extend my gratitude and admiration to fellow members of Cohort 19: Wallace, Bill, Devon, Melanie, Lashantah, Dawn, Brian, Luisa, Eileen, Renita, Keith, Ethel, Betsy, and, in particular, Tabby and Tammy (the other "T"s). We have shared laughter, celebrations, and even a few tears. For those 19ers who have not yet finished, please keep moving at whatever pace feels right, and do not stop until you are done. Remember that we have naysayers to prove wrong.

To my close friends and work colleagues, Amanda and Malinda, thank you for listening and challenging me intellectually. Your patience, love, forgiveness, and shared laughter have kept me sane and connected to humanity. If and when you are ready to further your travel down the academic road, I hope you will allow me to return the favor.

I would also like to thank several mentors and true leaders who believed in my abilities with such fervor that I had no choice but to start to believe in myself: Mrs. Bechtel from Julie Rohr, Dr. John Klauder, Faychone Durant, Dr. Kelly Coker-Daniel, Carol Daniels, Dr. Ellie Scheirer, Aaron Smith, Marge Hayden, Tim Ballentine, Dr. Kathe Kasten, and Andrew Post. Although I may no longer have you all in my daily life, I continue to learn about leadership by reflecting on your past actions and words.

I would also like to thank four special professors and mentors, Dr. Warren Hodge, Dr. Tracy Packiam Alloway, Dr. Terrence Cavanaugh, and especially my chair and biggest advocate, Dr. Larry Daniel. Thank you all for agreeing to offer your expertise and time through serving on my committee. I was told that the dissertation process is as much, if not more, of a learning experience as any of the doctoral courses. However, what I never expected was to learn how supportive and academically communal the experience would be. Whether strategy or luck led me to choose this specific combination of individuals with uniquely complimentary areas of expertise and personalities, I have a stellar committee. I hope that the final product is a testament to all that I have learned from each of you.

Finally, I would like to thank my family. In the great series of random events we call life, I ended up with the most unlikely and yet uniquely perfect combination of people to call *my family*. All of my life's past and future lessons are embodied in my sisters, brothers, parents, in-laws, grandparents, aunts, uncles, and cousins. Thank you all for loving me, despite everything. I wouldn't be here to write this if it weren't for each of you.

Table of Contents

Certificate of Approval ii

Dedication iii

Acknowledgements iv

List of Tables viii

List of Figures ix

Abstract of the Dissertation x

Chapter 1 - Introduction 1
- Purpose of the Study 1
- Significance of the Study 2
- Theoretical Framework 3
- Research Questions 6
- Key Terms and Definitions 7
- Organization of the Study 11

Chapter 2 - Literature Review 13
- Evolution of Standardization, Accountability, and Testing Policy in the U.S. 13
- Effects of High-Stakes Testing Accountability Policies 22
- Recent Test-Based Accountability in Florida 38
- District-Developed Interim Assessments 41
- Summary 46

Chapter 3 - Methods & Procedures 49
- Research Design 49
- Research Questions 50
- Hypotheses 51
- Population, Sample, and Data Sources 56
- Data Collection and Ethical Considerations 68
- Data Analysis Procedures 69
- Delimitations and Limitations of the Study 74
- Summary 78

Chapter 4 - Results 79
- Review of Methods and Procedures 80
- Sample 83
- Assumptions and Descriptive Statistics 84
- Analysis of Regression Models 92
- Analysis of Covariance (ANCOVA) 112
- Summary 118

Chapter 5 - Discussion 121
- Review of Study Design and Methods 121
- Research Question Answers and Discussion 123
- Analysis of Predictors Across Regression Models and Discussion 132
- Implications for Policymakers and Educational Leaders 139
- Recommendations for Future Studies 145
- Summary 147

Appendix A – University Instructional Review Board Waiver 149
Appendix B – District Instructional Review Board Approval 150
Appendix C – Correlation Coefficients for RQ2 151
References 156
Vita 168

List of Tables

Table 1. Sample Size by Number of IBA Tests Taken and Number of LSA Tests Taken 58
Table 2. Cases Removed from Regression Sample – Missing Interim Scores 59
Table 3. ANCOVA Sample Sizes – RQ3 60
Table 4. Sample Demographics – Regression and ANCOVA 60
Table 5. Comparison of Simultaneous 2012-2013 Interim Assessment Programs by Selected JCSEE Student Evaluation Standards 66
Table 6. District Mean 2013 Grade 6 Mathematics FCAT 2.0 Developmental Scale Score by Race/Ethnicity 72
Table 7. Overview of Study Design 75
Table 8. Descriptive Data for All Non-demographic Study Variables 86
Table 9. Demographics by 2012 Grade 5 Mathematics FCAT Achievement Level 89
Table 10. Multiple Regression Model Summary - RQ1 98
Table 11. Multiple Regression Model Summary for Block 5 of each 2012 FCAT Achievement Level – RQ2 99
Table 12. Regression Beta Weights – Block 5 for RQ1 (overall model) 100
Table 13. Regression Beta Weights and Structure Coefficients by 2012 FCAT Level – Block 5 for RQ2 (by Levels) 102
Table 14. Regression Beta Weights, Correlations with FCAT_SS_2013, and Structure Coefficients – Block 5 for RQ1 (overall model) 105
Table 15. Correlation Coefficients for RQ1 108
Table 16. Multicollinearity Diagnostic Measures – Block 5 for RQ1 110
Table 17. Multicollinearity Variance Inflation Factors (VIF) – Block 5 for RQ2 112
Table 18. Descriptive Statistics for RQ3 114
Table 19. ANCOVA Results for RQ3 116

List of Figures

Figure 1. Theoretical framework for the present study. 6

Figure 2. Types of accountability policies. 21

Figure 3. Types of educational assessments (Perie et al., 2009). 44

Figure 4. Concept map including key components in the literature review regarding U.S. educational policy and interim assessment. 47

Figure 5. Simplified Learning Schedule timeline by month and corresponding content tested by Interim Benchmark Assessments (IBA) and Learning Schedule Assessments (LSA). 63

Figure 6. Constructing Groups for ANCOVA Model – RQ3. 74

Figure 7. Illustrative example of a frequency distribution with a superimposed normal curve used to visually inspect normality. 85

Figure 8. Illustrative example of a scatterplot used to visually inspect homoscedasticity for metric independent variables. 87

Figure 9. Scatterplot depicting the heteroscedastic relationship between 2012 Grade 5 Mathematics FCAT 2.0 and 2013 Grade 6 Mathematics FCAT 2.0 developmental scale scores. 88

Figure 10. Illustrative example of a boxplot used to visually inspect homoscedasticity for categorical independent variables. 89

Figure 11. Illustrative example of a scatterplot with a superimposed line of best fit used to visually inspect linearity. 90

Figure 12. Illustrative example of a scatterplot with a superimposed line of best fit and a quadratic curve of best fit used to visually inspect whether a transformation might be appropriate. 91

Figure 13. Scatterplot of standardized predicted values versus standardized regression residuals. 93

Figure 14. Example scatterplot of standardized predicted values versus standardized regression residuals for students who achieved Level 2 on the 2012 Grade 5 Mathematics FCAT 2.0. 94

Figure 15. Bar graph of ANCOVA interactions. 115

Figure 16. Scatterplot matrix for ANCOVA. 117

Abstract of the Dissertation

Predictive and Instructional Utility of Two Simultaneous District Interim Assessment Programs

by

Tavymae Wells Chen

University of North Florida

Jacksonville, Florida

Dr. Larry G. Daniel, Dean, Chair

The purpose of the present correlational, ex post facto study was to evaluate the predictive ability and academic achievement criterion outcomes of two district-developed interim mathematics assessment programs for a sample of 5,801 grade 6 students in a large urban school district. Average scores for both interim assessment types were statistically significantly more related to 2013 FCAT 2.0 scores (r = .75 and .72; p < .001) than all other predictors (i.e., student demographics, Florida school grade, and student course GPA) except for 2012 FCAT 2.0 scores (r = .78; p < .001). Further, the newer interim assessment program with an instructional purpose and curriculum-based sequencing had slightly stronger overall predictive power (r_s = .88) and a higher criterion mean score (M = 218.08) than the older, state-test mirror interim assessment program (r_s = .85; M = 215.47). Regression models by prior year FCAT 2.0 Achievement Level yielded some predictor ranking discrepancies by prior achievement level. Although not statistically significant at the .01 level, groups of students with a more moderate total number of interim assessments outperformed groups with all or nothing.

Overall, the two types of interim assessment programs evaluated in the present study were good predictors of the state high-stakes test, 2012 Grade 6 Mathematics FCAT

2.0. However, more research must be done to identify with certainty whether or not the act of taking the interim tests and receiving feedback contributes to improved student learning.

CHAPTER 1

Introduction

Since the 1840s, students in the United States have been exposed to high-stakes educational assessment. Over the years and more recently with No Child Left Behind (NCLB) and Race to the Top (RTTT), the federal government has expanded its role in overseeing test-based accountability based on these educational assessments. Within the realm of educational test-based accountability and achievement-focused public schooling, interim assessments are nestled between the popularized formative assessments, with their lofty promises of instructional gains yet elusive definition, and stalwart summative assessments, omnipresent and routinely criticized. Unlike both formative and summative assessments, interim assessments may well be the "Goldilocks" of educational assessment.

Purpose of the Study

The purpose of the present study was to examine the outcomes of interim assessments and to help address the questions, "Do these interim assessments work the way they were expected to work?" and "Do they work better than what we already have?" Simply checking for alignment between purpose and results is insufficient for a full evaluation of interim assessments. Scriven, the scholar who first developed the current usage for the terms *formative* and *summative* evaluation (1967), emphasized that evaluation studies must examine side effects, consider cost effectiveness, and identify "critical competitors" (Scriven, 1974, p. 25). Researchers and evaluators must attempt to find alternative means that might accomplish the same or better ends at lower cost or with fewer negative side effects. For example, teacher-assigned grades, prior year test scores, and/or demographics might serve as alternatives to interim assessments.

The present study examined the end-of-year summative achievement outcomes of students who took either of two distinctly different types of interim assessment. The study compared the utility from a district-level perspective of a common form of interim assessment—the predictive whole-year summative version—with another, less common and more involved form of interim assessment—the instructional curriculum unit-based interim assessment. The study investigated whether these two types of interim assessments are valuable for predictive purposes above and beyond data widely available after NCLB, such as student demographics, prior year scores, and teacher-assigned grades.

Significance of the Study

The introduction of NCLB reporting requirements and subsequent consequences increased the sense of urgency among educators and policymakers to capture measures of performance prior to high-stakes testing dates. Research literature about formative assessment is quite common and detailed (Wiliam & Black, 1998a, 1998b); however, the literature relating to interim or benchmark testing is both sparse and inconclusive (Goertz , Olah, & Riggan, 2010; Henderson, Petrosino, Guckenburg, & Hamilton, 2008; Perie, Marion, & Gong, 2009; Shepard, 2010). Even so, evaluative and predictive data sought by educators and policymakers are usually not a product of the formative assessment process (Li, Marion, Perie, & Gong, 2010; Popham, 2008). As a result, local and national policymakers are leaning toward including interim assessments in required assessment systems.

Although not high-stakes, interim assessments are an integral component of next generation assessment systems. Interim assessment is uniquely situated between low-stakes classroom formative assessment and high-stakes external summative assessment. In the state of Florida, school districts are required by law to administer local assessments of some form to provide data on remedial student progress toward the state standardized tests (Fla. Stat. Ann. § 1008.25 (4)(a), 2013) and low-performing school improvement (Fla. Stat. Ann. § 1008.35 (1)(a), 2013). Many school districts, such as the district of focus in the present study, administer assessments to all students in all schools in lieu of simply satisfying these state requirements for remedial or low-performing students and schools.

Educators are spending more instructional time administering additional district-level tests and less time instructing students. The instructional time lost represents what economists would call an *opportunity cost* of test administration, meaning that educators must pay for testing in the form of the missed opportunity to teach. As this shift occurs, questions about whether interim assessments are valuable to educators, and if so, which kinds are more valuable, are key to evaluating policies and improving instructional practice as a whole.

Theoretical Framework

The three main types of assessment described by Perie, Marion, and Gong (2009) are summative, with a large scope and minimal frequency; formative, with a narrow scope and high frequency; and interim between the two, with a moderate scope and frequency. Each has a place within a K-12 district comprehensive assessment system. Although summative testing has been studied in the U.S. since the early 19th century, Wiliam and Black (1998a, 1998b) brought a positive light to ongoing, classroom formative assessment in the late 20th century. Interim assessment is newer still, and yet has been increasingly utilized for predictive, instructional, and evaluative purposes.

If educators are to forgo instructional time to administer interim assessments, the tests ought to have the most utility possible. One such characteristic of utility is the predictive ability of interim assessments versus other, less intrusive and time-consuming predictors such as teacher-assigned grades or demographics (Perie, et al., 2009). Another characteristic of utility is academic impact or instructional value.

The present study utilized Perie et al. (2009)'s framework for considering interim assessment programs. Included in their framework are evaluative criteria by which to consider different interim assessment systems, depending on which intended purpose the systems espouse. In particular, the first criterion for predictive interim assessments was used to help answer whether the tests provide predictive utility:

> The assessment should be highly correlated with the criterion measure (e.g., the end-of-the-year state assessment). The technical documentation should include evidence of the predictive link between the interim assessment and the criterion measure. However, in order to justify the additional testing and cost, the predictive assessment should be significantly more related to the criterion measure than other measures (e.g., teachers' grades) that could be used. (Perie et al., 2009, p. 10)

Perhaps another way to think about this is that data collected from a predictive interim test should have sufficient criterion related validity to justify the opportunity cost of administration. This thought is also reflected in a more general sense for all student assessments in the *Student Evaluation Standards*, developed by the Joint Committee on Standards for Educational Evaluation in 2001. The standard that applies here is standard A1 Validity Orientation, which recommends that student evaluations allow for valid interpretations (JCSEE, 2001).

In addition to predictive utility, instructional utility was evaluated using Perie et al.'s (2009) second criterion for instructional interim assessments:

> Ideally, the system should provide evidence, based on scientifically rigorous studies, demonstrating that the assessment system has contributed to improved student learning in settings similar to those in which it will be used. (p. 10)

JCSEE's first standard in the *Student Evaluation Standards*, P1 Service to Students, addresses the concept of improving student learning in a broader sense: that student evaluations should "promote sound principles, fulfillment of institutional missions, and effective student work, so that the educational needs of students are served." One method for determining whether or not the educational needs of students have been served, or in other words, that learning has occurred, is to use an existing summative test aligned to the subject taught.

All of these aspects of student assessment in general, as well as for interim assessments in particular, take place within the context of local and national policy (see Figure 1). Policy at the federal, state, and district levels greatly influences how and what assessment programs are implemented at the school level.

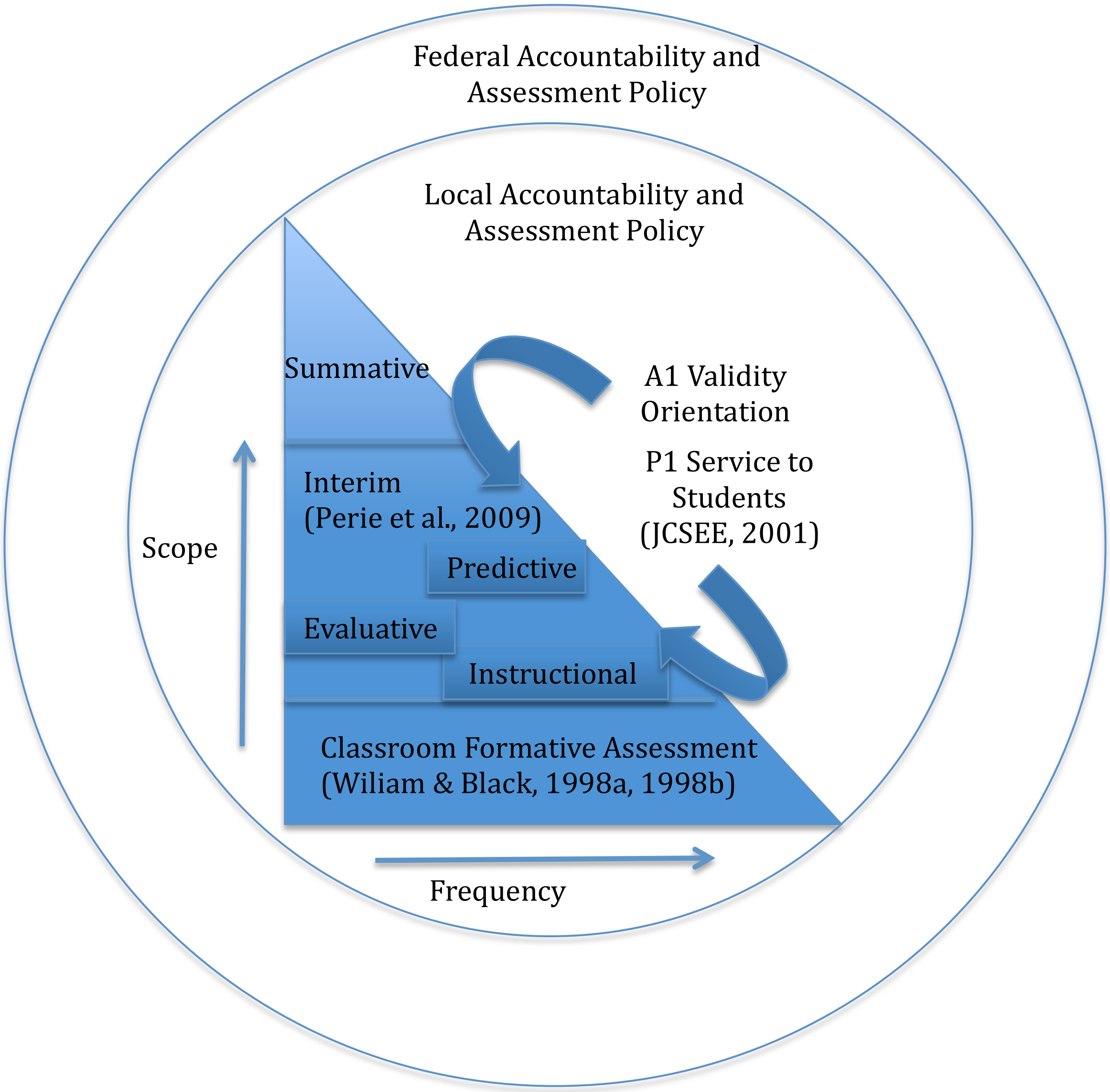

Federal Accountability and Assessment Policy
Local Accountability and Assessment Policy
Summative
Interim
(Perie et al., 2009)
Predictive
Evaluative
Instructional
Classroom Formative Assessment
(Wiliam & Black, 1998a, 1998b)
A1 Validity Orientation
P1 Service to Students
(JCSEE, 2001)
Scope
Frequency

Figure 1. Theoretical framework for the present study. Adapted with permission from "A framework for considering interim assessments" by M. Perie, S. Marion, and B. Gong, 2007. Retrieved from http://www.nciea.org.

Research Questions

The first two research questions in the present study were directed at predictive utility, and the last addressed instructional utility, of two simultaneous interim assessment programs implemented in a large, urban school district in Northeast Florida. The following questions provided a framework for study design, analysis and discussion in the present study:

1. To what extent can variance in middle school student scores on mathematics high-stakes state tests be explained by scores on district interim assessments after controlling for prior scores, student demographic variables, and teacher-assigned grades?
2. To what extent can variance in middle school student scores within achievement levels on mathematics high-stakes state tests be explained by scores on district interim assessments after controlling for prior scores, student demographic variables, and teacher-assigned grades?
3. To what degree does achievement, as measured by mathematics high-stakes state tests, of middle school students who have experienced less frequently administered, predictive interim assessments differ from the achievement of students who have experienced more frequently administered, instructional assessments, after controlling for prior scores?

Key Terms and Definitions

The following section presents definitions of terms germane to the present study.

Assessment or test – Assessment is “a comprehensive set of means for eliciting evidence of student performance” (National Research Council, 2001) or a “process of gathering information for the purpose of making judgments about a current state of affairs” (Pellegrino, 2002, p. 48). Assessments must necessarily include cognition, observations, and interpretation (Pellegrino, 2002). Although some consider *tests* a rigid subset of *assessments*, the terms *test* and *assessment* are used interchangeably throughout the present study, unless specifically differentiated in some way.

Interim Assessment - Interim assessment means “an assessment that is given at regular and specified intervals throughout the school year, is designed to evaluate students’ knowledge and skills relative to a specific set of academic standards, and produces results that can be aggregated (e.g., by course, grade level, school, or [Local Educational Agency] LEA) in order to inform teachers and administrators at the student, classroom, school, and LEA levels” (U.S. Department of Education, 2009, p. 15). Interim assessments are not high-stakes external assessments, nor are they low-stakes classroom assessments—they fall between these types. One key difference between interim assessments and formative assessments is that interim scores may be aggregated across teachers and schools, whereas formative scores tend to be unique to the teacher.

Some argue that when thinking about K-12 education as a whole, and graduation as the ultimate outcome, any assessment following the first test given in Kindergarten is an *interim* assessment. However, the present study adheres to the definition provided by the U.S. Department of Education above, which would exclude most state testing programs because they are not given at regular and specified intervals, and excludes informal and classroom tests because the results cannot necessarily be aggregated across grade levels, schools, or school districts.

Formative Assessment - “Assessments that pair the efforts of the student and teacher in order to develop an individual learning progression are said to be formative in nature” (Lile, 2012). Formative assessment is more informal than other types of assessment, and involves evaluation of student understanding while the student is still learning—prior to the end of the instructional unit or course. Interim assessments are not initially formative because they are typically developed external to the classroom and are more formal. However, depending on the type and security level of interim assessment, teachers may formatively use the data gathered by an interim assessment, and possibly the test itself ex post facto, to discuss details with students and develop an individual learning progression.

Summative Assessment - An assessment is summative if the main purpose is to determine what the student learned over the course of an instructional unit or year. Summative assessments are usually formal and are not intended for granular feedback purposes (e.g., evidence for mastery of specific skill areas). These tests typically are secure, meaning that students do not know the content of the test prior to administration, and in the case of high-stakes state tests, even school district faculty and staff are prohibited from viewing the contents.

Utility – In a general sense, utility or usefulness of assessment is defined by Herman and Baker (2009) as follows:

> Utility represents the extent to which intended users find the test results meaningful and are able to use them to improve teaching and learning. Benchmark tests with high utility provide information that administrators, teachers, and students can use to monitor student progress and take appropriate action. (p. 53)

The present study focused mainly on specific aspects of usefulness from the school district administrator perspective, namely predictive ability and end-results of achievement. Other perspectives are equally as important but were not addressed in the present study. For example, the *Student Evaluation Standards* contain seven utility standards, including information scope, evaluator qualifications, and effective reporting (JCSEE, 2001). These aspects are requirements for an overall useful student evaluation. However, for purposes of brevity in the present study, the word *utility* represented the usefulness in terms of administrators' use for prediction and final evaluation of instructional outcomes as measured by the state standardized test.

Educational Accountability - Within the realm of educational policy, "educational accountability" focuses upon transparency and the obligation to report assessment results to external evaluators and the public in an accurate manner. Accountability arose from external concerns about the efficacy of public school systems and specific actors within the system. Accountability systems are characterized by "increased real or perceived stakes of results for teachers and educational administrators" (Linn, 2000, p. 7).

High-stakes Testing - "High-stakes testing is the process of attaching significant consequences to a standardized test performance with the goal of incentivizing teacher effectiveness and student achievement" (Nichols, Glass, & Berliner, 2010, p. 3) or, conversely, exercising sanctions for lack of effectiveness or achievement.

Benchmark (standards) - Benchmarks, as defined in the present study, are operationalized as learning targets set by the Florida Department of Education (FDOE). Benchmarks are sub-components of *Big Ideas* or *Supporting Ideas* as delineated in the 2007 Mathematics Next Generation Sunshine State Standards for K-8. Each grade level in Florida has its own set of Big Ideas and Supporting Ideas, and each of these includes more granular benchmarks. M/J Math 1, the traditional sixth-grade mathematics course, includes 3 Big Ideas, 3 Supporting Ideas, and 19 benchmarks. For example, Big Idea 1 is "Develop an understanding of and fluency with multiplication and division of fractions and decimals," whereas one of the benchmarks under this Big Idea is “Explain and justify procedures for multiplying and dividing fractions and decimals” (FDOE, 2007).

Learning Schedule – The school district published a Learning Schedule for the majority of courses offered in the district during the 2012-2013 school year. Each Mathematics Learning Schedule included learning targets, assessment exemplars, a pacing guide for the district curriculum, and supplemental resource lists. Learning Schedules were developed by school-based educators and reviewed by district-level content specialists.

Module – For the 2012-2013 school year, a module was an instructional unit in any Mathematics course’s Learning Schedule. The Mathematics Learning Schedule authors wanted to use a term that would not be confused with a textbook *unit*, freeing the Learning Schedule authors to veer away from the sequence in the adopted textbooks and instead use the state’s benchmarks to guide sequencing.

Urban – The present study utilizes the United States Census Bureau’s definition of *urban* as “densely developed residential, commercial and other nonresidential areas” and as having 50,000 or more people in one area (U.S. Census Bureau, 2012). The school district of focus includes over 1 million residents, qualifying the school district as *urban*.

Organization of the Study

The present study is organized into five chapters. This first chapter has introduced the study, including context, research questions, significance, definition of key terms, and an overview of the organization of the study. Chapter 2 begins with a review of the historical literature leading up to and including current test-based accountability policy in the United States, followed by a discussion of literature relative to the effects of accountability policy as well as the application of the policy at the local level. The literature review also includes a basic assessment review of types of district-based interim assessments and a discussion regarding the differences between formative assessments and interim assessments. Together, these studies provided a conceptual and theoretical framework useful for the later analysis of data, discussion of findings, and conclusions. The purpose of Chapter 3 is to describe the design and methods of the study, including a description of the two interim assessment programs, the context surrounding the particular types of assessments used as data sources, sample selection criteria, study design, and the procedures used to collect and analyze data. Study delimitations and limitations are also included in Chapter 3. Chapter 4 presents the results of the study, and Chapter 5 includes a discussion of the major findings, implications of the study, and policy and future research recommendations.

CHAPTER 2

Literature Review

This chapter is an overview of literature regarding four main research areas relating to district-created interim assessments:

1. Evolution of Standardization, Accountability, and Testing Policy in the U.S.
2. Effects of High-Stakes Testing Accountability Policies
3. Recent Test-Based Accountability in Florida
4. District-Developed Interim Assessments

Each section in the chapter corresponds to one of these four research areas. The chapter begins with a broad historical perspective and ends with topics more particular to the present study, including major purposes for district-level interim assessments. A conceptual framework follows the literature review.

Evolution of Standardization, Accountability, and Testing Policy in the U.S.

A common misperception is that high-stakes testing is a new phenomenon. Interestingly, the original high-stakes test and standards-based instruction took place well before the No Child Left Behind Act and *A Nation at Risk*. Education-related accountability testing in the United States began during the 1840s in the same city that boasts the nation's first and oldest existing public school, Boston Latin School, and the nation's first high school, English High School (Kress, Zechmann, & Schmitten, 2011; Spring, 2005). Boston's first test to monitor their schools' effectiveness was very similar in purpose to what we now consider high-stakes tests in that it was used to support comparisons across classrooms and schools.

Just as high-stakes testing is not new, neither is the concept of educational standards. The first official standards began with a group of college presidents and professors who gathered in the early 1890s as the National Education Association's Committee of Ten on Secondary School Studies (Goertz, 2007, p. 4; Spring, 2005). The National Education Association, which is now the nation's largest teachers' union (Spring, 2011), and its Committee of Ten recommended a standardized curriculum for all high school students, including preparation for college or career, as well as how knowledge of the curriculum should be assessed. At the time, though, high schools only enrolled 10% of 14- to 17-year-olds, thus limiting the reach of these standards (Goertz, 2007; Kress et al., 2011; Spring, 2005).

Varied curriculum, standardized tests for tracking. Once students from working class and immigrant families began reaching high school in larger proportions, the National Education Association convened a Commission on the Reorganization of Secondary Education (CSRE) in 1918. The CSRE recommended a move away from the common curriculum previously in place (Goertz, 2007). Their report, *Cardinal Principles of Secondary Education*, called for a comprehensive high school with a differentiated curriculum including vocational courses (Spring, 2005).

Just as curriculum was becoming less standardized, educational testing was becoming more standardized by virtue of business and military influence. Frederick Winslow Taylor's concept of scientific management became popular among businesspeople around the turn of the 20th century. Taylor described efficient factory production that relied upon the managers' ability to study and gather a massive amount of data about the work, hire the best workers, provide incentives to do what was considered best practice, and divide the work appropriately (Taylor, 1916). Soon after, John Franklin Bobbitt applied Taylor's scientific management concept to educational management (Au, 2011).

The military also had a great deal of influence on the standardization of testing. In 1917, psychologists published the Army Alpha assessment to sort World War I recruits by their perceived ability, and popularized the idea that these multiple-choice standardized intelligence tests were superior to subjective, less scientific, constructed-response tests often seen in classrooms (Ravitch, 2010; Resnick, 1985). A few years later, educators began to sort students by their perceived ability by using the first series of standardized achievement tests, the Stanford Achievement Tests (Hamilton, 2003). Educators valued the efficiency and objectivity of machine-graded tests as student enrollment continued to increase and diversify; by 1930, approximately 70% of 14- to 17-year-olds were enrolled in high schools across the country (Kress et al., 2011; Ravitch, 2010; Spring, 2005).

Educational policymakers began to view the practice of implementing test policy as a favorable way to reform education for several reasons. Tests are relatively inexpensive (Linn, 2000; Ravitch, 2010), easily and quickly put into action or altered (Jennings, 2012; Linn, 2000), and produce visible results that will most likely demonstrate an increase within the first few years of implementation (Linn, 2000). Testing as a policy became even more popular following the 1947 speech by scientist and academic James B. Conant, who conceptualized an educational system which singled out and supported students with greater intellectual capacity for the benefit of the country's scientific community and military forces (Spring, 2005). The newly developed and highly controversial Selective Service College Qualification Test allowed college-bound men to choose whether to serve in the armed forces or go to college. Tests during this time were seen as tools for student selection and tracking (Linn, 2000).

After the war, however, public schools began to demand more funds and space as they tried to accommodate the Baby Boom generation and students from a wider variation in racial and ethnic backgrounds following the historic 1954 *Brown v. Board of Education* ruling and Civil Rights Act of 1964 (Ravitch, 2010; Spring, 2005; Gong, 2012). Warnings from academics and military critics that education had become anti-intellectual coincided with the Russian launch of the Sputnik I satellite. Responding to public outcries, Congress passed the National Defense Education Act (NDEA) in 1958. In a speech promoting NDEA, President Dwight D. Eisenhower called for a nationwide testing system to select, and incentives to persuade, students with high ability to seek scientific or professional career paths (Goertz, 2007; Spring, 2005). As part of NDEA, federal money was tied to the establishment of specific educational programs and curricular materials relating to math, science, and foreign languages. Few districts could refuse the much-needed money, increasing federal aid for and therefore intervention into education (Spring, 2011).

Title I and federal aid. Separate federal or "categorical" aid began with the passage of the 1965 Elementary and Secondary Education Act (ESEA) as part of Presidents John F. Kennedy and Lyndon B. Johnson's War on Poverty (Jennings, 2012; Spring, 2005, 2011). Title I of ESEA specified that federal funds would be made available to schools servicing children from low-income families as long as they met the national policy objectives (Mills, 2008; Spring, 2005, 2011). In addition to federal involvement in ongoing operations of local education for the first time (Kress et al., 2011), ESEA resulted in the first two formal national educational testing programs. Planning for the National Assessment of Educational Progress (NAEP) began in 1964 with a grant from the Carnegie Corporation, and the first national assessments were implemented in 1969 (National Center for Education Statistics [NCES], 2012). Some viewed the creation of NAEP and enactment of the Title I legislation as the "precursor to today's widespread use of tests as tools for holding educators accountable for student performance" (Hamilton, 2003, p. 27). The Title I Evaluation and Reporting System (TIERS) also linked the ESEA to commercial standardized tests for accountability purposes (Linn, 2000).

As the *Brown v. Board of Education* ruling of 1954 and Civil Rights Act of 1964 prompted wide-scale desegregation of schools across the country, testing for accountability and tracking continued to increase. New York was the first state to implement state-developed testing in 1965 (Ravitch, 2010). Florida began its state-developed testing program in 1971 (Herrington & MacDonald, 2000).

Equity-based policy continued into the 1970s with the passage of the Individuals with Disabilities Education Act (IDEA) of 1975; the Educational Amendments of 1972, in particular Title IX, which forbids gender discrimination in education or extracurricular programs; and the *Lau* remedies of 1974, which guaranteed children an opportunity to a meaningful education regardless of their language background (Jennings, 2012). As more students began to gain access to education and federal and state fiscal support increased, concerns about a lack of educational quality began to rise (Goertz, 2007). In September 1975, the New York Times published a story stating that SAT scores had fallen for over a decade, prompting many policymakers to revisit educational policy (Ravitch, 2010).

Test-based accountability. Around this time, test-based accountability policies became much more popular among state departments of education and were targeted to hold educators accountable for the operation and performance of their schools (Goertz, 2007). Tests designed to assess whether students had met minimum standards, "minimum competency tests," became widespread in the 1970s as well (Hamilton, 2003; Kress et al., 2011); the number of states requiring minimum competency tests increased from 2 states in 1973 to 34 states in 1983 (Linn, 2000). Teachers felt pressure to help their students achieve a passing score, and some purposely focused preparation on the specific test competencies (Goertz, 2007).

The U.S. federal government expanded its role through equity-based policies and categorical aid throughout the 1960s and 1970s. However, in 1979, the educational role of the federal government increased dramatically with the creation of the U.S. Department of Education and corresponding cabinet position of Secretary of Education by President Jimmy Carter (Spring, 2011). Although minimum competency tests were becoming ubiquitous and purportedly raising the number of students mastering basic skills, some argued a lack of emphasis on higher-order skills (e.g., critical thinking) within the school curriculum and pointed to international comparisons as a reason to reform yet again (Goertz, 2007).

Upon request of the newly formed U.S. Department of Education, *A Nation at Risk* was published in 1983 (National Commission on Excellence in Education [NCEE], 1983). As a matter of "national security," *A Nation at Risk* called for a change in national educational policy focus from basic competency to a commitment to excellence (Goertz, 2007; Kress et al., 2011; NCEE, 1983). The report cited numerous dismal national and international standardized test score comparisons and related facts, all of which were intended to shock the country out of mediocrity. Among the recommendations for developing what was called a "Learning Society," the report called for several input-related reforms: a set of common standards, more rigorous course offerings and progression, increased expectations of homework and effort, an expanded school day and school year, and improved quality of teacher preparation programs (NCEE, 1983). Because the report's alarming message used standardized test scores to point out the nation's problems, and because policymakers had grown fond of test-based policies, *A Nation at Risk* led to a nationwide increase in testing and test-based accountability policies that attached school-level incentives to test scores (Hamilton, 2003).

Standards-based reform. Shortly after *A Nation at Risk* was released, Secretary of Education T. H. Bell began publishing comparative state data. More states began creating their own tests and accountability policies in attempts to out-shine neighboring states, prompting physician John Cannell to point to what came to be known as the "Lake Wobegon" effect, wherein a majority of students in any given state were labeled "above average" (Ravitch, 2010). Beginning in the late 1980s, focus turned to the quality and rigor of standards (Goertz, 2007) as educators became more aware of the results-oriented approach adopted by the states (Kress et al., 2011). In 1989, the National Council of Teachers of Mathematics (NCTM) published a national set of standards, which detailed what all mathematics students should know and be able to do at different levels of schooling (Goertz, 2007). Later that year, President George H. W. Bush invited governors and other administrators to a national summit on education in Charlottesville, Virginia; the summit attendees established the National Education Goals Panel to address the problems with input-driven reform and the philosophical shift towards standards-based, outcomes-oriented reform (Ravitch, 2010).

Two years after the Charlottesville summit, President Bush recommended voluntary national standards and tests but was unsuccessful at convincing Congress to adopt them (Jennings, 2012; Ravitch, 2010; Spring, 2005). By that time, however, educators and other concerned citizens were pushing for more rigorous standards. In 1991, the U.S. Department of Education began awarding grants to consortia of professional groups of educators and academics to develop voluntary national standards in several subjects (Ravitch, 2010). Many of these consortia were well underway with writing efforts when, in the fall of 1994, Lynne Cheney criticized the national history standards for being biased (Ravitch, 2010). Ravitch claimed that, "the [national] standards movement died in 1995, when the controversy over the national history standards came to a high boil" (2010, p. 20).

Improving America's Schools Act. That same year, President Bill Clinton signed a reauthorization of the ESEA, entitled the Improving America's Schools Act (IASA). The IASA required states to develop their own standards, assessments aligned to these standards, and a method to identify low-performing schools (Kress et al., 2011). President Clinton also signed the Goals 2000 Educate America Act of 1994, which granted federal money to states so that they could write their own standards (Goertz, 2007; Ravitch, 2010). Even though all states complied with IASA, a large variance existed in state definitions of "success" and consequences for not achieving "success" (Goertz, 2007). For example, in 1999-2000, some states, such as Texas, North Carolina, and Florida, tied particularly strict repercussions to school accountability measures, while states such as Iowa, Colorado, Maine, and Montana had no school accountability measures at all (Carnoy & Loeb, 2002). Testing policy varied hugely from state to state, as well. Even though all states had tests that were aligned to their state standards, the grades tested, length, and types of tests were very different (Carnoy & Loeb, 2002).

Defining accountability. Part of the explanation for the large variance in accountability policies was that the word "accountability" was not very well defined. Linn (2000) described it this way: "accountability programs took a variety of forms, but shared the common characteristic that they increased real or perceived stakes of results for teachers and educational administrators" (p. 7). Overall, two main types of accountability systems exist (see Figure 2).

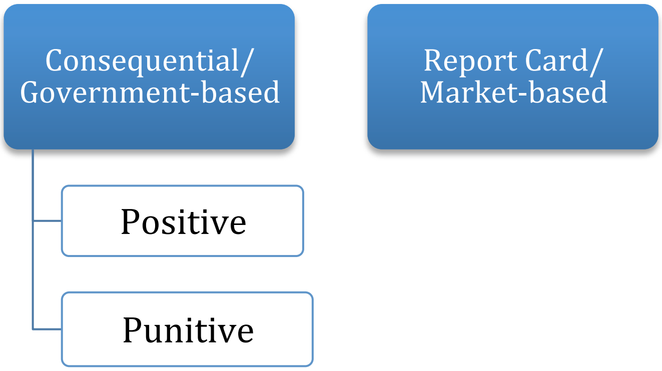


Figure 2. Types of accountability policies.

Although both types of policies in Figure 2 may have report card components, what the governing bodies do after the grades are reported determines the type of accountability policy. One type of accountability system provides explicit consequences to performance, called "consequential" by Hanushek and Raymond (2005, p. 306) or "government-based accountability" by Harris and Herrington (2006, p. 217). This type of consequential or government-based accountability is further split into what Diane Ravitch called "positive accountability," where states are mainly focused on helping schools, and "punitive accountability," where states are focused mainly on reconstitution or closing of schools (2010, p. 163). The other main type of accountability system only reports test results publicly, and allows the public to make the decision of what to do next. This approach has been dubbed "report card" accountability by Hanushek and Raymond (2005, p. 306) and "market-based accountability" by Harris and Herrington (2006, p. 221).

No Child Left Behind. In response to the wide variation in state policies, again a President (this time, Clinton in 1997), proposed voluntary national standards and national testing, and again the members of Congress refused to authorize it (Ravitch, 2010). However, these efforts would be revisited in just a few years with the inauguration of President George W. Bush. Shortly after taking office in 2001, President Bush signed into law the No Child Left Behind Act (NCLB), designed in part to address the variability in state accountability and testing policies (Goertz, 2007; Kress et al., 2011). NCLB required that states set more ambitious goals and more rigorous standards, increase the quality and quantity of testing, establish more serious consequences for poor test performance, and report test results by subgroups to expose and eliminate achievement gaps (Mills, 2008; Spring, 2011). Although the improvement goals were the same for each state (i.e., that all students would be "proficient" by 2014), NCLB allowed each state to once again develop its own standards and tests and to create its own definition for determining "proficiency" (Goertz, 2007). Spring (2011) described NCLB as a large step toward "a nationalized school system with state and local school authorities becoming conduits for federal policies" (p. 65).

The number of states with some version of a consequential accountability model increased from 12 states in 1996 to 39 states in 2000; by 2002, all states had a consequential accountability model in order to comply with NCLB (Hanushek & Raymond, 2005). Although the various states had different policies, they all had to participate in NAEP assessment. The effects of accountability and high-stakes testing policies, including NCLB, are discussed in the next section.

Effects of High-Stakes Testing Accountability Policies

NCLB's accountability plan placed a premium on "scientifically based research." As such, researchers were required to measure the intended effects of education on student learning and compare the results to the policy's stated goals, such as (a) increasing student test scores or (b) reducing the achievement gap. In addition to these things, researchers have also focused on some of the unintended consequences of NCLB and similar test-based accountability policies, including changes in (c) instructional practice, (d) data interpretation, (e) the role of testing, (f) teacher perspectives, (g) school and classroom climate, (h) motivation, (i) instructional leadership of principals, and (j) centralization.

Increasing student test scores. A clear measure of accountability systems is "whether they have a positive effect on test scores" (Carnoy & Loeb, 2002, p. 308). The academic achievement effects of accountability policies such as NCLB have been an area of great debate among educational researchers. NCLB was predicated in part on Grissmer and Flanagan's (1998) study, which highlighted dramatic NAEP gains in North Carolina and Texas following test-based accountability policies (Hamilton, 2003; Lee, 2008). However, soon after NCLB was enacted, Amrein and Berliner (2002) rebutted Grissmer and Flanagan's claims, stating that the gains in North Carolina and Texas were attributable to the increasing exclusion rates from testing, where school administrators suspended, retained, expelled, or reclassified selected students prior to the test.

Some researchers reported positive effects, along the lines of Grissmer and Flanagan (e.g., Bishop et al., 2001; Carnoy & Loeb, 2002; Hanushek & Raymond, 2005; Harris & Herrington, 2006; Stullich et al., 2006) whereas others (e.g., Lee, 2008; Lee & Wong, 2004; Nichols, Glass, & Berliner, 2006; Wei, 2012) found negative or no effects of accountability policies. Some studies reported mixed or diminishing effects (e.g., Chudowsky, Chudowsky, & Kober, 2007; Dee & Jacob, 2011; Fuller, Gesicki, Kang, & Wright, 2007; Jacob, 2005; Lee & Reeves, 2012; Nichols, Glass, & Berliner, 2012). Most studies with mixed results showed an improvement in mathematics NAEP scores—particularly in 4th grade—but not in reading. In a meta-analysis about test-driven external accountability policy, Lee (2008) noted that the average Cohen's *d* effect size for studies on academic achievement in the late 1990s (*M* = .47) was significantly larger than the average effect size from either the 1980s (*M* = .08) or the early 1990s (*M* = −.13), prior to the full implementation of IASA and NCLB. This may mean that overall, test-based accountability does have a positive effect, however slight and varied, at least in terms of scores on large-scale achievement tests.

Methodologies for these studies have varied widely, but for the most part, researchers have used NAEP as an external evaluative measure, although some older studies used state test scores. In their literature review, Nichols, Glass, and Berliner (2012) sorted the extant literature into three groups, based on methodology: (a) comparisons of achievement in states with a longer history of accountability policy to those with a shorter history (e.g., Amrein & Berliner, 2002; Dee & Jacob, 2009, 2011; Lee, 2008), (b) correlation or regression techniques to determine the relationship between some ranking based on accountability stringency and achievement (e.g., Carnoy & Loeb, 2002; Hanushek & Raymond, 2005), and (c) focused study on one particular aspect of policy and its impact on specific areas of the country (e.g., Neal & Schanzenbach, 2010). Figlio and Ladd (2008) pointed to three studies that they considered the most methodologically sound (i.e., Carnoy & Loeb, 2002; Jacob, 2005; Hanushek & Raymond, 2005). However, Hamilton (2003) noted that all research on effects of large-scale assessment has major limitations, including difficulty of obtaining permission to use data, difficulty of devising a true causal experimental study, diversity of state policy and programs, and poor measurement of the "construct of interest" resulting from non-representative sampling or non-random refusal (p. 32).

Reducing the achievement gap. After the passage of NCLB, educational agencies could "no longer ignore students with learning disabilities, those who have limited English proficiency, racial and ethnic minorities, and those who come from low-income families" (Wong, 2013, p. 411). However, more evidence exists to support no change, or even an increase, in the achievement gap since NCLB. Harris and Herrington (2006) referred to an increasing achievement gap in NAEP scores, "reversing decades of steady improvement in outcome equity" (p. 209). Others (Hanushek & Raymond, 2005; Watanabe, 2008) also documented an increase in the Black-White achievement gap via re-segregation and superficial teaching. Lee found no statistically significant effect on the racial achievement gap in his 2004 study with Wong or his 2008 meta-analysis on test-driven external accountability policy, but later reported with Reeves (2012) a narrowing achievement gap "associated with long-term statewide instructional capacity and teacher resources rather than short-term NCLB implementation" (p. 209).

Some evidence exists to support the claim that accountability policies have not changed or have jeopardized minority student achievement. Wei (2012) concluded in her study, "NCLB state accountability policy has not created equal outcomes across different academic subjects and racial groups" (p. 297). Some studies (e.g., Diamond & Spillane, 2004; Watanabe, 2008) have documented higher proportions of Black students in schools or courses focused solely on superficial test preparation.

Instructional practice. Test-based accountability policy has had consequences that were outside of the scope of the original stated intents. Written by a joint committee including the American Educational Research Association (AERA), American Psychological Association (APA), & the National Council on Measurement in Education (NCME), the *Standards for Educational and Psychological Testing* specifically indicate that consequences should be studied as part of any validity investigation (Joint Committee on Standards for Educational and Psychological Testing [JCSEPT], 1999). Further, validity considerations become critically important in high-stakes environments: "An examination of consequences is especially important for testing programs that are intended to serve as policy tools" (Hamilton, 2003, p. 26).

With increased consequences, testing can become a "central preoccupation in the schools" and "not just a measure but an end in itself" (Ravitch, 2010, pp. 12-13). Tests may "seem to exert a more powerful influence than standards" (Hamilton, 2003, p. 36). Educators and students feel pressure under a consequential accountability system to perform by any means necessary, perhaps to the exclusion of other important goals of the organization (Ravitch, 2010). For many teachers, this translates to a combination of narrowing the curriculum to only what is assessed on high-stakes tests and coaching students about how to answer quickly and accurately, given the testing method (Darling-Hammond, 2004; Goertz, 2007; Hamilton, 2003; Kim, 2010; Linn, 2000; Ravitch, 2010). However, these practices can contribute to unfair consequences based on inflated scores and corresponding invalid inferences.

Superficial test preparation and familiarity of test format result in a predictable pattern of test scores: a somewhat low minimum score after the first year of implementation, followed by a spike in each of the next several years as students and educators learn the specifics of the test (Carnoy & Loeb, 2002; Jacob, 2005). These negative consequences of testing are not new; a narrowing of the curricular focus was documented during the minimum competency era of the 1970s, when consequential accountability pertaining to graduation became widespread, and has since followed with other high-stakes testing programs (Hamilton, 2003; Linn, 2000). However, Hanushek and Raymond (2005) contended that reports of negative impacts such as narrowing of the curriculum are likely to be overstated.

Similar to narrowing of the curriculum or teaching to the test, another phenomenon arising out of making test scores the "primary measure of school quality" (Ravitch, 2010, p. 15) is the tendency to focus on *bubble students*, or students who have scored close to but just below the proficiency cut score (Booher-Jennings, 2005; Jacob, 2005). Neal and Schanzenbach, among the first to document this "educational triage" on a large scale over time, found that as proficiency requirements became more stringent, they observed "noteworthy increases in reading and math scores among students in the middle of the achievement distribution but not among the least academically advantaged students" (2010, p. 263). Thus, within high-stakes testing environments, both students who are struggling most with the content and those who are most capable and need to be challenged may actually receive less attention from their teacher (Moon, Brighton, Jarvis, & Hall, 2007).

Data interpretation. Data are the focus of accountability systems, and "accountability occurs only when a useful set of processes exists for interpreting and acting on information in educationally productive ways" (Darling-Hammond, 2010, p. 1081). One of the most common criticisms about the implementation of high-stakes testing policies originates from the field of psychometrics. Tests are limited to the extent that they provide "a *sample* of examinee behavior under certain, very specific conditions" (Hamilton, 2003, p. 26). Professional research communities are clear that a single test score alone should not be used to make decisions about promotion or retention (National Research Council [NRC], 1999), or any other high-stakes decision (JCSEPT, 1999; APA, 2013). Many other types of measurements can be used for these purposes; for example, Ravitch (2010) suggested teacher-derived performance measures such as grades, participation, and homework.

Although NCLB allowed for multiple assessments, problems with technical quality of informal assessments, coupled with curricular alignment requirements, prohibited most states from finding cost-effective methods of production (Lee, 2008). However, the unseen cost of basing important educational judgments on limited evidence is a loss of validity. Worse, the cost-effective practice of administering the same testing instruments repeatedly can “distort instruction and lead to inflated and non-generalizable estimates of student gains in achievement” (Linn, 2000, p. 6). In other words, repeated use of the same test encourages superficial recall-based instructional methods, score inflation, and “teaching to the test.” This practice hampers the ability of the test results to show accurate (i.e., valid) trends in student learning over time.

Score inflation, a serious threat to high-stakes accountability assumptions, occurs when test scores over-estimate what students actually know or can do. Score inflation is similar to monetary inflation, where a higher price is associated with the same amount of goods. Those not aware of score inflation and its effects on validity may be tempted to generalize a test score to a broader domain, beyond the scope of the test. However, how much value would they place on a score if they knew it carried less meaning? Hamilton (2003) credits John Cannell as the first to bring nationwide attention to the problem of score inflation when he discovered that most districts and states were reporting higher-than-national average scores. It is statistically impossible for a majority of scores in any group to be above the average of that group, so how could this happen? Linn (2000) offered some possible explanations:

> the use of old norms, the repeated use of the same test from year after year, the exclusion of students from participation in accountability testing programs at a higher rate than they are excluded from norming studies, and the narrow focusing of instruction on the skills and question types used on the test. (p. 7)

To reduce the likelihood of score inflation, Koretz and Béguin (2010) made the recommendation to minimize repeated use of tests and test items, expand the variety of types of tasks used on tests, and reduce the stakes for the tests by allowing for other types of accountability information sources. Another way to audit tests for score inflation is to calculate correlations between and mean trends among the high-stakes test in question and another, lower stakes test (Hamilton, 2003).

Another issue is the limited information evaluators may derive from multiple-choice responses. Indeed, educators may feel erroneously that "selecting a correct answer among limited options on a timed test is regarded as the only valid way of demonstrating knowledge" (Kim, 2010, p. 17). By contrast, performance-based assessments that involve more open-ended questions and require extended responses are favored in the literature (e.g., Kim, 2010; Perie et al., 2009) as a more accurate way to measure the breadth of student learning. Nevertheless, the high cost of developing such tests prohibits their use (Pilotin, 2013). This limitation has led some states to drop highly complex material from assessments, and therefore curriculum, because it cannot be tested using multiple-choice questions (Pilotin, 2013).

Finally, states have different definitions of "proficiency" as required by NCLB. The tests are different for each state, and the process of setting cut scores for different performance levels is different for each state. In practice, the performance setting process is "inherently judgmental," but that fact "is rarely communicated to the public" (Hamilton, 2003, p. 46). As a result, the illusion of objectivity remains, and high-stakes decisions continue to be made based on limited or missing information.

Policymakers' ongoing penchant for overgeneralizing results of one test is perhaps the result of ignorance about basic psychometric principles, paired with pressure from the public to produce affordable accountability systems. As Ravitch put it, "Elected officials assumed the tests were good enough to do what they were supposed to do—measure student performance—and that a test is a test; they did not give much thought to such technical issues as validity or reliability. Everyone, it seemed, wanted 'accountability'" (2010, p. 95).

School and classroom climate. Several negative effects of high-stakes accountability have been prevalently noted in the research on school and classroom climate. For example, Hamilton (2003) noted a series of studies that indicated declining teacher and student morale and increasing stress because of high-stakes testing. Some researchers have argued that test-based accountability has created such unreachable goals that students are dropping out of school at the highest rates ever (McNeil, Coppola, Radigan, & Vasquez Heilig, 2008). Unrealistic goals set by external entities can also contribute to score inflation as students and educators "take shortcuts when they believe goals are unattainable" (Hamilton, 2003, p. 46). Cheating by having a teacher or administrator change student answers after testing is an extreme, but present, form of shortcut taken by desperate educators (Jacob & Levitt, 2003). In Atlanta, 35 educators from 58 schools and the district office, including administrators, teachers, a school secretary, and former superintendent, Beverly Hall, were recently indicted for racketeering, conspiracy, and making false statements (USA Today, 2013). Another popular discussion involves diminishing instructional time dedicated to tested subjects. Many researchers have reported that while the amount of instructional time dedicated to math and English/language arts has increased, time devoted to other subjects has diminished (e.g., Dee, Jacob, & Schwartz, 2013; McMurrer, 2008; Rentner et al., 2006), leading to the conclusion that what is assessed is what is taught.

However, not all studies include evidence of negative effects on schools and classrooms. In a recent study, Dee, Jacob, and Schwartz (2013) used financial district data and pooled cross sections of teacher and principal surveys to examine the effects of accountability reforms such as NCLB on education policies and practices in schools. They found that per-pupil spending increased after NCLB by almost $600, teacher compensation increased, as did the number of elementary teachers with advanced degrees. In the same study, Dee et al. (2013) reported evidence that NCLB led to improvements in teacher-reported absenteeism, tardiness, and apathy.

Centralization. One of the most traditional and indoctrinated principles in the U.S. educational system is the concept of local control (Pilotin, 2013). How, then, did the U.S. develop this increasingly standardized and centralized state of education? One explanation is that the emphasis on the competitive nature of test-based accountability increased around the world after reports such as *A Nation at Risk* highlighted international math and science tests and rankings, leading to the "global educational reform movement" (Sahlberg, 2008, p. 47). Another explanation is that education has also long been offered as *the* method for nations to regain economic strength in what Spring (2011) called the "human capital paradigm." This link between education and economic strength was emphasized in *A Nation at Risk*. Both the global competitiveness and economic strength arguments introduce external factors; however, the former encourages strict adherence to a centralized curriculum whereas the latter encourages notions of innovation and risk-taking (Sahlberg, 2008). Yet another explanation for increased centralization is that federal categorical aid for education, which began in the 1950s, has always been tied to requirements of local and state educational agencies (Spring, 2011).

Because test-based accountability policies such as NCLB have resulted in increased centralization, decisions about things such as what to teach and assess are increasingly made outside of the local arena (Spring, 2011), posing a threat to traditional instructional jurisdiction within local schools (Rutledge, 2010) and earning the title, the "New Talyorism" (Au, 2011, p. 25). However, some argue that today's increased global mobility, communication, and competition necessitate a movement toward a more centralized educational system that supports state experimentation and dissemination of best practices (e.g., Pilotin, 2013).

A key player in the centralization movement of test-based accountability policy has been the local school district. District personnel have gained authority to initiate strategies and programs to achieve the educational expectations of state and federal agencies. One of the first ways that districts became involved in this work was to align curriculum and instruction to state standards and tests, pupil progression plans, and other district policies (Goertz, 2007). Many districts have responded to the pressure of test-based accountability by providing instructional assistance to schools in the form of professional development and coaching, generating student data reports, and publishing curriculum guides and other documents to support aligned instruction in the schools (Chudowsky et al., 2007; Goertz, 2007).

Teacher perspectives. Teachers have voiced concerns about the quality of state tests (Abrams, Pedulla, & Madaus, 2003; Goertz, 2007; Lile, 2012; Pedulla, Abrams, Madaus, Russell, Ramos, & Miao, 2003). In 2012, Scholastic, with support from the Bill and Melinda Gates Foundation, conducted a survey of more than 10,000 teachers. Scholastic found that teachers "say standardized tests alone cannot provide a complete understanding of either student achievement or teacher performance. They are clear in their call for multiple, more frequent measures of teaching and learning" (2012, p. 25). Many teachers despise the use of a solitary test to measure student learning (Spring, 2011). Test-based accountability policies have made some teachers feel wedged between two seemingly disparate forces: the socio-moral purposes of teaching and efficiency-driven education based on achievement (Sahlberg, 2008). As pressure to perform increases, so do incentives for teachers to teach in schools where “students are easy to teach and school stability is high” (Darling-Hammond, 2004, p. 1058). As a result, a sizable proportion of students with the greatest needs are attending schools that, by their nature, disincentive teachers who choose to teach there (Kozol, 1991).

However, in light of recent changes to educational policy, teachers have reported that they are concentrating more on struggling students, attempting novel teaching methods to reach them, and raising their expectations for these students (Desimone, 2013). Further, in her study of educators in 32 schools across five states, Desimone (2013) reported that although the respondents felt stress and pressure, they also felt an increased personal responsibility for their students’ learning. Likewise, Goertz (2007) reported that teachers tend to align their instruction to the standards assessed on tests and use the data generated by these tests to “identify students who need additional help, topics requiring more emphasis, and gaps in curriculum and instruction” (p. 11).

Many educators are motivated by the fear of punitive actions due to low student performance against themselves and their school (Diamond & Spillane, 2004; Jennings, 2012; Rutledge, 2010). Teachers may respond to accountability policies by focusing more on achievement and working harder; however, if tests do not have clear, meaningful consequences attached to them, “teachers pay little attention” (Hamilton, 2003, p. 33). Testing “may also improve teachers’ motivation and morale if it is accompanied by efforts on the part of the school administration to provide appropriate learning opportunities” (Hamilton, 2003, p. 38).

Instructional leadership of principals. Principals have focused more on external goals based on standards and test scores as a result of test-based accountability policies (Diamond & Spillane, 2004; Rutledge, 2010). Many principals reported that they increased professional development opportunities at their schools, added extra sections after-school or during the summer for remediation, and led curriculum revision projects (Hamilton, 2003). On the other hand, principals have also reported that they sometimes reassign teachers to improve the instructional quality for students in tested grades and subjects, focused more on short-term goals than long-term instructional change, and used field trips and parties as incentives for academic achievement (Hamilton, 2003).

The role of testing. Following the enactment of NCLB, the role of testing became more prominent in accountability policy. Higher stakes were attached to test scores, including state sanctions and rewards, as well as public shame or glory. Even though public perception held that tests were problematic, some researchers were able to separate the test development process from the high-stakes that were now placed upon test scores (Clune, 1993; Goertz, 2007; Hamilton, 2003; Jennings, 2012; Linn, 2000; Ravitch, 2010; Spring, 2011). The sentiment expressed by Diane Ravitch (2010) is repeated throughout the literature:

> The anti-testing forces lashed out against the wrong target. Testing was not the problem. Tests can be designed and used well or badly. The problem was the misuse of testing for high-stakes purposes, the belief that tests could identify with certainty which students should be held back, which teachers and principals should be fired or rewarded, and which schools should be closed—and the idea that these changes would inevitably produce better education. Policy decisions that were

> momentous for students and educators came down from elected officials who did not understand the limitations of testing. (p. 150)

Test data are useful components of an accountability system to the extent that they are “relevant, valid, timely, and useful” (Darling-Hammond, 2004, p. 1080). However, test data should not represent the entirety of the system. Campbell’s Law is often cited by opponents of high-stakes accountability: "The more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor" (Campbell, 1976, p. 49). Current test-based accountability policies are seen by some researchers as too narrowly focused, failing to consider fundamental components such as curriculum development, local leadership, and community contexts (Darling-Hammond, 2004; Hamilton, 2003; Ravitch, 2010; Sahlberg, 2008).

Tests are also frequently used for more than one purpose, a practice advised against by Perie, Marion, and Gong (2009) and the American Psychological Association (2013), among others. The properties of a test that make it appropriate for providing instructional feedback will tend to make it unsuitable for accountability purposes, and vice versa (Hamilton, 2003). There is no “royal road” to an assessment system that effectively serves all functions (Black & Wiliam, 2005, p. 260). The current challenge for policy makers and educators is to find alternative accountability frameworks and comprehensive assessment systems that include varying types of assessments intended for improving classroom practice and student achievement, while also avoiding some of the negative effects of using any one single low-level test (Volante & Ben Jaafar, 2010).

Race to the Top. In response to a recent national recession, President Barack Obama signed into law the American Recovery and Reinvestment Act of 2009 (ARRA), which was designed to stimulate the economy by investing in key areas (U.S. Department of Education, 2009). ARRA included $4.35 billion for the "Race to the Top" Fund, a competitive federal grant program that encouraged states to apply for funding in two phases. To win part of the grant money, states had to provide evidence that they had been successful in raising student achievement in the past and had an innovative plan to accelerate educational reforms in the future. States were required to include plans for "adopting standards and assessments that prepare students to succeed in college and the workplace and to compete in the global economy" and "building data systems that measure student growth and success, and inform teachers and principals about how they can improve instruction" (U.S. Department of Education, 2009, p. 2). In more concrete terms, those states competing for grant money were required to become a member of at least one national consortium of a "substantial amount of states" and adopt a set of common standards and common assessments (U.S. Department of Education, 2009, (B)(1)(i)(b) and (B)(2)(i)(b)).

Improving testing policy. During the 1990s, a governor told Jennings, "tests would be the lever that would bring about broad school improvement. Raising the quality of the teaching force and more equitably distributing good teachers were seen as more difficult than requiring testing and making public the results" (2012, p. 6). However, as discussed previously, placing high stakes on a single test is not the ideal basis for accountability policy. So what can be done to improve testing policy? The first step is to learn from the effects of past policies, such as NCLB.

The U.S. is at the beginning of its next phase of accountability; therefore, most of the studies on NCLB can be considered a baseline for RTTT (Wong, 2013). There is hope for significant change in the next decade's transition to the next phase of accountability (Pilotin, 2013), and it is likely that new hybrid accountability structures will emerge (Wong, 2013). Although it may be impossible to design the perfect test-based accountability system, it is still worthwhile to pursue a system that maximizes intended benefits and minimizes negative unintended consequences (Feuer, 2010).

Volante and BenJaafar (2010) offered three ways to modify existing test-based accountability systems; states can either (a) include assessment items that elicit critical and higher-order thinking skills, or (b) incorporate classroom-based assessments, or (c) use classroom-based assessments alone. Certainly, utilizing teacher-given grades for accountability purposes is one way of implementing the last option. In a study comparing teacher-assigned grades and standardized scores from the National Education Longitudinal Study of 1988 (NELS), Willingham, Pollack, and Lewis (2002) reported a moderate correlation (R = .62) between course grades and standardized scores. Other studies have also reported a correlation between teacher-assigned grades and standardized test scores of approximately .5 to .6 (Bowers, 2010; Brennan, Kim, Wenz-Gross, & Siperstein, 2001; Linn, 2000; Woodruff & Ziomek, 2004).

Darling-Hammond (2004) noted that more of the accountability success stories came from areas that "have focused on broader notions of accountability, including investments in teacher knowledge and skill, organization of schools to support teacher and student learning, and systems of assessments that drive curriculum reform and teaching improvements" (p. 1047). For instance, in Finland, one of the top scoring countries on international tests such as the OECD Programme for International Student Assessment

(PISA) and IEA Trends in International Mathematics and Science Study (TIMSS), teachers go through an intense apprenticeship and ongoing professional development for how to assess their students (Sahlberg, 2011). Finnish teachers are given the ultimate responsibility for assessing their own students—and that evaluation is trusted above all else (Sahlberg, 2011). Although this is an extreme example for comparison with the current state of the United States' educational system, it serves as one possibility for how the U.S. might transition in the future.

Recent Test-Based Accountability in Florida

The state of Florida placed fourth in phase 2 of the Race to the Top competition in August 2010 and received $700 million to implement educational reforms. As Florida is one of the largest and most diverse states in the country, its efforts in reshaping its public school system are likely to have implications for similar reforms in other states. According to the Florida Department of Education (FDOE), a "key component" of Florida's successful bid for the Race to the Top funds focused on content standards and the creation of balanced assessment approaches (FDOE, 2013c). Subsequently, with $20 million of RTTT funds and through partnering with districts, the State has commissioned tests for what are known as "hard-to-measure" subjects such as Arts or World Languages (Fla. Stat. Ann. § 1008.25 (6), 2013). The bid-winning proposal from Florida pointed to its long-standing history of test-based accountability policies and willingness to participate in the Differentiated Accountability pilot, which included a tiered support system for schools and districts that did not meet achievement goals set by the state (FDOE, 2013c).

The state standardized testing program for accountability grew out of the Educational Accountability Act (Section 229.57, F.S.) of 1971, which mandated a statewide testing program. During the period between 1974 and 2011, the Florida state accountability system experienced substantive revision in scope and depth. In 1976, legislators modified and expanded the Educational Accountability Act to include reading, writing, and mathematics for students in grades 3, 5, 8, and 11, as well as a graduation exam for the graduating class of 1979 (Herrington & MacDonald, 2000). Although control shifted back to local districts and schools with Blueprint 2000 enacted in 1991 for implementing programs and setting standards, schools in Florida operated under stricter accountability and testing provisions designed to identify and reward high performing schools while prescribing interventions for lower performing schools (Herrington & MacDonald, 2000). Following the adoption of the Sunshine State Standards in 1996, FDOE administered a new series of assessments entitled the Florida Comprehensive Assessment Test (FCAT). Florida state legislators took control once again in 1999 when they enacted the Accountability Plus (A+) plan and the state assessment program grew to include science as well as grades 3-10 for reading and mathematics (Herrington & MacDonald, 2000).

Florida's Department of Education began revising the Florida Sunshine State Standards once again in 2007, and named the finished revision the Next Generation Sunshine State Standards to reflect the change (FDOE, 2007). Shortly thereafter, FCAT became FCAT 2.0 as a new assessment was required for the new Next Generation standards. The newly aligned Mathematics and Reading FCAT 2.0 assessments were first administered in the 2010-11 school year. That same school year, however, Florida revised the standards yet again following the decision to adopt 85% or more of the Common Core Standards. This was done to comply with the application requirements for the national Race to the Top grant. The next iteration of standards was dubbed the "2010 Next Generation Sunshine State Standards," and remains in effect at the time of the present study. FDOE is currently deciding which assessment will replace FCAT 2.0 for the 2014-15 school year. In a recent media advisory, FDOE announced a three-day education accountability summit to discuss, among other things, the next statewide accountability assessment (FDOE, 2013b).

The results for the Grade 6 Mathematics FCAT 2.0 were reported in two forms for the 2012-2013 school year: (a) a developmental scale score (SS) from 170-284 that provides the ability to track student growth and progress over time and allows comparison from year to year to identify growth, and (b) a corresponding performance level of 1, being the lowest, to 5, the highest (FDOE, 2013e). A performance level of 3 or above is considered proficient in the state of Florida. For the purposes of the present study, the developmental scale score was used to retain accuracy and maximally model variation among student scores, instead of using the categorical performance level, which truncates differences measured by the developmental scores.

The FCAT 2.0 is not just a high-stakes test for individual students in Florida. Schools across Florida receive a grade based mostly on how certain groups of students perform on FCAT 2.0. Also, Florida law (Fla. Stat. Ann. § 1012.34, 2013) stipulates that 50% of a teacher's evaluation is based on observation, while the other 50% is based on the performance of the teacher's students, including up to 30% based on statewide assessment data for evaluation purposes and the other 20% can be student outcome data specific to the job responsibility. Value-Added Models (VAM) required to fulfill laws such as this are not unique to Florida; Tennessee pioneered the concept of assessing teachers based on student scores, and it has since been picking up favor and criticism around the country (Amrein-Beardsley, 2008; Hill, Kapitula, & Umland, 2010; National Council on Teacher Quality [NCTQ], 2012; Pullin, 2013; Ready, 2013).

District-Developed Interim Assessments

Districts and schools have supplemented state standardized tests with their own assessments to measure progress and give feedback since the mid-1990s (Goertz, 2007). These localized assessment programs were considered a "promising approach to helping teachers make better use of assessment information" (Hamilton, 2003, p. 49). However, after the enactment of NCLB in 2001, many school districts reacted to the increased pressure to raise test scores and close achievement gaps by developing or purchasing interim tests (Shepard, Davidson, & Bowman, 2011). These interim, or benchmark, tests were administered periodically throughout the school year and indicated progress toward statewide test scores. They are now fairly widespread; in a national survey, 82% of large urban school districts reported that they had instituted some form of interim assessment, and 69% of these districts had done so following the passage of NCLB (Burch, 2010).

Research on interim or benchmark assessments is not as comprehensive as research on other types of assessments, possibly because of the meteoric rise of benchmark assessments to popularity among school districts across the United States following NCLB (Shepard et al., 2011). Although interim assessments were originally intended as an "early warning system" for state accountability tests (Wiliam, 2004), some vendors and developers capitalized on the increasingly positive research about formative assessments and sold their interim assessments as "formative assessments" or "formative assessment systems" (Heritage, 2010; Li et al., 2010; Popham, 2008; Shepard et al., 2011). This contributed to confusion between the terms *formative assessment* and *interim assessment* (Chappius, 2005; Goertz, Olah, & Riggan, 2010; Herman, Osmundson, & Dietel, 2010; Perie et al., 2009).

Formative assessment versus interim assessment. In October 2006, after an extensive review of the formative assessment literature and consultation with internationally recognized assessment experts, a state collaborative sponsored by the Council of Chief State School Officers (CCSSO) attempted to resolve the confusion about the use of the term *formative* by issuing this definition (McManus, 2008):

> Formative assessment is a process used by teachers and students during instruction that provides feedback to adjust ongoing teaching and learning to improve students' achievement of intended instructional outcomes. (p. 3)

The process of formative assessment is not effective simply because an instrument is implemented or because periodic feedback is given to students. For example, in a meta-analysis of 607 effect sizes about feedback interventions (FIs), Kluger and DeNisi (1996) found that even though FIs improved performance on average (*d* = .41), over one-third of the FIs actually *decreased* performance. Kluger and DeNisi (1996) reported negative mean effect sizes when the FI was discouraging (*d* = -.14) and/or when the task was physical (*d* = -.11). In Sadler's (1989) early model of formative assessment, the fundamental purpose was to enable the student to come to an understanding of the teacher's concept of quality, ultimately leading to student self-monitoring. At a minimum, effective formative assessment requires high-quality instruments, task-related specific feedback to students, and student participation in the assessment process (Black & Wiliam, 1998b; Shute, 2008). Other essential elements include clear learning targets (Brookhart, 2011; Sadler, 1989; Wiliam & Thompson, 2008); learning progressions usable as student performance maps and planning tools (Heritage, 2008; McManus, Wilson, & Draney, 2004); instructionally meaningful, curriculum-embedded assessment tasks that reveal students' thinking processes (Shepard, 2006); and timely availability of results (Popham, 2006).

Black and Wiliam (1998a, 1998b) famously characterized the then implicit domain of formative assessment by synthesizing research on diverse fields, including teachers' assessment practices, feedback (Kluger & DeNisi, 1996; Sadler, 1989), self-assessment, selfperception, achievement motivation, and quality of tasks. From their review, Black and Wiliam (1998a, 1998b) concluded that formative assessment had the potential to increase student learning by .4 to .7 standard deviations—much greater than what typical educational interventions would produce, and that low achievers realized the largest gains. To give perspective, they noted that if a country were to gain .7 standard deviations, it would move in ranking from the middle to among the top five nations on any international test.

Several contradictory studies have recently been published regarding the efficacy of formative assessment. For example, Bennett (2011) noted that Black and Wiliam (1998a, 1998b) did not do a meta-analysis and only found 20 or so exemplary studies among those included in their synthesis. Kingston and Nash (2011) also investigated the relationship between formative assessment practices and student achievement, and concluded that the median effect size of this relationship is closer to .20 rather than the .40 - .70 range reported by Black and Wiliam (1998a, 1998b). However, McMillan, Venable, and Varier (2013) recently rebutted Kingston and Nash's (2011) meta-analysis, finding fault in the selection criteria and quality of the 13 studies that were included in their study. In a similar fashion, Dunn and Mulvenon (2009) took issue with the apparent lack of consensus about the definition of formative assessment and limited empirical evidence, yet were contradicted on both counts by Filsecker and Kerres (2012). Filsecker and Kerres (2012) provided a synopsis of 11 definitions of formative assessment and synthesized them into their own definition: "a series of informed and informing actions that change the current state of the reciprocal teaching-learning relationship toward a more knowledgeable one" (p. 4).

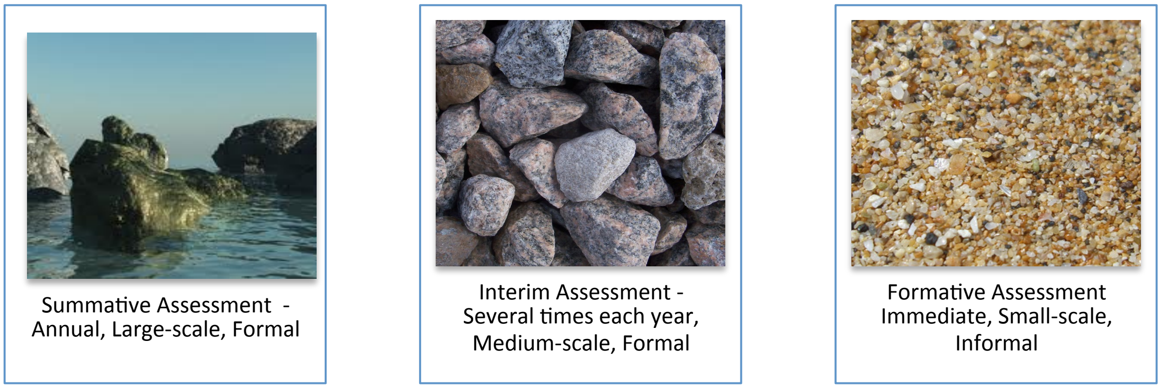


Figure 3. Types of educational assessments (Perie et al., 2009).

Although the promises of formative assessment are abundant, *formative* and *interim* assessments are not the same. Wiliam referred to interim assessment as not formative but "early-warning summative" (2004, p. 4), and Shepard (2005) clarified that the individual profile data from interim assessments are not directly formative because (a) the data available are at too gross a level of generality, and (b) feedback for improvement is not part of the process. Perie et al. (2009) further clarified the distinctions between these types of assessments (see Figure 3) and helped to establish *interim* and *benchmark* as more appropriate labels for longer-term, periodic tests.

While some evidence points to the effectiveness of formative assessment, studies about the effects of interim or benchmark assessment have been inconclusive (Goertz et al., 2010; Henderson, Petrosino, Guckenburg, & Hamilton, 2008; Shepard, 2010). Some researchers have provided evidence that interim assessment systems support desirable classroom and professional collaboration effects when adequate instructional leadership is provided at the school level (Bulkley, Christman, Goertz, & Lawrence, 2010; Bulkley, Olah, & Blanc et al., 2010; Crane, 2008; Downey, Steffy, Poston, & English, 2009; Goertz et al., 2010; Kerr, Marsh, Ikemoto, Darilek, & Barney, 2006; Shepard, 2010).

Purposes of interim assessment. Perie et al. (2009) offered a framework for considering how interim assessments might be used as part of a comprehensive assessment system and defined interim assessments, in part, on their middle-range time-scale location somewhere along a continuum between (a) once-per-year and (b) instantaneous and ongoing. Three main purposes for interim assessment were discussed in Perie et al.'s 2009 article: *instructional,* wherein the primary goal is to adapt instruction to meet student needs; *predictive*, which means that the assessments are designed to report a likelihood of achievement on some future summative test; and *evaluative*, mostly for external audiences outside of the school site potentially eliciting large, future changes system-wide rather than a local audience at a single school.

Predictive utility of interim assessment. Initially, many interim assessment programs were implemented to predict scores on high-stakes state assessments. Some studies have shown a strong relationship between these interim assessment scores and the corresponding summative high-stakes assessment scores. Brown and Coughlin (2007) reported correlations ranging from .7 to .8 between TerraNova tests and the Pennsylvania state assessment; Williams (2008) reported correlations of .6 between statewide Grade 4 interim tests and the Texas Assessment of Knowledge and Skills; Underwood (2010) reported correlations of .6 between Grade 10 district interim assessments and the FCAT; Kingston et al. (2011) reported correlations ranging from .6 to .8 between the Kansas Interim Assessment System and their statewide high-stakes test; and Chen (2011) reported correlations ranging from .6 to .8 between district interim assessments and FCAT for Grades 3-10 reading, Grades 3-8 math, Algebra, Grades 5, 8, and 11 Science.

However, it should be noted that in studies where multiple administrations of the same interim test are used as predictor variables in multiple regression and other similar correlational models, predictor variable collinearity is a high possibility (Chen, 2011; Linn, 2000). Collinearity exists when two predictor variables, such as interim benchmark scores, are highly correlated and are used together in a multiple regression; multicollinearity is the term used for many highly correlated predictors (Hinkle, Wiersma, & Jurs, 2003). If not controlled or explained, collinearity can become an issue when interpreting results because it may be difficult to determine which predictor is having the effect. There are several methods for addressing collinearity, including selecting the variables purposefully to prevent redundant information. The present study utilized averaging in the form of a grade point average to control for collinearity among teacher-assigned grades. Similarly, interim test scores were averaged by type instead of being included as separate variables for each individual interim test administration.

Summary

A review of historical events was provided to explain the complexity of educational reform and test-based accountability. For over a century, test-based accountability has been fundamental to educational reform movements. Current attempts to change educational testing policy are still grounded in the existing underlying assumptions about accountability, tempered by the history of assessment in the United States and the international push to compete. Consequently, lack of consistency and clear conceptualization of terms such as *formative* and *interim* assessment impedes policymakers and researchers interested in the phenomenon of assessment. Despite the inconsistent and fluid nature of the concept of assessment, it remains an important consideration within today's educational zeitgeist of competition and achievement-driven focus (see Figure 4).

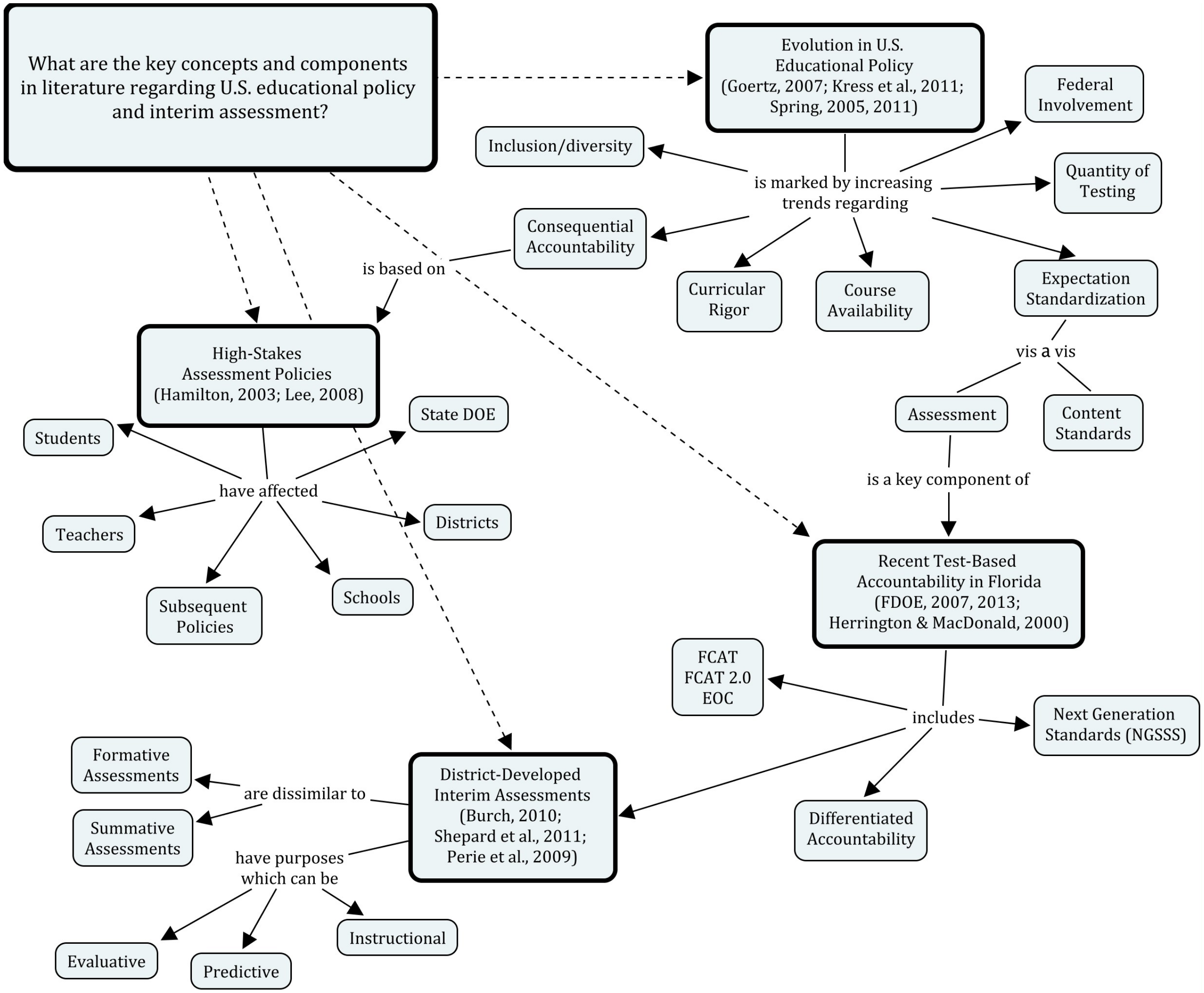

What are the key concepts and components in literature regarding U.S. educational policy and interim assessment?
Evolution in U.S. Educational Policy (Goertz, 2007; Kress et al., 2011; Spring, 2005, 2011)
is marked by increasing trends regarding
Federal Involvement
Inclusion/diversity
Quantity of Testing
Consequential Accountability
Curricular Rigor
Course Availability
Expectation Standardization
vis a vis
Assessment
Content Standards
is a key component of
is based on
High-Stakes Assessment Policies (Hamilton, 2003; Lee, 2008)
have affected
Students
State DOE
Teachers
Districts
Subsequent Policies
Schools
Recent Test-Based Accountability in Florida (FDOE, 2007, 2013; Herrington &amp; MacDonald, 2000)
includes
FCAT
FCAT 2.0
EOC
Next Generation Standards (NGSSS)
Differentiated Accountability
District-Developed Interim Assessments (Burch, 2010; Shepard et al., 2011; Perie et al., 2009)
are dissimilar to
Formative Assessments
Summative Assessments
have purposes which can be
Evaluative
Predictive
Instructional

Figure 4. Concept map including key components in the literature review regarding U.S. educational policy and interim assessment.

In conclusion, the present study sought to capture whether and what types of interim assessment work best. Interim assessment is worth exploring because interim assessment policy is changing pedagogical and managerial practices. Further, interim assessment is uniquely situated between low-stakes classroom formative assessment and high-stakes external summative assessment. As long as test-based accountability policy is in place, studies that can further knowledge in the area are warranted. Chapter 3 provides a description of the design, procedures, data collection, and analyses used to answer the two research questions posited for the present study.

CHAPTER 3

Methods & Procedures

Test-based accountability and the increased pressure from high-stakes national and state policies have caused school districts to struggle to find ways to assess students prior to the state summative test (Blanc et al., 2010; Simpson, LaCava, & Graner, 2004). As school districts attempt to find the most appropriate type of interim assessment within the traditional confines of money and time, it is vital to study what is currently being done and whether it is working or not, according to the original stated purpose (Brown & Coughlin, 2007; Perie et al., 2009). This chapter presents the methods and procedures used for the present study on two of these types of interim assessments given over the course of the same school year.

Research Design

The present retrospective, predictive study employed a non-experimental, correlational research design. The research design was constructed to answer questions relating to utility of interim assessment programs ex post facto using an archived data set. Two analytic procedures were employed to address the research questions: multiple linear regression to address predictive utility, and analysis of covariance (ANCOVA) to address instructional utility. Prior year test performance was controlled statistically using regression blocks within the multiple regression procedure, and a covariate within ANCOVA.

Research Questions

The main purposes for the two different types of interim assessment programs examined in the present study are predictive (i.e., the Interim Benchmark Assessment [IBA]) and instructional (i.e., the Learning Schedule Assessment [LSA]). Neither of these stated purposes are high-stakes; however, some educators have used the results from the interim assessments to adjust grouping within classrooms, alter instructional methods, or administer educational interventions such as tutoring or one-on-one help during class. The research questions that guided the present study were intended to evaluate the alignment of the district interim assessment program to the intended purpose of the assessment programs, while comparing other factors for prediction, and accounting for previous student performance:

1. To what extent can variance in middle school student scores on mathematics high-stakes state tests be explained by scores on district interim assessments after controlling for prior scores, student demographic variables, and teacher-assigned grades?
2. To what extent can variance in middle school student scores within achievement levels on mathematics high-stakes state tests be explained by scores on district interim assessments after controlling for prior scores, student demographic variables, and teacher-assigned grades?
3. To what degree does achievement, as measured by mathematics high-stakes state tests, of middle school students who have experienced less frequently administered, predictive interim assessments differ from the achievement of students who have experienced more frequently administered, instructional assessments, after controlling for prior scores?

Hypotheses

The purpose of the first two research questions was to determine, after accounting for variables that are known to predict future performance, how much variance in high-stakes state test scores can be explained by district interim assessment scores. The first research question addressed the whole sample, while the second research question investigated variance by FCAT achievement level.

The initial null hypothesis for research question 1 was:

H_{01}: The effect size, R^2, for a model containing demographic variables as predictors of high-stakes mathematics scores will be zero

or, symbolically,

H_{01}: $R^2_{\text{BLACK, HISP, WHITE, MULTI, AMER\_IN, GENDER, ELL, ESE, FREE\_RED*FCAT\_SS\_2013}} = 0.$

The corresponding alternate hypothesis was:

H_{a1}: The effect size, R^2, for a model containing demographic variables as predictors of high-stakes mathematics scores will be greater than zero

or, symbolically,

H_{a1}: $R^2_{\text{BLACK, HISP, WHITE, MULTI, AMER\_IN, GENDER, ELL, ESE, FREE\_RED*FCAT\_SS\_2013}} > 0.$

Similarly, beginning with Block 2, the next hypothesis is that there will be no change in effect size from the prior block:

H_{02}: The difference in effect size, ΔR^2, between a model containing school grade and demographic variables and a model containing only demographic variables as predictors of high-stakes mathematics scores will be zero

or, symbolically,

H_{02}: $R^2_{\text{BLACK, HISP, WHITE, MULTI, AMER\_IN, GENDER, ELL, ESE, FREE\_RED, SCHOOL\_GRADE*FCAT\_SS\_2013}} - R^2_{\text{BLACK, HISP, WHITE, MULTI, AMER\_IN, GENDER, ELL, ESE, FREE\_RED*FCAT\_SS\_2013}} = 0$.

The corresponding alternate hypothesis is:

H_{a2}: The difference in the effect size, ΔR^2, between a model containing school grade and demographic variables and a model containing only demographic variables as predictors of high-stakes mathematics scores will be greater than zero

or, symbolically,

H_{a2}: $R^2_{\text{BLACK, HISP, WHITE, MULTI, AMER\_IN, GENDER, ELL, ESE, FREE\_RED, SCHOOL\_GRADE*FCAT\_SS\_2013}} - R^2_{\text{BLACK, HISP, WHITE, MULTI, AMER\_IN, GENDER, ELL, ESE, FREE\_RED*FCAT\_SS\_2013}} > 0$.

The rest of the hypotheses follow the same pattern as H_{02} and H_{a2}:

H_{03}: $R^2_{\text{BLACK, HISP, WHITE, MULTI, AMER\_IN, GENDER, ELL, ESE, FREE\_RED, SCHOOL\_GRADE, GPA*FCAT\_SS\_2013}} - R^2_{\text{BLACK, HISP, WHITE, MULTI, AMER\_IN, GENDER, ELL, ESE, FREE\_RED, SCHOOL\_GRADE*FCAT\_SS\_2013}} = 0$.

H_{a3}: $R^2_{\text{BLACK, HISP, WHITE, MULTI, AMER\_IN, GENDER, ELL, ESE, FREE\_RED, SCHOOL\_GRADE, GPA*FCAT\_SS\_2013}} - R^2_{\text{BLACK, HISP, WHITE, MULTI, AMER\_IN, GENDER, ELL, ESE, FREE\_RED, SCHOOL\_GRADE*FCAT\_SS\_2013}} > 0$.

H_{04}: $R^2_{\text{BLACK, HISP, WHITE, MULTI, AMER\_IN, GENDER, ELL, ESE, FREE\_RED, SCHOOL\_GRADE, GPA, FCAT\_SS\_12*FCAT\_SS\_2013}} - R^2_{\text{BLACK, HISP, WHITE, MULTI, AMER\_IN, GENDER, ELL, ESE, FREE\_RED, SCHOOL\_GRADE, GPA*FCAT\_SS\_2013}} = 0.$

H_{a4}: $R^2_{\text{BLACK, HISP, WHITE, MULTI, AMER\_IN, GENDER, ELL, ESE, FREE\_RED, SCHOOL\_GRADE, GPA, FCAT\_SS\_12*FCAT\_SS\_2013}} - R^2_{\text{BLACK, HISP, WHITE, MULTI, AMER\_IN, GENDER, ELL, ESE, FREE\_RED, SCHOOL\_GRADE, GPA*FCAT\_SS\_2013}} > 0.$

H_{05}: $R^2_{\text{BLACK, HISP, WHITE, MULTI, AMER\_IN, GENDER, ELL, ESE, FREE\_RED, SCHOOL\_GRADE, GPA, FCAT\_SS\_12, AVG\_IBA, AVG\_LSA*FCAT\_SS\_2013}} - R^2_{\text{BLACK, HISP, WHITE, MULTI, AMER\_IN, GENDER, ELL, ESE, FREE\_RED, SCHOOL\_GRADE, GPA, FCAT\_SS\_2012*FCAT\_SS\_2013}} = 0.$

H_{a5}: $R^2_{\text{BLACK, HISP, WHITE, MULTI, AMER\_IN, GENDER, ELL, ESE, FREE\_RED, SCHOOL\_GRADE, GPA, FCAT\_SS\_12, AVG\_IBA, AVG\_LSA*FCAT\_SS\_2013}} - R^2_{\text{BLACK, HISP, WHITE, MULTI, AMER\_IN, GENDER, ELL, ESE, FREE\_RED, SCHOOL\_GRADE, GPA, FCAT\_SS\_2012*FCAT\_SS\_2013}} > 0.$

The second research question has the same hypotheses as the first question, with a modified (restricted) domain. Using the 2012 Grade 5 Mathematics FCAT 2.0 achievement level as a sorting variable (FCAT_AL_2012), the sample was parsed into five groups corresponding to the five achievement levels, where Level 1 is the lowest and Level 5 is the highest. Students must make a Level 3 to pass the FCAT 2.0. The same multiple regression tests used to answer research question one were then run for each FCAT 2.0 achievement level. The achievement levels are calculated based on the developmental scale scores. In other words, FCAT 2.0 Level 1 had five sets of hypotheses identical to those of research question 1, FCAT 2.0 Level 2 had five sets, etc. The hypotheses of most interest are those of the final block and are listed below, by sample population group (achievement level).

2012 Grade 5 Mathematics FCAT 2.0 - Level 1 (FCAT_AL_2012 = 1)

H_{06}: $R^2_{\text{BLACK, HISP, WHITE, MULTI, AMER\_IN, GENDER, ELL, ESE, FREE\_RED, SCHOOL\_GRADE, GPA, AVG\_IBA, AVG\_LSA*FCAT\_SS\_2013}} - R^2_{\text{BLACK, HISP, WHITE, MULTI, AMER\_IN, GENDER, ELL, ESE, FREE\_RED, SCHOOL\_GRADE, GPA, FCAT\_SS\_2012*FCAT\_SS\_2013}} = 0$.

H_{a6}: $R^2_{\text{BLACK, HISP, WHITE, MULTI, AMER\_IN, GENDER, ELL, ESE, FREE\_RED, SCHOOL\_GRADE, GPA, AVG\_IBA, AVG\_LSA*FCAT\_SS\_2013}} - R^2_{\text{BLACK, HISP, WHITE, MULTI, AMER\_IN, GENDER, ELL, ESE, FREE\_RED, SCHOOL\_GRADE, GPA, FCAT\_SS\_2012*FCAT\_SS\_2013}} > 0$.

2012 Grade 5 Mathematics FCAT 2.0 - Level 2 (FCAT_AL_2012 = 2)

H_{07}: $R^2_{\text{BLACK, HISP, WHITE, MULTI, AMER\_IN, GENDER, ELL, ESE, FREE\_RED, SCHOOL\_GRADE, GPA, AVG\_IBA, AVG\_LSA*FCAT\_SS\_2013}} - R^2_{\text{BLACK, HISP, WHITE, MULTI, AMER\_IN, GENDER, ELL, ESE, FREE\_RED, SCHOOL\_GRADE, GPA, FCAT\_SS\_2012*FCAT\_SS\_2013}} = 0$.

H_{a7}: $R^2_{\text{BLACK, HISP, WHITE, MULTI, AMER\_IN, GENDER, ELL, ESE, FREE\_RED, SCHOOL\_GRADE, GPA, AVG\_IBA, AVG\_LSA*FCAT\_SS\_2013}} - R^2_{\text{BLACK, HISP, WHITE, MULTI, AMER\_IN, GENDER, ELL, ESE, FREE\_RED, SCHOOL\_GRADE, GPA, FCAT\_SS\_2012*FCAT\_SS\_2013}} > 0$.

2012 Grade 5 Mathematics FCAT 2.0 - Level 3 (FCAT_AL_2012 = 3)

H_{08}: $R^2_{\text{BLACK, HISP, WHITE, MULTI, AMER\_IN, GENDER, ELL, ESE, FREE\_RED, SCHOOL\_GRADE, GPA, AVG\_IBA, AVG\_LSA*FCAT\_SS\_2013}} - R^2_{\text{BLACK, HISP, WHITE, MULTI, AMER\_IN, GENDER, ELL, ESE, FREE\_RED, SCHOOL\_GRADE, GPA, FCAT\_SS\_2012*FCAT\_SS\_2013}} = 0$.

H_{a8}: $R^2_{\text{BLACK, HISP, WHITE, MULTI, AMER\_IN, GENDER, ELL, ESE, FREE\_RED, SCHOOL\_GRADE, GPA, AVG\_IBA, AVG\_LSA*FCAT\_SS\_2013}} - R^2_{\text{BLACK, HISP, WHITE, MULTI, AMER\_IN, GENDER, ELL, ESE, FREE\_RED, SCHOOL\_GRADE, GPA, FCAT\_SS\_2012*FCAT\_SS\_2013}} > 0$.

2012 Grade 5 Mathematics FCAT 2.0 - Level 4 (FCAT_AL_2012 = 4)

H_{09}: $R^2_{\text{BLACK, HISP, WHITE, MULTI, AMER\_IN, GENDER, ELL, ESE, FREE\_RED, SCHOOL\_GRADE, GPA, AVG\_IBA, AVG\_LSA*FCAT\_SS\_2013}} - R^2_{\text{BLACK, HISP, WHITE, MULTI, AMER\_IN, GENDER, ELL, ESE, FREE\_RED, SCHOOL\_GRADE, GPA, FCAT\_SS\_2012*FCAT\_SS\_2013}} = 0.$

H_{a9}: $R^2_{\text{BLACK, HISP, WHITE, MULTI, AMER\_IN, GENDER, ELL, ESE, FREE\_RED, SCHOOL\_GRADE, GPA, AVG\_IBA, AVG\_LSA*FCAT\_SS\_2013}} - R^2_{\text{BLACK, HISP, WHITE, MULTI, AMER\_IN, GENDER, ELL, ESE, FREE\_RED, SCHOOL\_GRADE, GPA, FCAT\_SS\_2012*FCAT\_SS\_2013}} > 0.$

2012 Grade 5 Mathematics FCAT 2.0 - Level 5 (FCAT_AL_2012 = 5)

H_{010}: $R^2_{\text{BLACK, HISP, WHITE, MULTI, AMER\_IN, GENDER, ELL, ESE, FREE\_RED, SCHOOL\_GRADE, GPA, AVG\_IBA, AVG\_LSA*FCAT\_SS\_2013}} - R^2_{\text{BLACK, HISP, WHITE, MULTI, AMER\_IN, GENDER, ELL, ESE, FREE\_RED, SCHOOL\_GRADE, GPA, FCAT\_SS\_2012*FCAT\_SS\_2013}} = 0.$

H_{a10}: $R^2_{\text{BLACK, HISP, WHITE, MULTI, AMER\_IN, GENDER, ELL, ESE, FREE\_RED, SCHOOL\_GRADE, GPA, AVG\_IBA, AVG\_LSA*FCAT\_SS\_2013}} - R^2_{\text{BLACK, HISP, WHITE, MULTI, AMER\_IN, GENDER, ELL, ESE, FREE\_RED, SCHOOL\_GRADE, GPA, FCAT\_SS\_2012*FCAT\_SS\_2013}} > 0.$

To address the third research question about the differences in interim assessment effects on achievement, the study utilized an analysis of covariance (ANCOVA) model, with FCAT_SS_2013 again serving as the dependent variable. ANCOVA allows the researcher to increase the precision of a statistical comparison of group means by partitioning out variance attributed to a covariate, which ideally results in a smaller error variance (Hinkle, Wiersma, & Jurs, 2003). Because the covariate should be in the same domain as the dependent variable, it is not reasonable to attempt to statistically control for variances in demographics or teacher-assigned grades in this model.

The independent variables included two categorical designations (Low or High) for the number of 2013 LSA tests taken (NUM_LSA) and the number of 2013 IBA tests taken (NUM_IBA). The best predictor of future performance is past performance; therefore, the present study statistically controlled for this with a covariate, the 2012 Grade 5 Mathematics FCAT 2.0 developmental scale score (FCAT_SS_2012).

The null hypothesis for research question 3 is:

H_{011}: After adjusting for 2012 Grade 5 Mathematics FCAT 2.0 developmental scale scores, the means of 2013 Grade 6 Mathematics FCAT 2.0 developmental scale scores in all four groups will be equal to each other ($\mu'_1 = \mu'_2 = \mu'_3 = \mu'_4$).

The corresponding alternate hypothesis is:

H_{a11}: After adjusting for 2012 Grade 5 Mathematics FCAT 2.0 developmental scale scores, the means of 2013 Grade 6 Mathematics FCAT 2.0 developmental scale scores in all four groups will not be equal to each other ($\mu'_i \neq \mu'_k$ for some i, k).

Population, Sample, and Data Sources

Population. Students who were enrolled in sixth grade mathematics during the 2012-2013 school year make up the selected population from which samples were drawn. The school district's online data analysis tool was the primary source of information pertaining to students, and the Florida state Department of Education (FDOE) website was the primary source of information pertaining to schools in the district.

In Florida, all students enrolled in fifth grade during the 2011-12 school year were required to take either the Grade 5 Mathematics FCAT 2.0 or Florida Alternative

Assessment (FAA), and all students enrolled in sixth grade during the 2012-2013 school year were required to take the Grade 6 Mathematics FCAT 2.0 or FAA. The majority of students enrolled in sixth grade were scheduled into the courses M/J Math 1 or M/J Math 1 Adv for mathematics. A very small percentage of students in the sixth grade were enrolled in an accelerated mathematics course such as M/J Math 2 Adv or an Access Points Math course designed for students with severe cognitive disabilities.

However, beginning in seventh grade, the variation in mathematics course enrollment increases dramatically, which would unnecessarily confound predictive analysis. Therefore, the present study focused on students enrolled in the sixth grade alone. Further, the Superintendent discontinued the Reading Interim Benchmark Assessment (IBA) after the first administration in Fall 2012, thus reducing the power of an analysis on reading interim assessments. For this reason, mathematics alone was analyzed.

Sample. The present study involved analyzing an archival data set from sixth grade students (originally *n* =9,038) enrolled in either M/J Math 1 or M/J Math 1 Adv during the 2012-2013 school year. Although over 9,000 students were enrolled for a portion of their sixth grade year during 2012-2013 at the focus district, not all of these students were enrolled for the majority of the school year. Teacher-given grades and FCAT scores were utilized as a filter to ensure that only students enrolled in the focus district for the majority of the 2012-2013 school year were included in the sample. The following inclusion criteria were used to determine the initial study sample:

1. had a response for each demographic variable,
2. attended a school that received a school grade from the state of Florida,

3. received a teacher-given grade other than "incomplete" for the first three quarters of the 2012-2013 school year in either M/J Math 1 or M/J Math 1 Adv, and
4. had a score for the 2012 Grade 5 Mathematics FCAT 2.0 and 2013 Grade 6 Mathematics FCAT 2.0.

After applying each of these four criteria, the study sample size was 5,801 (see Table 1). The majority of students in the study sample took a combination of 2 IBA tests and 7 LSA tests (*n* = 1, 173). Note that some students had no IBA test scores and/or no LSA test scores.

Table 1

Sample Size by Number of IBA Tests Taken and Number of LSA Tests Taken

		NUM_IBA				
		3	2	1	0	*Total*
NUM_LSA	8	248	419	41	0	*708*
	7	194	1,173	54	3	*1,424*
	6	134	979	36	1	*1,150*
	5	99	588	27	5	*719*
	4	42	437	26	5	*510*
	3	31	414	21	2	*468*
	2	25	133	17	1	*176*
	1	6	191	13	3	*213*
	0	39	149	11	234	*433*
	Total	*818*	*4,483*	*246*	*254*	***5,801***

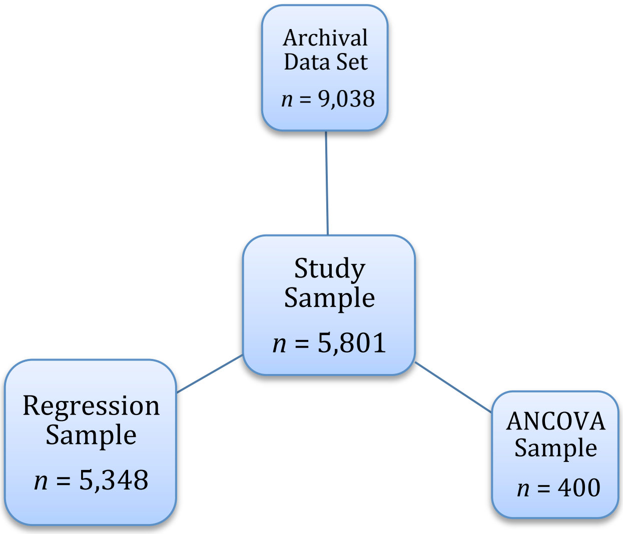


Figure 19. Sample sizes for regression and ANCOVA samples.

Each type of statistical test required a further modification to this original study sample (see Figure 19). The multiple regression models included two variables for average interim test scores, AVG_IBA and AVG_LSA, which required an additional regression-only sample restriction:

1. had scores for at least one IBA and at least one LSA so that the AVG_IBA and AVG_LSA values were not blank.

Cases with missing scores for one or both type of interim assessment (*n* = 453) were dropped from the already reduced sample of 5,801, resulting in a regression-specific sample of 5,348 (see Table 2).

Table 2

Cases Removed from Regression Sample – Missing Interim Scores

Interim Average Variables	Students Missing Scores *n*
AVG_IBA	254
AVG_LSA	433
Either AVG_IBA or AVG_LSA	**453**

Because the ANCOVA model requires balanced cells, and the selected groups (see Table 3) had no less than 101 cases, 100 cases were selected at random from each of four groups. Group construction method and details appear in the Data Analysis Procedures section later in this chapter. The original *n* came from the study sample (*n* = 5,801) because the regression sample (*n* = 5,348) excluded cases with 0 scores for either IBA or LSA. The additional selection criteria for the ANCOVA-only sample (*n* = 400) were:

1. had either 0, 1, or 3 IBA test scores, and
2. had either 0, 1, 2, 3, 6, 7, or 8 LSA test scores.

Table 3

ANCOVA Sample Sizes – RQ3

Group	NUM_IBA	NUM_LSA	Original *n*	Sample *n*
1 – High both	High – 3	High – 6, 7, or 8	576	100
2 – High IBA, Low LSA	High – 3	Low – 0, 1, 2, or 3	101	100
3 – Low IBA, High LSA	Low – 0 or 1	High – 6, 7, or 8	135	100
4 – Low both	Low – 0 or 1	Low – 0, 1, 2, or 3	302	100

Summary demographic statistics for the 5,348 students who were included in the regression models and the 400 students who were included in the ANCOVA model of the present study are shown in Table 4.

Table 4

Sample Demographics – Regression and ANCOVA

	Regression (*n* = 5,348)		ANCOVA (*n* = 400)	
Demographics	Frequency	Percentage	Frequency	Percentage
English Language Learner*				
Yes	145	2.7	6	1.5
No	5,203	97.3	394	98.5

Exceptional Student Education**				
Yes	590	11.0	102	25.5
No	4,758	89.0	298	74.5
Free/Reduced Lunch				
Yes	2,968	55.5	258	64.5
No	2,380	44.5	142	35.5
Gender				
Male	2,633	49.2	212	53.0
Female	2,715	50.8	188	47.0
Race/Ethnicity				
Black or African American, Non-Hispanic	2,588	48.4	254	63.5
White, Non-Hispanic	1,845	34.5	111	27.8
Hispanic/Latino	479	9.0	17	4.3
Multiracial, Non-Hispanic	214	4.0	15	3.8
Asian, Non-Hispanic	208	3.8	3	0.8
American Indian or Alaska Native, Non-Hispanic	14	.3	0	0
School Grade				
Attended an "A" school	1,842	34.4	68	17.0
Attended a "B" school	464	8.7	22	5.5
Attended a "C" school	1,682	31.5	142	35.5
Attended a "D" school	877	16.4	126	31.5
Attended an "F" school	483	9.0	42	10.5
2012 FCAT 2.0 Grade 5 Mathematics Achievement Level				
Level 5 – Mastery of NGSSS	162	3.0	4	1.0
Level 4 – Above satisfactory	696	13.0	29	7.3
Level 3 – Satisfactory	1,754	32.8	109	27.3
Level 2 – Below satisfactory	1,629	30.5	135	33.8
Level 1 – Inadequate	1,107	20.7	123	30.8
2013 FCAT 2.0 Grade 6 Mathematics Achievement Level				
Level 5 – Mastery of NGSSS	150	2.8	6	1.5
Level 4 – Above satisfactory	673	12.6	29	7.3
Level 3 – Satisfactory	1,523	28.5	95	23.8
Level 2 – Below satisfactory	1,689	31.6	112	28.0
Level 1 – Inadequate	1,313	24.6	158	39.5

* The focus district used LAS Links to determine ELL status. LAS Links is a standardized test of English language skills commonly used in Florida.

** All students with any primary exceptionality code were included in ESE.

Data sources. In 2012, the school district of focus administered two distinctly different interim assessment programs in an attempt to develop a comprehensive assessment system (see Figure 5). The first program, the Interim Benchmark Assessment (IBA) program, was intended to be a comprehensive and predictive measure of students' progress towards mastery of skills assessed on the FCAT. In a 2010 paper, Brian Gong of the Center for Assessment referred to this type of interim assessment as the *state-test mirror* design. The IBA program was older—it had been in place since 2004—and was essentially the same test administered at the beginning, middle, and toward the end of each school year.

The more recent program, the Learning Schedule Assessment (LSA) program, was intended to address the needs of educators who wanted assessments with an instructional purpose, and included a different pre and post construct for each instructional module. In other words, instead of one large comprehensive test, the LSAs were intended to assess only what was taught in one module. Gong referred to this type of assessment as the *non-cumulative instructional mirror* design (2010). See Figure 5 for a timeline/content comparison between the two types of interim assessment programs.

Month	Aug	Sep	Oct	Nov	Dec	Jan	Feb	Mar	*April*
Learning Schedule	Mod A	Mod B	Mod C	Mod D	Mod E	Mod F	Mod G	Mod H	*FCAT 2.0*

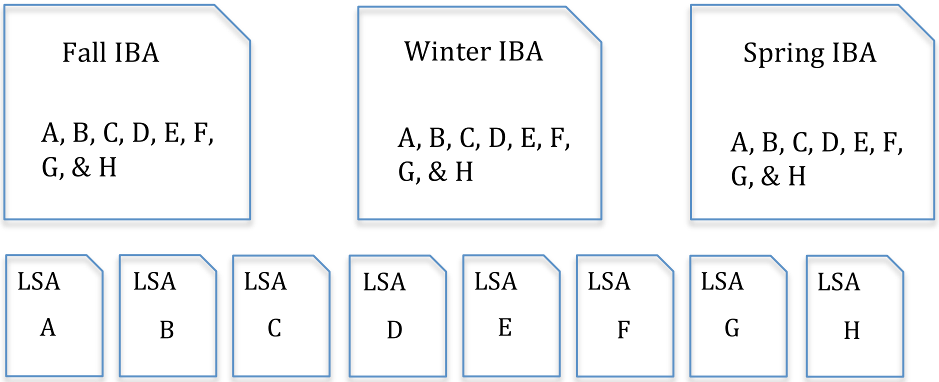


Figure 5. Simplified Learning Schedule timeline by month and corresponding content tested by Interim Benchmark Assessments (IBA) and Learning Schedule Assessments (LSA). Note that the three IBA administrations are the same test. The eight LSA tests are different and only assess the benchmarks in one module each.

Differences between assessment programs. Multiple differences exist in the design, purpose, and administration of the two types of assessments. Security for the Interim Benchmark Assessments (IBAs) was paramount as the same test was used in each administration and over multiple years. Also, each IBA was viewed as practice for the FCAT 2.0; as such, all course benchmarks were included in the 48-item test and all aspects of the IBA administration were purposefully as similar as possible to the strict FCAT 2.0 administration. The Learning Schedule Assessments (LSAs) were intended to measure only what was in one instructional unit each. LSAs were administered over the computer or on paper, depending on whether the teacher had access to computers and reliable Internet connections or not, through LearningStation's Insight product. Teachers also had control over when their students took the LSAs, administering the post-tests only after the instructional unit was taught, whereas the IBAs were administered within certain time frames three times in the year.

In both cases, all teachers could see their students' scores through the online data analysis tool, Pearson Inform. However, the amount of time between when students took the test and when the test scores were displayed varied. The school testing coordinator for each school organized the process for securely collecting IBA bubble sheets and shipping them to the district's main data warehouse, where testing department staff scanned the bubble sheets and cleaned the resulting data file prior to uploading to Inform. Cleaning involved processes such as removing duplicate or blank student sheets and correcting human errors such as incorrectly marked school numbers or grade levels. However, because the LSAs could be scanned at the school sites or collected via students taking the test on a computer, resulting data were available instantaneously online via LearningStation Insight, and shortly thereafter in Pearson Inform. Variable security settings on both Pearson Inform and LearningStation Insight allowed appropriate administrators and support staff to see applicable student, teacher, and/or school data as well.

Professional development for the IBA program was provided to testing coordinators at the school sites, and included only the protocols for how to securely proctor the tests. On the other hand, district academic services staff and testing department staff trained teachers, instructional coaches, and principals on two aspects of the LSAs: (a) how to administer the tests on the online assessment platform, LearningStation Insight; and (b) how to appropriately interpret data from the assessment. Principals were also strongly encouraged to provide local, personalized professional development for their teachers on how to understand their data. In addition, a district-wide professional learning community was formed to study each assessment prior to administration and again after it was given to students. At these learning community meetings, teachers discussed the benchmarks addressed in each instructional module, whether their students met each benchmark, common misunderstandings in students' responses, and ideas for how to adjust their instruction in the next module.

Table 5 displays a comparison between the two interim assessment programs using selected standards from the *Student Evaluation Standards,* developed by the Joint Committee on Standards for Educational Evaluation, as an organizational framework (JCSEE, 2001). The present study addressed two standards, validity orientation and service to students. The present study sought to evaluate the alignment to the assessment program purposes of state test score prediction and instructional use through analyzing interim assessment scores. The validity orientation standard is that "student evaluations should be developed and implemented so that interpretations made about the performance of a student are valid and not open to misinterpretation" (JCSEE, 2001, p. 2). Through using JCSEE (2001)'s *Student Evaluation Standards* and Perie et al. (2009)'s evaluative criteria within their framework for considering interim assessment programs, this study sought to demonstrate one possible method to evaluate predictive validity for district-developed interim benchmark testing programs.

JCSEE places standard P1 Service to Students, that evaluations should "promote sound education principles…so that educational needs of students are served," at the top of their list of standards (2001, p. 1). One method to evaluate whether student needs have been met is to analyze the resulting end-of-year state standardized test scores.

It should be noted that although JCSEE includes a section entitled, "Utility Standards," the standards are addressing usefulness from a different perspective than in the present study. Additionally, educational evaluation standards were also developed by the Joint Committee on Standards for Educational and Psychological Testing (JCSEPT, 1999) and are similar in many ways; the three organizations comprising JCSEPT, namely AERA, APA, and NCME, are also organizational members of JCSEE. For purposes of clarity, however, the comparison in Table 5 utilizes the more recent JCSEE *Student Evaluation Standards* (2001).

Table 5

Comparison of Simultaneous 2012-2013 Interim Assessment Programs by Selected JCSEE Student Evaluation Standards

JCSEE Student Evaluation Standard	Interim Benchmark Assessments (IBA)	Learning Schedule Assessments (LSA)
Propriety Standards		
Service to Students[a]	Predictive purpose	Instructional purpose
Appropriate Policies & Procedures	Test Administration Manual; Test proctor may adjust time or use other modifications to accommodate ELL/ESE needs	No manual; Test proctor may adjust time or use other modifications to accommodate ELL/ESE needs
Access to Evaluation Information	Job role-specific permission to access data	Job role-specific permission to access data
Utility Standards		
Defined Users & Uses	Course-specific; scores used to predict performance on FCAT 2.0	Course- and module-specific; scores used to judge growth over one module
Information Scope	48 MC items covering the whole course	10-20 MC items covering one instructional module
Evaluator Qualifications	Teachers, district content specialists, test specialists	Teachers, district content specialists, test specialists
Explicit Values	Reported using % correct and likelihood of passing, normed using all prior year students in district	Reported using % correct; standardized scores available using class norms
Effective Reporting	Reported on Inform after approx. 2 weeks	Reported on Insight immediately and Inform after approx. 1 week

Feasibility Standards		
Practical Orientation	Paper-based only testing platform; scored using Scantron machines	Paper- and computer-based testing platforms; scored using Scantron machines, plain-paper scanners, or automatically by computer
Political Viability	September, December, and February; district-wide dates	Given at the end of each module; teacher decides
Evaluation Support	Internal test reliability report; Test Coordinator training materials	Insight training materials for all educators
Accuracy Standards		
Validity Orientation[a]	Content expert review, ex post facto item discrimination; no criterion-related validity measures	Content expert review; no construct or criterion-related validity measures
Defined Expectations for Students	State standards for each course are public knowledge; IBA aligns to FCAT-tested standards	State standards for each course are public knowledge; LSAs align to Learning Schedule modules
Reliable Information	Using a 2011 sample (*n* = 8,737), reliability was moderate (KR-20 = .745); Standard error of measurement = 3.10; Range 0-48	Point-biserial correlation by item for each test, visible to teachers; no overall reliability coefficient available
Bias Identification and Management	Test item specifications, Outcomes comparison	Test item specifications, Outcomes comparison
Handling Information and Quality Control	Secure test – no access to item results	Non-secure test – can see item results after completion
Analysis of Information	Internal review performed once per test version	Informal review available via Insight reporting tools
Metaevaluation	Not done by district	Not done by district

a. The present study addressed these student evaluation standards.

Reliability and validity. The school district of focus calculated a KR-20 reliability coefficient of .745 for the 2011 cohort of students enrolled in either M/J Math 1 or M/J Math 1 Adv and taking the 2011 IBA. This reliability measure passes the .74 criterion (Li et al.,

2010) for a low- or moderate-stakes tests. However, as a new test program, the LSA tests did not yet have an available report including an overall reliability coefficient; rather, the LearningStation Insight platform included point-biserial values for each item on each test, calculated by one classroom of students at a time. This was visible to teachers and administrators, but did not have much utility at the district level to analyze the entire body of data district-wide. It should be noted that the sample for the present study was a subset of the overall dataset collected by the school district. As data were not available at the item level, it was not possible to compute an internal consistency reliability coefficient for the specific set of data subset included herein. Hence, the district's internal consistency estimate of .745 served as an approximate benchmark for assessing reliability for the data used herein; the actual reliability estimate for the sample may have varied based on range of variation, amount of measurement error, or other factors related to the data subset.

Content experts including course-specific teachers, school-based instructional coaches, and district-based content specialists reviewed all tests in both assessment programs, contributing to content validity. In addition, the IBA tests were reviewed for item discrimination after the 2011 administrations were complete, also contributing to the analysis of construct validity for that group of scores. However, no criterion-related validity measures were taken or assessed for any administration of either program. The present study analyzed criterion-related validity of the 2013 data set by assessing predictive validity using the 2013 FCAT 2.0 as a the criterion measure.

Data Collection and Ethical Considerations

Archived data were collected from the Pearson Inform online data analysis tool in separate files, aggregated and linked using Microsoft Excel and Access, and then deidentified prior to the study using a four-digit random number. School grades were downloaded from the Florida Department of Education website dedicated to reporting school grades (FDOE, 2013a) and linked using school numbers in the Excel data file. No student identifiers remained in the data set, which was stored on an external hard drive. The files on the jump drive were destroyed at the conclusion of the study. A waiver was received from the University of North Florida's Institutional Review Board and permission was granted by the school district's Institutional Review Board to use the data set.

Data Analysis Procedures

The first criterion for predictive interim assessments in Perie et al. (2009)'s framework for evaluating interim benchmark assessments was used in to select an analysis procedure to determine whether the interim tests provide predictive utility:

> The assessment should be highly correlated with the criterion measure (e.g., the end-of-the-year state assessment). The technical documentation should include evidence of the predictive link between the interim assessment and the criterion measure. However, in order to justify the additional testing and cost, the predictive assessment should be significantly more related to the criterion measure than other measures (e.g., teachers' grades) that could be used. (p. 10)

The present study employed a multiple regression model using data from students enrolled in Grade 6 during the 2012-2013 school year to answer the first and second research question about predictive ability of interim assessments in general while statistically controlling other measures such as teacher-assigned grades. Multiple regression analysis produced correlations among and between all variables, including with the criterion measure in this case, the Grade 6 Mathematics FCAT 2.0. In addition, the regression variate quantifies the combined predictive power of all variables in each block.

Instructional utility was evaluated using Perie et al.'s (2009) second criterion for instructional interim assessments:

> Ideally, the system should provide evidence, based on scientifically rigorous studies, demonstrating that the assessment system has contributed to improved student learning in settings similar to those in which it will be used. (p. 10)

Analysis of covariance (ANCOVA) was utilized to answer the third and final research question about the difference in groups of students that took either of two types of interim assessment. The differences between the groups will determine which group of students experienced greater learning over the course of the 2012-2013 school year, as measured by the 2013 Grade 6 Mathematics FCAT 2.0 state assessment.

SPSS® software version 21 was used to perform the statistical analysis and develop some of the visual displays of data. Some preventative statistical measures were taken prior to performing the tests. A conservative significance value of .01 (α = .01) was utilized when evaluating the results to account for the possibility of family-wise Type 1 error as a result of 31 statistical tests performed on the same data set. To reduce possibility of collinearity, student grade point averages (GPA) from the first three academic quarters directly preceding the 2013 FCAT 2.0 administration window were used. Further, IBA score averages and LSA score averages were used in an attempt to capture the performance on these tests using a single variable each. Gain scores were considered and rejected because they result in doubling any measurement error presented by the assessment.

Dependent variable. The dependent variable in both procedures was the 2013 Grade 6 Mathematics FCAT 2.0 developmental scale score. Developmental scale scores are comparable across grade levels, whereas scale scores are unique to each grade level. This standardized state test was administered over the week of April 15 - 19, 2013. Therefore, grades and scores assigned after this date were not included in the analysis.

Demographic variables. Statistical controls were employed through a block entry method in the multiple regression model and with a covariate in the ANCOVA procedure. In the first multiple regression block, predictor variables included demographic variables identifying race/ethnicity, English Language Learner (ELL) status, Exceptional Student Education (ESE) status, gender, and Free/Reduced Lunch status (FREE_RED)—a proxy for Socioeconomic Status—of each student.

Dummy variables were used to code available race/ethnicity data, with a variable each for Non-Hispanic Black or African American (BLACK), Hispanic/Latino (HISP), Non-Hispanic White (WHITE), Non-Hispanic Multiracial (MULTI), and Non-Hispanic American Indian or Alaska Native (AMER_IN). Although the state of Florida collected data for identifying Hawaiian or Other Pacific Islander students in 2013, the focus district did not include that classification in the data set. Indicator coding was used to transform the raw categorical data available; 1 or 0 indicated membership or non-membership in each category, respectively. Asian was the reference category, or the omitted group receiving all zeros. The choice for reference category was purposeful; the mean developmental scale score for Asian, Non-Hispanic students was higher than the mean for any other race/ethnicity group on the 2013 Grade 6 Mathematics FCAT 2.0 in the focus district (Table 6). Regression coefficients for the other racial/ethnic dummy variables represent deviations from the Asian, Non-Hispanic group (Hair, Black, Babin, & Anderson, 2010), equivalent to the achievement gap concept from NCLB literature (e.g., Harris & Herrington, 2006; Lee & Reeves, 2012; Wong, 2013).

Table 6

District Mean 2013 Grade 6 Mathematics FCAT 2.0 Developmental Scale Score by Race/Ethnicity

Race/Ethnicity	Mean Score
White, Non-Hispanic	231
Black or African American, Non-Hispanic	219
Hispanic/Latino	223
Asian, Non-Hispanic	**240**
American Indian or Alaska Native, Non-Hispanic	231
Native Hawaiian or Other Pacific Islander, Non-Hispanic	229
Multiracial, Non-Hispanic	224
Total	225

Note. Public data retrieved January 2014 from "Student Performance Results: District Math Demographic Report" by Florida Department of Education (FDOE], http://app1.fldoe.org/fcatdemographics.

Other predictor variables included in the remaining four blocks were Florida school grade, student grade point average (GPA), prior year FCAT 2.0 developmental scale score, and average interim benchmark scores. By entering the interim benchmark scores in the last block, the change in effect size (R^2) between Block 4 and Block 5 reflected predictive power of those scores, independent of the effect of the other predictor variables.

School grades. In Florida, most public and charter schools are assigned a grade between A and F based on student growth and performance, as well as rigorous coursework and advancement opportunities. This school grade is intended to provide an easily interpreted general statement about the academic strength of the school environment. Although FCAT 2.0 and state End of Course scores are the basis of the majority of the calculations, the populations vary. For example, a portion of the school grade is strictly based on gains made by students who earned the lowest 25% of scores in the prior year (FDOE, 2013a). Therefore, the independent variable SCHOOL_GRADE was included as a measure of holistic school environment and collinearity diagnostics were included in the analysis to identify possible collinearity between school grade and the other FCAT-based variables. A typical grade scale was used to code school grades: A was coded as 4, B was coded as 3, C was coded as 2, D was coded as 1, and F was coded as 0.

GPA. Teacher-assigned student grades for the first three quarters of the M/J Math 1 or M/J Math 1 Adv course were also coded using the same scheme, and then averaged into a student grade point average (GPA). Fourth quarter student grades were not included in this average because these grades were assigned by teachers after the 2013 FCAT 2.0 administration date. Students with incomplete or otherwise missing teacher-assigned grades were excluded from the study sample.

Interim scores. Interim scores were also combined into two independent variables: one for average IBA scores, AVG_IBA, and one for average LSA scores, AVG_LSA. Averages were calculated differently based on the type of assessment. Each student had a maximum of three IBA scores, but each of those scores was based on identical forms of the test with the same items. As a result, averaging the three IBA scores was appropriate. However, the eight LSAs varied in length, so a weighted average was calculated based on the number of items. Although eight LSAs were included in the present study, three LSAs were excluded because the administration dates followed the 2013 FCAT 2.0 administration date.

FCAT 2.0 achievement levels. The second research question addresses the predictive utility of the interim assessments for students stratified by FCAT achievement level. Five models were necessary to reflect each of five FCAT achievement levels. The achievement levels are based on the developmental scale score and have a range of 1 to 5, where 5 is the highest achievement level.

Number of interim scores. To answer the third research question about academic performance differences between students who took IBAs and LSAs, an analysis of variance

(ANCOVA) model was employed. In the ANCOVA model, the two independent variables were categorical descriptors (either Low or High) regarding the number of IBA scores, NUM_IBA, and the number of LSA scores, NUM_LSA, for each student (see Figure 6). Four groups were constructed from the four cells in the Low-High matrix. In addition to these two variables, the prior year FCAT 2.0 score was utilized as a covariate. An overview of all variables used in the study design is displayed in Table 7.

		NUM_IBA	
		High – 3	Low – 0 or 1
NUM_LSA	High 6, 7, or 8	Group 1	Group 3
	Low 0, 1, 2, or 3	Group 2	Group 4

Figure 6. Constructing Groups for ANCOVA Model – RQ3.

Delimitations and Limitations of the Study

Delimitations. The present study was delimited to one large, urban school district in Northeast Florida due to its ease of availability to the researcher. The 5,801 students can be regarded as a sample of the population of the students in the state of Florida but did not necessarily represent a stratified sample of the entire population of students in the state. Thus, the relatively small size and geographic limits potentially compromised generalizability of the findings to other school systems around the state. In addition, this school district was unique in that many of the students participated in two simultaneous interim assessment programs during the 2012-2013 school year, making it an appropriate

Table 7

Overview of Study Design

		Variables		
Research Question	Model	Dependent	Independent	Covariate
To what extent can variance in high-stakes scores be explained by scores on interim assessments?	Multiple Regression with 5 blocks	FCAT_SS_2013	*Block 1* ELL, ESE, FREE_RED, GENDER, BLACK, HISP, WHITE, MULTI, AMER_IN *Block 2* SCHOOL_GRADE *Block 3* GPA *Block 4* FCAT_SS_2012 *Block 5* AVG_IBA, AVG_LSA	
To what extent can variance within achievement levels be explained by scores on interim assessments?	Restricted-Domain[a] Multiple Regression with 5 blocks	FCAT_SS_2013	*Block 1* ELL, ESE, FREE_RED, GENDER, BLACK, HISP, WHITE, MULTI, AMER_IN *Block 2* SCHOOL_GRADE *Block 3* GPA *Block 4* FCAT_SS_2012 *Block 5* AVG_IBA, AVG_LSA	
How does the achievement of students taking IBAs differ from LSAs?	ANCOVA	FCAT_SS_2013	NUM_IBA, NUM_LSA	FCAT_SS_2012

[a] Multiple regression was performed once for each 2012 Grade 5 Mathematics FCAT 2.0 Achievement Level.

setting for the present study; by contrast, many school districts administer only one type of interim assessment in any given school year.

The study was also delimited to two purposes for interim testing, namely predictive and instructional, and did not address any evaluative purposes (Perie et al., 2009). Further research is necessary to investigate how interim assessments are used to evaluate programs, professional development practices, and district initiatives. Additionally, other methods for gathering data (e.g., surveys, interviews) could be employed to assess use of data.

Although topics such as students' beliefs about the nature of mathematics, purposes for learning, and motivation are extremely important to the field of assessment, the present study did not address these. Further, although the two assessment programs in the present studied varied by platform—one was only available on paper and the other was available either on paper or computer, the differences in outcome attributed to platform were not analyzed. This, along with other aspects of both tests, such as

Grade 6 was chosen because it is the last school grade where the majority of students are still enrolled in the same course: M/J Math 1 (or the advanced version of the same course, M/J Math 1 Adv). Beginning in grade 7, the variance in mathematics course enrollment is much larger, contributing one more potentially confounding variable: instruction level.

Although the statistical techniques involved in the present study (multiple regression and ANCOVA) are appropriate for answering the research questions, other factors were not included in the analysis or accounted for explicitly in the results. These factors, such as test quality, professional development practices, teacher pedagogical content knowledge, professional collaborative efforts between teachers, student-teacher interactions, school leadership, curricular programs, technology usage, and school culture potentially could impact the results of the present study. These all represent areas for possible future research.

Limitations. About halfway through the 2012-2013 school year, a newly elected superintendent discontinued one type of interim assessment (IBA) for reading in favor of a statewide test (Florida Assessments for Instruction in Reading—FAIR). As a result, insufficient data existed for a similar analysis regarding reading. Additionally, other state-tested subjects such as science, writing, and social studies were not administered to every grade, making it impossible to use prior year test scores for these subjects. Therefore, the present study included mathematics alone.

Another limitation is that the present study was a correlational study and therefore did not allow for any causal inferences based upon the variables involved. Additionally, as Hamilton (2003) noted, all research involving large-scale assessment such as the dependent variable in the present study, has limitations involving causality and random sampling.

The nature of learning is such that knowledge is acquired over time. In the present study, predictive utility of interim tests was evaluated. However, to the extent that students learn during the time between the interim test and the state high-stakes test, the predicted score will fall below what the student actually earns on the summative test, thereby diminishing the predictive power of the interim test. The opposite can and does happen often: students perform well on an interim test but do poorly on the high-stakes test at the end of the year. Fatigue, illness, negative emotions, literacy-related mistakes, guessing, and miscoding on answer documents all may affect high-stakes testing outcomes. Other issues within the tests, such as low item discrimination, and outside of the test, such as with the testing environment or inaccurate scoring, introduce further error in studies focusing on assessment.

Summary

This chapter included an overview of the research design; a review of research questions; details about the study population, sample, and data sources; a comparison of the interim assessment programs using JCSEE standards; an overview of data analysis procedures, data collection procedures, and ethical considerations; and study delimitations and limitations. The next chapter includes details about the statistical analysis performed and the results.

CHAPTER 4

Results

The purpose of the present study was to evaluate the utility of two different interim assessment programs administered over the course of one school year, Learning Schedule Assessments (LSA) and Interim Benchmark Assessments (IBA). Three research questions, two predictive and one comparison, provided a focus to the study, served as a framework for data analysis, and were used to organize the subsequent discussion of findings.

Research questions guiding the efforts of this dual-technique quantitative study were: (a) To what extent can variance in middle school student scores on mathematics high-stakes state tests be explained by scores on district interim assessments after controlling for prior scores, student demographic variables, and teacher-assigned grades? (b) To what extent can variance in middle school student scores within achievement levels on mathematics high-stakes state tests be explained by scores on district interim assessments after controlling for prior scores, student demographic variables, and teacher-assigned grades? and (c) To what degree does achievement, as measured by mathematics high-stakes state tests, of middle school students who have experienced less frequently administered, predictive interim assessments differ from the achievement of students who have experienced more frequently administered, instructional assessments, after controlling for prior scores? See Table 7 for a study outline review.

Chapter 4 includes a brief overview of the methods and procedures, discussion about characteristics of the sample and significance level, descriptive statistics about and between the variables for the present study, a discussion of multiple regression and ANCOVA analyses results to determine the degree of statistical significance, and answers to research questions developed in Chapter 3. The chapter concludes with a summary.

Review of Methods and Procedures

Two main analytic protocols were utilized to address the three research questions. Six multiple regression models were used to answer the first (for all *n*) and second (2012 Grade 5 Mathematics FCAT 2.0 Achievement Level 1, Level 2, Level 3, Level 4, and Level 5) research questions pertaining to predictive power of interim assessments versus other predictors such as student GPA. ANCOVA was used to answer the third question about the instructional utility of the two different types of interim assessments, with prior year performance acting as the covariate. Both modeling procedures require an initial review of the data including descriptive statistics and correlations, tests of underlying assumptions for individual variables as well as the overall model, and analysis of the individual components.

Regression. To address the predictive utility of the two interim assessment programs, IBAs and LSAs, this study utilized multiple regression with an explanatory objective. Multiple regression is appropriate for modeling one dependent relationship between one metric dependent variable and two or more metric predictor variables (Hair et al., 2010). Although multiple regression is widely used as a tool to optimize dependent variable prediction, multiple regression can also be used for explanation. In the present study, the main focus of regression was to determine the relative predictive power of each of the independent predictor variables.

To address the first research question regarding predictive utility of interim assessments versus other known predictors, the present study used a multiple regression model with five blocks. The dependent variable was the 2013 Grade 6 Mathematics FCAT 2.0 developmental scale score (FCAT_SS_2013). Independent variables included, by block:

- Block 1: demographic variables in Table 4 (BLACK, HISP, WHITE, MULTI, AMER_IN, GENDER, ELL, ESE, FREE_RED);
- Block 2: the school grade for which the student was enrolled (SCHOOL_GRADE);
- Block 3: the average of teacher-assigned grades for M/J Math 1 or M/J Math 1 Adv from the first three quarters of the 2012-2013 school year (GPA);
- Block 4: the 2012 Grade 5 Mathematics FCAT 2.0 developmental scale score (FCAT_SS_2012); and
- Block 5: the average percent correct for IBAs and LSAs (AVG_IBA, AVG_LSA).

The second research question sought to identify major differences in predictive utility of interim assessments by 2012 FCAT 2.0 achievement levels. Five multiple regression models, corresponding to the five achievement levels, with five blocks each were analyzed. The dependent variable was the 2013 Grade 6 Mathematics FCAT 2.0 developmental scale score (FCAT_SS_2013). Independent variables included, by block:

- Block 1: demographic variables in Table 4 (BLACK, HISP, WHITE, MULTI, AMER_IN, GENDER, ELL, ESE, FREE_RED);

- Block 2: the school grade for which the student was enrolled (SCHOOL_GRADE);
- Block 3: the average of teacher-assigned grades for M/J Math 1 or M/J Math 1 Adv from the first three quarters of the 2012-2013 school year (GPA); and
- Block 4: the 2012 Grade 5 Mathematics FCAT 2.0 developmental scale score (FCAT_SS_2012); and
- Block 5: the average percent correct for IBAs and LSAs (AVG_IBA, AVG_LSA).

ANCOVA. The third research question asked about the nature of the difference in FCAT scores between students who have taken two different types of interim assessments, IBA and LSA. However, prior year test scores (r = .78) would threaten internal validity of the findings for a traditional analysis of variance (ANOVA). Experimental control was out of reach as the design was ex post facto and non-experimental. Therefore, in an attempt to minimize this threat of potentially confounding 2012 Grade 5 Mathematics FCAT 2.0 developmental scale scores, analysis of covariance (ANCOVA) was used to statistically control the FCAT_SS_2012 covariate.

ANCOVA combines regression analysis and ANOVA and is developed by adjusting a conventional ANOVA for the regression of the dependent variable on the covariate. The variation of the dependent variable, in this case 2013 FCAT scores, is partitioned out so that the researcher is better able to analyze the effects of the primary independent variables (Hinkle, Weirsma, & Jurs, 2003). This partitioning of the variance culminates in a reduction in the sum of squared errors and, consequently, the mean square error (Onwuegbuzie & Daniel, 2001).

The dependent variable in the ANCOVA model was the same as the dependent variable in the multiple regression models: 2013 Grade 6 Mathematics FCAT 2.0 developmental scale score, FCAT_SS_2013. Two independent variables were used: a categorical variable describing the number of Interim Benchmark Assessments scores, NUM_IBA, and a categorical variable describing the number of Learning Schedule Assessment scores, NUM_LSA. The covariate was 2012 Grade 5 Mathematics FCAT 2.0 developmental scale score, FCAT_SS_2012.

Sample

The present study involved analyzing an archival data set from sixth grade students (originally *n* = 9,038) enrolled in either M/J Math 1 or M/J Math 1 Adv during the 2012-2013 school year. After filtering out students without teacher-given grades for the first three quarters, a school grade, all demographics, or scores for both the 2012 Grade 5 and 2013 Grade 6 Mathematics FCAT 2.0 tests, the sample size was reduced to 5,801. The multiple regression models included two variables, AVG_IBA and AVG_LSA, which required an additional sample restriction. 453 cases were lacking interim assessment scores for all of one or both types of interim assessment, and therefore had at least one missing interim average. As the missing data processes were nonrandom, a modeling-based approach was a logical remedy. Cases with missing averages for one or both type of interim assessment (*n* = 453) were dropped from the already reduced sample, resulting in a sample of 5,348 for regression models.

The cases with missing averages were included for the ANCOVA sample, however, because the students who took no tests of either type were part of the “Low” categorical groups, referring to a low number of tests taken. ANCOVA required a balanced cell design, thus a random sample of 100 cases was selected from each of four groups (See Table 3), for a total of *n* = 400. The four groups were based on combinations of categorical values for the number of interim assessment scores for each type of program (High IBA, High LSA; High IBA, Low LSA; Low IBA, High LSA; and Low IBA, Low LSA).

Statistical Significance Level

Both sample sizes for regression (*n* = 5,348) and ANCOVA (*n* = 400) are adequate to ensure statistical power and limited generalizability to other school systems around the state. The sample did not necessarily represent a stratified sample of the entire population of students in the state. A statistical significance level (α) of .01 was used for the present study to compensate for the increased sensitivity of statistical significance tests due to large sample sizes and to prevent family-wise Type I error. In addition, both sample sizes exceeded the desired generalizability ratio of 20 observations to each independent variable (Hair et al., 2010). However, the samples may not be representative of populations outside of large, urban school districts in Florida.

Assumptions and Descriptive Statistics

Four assumptions underlying the statistical techniques of multiple regression and ANCOVA were tested for individual variables prior to model estimation, as well as the variates after model estimation. Individual variable assumption tests are discussed in this section, while the variate assumption tests are discussed in the section regarding model analyses results. The four assumptions commonly tested prior to inferential statistical analysis are (Hair et al., 2010):

1. Normality

2. Homoscedasticity
3. Linearity
4. Absence of correlated errors

Normality. For every independent variable as well as the dependent variable, frequency histograms with superimposed normal curves were visually inspected for the recognizable bell-shape indicative of normal distributions (see Figure 7). Dichotomous demographic predictor variables (i.e., BLACK, HISP, WHITE, MULTI, AMER_IN, GENDER, ELL, ESE, FREE_RED) were excluded from further analysis, as they were not expected to have a normal distribution (Williams, Grajales, & Kurkiewicz, 2013).

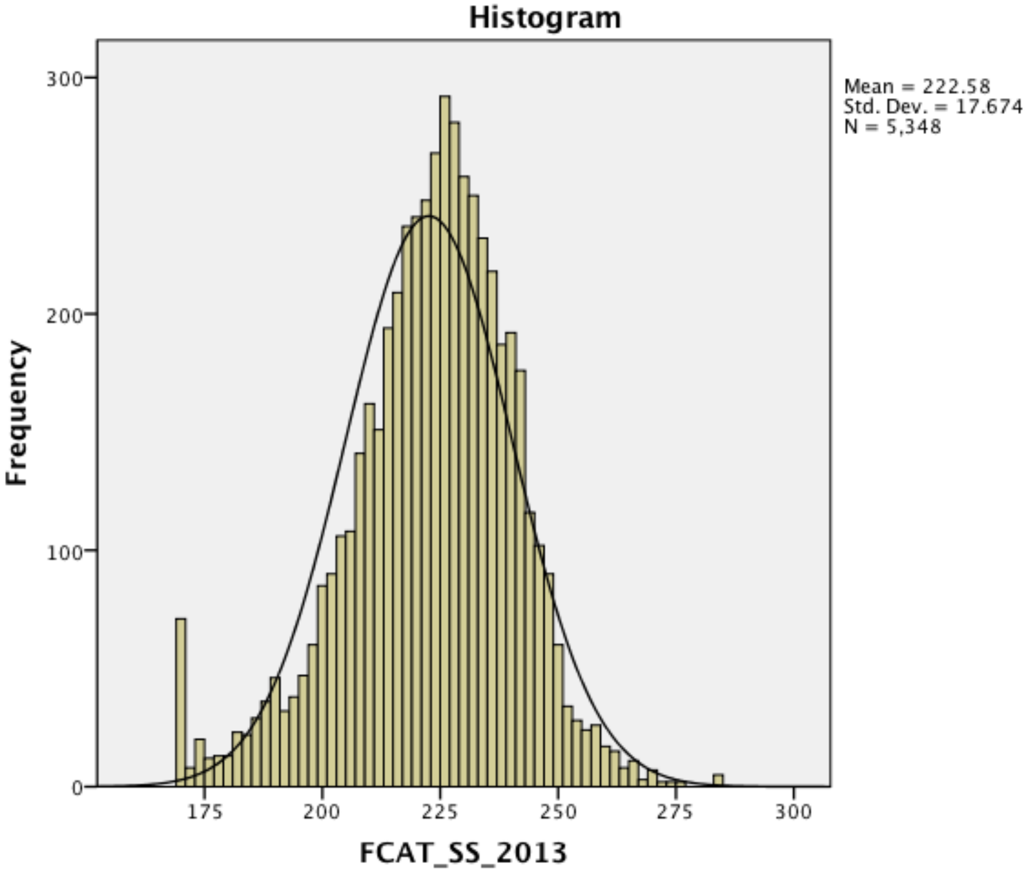


Figure 7. Illustrative example of a frequency distribution with a superimposed normal curve used to visually inspect normality – Dependent variable FCAT_SS_2013

Errors of the predictor models including these demographic variables were examined after regression analyses were performed to verify trustworthiness of the results. Skewness and kurtosis were evaluated for the remaining non-demographic predictor variables as statistical measures of normality (see Table 8). Some possible violations existed among the non-demographic variables SCHOOL_GRADE and AVG_IBA. SCHOOL_GRADE was slightly platykurtic (flatter than normal), and AVG_IBA was skewed right (shifted to the left). However, regression is robust to non-normal variables with a large sample size (Hair et al., 2010).

Table 8

Descriptive Data for All Non-demographic Study Variables

	Range				Shape Descriptors		
Variable	Potential	Actual	*M*	*SD*	*Skewness*	*Kurtosis*	*n*
FCAT_SS_2012	163 - 279	163 - 279	217.69	17.48	−.39	.52	5,348
FCAT_SS_2013	170 - 284	170 - 284	222.59	17.67	−.52	.65	5,348
SCHOOL_GRADE	0 - 4	0 - 4	2.43	1.34	−.23	**−1.19**	5,348
GPA	0 - 4.0	0 - 4.0	2.51	.91	−.32	−.35	5,348
AVG_IBA	0 - 100.0	8.7 - 88.1	36.87	10.48	**.79**	.84	5,348
AVG_LSA	0 - 100.0	5.6 - 100.0	54.55	17.40	.16	−.61	5,348

Note. Some students had no scores for either or both IBAs and/or LSAs. Therefore, *n* was lower for the averages. Bolded values are possible violations of normality assumptions.

Homoscedasticity. Scatterplots (see Figures 8 & 9) and boxplots (see Figure 10) were visually inspected to determine whether the relationship between each independent variable and the dependent variable was homoscedastic, or evenly dispersed. In other words, this assumption tests whether the variance in the dependent variable is spread across the range of values for each independent variable. This is important because predictions in multiple regression are based on variance in the dependent variable, and where heteroscedasticity exists, predictions will be better at some levels of the independent variable than at others (Hair et al., 2010). Scatterplots were graphed for metric variables such as GPA and checked for an elliptical distribution of points, indicating an equal dispersion of observations. All of the scatterplots roughly resembled ellipses.

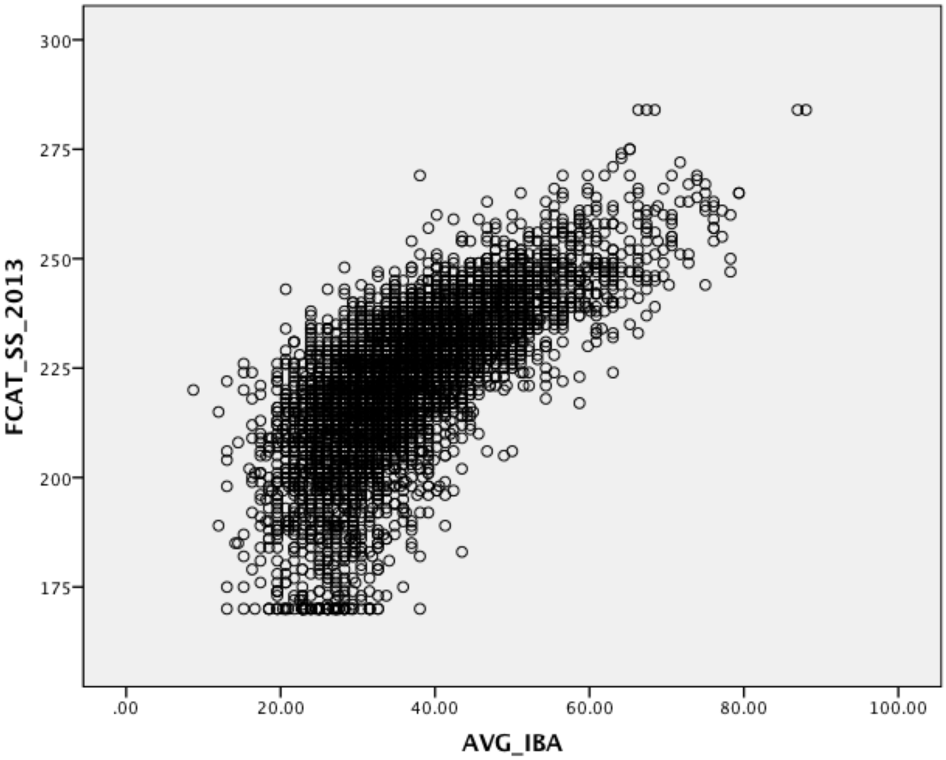


Figure 8. Illustrative example of a scatterplot used to visually inspect homoscedasticity for metric independent variables. This scatterplot depicts the ellipsoid-shaped relationship between average IBA scores and 2013 Grade 6 Mathematics FCAT 2.0 developmental scale scores.

The present research model accounted for the heteroscedastic relationship between FCAT_SS_2012 and FCAT_SS_2013 (see Figure 9) by including the second research question, which examined the regression one FCAT achievement level at a time.

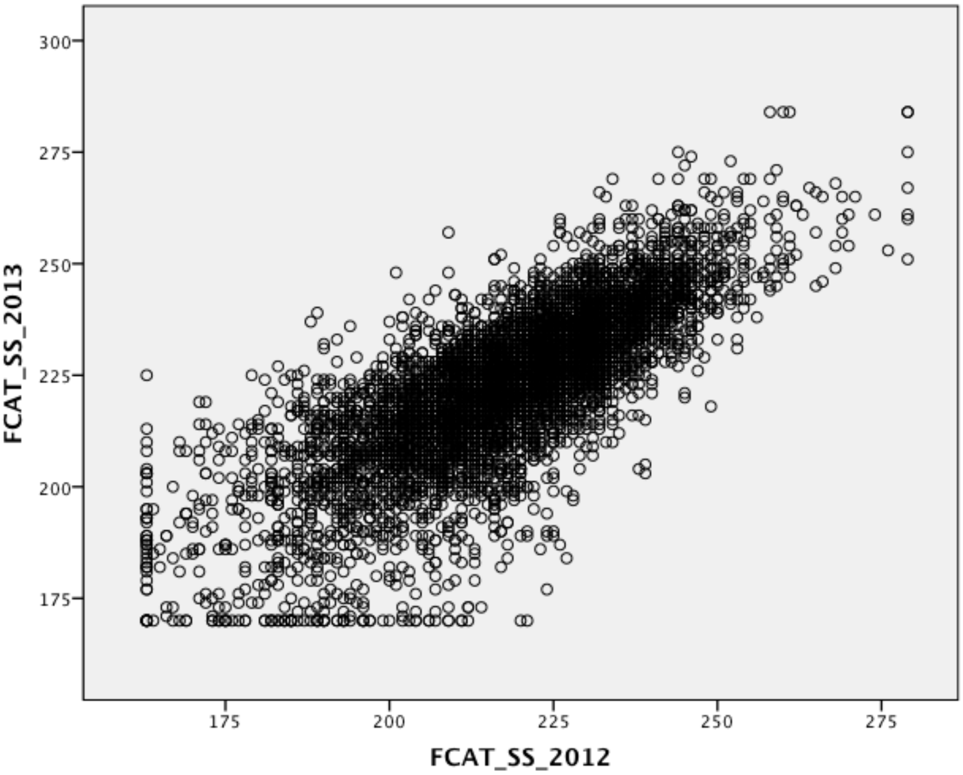


Figure 9. Scatterplot depicting the heteroscedastic relationship between 2012 Grade 5 Mathematics FCAT 2.0 and 2013 Grade 6 Mathematics FCAT 2.0 developmental scale scores. Note the cone shape characterized by a large dispersion closer to the origin and a smaller dispersion at the opposite side. Research question two analyzed the relationship between the predictor variables in research question one and the dependent variable, 2013 Grade 6 Mathematics FCAT 2.0 scale scores one 2012 Grade 5 Mathematics FCAT achievement level at a time.

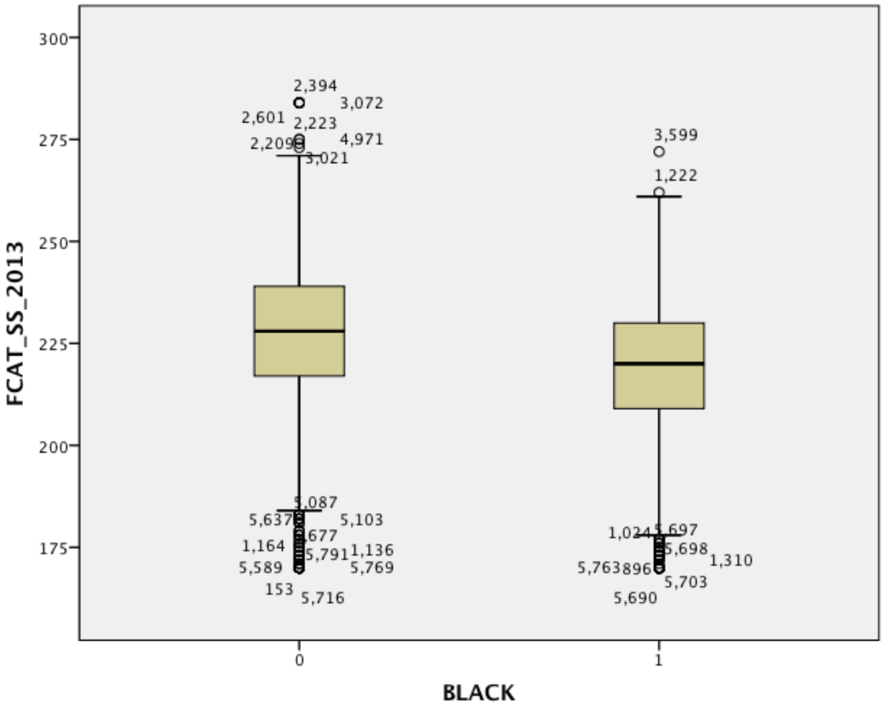


Figure 10. Illustrative example of a boxplot used to visually inspect homoscedasticity for categorical independent variables. Outliers are identified by case numbers, demonstrating the floor and ceiling effects common among test score distributions.

Boxplots (see Figure 10) were graphed for nominal and ordinal variables such as demographics and examined for similar lengths in boxes and whiskers between groups. All of the boxplots were similar in lengths except for AMER_IN, as only 14 out of 5,801 students, or .2% of the sample, were classified as Non-Hispanic American Indian or Alaska Native (see Table 9).

Table 9

Demographics by 2012 Grade 5 Mathematics FCAT Achievement Level

2012 FCAT Level	ELL	ESE	FREE_ RED	GENDER		BLACK	HISP	WHITE	MULTI	AMER _IN
				Female	Male					
1	97	252	848	624	566	750	109	259	40	1
2	33	180	1032	909	845	987	157	496	63	3
3	23	153	942	970	929	784	200	743	88	9
4	6	48	297	365	413	216	65	410	33	1
5	1	31	43	84	96	26	7	122	5	0
Total	160	664	3162	2952	2849	2763	538	2030	229	14

Linearity. Scatterplots with superimposed lines of best fit were also examined for linearity for each of the scaled metric independent variables (see Figure 11). In cases where linearity was questionable, quadratic curves of best fit were graphed and the difference in fit was noted (see Figure 12). Another consideration was that keeping the original variables allows for easier interpretation of the statistical model results. Examination of the scatterplots did not reveal any apparent nonlinear relationships, thus preserving the assumption of linearity for the individual variables.

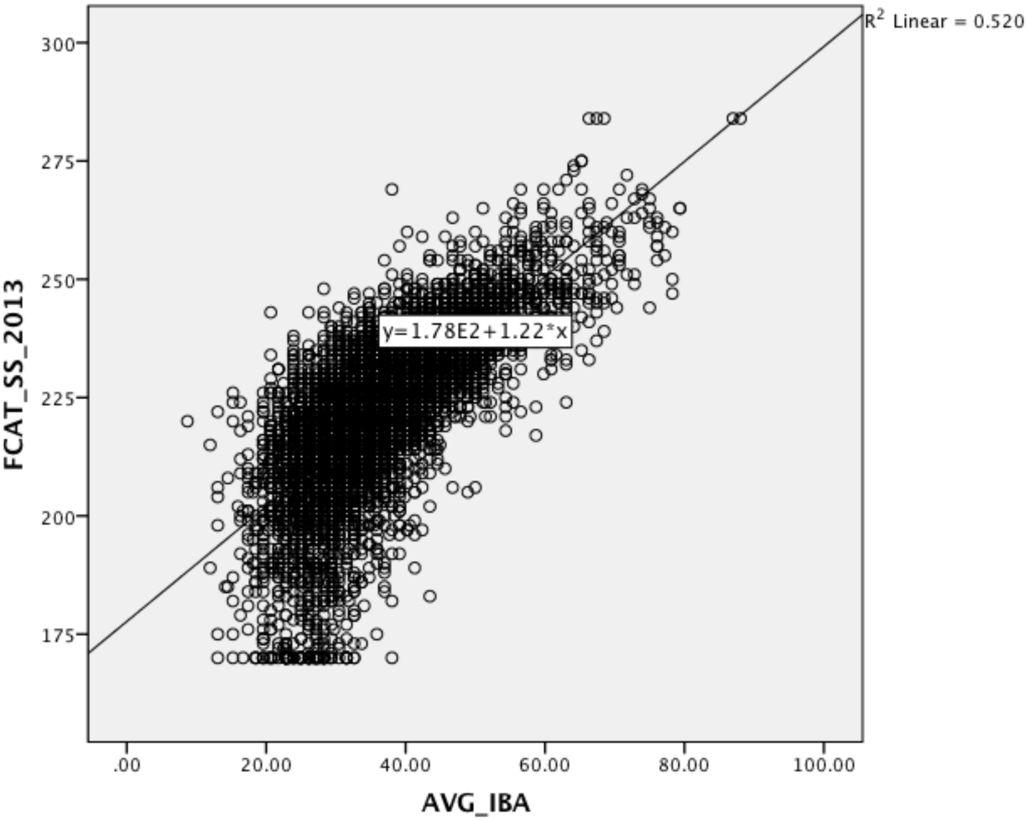


Figure 11. Illustrative example of a scatterplot with a superimposed line of best fit used to visually inspect linearity.

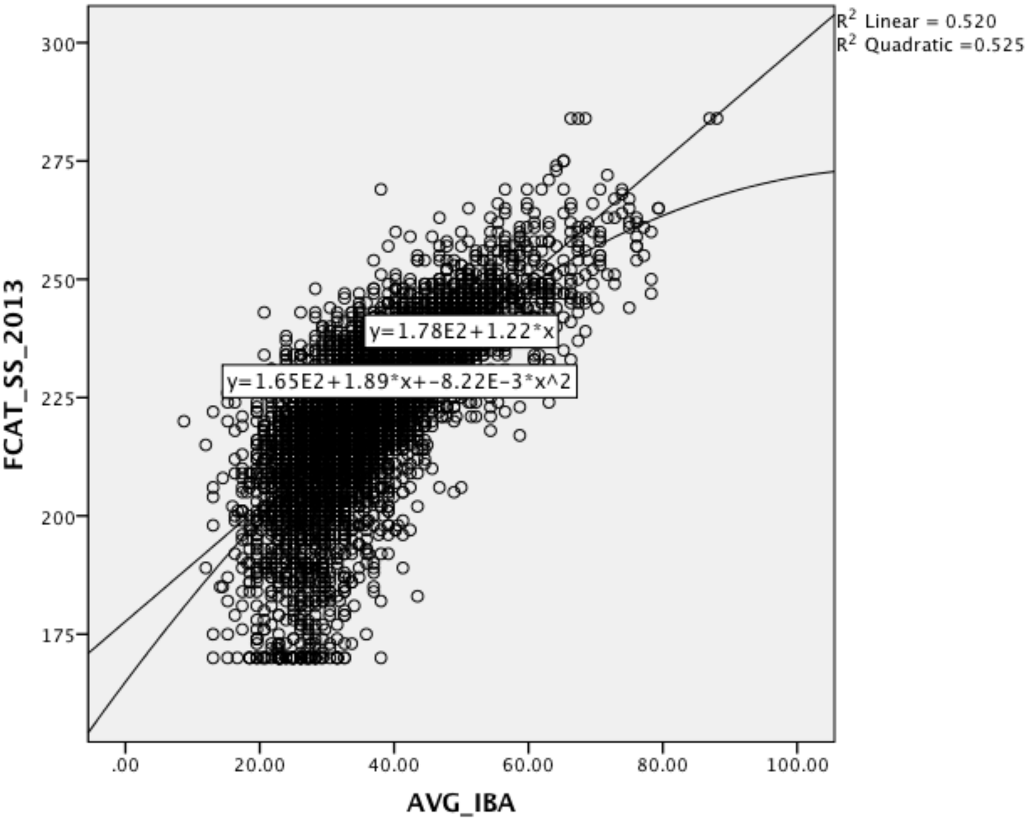


Figure 12. Illustrative example of a scatterplot with a superimposed line of best fit and a quadratic curve of best fit used to visually inspect whether a transformation might be appropriate. In this case, the curve's fit is negligibly better ($\Delta R^2 = .005$) than the line, so the AVG_IBA data remained untransformed in the model.

Absence of correlated errors. The practice of inferential statistics has an inherent amount of measurement error. However, errors present in relationships between variables should not be correlated; otherwise, an unaccounted for factor may be affecting the results. This basic assumption of independence of errors was addressed with residuals after the multiple regression model was estimated, and the discussion occurs in that later section. Graphical and statistical tests of assumptions revealed few violations. Where violations did exist, they were relatively minor and should not present any serious problems in the course of the data analysis.

Analysis of Regression Models

A confirmatory approach was employed when constructing the regression model. Independent predictor variables were entered manually into five blocks for each of the six regression models. In each case, the enter method was used for each block. This method was chosen over a stepwise process because a stepwise estimation method increases the probability that the regression model will be affected by multicollinearity (Hair et al., 2010). Additionally, interpretation of the variate is more straightforward when using the enter method.

Testing variates for assumptions. The regression variates were then tested for meeting underlying regression assumptions. Although individual variables were tested for normality, homoscedasticity, linearity, and the absence of correlated errors, the variate includes the collective effect of variables and therefore must also meet these same assumptions. To test assumptions for the variates, residuals were examined. Residuals, or errors, in this case were the differences between the predicted developmental scale scores for FCAT_SS_2013, and the actual developmental scale scores.

Standardized residuals for the overall model were plotted against the predicted FCAT_SS_2013 scores (see Figure 13) and examined for violations of assumptions. The scatterplot resembled the null plot, an indicator that the four basic assumptions were met. Residuals fell randomly for the most part, with fairly equal dispersion and no strong tendency to be either greater or less than zero. The floor effect of standardized testing can be seen in the linear limit on the bottom left corner of the scatterplot. The points along this line represent scores that were predicted to be higher than they were. Variation for lower predicted scores was slightly greater than variation for higher predicted scores, indicating slight heteroscedasticity.

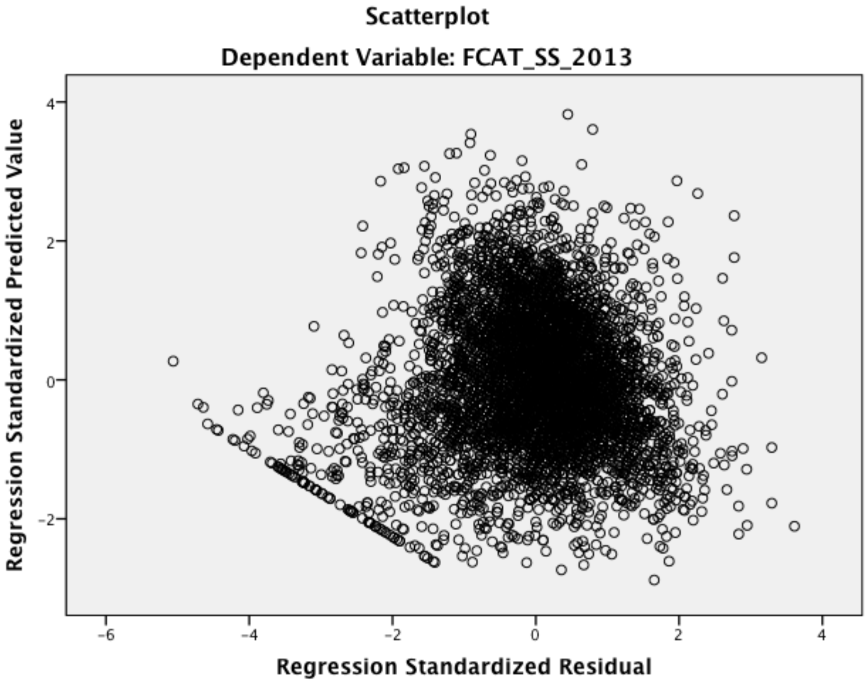


Figure 13. Scatterplot of standardized predicted values versus standardized regression residuals.

Standardized residuals for the other regression models were also plotted against the predicted FCAT_SS_2013 scores, sorted by 2012 Grade 5 Mathematics FCAT achievement levels (see Figure 14 for an example depicting 2012 FCAT Level 2). At first glance, these five scatterplots seemed to display greater variation in residuals than in predicted values. However, upon further inspection of the scales in the scatterplot, it was determined that the scale for predicted values was smaller than that of the residual, leading to a visual "lengthening" of the graph. Another observation was that each model included a small group of outliers (see the arrow in Figure 14) with large negative residuals, indicating that the actual 2013 Grade 6 Mathematics FCAT 2.0 developmental scale score was much lower than the predicted score, based on the independent variables in the model. This could be attributed to test anxiety, unpreparedness, illness on the day of the test, or some other socio-emotional reaction to the state testing process.

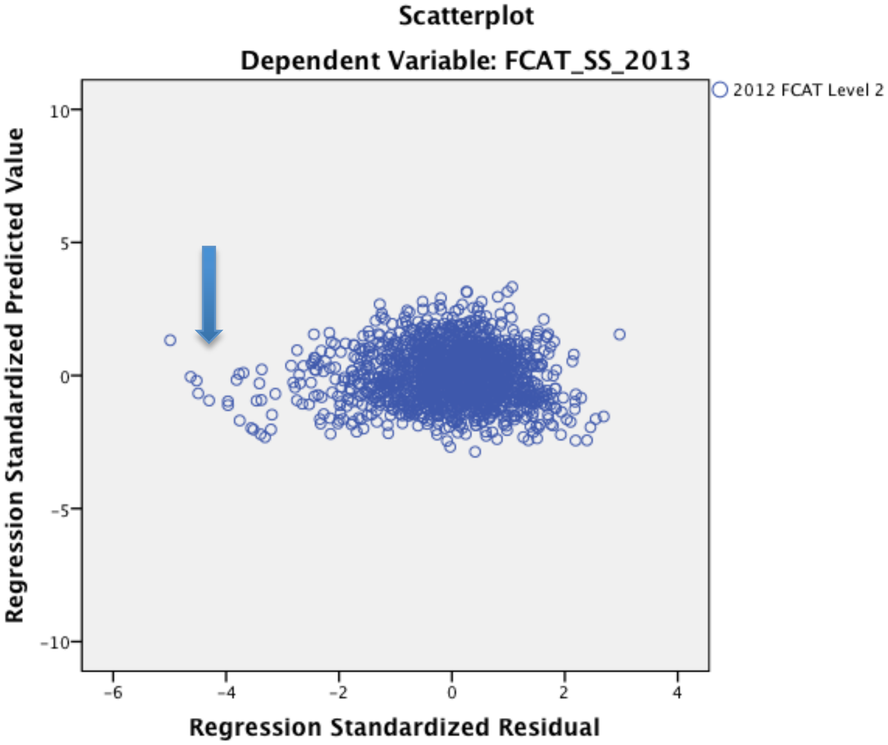


Figure 14. Example scatterplot of standardized predicted values versus standardized regression residuals for students who achieved Level 2 on the 2012 Grade 5 Mathematics FCAT 2.0.

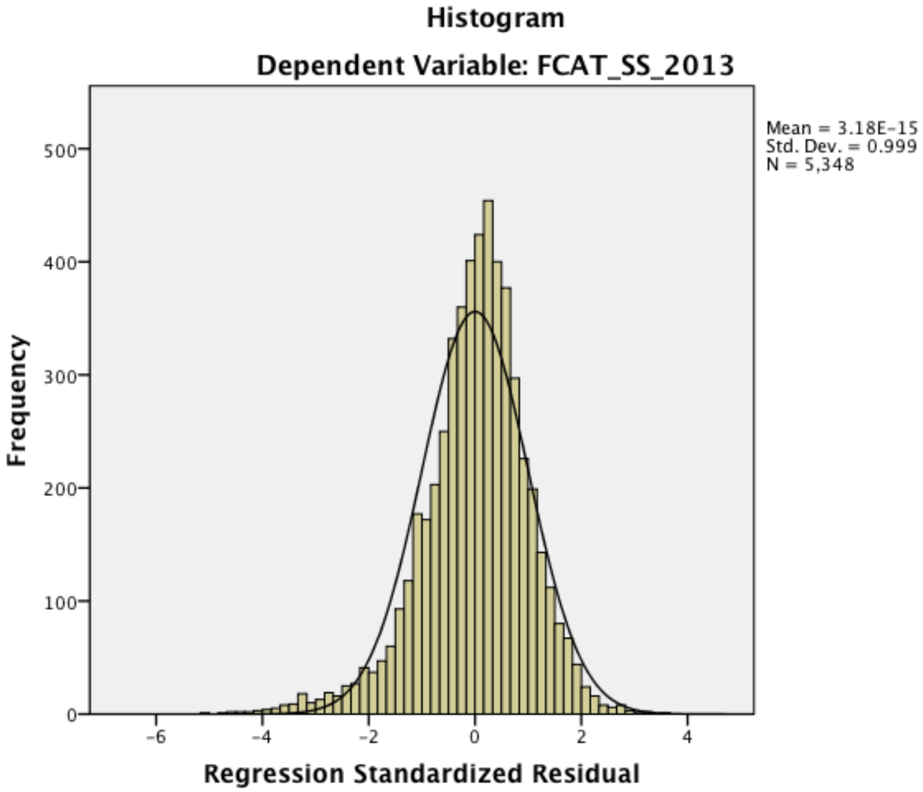


Figure 11. Standardized residual histogram with superimposed normal curve.

To test normality of the residuals, frequency histograms with superimposed normal curves were visually inspected. All six of the histograms depicted distributions approximating the normal distribution (see Figure 11 for an example). The model for 2012 FCAT Level 5 students was the least normal, due to a small sample (*n* = 162) and a resulting ill-formed distribution. In that case, the normal probability plot was inspected for a match between the plotted residuals and the straight line depicting the normal probability distribution (see Figure 12). The normal probability plot depicts residuals falling below the normal line at the top scores of Level 5. This means that at the highest possible scores, the distribution is more leptokurtic (flatter) than expected, which is equivalent to the ceiling effect commonly seen on standardized tests. In other words, the test is limited in that it cannot measure the highest levels of achievement so we see more students with the top score than scores below.

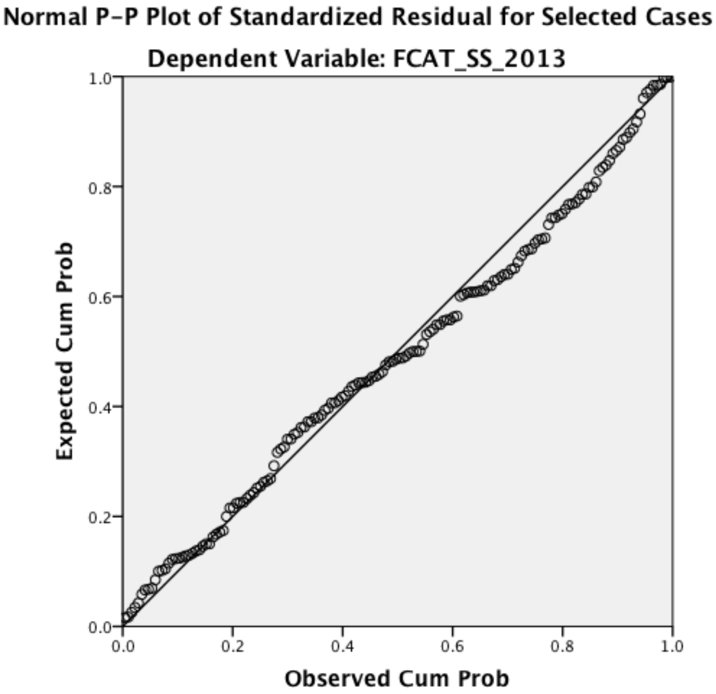


Figure 12. Normal probability plot of the standardized residuals for the 2012 FCAT Level 5 regression model.

Statistical significance of the regression models. The coefficient of determination, R^2, is used to test the statistical significance of the overall model. R^2, also known as the effect size or coefficient of determination, describes the amount of variation explained by the regression model. In a regression with blocks, each block can be considered a separate regression model, where each block is cumulative. The change in the coefficients of determination, ΔR^2, can then be interpreted as the amount of variation explained by each subset of variables added to the prior block. In this way, potential moderating variables are statistically controlled and ΔR^2 for the last block represents only the effect of the interest variables.

The first study null hypothesis, H_{01}, was that $R^2 = 0$ for the first block of the regression model corresponding to research question one (RQ1). This hypothesis was rejected in favor of the alternate hypothesis, as demographic variables account for 16% of the variance in 2013 Grade 6 Mathematics FCAT 2.0 developmental scale scores (see Table 10). The *F* ratio tests the statistical significance of this hypothesis; because $F = 110.75$ and $p < .001$ for the first block, the first block of this regression model is expected to be significant in multiple samples from the population and not just the sample in the study (Hair et al., 2010).

The same was true for all of the null hypotheses referring to the first research question. The second null hypothesis, H_{02}, that school grade would have no additional effect on the variance, was rejected because school grade explains an additional 6% of the variance ($\Delta R^2 = .06$) and was statistically significant ($\Delta F = 435.19$, $p < .001$). Student GPA contributed a considerable and statistically significant 18% to the explanatory power ($\Delta R^2 = .18$, $\Delta F = 1{,}575.63$, $p < .001$), resulting in the rejection of H_{03}. The largest amount of variance, 25%, was explained by the prior year's FCAT score ($\Delta R^2 = .25$, $\Delta F = 3{,}868.44$, $p < .001$) and resulted in the rejection of H_{04}. Finally, the variables of interest, interim averages, explained another 7% of the variance in 2013 FCAT scores ($\Delta R^2 = .07$, $\Delta F = 623.93$, $p < .001$), resulting in the rejection of H_{05}. Of the total variance in 2013 Grade 6 Mathematics FCAT 2.0 developmental scale scores, 72% was explained by the combination of demographic variables, school grade, student GPA, prior year test score, and interim averages.

Table 10

Multiple Regression Model Summary - RQ1

Block	Predictors	R^2	Standard Error of Estimate	ΔR^2	ΔF
1	Demographics[a]	.16[c]	16.24	.16	110.75**
2	School Grade	.22	15.62	.06	435.19**
3	GPA	.40	13.72	**.18**	1,575.63**
4	Grade 5 FCAT	.65	10.45	**.25**	3,868.44**
5	Interim Averages[b]	.72	9.41	.07	623.93**

[a]Demographic variables included race/ethnicity, ELL status, ESE status, gender, and Free/Reduced lunch status.
[b]Interim averages included the average of IBA scores and the weighted average of LSA scores.
[c]Adjusted R^2 values were equivalent to all R^2 values.
** $p < .001$.

Null hypotheses relating to research question two (RQ2) were similar to the previous five hypotheses from research question one. Five models were constructed to address research question two (see Table 11). The 2012 Grade 5 Mathematics FCAT 2.0 achievement level (FCAT_AL_2012) was used as a selection variable, resulting in one regression model for each of the five achievement levels. All null hypotheses regarding a zero effect size were rejected. The effect size of the Level 5 model was high (R^2 = .53), indicating that for students who scored a Level 5 in the previous year, the complete set of predictors explain the variation in FCAT developmental scale scores very well. The other achievement levels also had moderate explanatory power, with decreasing effect sizes as the achievement levels approached 1. The effect size of the Level 1 model was the lowest, but was still moderate (R^2 = .30).

Table 11

Multiple Regression Model Summary for Block 5 of each 2012 FCAT Achievement Level – RQ2

2012 FCAT Achievement Level	R^2	Adjusted R^2	Standard Error of Estimate	ΔR^2	ΔF
1	.30	.29	12.51	.11	89.59**
2	.34	.34	9.61	.16	197.57**
3	.47	.47	7.56	.19	307.55**
4	.49	.48	6.78	.21	136.19**
5	.53	.49	7.95	.13	19.75**

** $p < .001$.

The change in R^2 from Block 4 to Block 5 was analyzed for each of the five models to determine the effect of including interim averages. The FCAT Level 4 group had the highest effect change ($\Delta R^2 = .21$), followed by Level 3 ($\Delta R^2 = .19$), Level 2 ($\Delta R^2 = .16$), Level 5 ($\Delta R^2 = .13$), and Level 1 ($\Delta R^2 = .11$). In other words, the interim scores had less of an effect on predicting FCAT scores for students who scored at the extremes (Levels 1 & 5) than for those that were in the middle (Levels 2, 3, & 4). All of these changes were statistically significant at the .001 significance level.

The adjusted coefficient of determination (adjusted R^2) accounts for the natural rise in R^2 as a result of additional, even nonsignificant, predictor variables. The adjusted R^2 measure is a way to relate the level of overfitting in a regression model. In these models, the adjusted R^2 values are within .01 of the R^2 values except for in the Level 5 model (R^2 - adjusted $R^2 = .04$). Statistical significance of beta weights for individual predictors in the Level 5 model were reviewed for further analysis, which is included in a later section.

Interpreting the regression variate and coefficients. When a statistically significant effect size exists among predictors, regression coefficients can be used to interpret the type (positive or negative) and the strength of the relationship between independent and dependent variables in the regression variate (Hair et al., 2010). However, coefficients are calculated using the original scale of the data. To allow for comparison between coefficients of various units of measurement, SPSS also calculates beta weights, β, which are standardized regression coefficients.

Standardized beta weights. Beta weights are shown in Table 12 for the last block in the overall regression model, Block 5. For interpretative purposes, beta weights with an absolute value of .2 or more are displayed in bold typeface. The largest contribution to the predicted value came from prior year test scores, FCAT_SS_2012 (β = .38, t = 32.62, p < .001), followed by the weighted average of LSA scores, AVG_LSA (β = .28, t = 22.51, p < .001), and the average of IBA scores, AVG_IBA (β = .22, t = 19.23, p < .001). Each type of average interim score contributed more to the predictive equation than student GPA, school grade, or any individual student demographic.

Table 12

Regression Beta Weights – Block 5 for RQ1 (overall model)

Predictor	β
ELL	−.02*
ESE	−.07**
FREE_RED	.00
GENDER	.00
BLACK	.05
HISP	.02
WHITE	.04
MULTI	.02
AMER_IN	.00
SCHOOL_GRADE	.03*
GPA	.05**
FCAT_SS_2012	**.38****
AVG_IBA	**.22****
AVG_LSA	**.28****

* p < .01. ** p < .001.

According to beta weights, race/ethnicity, gender, and free/reduced lunch status were not a statistically significant portion of the prediction equation for Block 5 of the initial regression model. The remaining factors, ELL and ESE status, school grade, and student GPA were statistically significant but were not practically significant relative to the largest three contributors. Three predictor variables, free/reduced lunch status, gender, and American Indian, had a standardized coefficient of approximately zero. The standardized regression equation for Block 5 of the regression model addressing research question one was:

$$\widehat{Z}_{Y_{All}} = (-.02 * Z_{ELL}) + (-.07 * Z_{ESE}) + (.05 * Z_{BLACK}) + (.02 * Z_{HISP}) + (.04 * Z_{WHITE}) \\ + (.03 * Z_{SCHOOL\ GRADE}) + (.05 * Z_{GPA}) + (.38 * Z_{FCAT\ SS\ 2012}) \\ + (.22 * Z_{AVG\ IBA}) + (.28 * Z_{AVG\ LSA}) + \varepsilon_{All}$$

For the second research question, five models were analyzed. Beta weights are shown in Table 13 for the last block in each regression model, Block 5. For interpretative purposes, beta weights with an absolute value of .2 or more are displayed in bold typeface within Table 13. These included FCAT_SS_2012 (β = .27, p < .001), AVG_LSA (β = .26, p < .001), and AVG_IBA (β = .22, p < .001) for the 2012 Level 1 model, similar to the initial regression model of all students. However, the prior year test scores, FCAT_SS_2012, had relatively smaller beta weights for the Level 2(β = .14, p < .001), 3(β = .13, p < .001), and 4 (β = .10, p < .001) models. The highest beta weights for each of these models were the interim averages, AVG_LSA ($\beta_{Level\ 2}$ = .35, p < .001; $\beta_{Level\ 3}$ = .36, p < .001; $\beta_{Level\ 4}$ = .32, p < .001) and AVG_IBA ($\beta_{Level\ 2}$ = .24, p < .001; $\beta_{Level\ 3}$ = .29, p < .001; $\beta_{Level\ 4}$ = .32, p < .001). Interestingly, in the Level 5 model, the largest beta weight was AVG_IBA (β = .40, p < .01), followed by FCAT_SS_2012 (β = .19, p < .001). AVG_LSA (β = .07, p < .001) did not contribute much to the 2013 prediction for students who scored a Level 5 in 2012.

Table 13

Regression Beta Weights and Structure Coefficients by 2012 FCAT Level – Block 5 for RQ2 (by Levels)

	2012 Grade 5 Mathematics FCAT 2.0 Achievement Level									
	1		2		3		4		5	
Predictor	β	r_s	β	r_s	β	r_s	β	r_s	β	r_s
ELL	-.04	-.11	-.01	.00	-.03	-.07	-.05	-.09	.01	-.01
ESE	-.12**	-.40	-.08**	-.25	-.08**	-.20	-.05	-.12	.06	.35
FREE_RED	-.02	-.11	.01	-.06	-.03	-.19	.02	-.11	.00	-.22
GENDER	-.01	.07	-.01	.04	.02	.09	.01	.10	-.01	-.04
BLACK	-.07	-.04	-.02	-.12	-.11	-.21	-.11	-.13	-.08	-.28
HISP	-.02	.00	-.02	.02	-.05	-.04	-.07	-.02	-.03	-.11
WHITE	-.04	.05	-.02	.11	-.09	.15	-.12	.04	-.05	.19
MULTI	.00	.05	-.03	-.04	-.05	.00	-.07	-.05	-.02	-.18
AMER_IN	.01	.04	-.01	-.01	.00	.03	.00	.03	[a]	[a]
SCHOOL_GRADE	.05	.29	.02	.25	.06	.32	.05	.26	.03	.22
GPA	.05	.35	.06*	**.50**	.07**	**.57**	.12**	**.62**	.30	**.60**
FCAT_SS_2012	**.27****	**.73**	.14**	**.51**	.13**	**.53**	.10**	**.51**	.19**	**.66**
AVG_IBA	**.22****	**.63**	**.24****	**.69**	**.29****	**.75**	**.32****	**.82**	**.40***	**.84**
AVG_LSA	**.26****	**.71**	**.35****	**.84**	**.36****	**.86**	**.32****	**.87**	.07**	**.70**

[a] n = 0 for Level 5 American Indian

* $p < .01$. ** $p < .001$.

ELL status, race/ethnicity, gender, and free/reduced lunch status were not a statistically significant portion of the prediction equation for Block 5 of each regression model. The remaining factors, ESE status, school grade, and student GPA were statistically significant but were not practically significant relative to the largest three contributors. Three predictor variables, free/reduced lunch status, gender, and American Indian, had a standardized coefficient of approximately zero. The standardized equations for Block 5 of each regression model addressing research question two were:

$$\begin{aligned}\widehat{Z}_{Y_{Level\ 1}} = {} & (-.04 * Z_{ELL}) + (-.12 * Z_{ESE}) + (-.02 * Z_{FREE\ RED}) + (-.01 * Z_{GENDER}) \\ & + (-.07 * Z_{BLACK}) + (-.02 * Z_{HISP}) + (-.04 * Z_{WHITE}) + (.01 * Z_{AMER\ IN}) \\ & + (.05 * Z_{SCHOOL\ GRADE}) + (.05 * Z_{GPA}) + (.27 * Z_{FCAT\ SS\ 2012}) \\ & + (.22 * Z_{AVG\ IBA}) + (.26 * Z_{AVG\ LSA}) + \varepsilon_{Level\ 1}\end{aligned}$$

$$\begin{aligned}\widehat{Z}_{Y_{Level\ 2}} = {} & (-.01 * Z_{ELL}) + (-.08 * Z_{ESE}) + (.01 * Z_{FREE\ RED}) + (-.01 * Z_{GENDER}) \\ & + (-.02 * Z_{BLACK}) + (-.02 * Z_{HISP}) + (-.02 * Z_{WHITE}) + (-.03 * Z_{MULTI}) \\ & + (-.01 * Z_{AMER\ IN}) + (.02 * Z_{SCHOOL\ GRADE}) + (.06 * Z_{GPA}) \\ & + (.14 * Z_{FCAT\ SS\ 2012}) + (.24 * Z_{AVG\ IBA}) + (.35 * Z_{AVG\ LSA}) + \varepsilon_{Level\ 2}\end{aligned}$$

$$\begin{aligned}\widehat{Z}_{Y_{Level\ 3}} = {} & (-.03 * Z_{ELL}) + (-.08 * Z_{ESE}) + (-.03 * Z_{FREE\ RED}) + (.02 * Z_{GENDER}) \\ & + (-.11 * Z_{BLACK}) + (-.05 * Z_{HISP}) + (-.09 * Z_{WHITE}) + (-.05 * Z_{MULTI}) \\ & + (.06 * Z_{SCHOOL\ GRADE}) + (.07 * Z_{GPA}) + (.13 * Z_{FCAT\ SS\ 2012}) \\ & + (.29 * Z_{AVG\ IBA}) + (.36 * Z_{AVG\ LSA}) + \varepsilon_{Level\ 3}\end{aligned}$$

$$\begin{aligned}\widehat{Z}_{Y_{Level\ 4}} = {} & (-.05 * Z_{ELL}) + (-.05 * Z_{ESE}) + (-.02 * Z_{FREE\ RED}) + (.01 * Z_{GENDER}) \\ & + (-.11 * Z_{BLACK}) + (-.07 * Z_{HISP}) + (-.12 * Z_{WHITE}) + (.07 * Z_{MULTI}) \\ & + (.05 * Z_{SCHOOL\ GRADE}) + (.12 * Z_{GPA}) + (.10 * Z_{FCAT\ SS\ 2012}) \\ & + (.32 * Z_{AVG\ IBA}) + (.32 * Z_{AVG\ LSA}) + \varepsilon_{Level\ 4}\end{aligned}$$

$$\begin{aligned}\widehat{Z}_{Y_{Level\ 5}} = {} & (.01 * Z_{ELL}) + (.06 * Z_{ESE}) + (-.01 * Z_{GENDER}) + (-.08 * Z_{BLACK}) \\ & + (-.03 * Z_{HISP}) + (-.05 * Z_{WHITE}) + (-.02 * Z_{MULTI}) + (.03 * Z_{AMER\ IN}) \\ & + (.30 * Z_{SCHOOL\ GRADE}) + (.19 * Z_{GPA}) + (.40 * Z_{FCAT\ SS\ 2012}) \\ & + (.07 * Z_{AVG\ IBA}) + (.01 * Z_{AVG\ LSA}) + \varepsilon_{Level\ 5}\end{aligned}$$

Structure coefficients. Again, assuming a sufficient omnibus effect size, it is appropriate to investigate the contribution from individual predictors. Because beta weights are context-specific and sensitive to collinearity, some researchers have argued for the interpretation of structure coefficients in addition to beta weights (e.g., Courville & Thompson, 2001; Henson, 2002; Onwuegbuzie & Daniel, 2001; Thompson, 2006; Thompson & Borrello, 1985). Conceptually, a structure coefficient, r_S, is a correlation between a predictor variable and the predicted values ($\hat{Y}$) of the dependent variable. When squared, structure coefficients determine how much variance each predictor contributes to $\hat{Y}$. Structure coefficients can be calculated by dividing the correlation between each predictor variable and the dependent variable ($r_{X,Y}$ in Table 14) by the correlation between the actual and predicted values of the dependent variable (*R* = .847 for Block 5 of the overall regression model). The resulting structure coefficient values are displayed in Tables 14 and 15 in the column titled r_S. The equation used to calculate each structure coefficient was (Courville & Thompson, 2001, p. 238):

$$r_S = \frac{r_{X,Y}}{R}$$

Table 14

Regression Beta Weights, Correlations with FCAT_SS_2013, and Structure Coefficients – Block 5 for RQ1 (overall model)

Predictor	β	$r_{X,Y}$	r_s	r_s^2
ELL	−.02*	−.13**	−.15	.02
ESE	−.07**	−.19**	−.23	.05
FREE_RED	.00	−.22**	−.26	.07
GENDER	.00	.00	.00	.00
BLACK	.05	−.24**	−.29	.08
HISP	.02	−.01	−.01	.00
WHITE	.04	.21**	.25	.06
MULTI	.02	.01	.01	.00
AMER_IN	.00	.02	.02	.00
SCHOOL_GRADE	.03*	.36**	.42	.18
GPA	.05**	.53**	**.62**	.38
FCAT_SS_2012	**.38****	**.78****	**.92**	.85
AVG_IBA	**.22****	**.72****	**.85**	.72
AVG_LSA	**.28****	**.75****	**.88**	.77

* $p < .01$. ** $p < .001$.

The largest structure coefficients in the overall model were the same as the largest beta weights: 2012 Grade 5 FCAT scores (r_s = .92, β = .38), the weighted average of LSA scores (r_s = .88, β = .28), and average IBA score (r_s = .85, β = .22). This was also true for the Level 1 and Level 2 models. However, GPA in Level 3 had a larger structure coefficient (r_s = .57, β = .07) than FCAT_SS_2012 (r_s = .53, β = .13). The same was true in the Level 4 model: GPA (r_s = .62, β = .12) had a larger structure coefficient than FCAT_SS_2012 (r_s = .51, β = .10). For Level 5, the largest structure coefficient and beta weight was average IBA score (r_s = .84, β = .40).

Collinearity was detected in three areas by observing a near-zero beta weight with a considerably larger structure coefficient (Courville & Thompson, 2001, p. 239): (a) GPA in the overall model (r_s = .62, β = .05), (b) SCHOOL_GRADE in the overall model (r_s = .42, β = .03), and (c) AVG_LSA in the Level 5 model (r_s = .70, β = .07). In all of these cases, the independent variable plays a role in explaining the variance in 2013 FCAT scores, but the variable is collinear or multicollinear with at least one additional predictor variable. In other words, these predictors were useful in predicting 2013 FCAT scores, but any shared predictive power was arbitrarily assigned to another predictor when calculating beta weights. This finding is not surprising considering that a number of the predictor variables were, in effect, some measure of previous achievement; hence, these several predictors were all measuring within the same domain.

Beta weights and structure coefficients of GENDER, HISP, MULTI, and AMER_IN are close to zero in every model, so these variables are not practical predictors of 2013 FCAT Grade 6 Math scores, regardless of 2012 FCAT Level. There is one exception: MULTI in the FCAT Level 5 model (r_s = -.18, β = -.02). This is possibly explained by the small cell size (n = 5; see Table 7).

No suppressor variables were noted. These variables "improve the predictive power of the other independent variables in the model by suppressing variance that is irrelevant to the prediction, as a result of the suppressor variable's relationship with the other independent variables" (Onwuegbuzie & Daniel, 2001, Multiple regression section, para. 7).

Assessing multicollinearity. Multicollinearity among three or more predictor variables, or, similarly, collinearity between two predictor variables, is an issue with data and not necessarily the regression model. However, because either can interfere with interpretation of the variate, it is important to identify any collinearity and its impact on the results so that remedies can be applied where necessary. Examining correlation coefficients between pairs of predictors is the quickest means for identifying collinearity (not multicollinearity) among independent variables in the data set. Pearson correlation coefficients were reviewed (Table 15) for independent variables that were highly correlated with the dependent variable, a desirable trait for a predictor variable, but also moderately or highly correlated with another independent variable, indicating collinearity. Using the rule of thumb in Hinkle, Weirsma, and Jurs (2003, p. 109), correlation coefficients of |.70| or greater were identified as high correlations and are represented in bold typeface on Table 15.

Table 15

Correlation Coefficients for RQ1

Variable	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1. FCAT_SS_2013 (dependent)	--														
2. ELL	-.13**	--													
3. ESE	-.19**	-.03	--												
4. FREE_RED	-.22**	.04**	.02	--											
5. GENDER	.00	.01	-.12**	.00	--										
6. BLACK	-.24**	-.13**	-.03	.28**	.02	--									
7. HISP	-.01	.25**	-.01	.00	.01	-.30**	--								
8. WHITE	.21**	-.06**	.03	-.25**	-.02	-.70**	-.23**	--							
9. MULTI	.01	-.02	.04**	-.02	-.01	-.20**	-.06**	-.15**	--						
10. AMER_IN	.02	-.01	.01	-.04*	.01	-.05**	-.02	-.04*	-.01	--					
11. SCHOOL_GRADE	.36**	-.07**	-.02	-.26**	-.02	-.29**	.01	.26**	.01	.02	--				
12. GPA	.53**	-.03	-.10**	-.20**	.12**	-.21**	.02	.13**	-.03	-.03	.22**	--			
13. FCAT_SS_2012	.78**	-.16**	-.15**	-.24**	-.04*	-.26**	-.01	.24**	.02	.01	.34**	.48**	--		
14. AVG_IBA	.72**	-.12**	-.12**	-.21**	-.05**	-.26**	-.03	.24**	.01	.01	.30**	.46**	.71**	--	
15. AVG_LSA	.75**	-.09**	-.13**	-.23**	.03	-.25**	.01	.20**	.02	.02	.43**	.64**	.69**	.67**	--

*$p < .01$. ** $p < .001$.

Three predictor variables, FCAT_SS_2012 ($r = .78$, $p < .001$), AVG_LSA ($r = .75$, $p < .01$), and AVG_IBA ($r = .72$, $p < .001$), had high positive correlations with the dependent variable, FCAT_SS_2013. This was expected as each of these variables had the largest contribution to the regression equation. However, there were two high correlations between (a) FCAT_SS_2012 and AVG_IBA ($r = .71$) and (b) WHITE and BLACK ($r = -.70$). The high positive correlation between the 2012 FCAT scores and the average IBA scores represents one measure of IBA data reliability, considering the predictive purpose of the IBA test and the structure mimicking FCAT format. The high correlation also indicated collinearity in the data set from these two predictor variables. The high negative correlation between WHITE ($n = 2{,}030$) and BLACK ($n = 2{,}763$) was attributed to the fact that these were the largest racial demographic groups, and the variables are mutually exclusive (students are not able to select more than one racial category; multiracial students select “MULTI”). Another consideration was that ASIAN, a much smaller proportion (3.8%) of the student sample, was the reference category for “dummy” coding of the race/ethnicity variable. Perhaps if either WHITE or BLACK were used as the reference category, this collinearity would not have been an issue. Further, dummy coded variables are not linear by nature, and, therefore, correlations among coded columns for the same variable generally indicate differences more so than statistical relationship.

Correlations were also examined for the five models pertaining to research question two (see Appendix C). In all five models, only one statistically significant high correlation existed: BLACK and WHITE in the Level 2 model ($r = -.72$). The explanation is similar to the one for the overall model. All other observations mirrored the correlations for the overall model as well.

Tolerance and VIF. Correlation coefficients between two predictor variables are not sufficient to detect multicollinearity, which includes the combined effect of three or more predictor variables. Tolerance, and its inverse, the variance inflation factor (VIF) are frequently used to identify multicollinearity (Hair et al., 2010). Tolerance is the amount of variance of the predictor variable that is not explained by the other predictors. For example, GENDER had a high tolerance of .95 (see Table 16), meaning other predictors such as free/reduced lunch status or average IBA scores explained only 5% of the variance in gender. This means that gender did not contribute to multicollinearity in the model.

Table 16

Multicollinearity Diagnostic Measures – Block 5 for RQ1

Predictor	Tolerance	*VIF*	$\sqrt{VIF}$
ELL	.86	1.16	1.08
ESE	.95	1.05	1.03
FREE_RED	.86	1.16	1.08
GENDER	.95	1.05	1.02
BLACK	**.13**	**7.63**	**2.76**
HISP	.33	3.07	1.75
WHITE	**.15**	**6.73**	**2.59**
MULTI	.50	2.01	1.42
AMER_IN	.94	1.07	1.03
SCHOOL_GRADE	.75	1.33	1.15
GPA	.56	1.80	1.34
FCAT_SS_2012	.39	2.54	1.59
AVG_IBA	.42	2.37	1.54
AVG_LSA	.34	2.96	1.72

Hair, Black, Babin, and Anderson (2010) recommend a .10 cutoff for tolerance values and a corresponding VIF cutoff of 10 for multicollinearity diagnosis. None of the values met those cutoffs; however, the variables BLACK and WHITE had tolerance levels of .13 and .15 respectively, meaning 87% of the variance in Black designation and 85% of the variance in White designation were explained by other predictors. These two predictors also had larger *VIF* values of 7.63 and 6.73, respectively. The correlation between BLACK and WHITE was high and negative (*r* = -.70), whereas the correlation between each of these and the dependent variable were low (*r* = -.24 and .21). In other words, these two variables have more of a relation to each other than they do to the 2013 FCAT scores, a potential indicator of multicollinearity in the data.

The square root of VIF indicates the factor by which the standard error is increased as a result of multicollinearity (Hair et al., 2010). The $\sqrt{VIF}$ measures for BLACK and WHITE were over 2, indicating a more than doubled standard error. Again, the majority of students (4,793 out of 5,801) in the data set were either categorized as Black (*n* = 2,763) or White (*n* = 2,030), and the groups were mutually exclusive.

Similar multicollinearity diagnostics for the five models addressing research question two are shown in Table 17. There were two *VIF* values exceeding 10, and both were for the independent variable BLACK in the Level 1 and Level 2 models. The predictor variable WHITE also had high *VIF* values for Level 1 and Level 2. Because the majority of students were in these racial/FCAT categories (see Table 4), the chances for collinearity are also high in these pairings.

Table 17

Multicollinearity Variance Inflation Factors (VIF) – Block 5 for RQ2

	FCAT	FCAT	FCAT	FCAT	FCAT
	Level 1	Level 2	Level 3	Level 4	Level 5
Predictor	*VIF*	*VIF*	*VIF*	*VIF*	*VIF*
ELL	1.69	1.08	1.05	1.03	1.10
ESE	1.13	1.06	1.04	1.02	1.18
FREE_RED	1.03	1.11	1.11	1.19	1.27
GENDER	1.07	1.05	1.05	1.09	1.19
BLACK	**12.13**	**10.39**	7.03	3.86	2.31
HISP	4.44	3.95	3.23	2.04	1.30
WHITE	**9.13**	**8.83**	6.63	4.10	2.50
MULTI	2.59	2.39	2.10	1.56	1.36
AMER_IN	1.06	1.08	1.13	1.04	1.16
SCHOOL_GRADE	1.13	1.20	1.22	1.19	1.32
GPA	1.29	1.38	1.51	1.58	1.48
FCAT_SS_2012	1.21	1.12	1.20	1.20	1.74
AVG_IBA	1.14	1.18	1.33	1.49	1.94
AVG_LSA	1.35	1.58	1.84	1.87	1.10

Analysis of Covariance (ANCOVA)

To address the third research question dealing with how the differences in number of interim assessment scores effected FCAT achievement, an analysis of covariance (ANCOVA) model was utilized. FCAT_SS_2013 was again the dependent variable, whereas the independent variables included two categorical descriptors of the number of 2013 IBA tests taken (NUM_IBA) and the number of 2013 LSA tests taken (NUM_LSA). Students who had 0 or 1 IBA score were categorized as “Low” for NUM_IBA, whereas students who had 3 IBA scores were categorized “High.” Students who had 0, 1, 2, or 3 LSA scores were categorized as “Low” for NUM_LSA, whereas students who had 6, 7, or 8 LSA scores were categorized “High.”

In addition, the present study statistically controlled for a covariate, the 2012 Grade 5 Mathematics FCAT 2.0 developmental scale score (FCAT_SS_2012). ANCOVA is appropriate for analyzing intact groups of cases. Further, ANCOVA allows the researcher to increase the precision of a statistical comparison of group means by partitioning out variance attributed to a covariate, which ideally results in a smaller error variance (Hinkle, Wiersma, & Jurs, 2003). The null hypothesis for research question 3, H_{011}, was: After adjusting for 2012 Grade 5 Mathematics FCAT 2.0 developmental scale scores, the means of 2013 Grade 6 Mathematics FCAT 2.0 developmental scale scores will be equal to each other ($\mu'_1 = \mu'_2 = \mu'_3 = \mu'_4$).

Testing assumptions. Several assumptions were tested prior to using ANCOVA. First, assumptions for ANOVA were tested: (a) the observations are random and independent, (b) the dependent variable is normally distributed, and (c) homogeneity of variance. Although the students were not randomly assigned to each group, the cases selected out of each group were random and the cells were balanced. This could have potentially introduced Type I error; however, "ANOVA is robust with respect to the violation of the assumptions, particularly when there are large and equal numbers of observations in each cell of the factorial [model]" (Hinkle et al., 2003, p. 409). The dependent variable, FCAT_SS_2013 was normally distributed (see Figure 7). Finally, a Levene's test indicated that the error variance of the dependent variable was homogeneous across groups, F (3,396)= .54, p = .66, which met the homogeneity of variance assumption at the α = .01 statistical significance level used in the ANCOVA model.

Assumptions pertaining to the covariate in ANCOVA were tested in addition to the assumptions of ANOVA. The covariate was tested for linearity using a scatterplot (see Figure 9). The ellipsoid shape indicated that the relationship between the covariate, FCAT_SS_2012, and the dependent variable, FCAT_SS_2013, would not be better described as nonlinear. Second, the correlation between the covariate and the dependent variable, FCAT_SS_2013, was examined and found to be high (r = .78, p < .001). Further, because the

IBA and LSA data came from the 2012-2013 school year and the covariate, FCAT_SS_2012, was collected in April of 2012, the covariate was unaffected by the independent variables, NUM_IBA and NUM_LSA.

Descriptive statistics. Several observations were made from the descriptive statistics (see Table 18). The 100 students in Group 3, with a low number of IBA scores and a high number of LSA scores, had the highest mean 2013 Grade 6 Mathematics FCAT 2.0 developmental scale score (*M* = 222.63, *SD* = 18.07). The lowest mean belonged to Group 4, the group with the low number of both types of interim scores (*M* = 212.09, *SD* = 20.00). With a high number of both IBA and LSA test scores, Group 1 had better average 2013 FCAT scores (*M* = 213.52, *SD* = 16.74) than the group with a low number of both tests, but worse than Groups 1 or 2 with a combination of low and high numbers of interim scores.

Further, Group 1 with a high number of both types of interim assessment had the highest average 2012 Grade 5 score (*M* = 219.07, *SD* = 14.65) and dropped to the second worst performing group. Groups 2 and 4 stayed relatively flat. Group 3 had the most improvement from 2012 to 2013 (*M* = 208.33 to 222.63).

Table 18

Descriptive Statistics for RQ3

Group	NUM_IBA	NUM_LSA	2012 Grade 5 Math FCAT 2.0 Scores		2013 Grade 6 Math FCAT 2.0 Scores		*n*
			M	*SD*	*M*	*SD*	
1	High	High	**219.07**	14.65	213.52	16.74	100
2		Low	217.96	23.40	217.41	21.17	100
		Total	213.42	20.48	215.47	19.13	200
3	Low	High	208.33	18.78	**222.63**	18.07	100
4		Low	208.87	15.92	212.09	20.00	100
		Total	213.70	17.64	217.36	19.73	200
	Total	High	213.97	16.09	218.08	17.97	200
		Low	213.15	21.71	214.75	20.71	200
		Total	213.56	19.09	216.41	19.43	400

Overall (see Figure 15), students who took a lower number of IBAs scored better on average (M = 217.36, SD = 19.73) than those who took a higher number (M = 215.47, SD = 19.13), and students who took a higher number of LSAs scored better on average (M = 218.08, SD = 17.97) than those who took a lower number (M = 214.75, SD = 20.71).

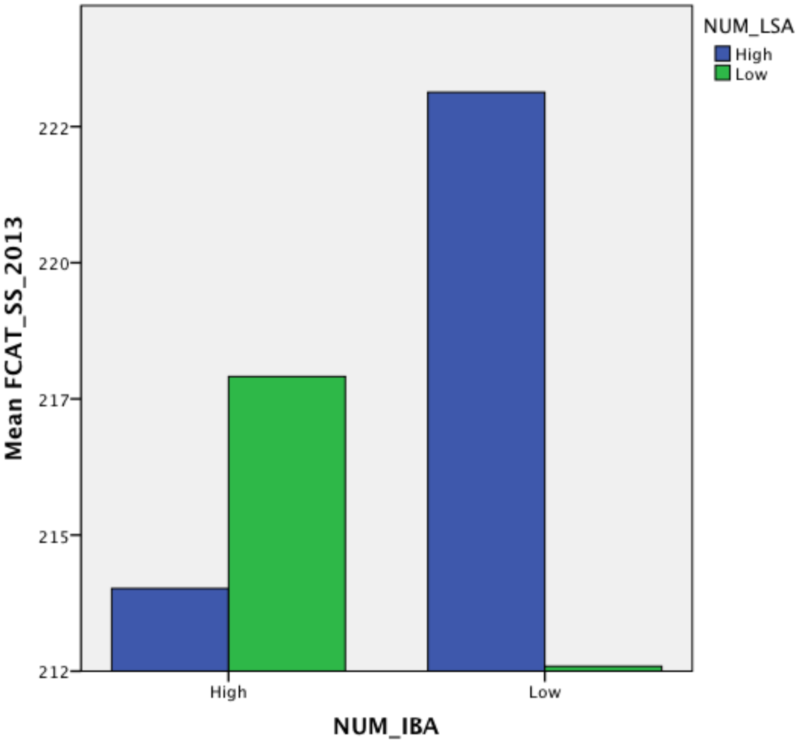


Figure 15. Bar graph of ANCOVA interactions. The *y*-axis scale is based on the range of 2013 Grade 6 Mathematics FCAT 2.0 developmental scale scores in the data set. Note that the highest average was the group with the Low number of IBA scores and High number of LSA scores.

Statistical significance of the ANCOVA model. The last null hypothesis, H_{011}, was rejected as the means are not the same. Effect size (see Table 19) for the covariate (η^2 = .599, F(1,395) = 590.052, p < .001) was statistically significant at the α = .01 level, yet the effect sizes for NUM_LSA (η^2 = .012, F(1,395) = 4.86, p < .05), NUM_IBA (η^2 = .005, F(1,395) = 1.90, p = .17), and the interaction between numbers of IBA and LSA scores (η^2 = .001, F(1,395) = .31, p = .58) were very low and not statistically significant. Further, the observed power for NUM_IBA (.12) and the interaction between NUM_IBA and NUM_LSA (.02) mean that there would only be a 12% and 2% chance, respectively, of finding a statistically significant difference in any particular sample of 400 students, assuming that the variance in this sample is demonstrative of other samples in the population. However, there is a 35% chance that another sample including NUM_LSA would find a statistically significant difference between groups. Overall, the model explained 62% of the variance in 2013 FCAT scores.

Table 19

ANCOVA Results for RQ3

Source	Sum of Squares	*df*	Mean Square	*F*	η^2	Power
Corrected Model	92,940.74[a]	4	23,235.19	158.92**	.62	1.00
Intercept	6,233.51	1	6,233.51	42.63**	.10	1.00
FCAT_SS_2012	86,270.45	1	86,270.45	590.05**	.60	1.00
NUM_IBA	278.06	1	278.06	1.90	.01	.12
NUM_LSA	710.72	1	710.72	4.861*	.01	.35
NUM_IBA * NUM_LSA	45.39	1	45.39	.31	.00	.02
Error	57,752.20	395	146.21			
Total	18,884,441.0	400				
Corrected Total	150,692.94	399				

* $p < .05$. ** $p < .001$.

[a] $R^2 = .62$. Adjusted $R^2 = .61$.

In addition to numerical analysis, a matrix of scatterplots (see Figure 16) was examined to visually verify that the ANCOVA model demonstrated a linear correlation and was relatively free of assumption violations. The scatterplots depicting the predicted and observed values represents a strong positive correlation and is elliptical in shape, supporting the linear function used in the model. A nonlinear model would not be more suitable to the data set. Standardized residuals for the model were also plotted against the predicted FCAT_SS_2013 scores and examined for violations of ANOVA assumptions. The scatterplots resembled the null plot, an indicator that the three basic assumptions were met. Residuals fell randomly for the most part, with fairly equal dispersion and no strong tendency to be either greater or less than zero.

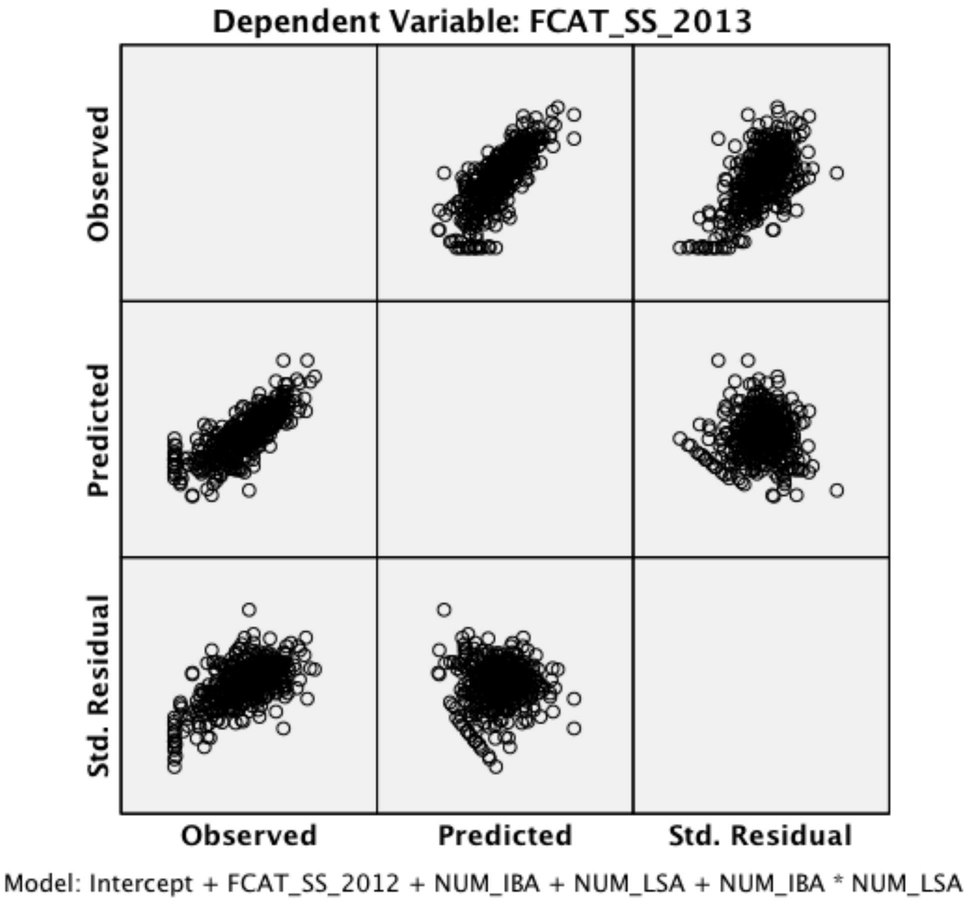


Figure 16. Scatterplot matrix for ANCOVA. Includes standardized residuals, predicted values, and observed values.

In summary, after statistically controlling for the 2012 Grade 5 Mathematics FCAT 2.0 developmental scale scores, which explained 60% out of the 62% variance in 2013 FCAT scores, there were no statistically significant main or interaction effects for NUM_IBA or NUM_LSA. Although students who had a low number of IBA scores (0 or 1 score) scored higher than students who had a high number of IBA scores (3 scores), they did not score significantly higher. Results showed that those who had a high number of LSA scores (6, 7, or 8 scores) scored higher on the 2013 FCAT than students with a low number of LSA scores (0, 1, 2, or 3); however, again this difference was not statistically significant at the .01 level. Additionally, there was no statistically significant NUM_IBA by NUM_LSA interaction. Despite the lack of statistical significance, scores in Group 3, with a low number of IBA scores and a high number of LSA scores, had the highest overall average 2013 Grade 6 Mathematics FCAT developmental scale score.

Summary

Multiple regression results and discussion. The first regression model including all students in the regression sample (n = 5,348) had the highest predictive power (R^2 = .72). In other words, a combination of the 14 predictor variables including student demographics, school grade, GPA, prior year FCAT, and interim averages, were able to explain 72% of the variance in 2013 FCAT scores. The other models were not as powerful, in part because of the reduced sample size. The largest explained variance out of the FCAT achievement level models was the Level 5, with a large effect size of .53, followed by moderate effect sizes in the Level 4 (R^2 = .49), Level 3 (R^2 = .47), Level 1 (R^2 = .36), and Level 2 (R^2 = .34) models.

Collinearity and multicollinearity are likely issues within all models, particularly between BLACK and WHITE predictors, as well as between the AVG_LSA, AVG_LSA, and FCAT_SS_2012 predictors. Beta and structure coefficients were larger for LSA scores than for IBA scores in all models except for Level 5, implying that LSA scores had a larger capacity to explain variance in FCAT scores for all groups except students who scored a Level 5 on the 2012 FCAT.

For the overall model, 65% of the variance in 2013 FCAT scores was explained by the combination of demographic variables, student GPA, and prior year scores. Adding interim scores to the model resulted in an additional 7% of explanatory power ($\Delta R^2 = .07$), which was statistically significant. Interim scores added the most predictive power to the achievement level models for Levels 3 ($\Delta R^2 = .19$) and 4 ($\Delta R^2 = .21$), and added the least to models for the extremes – FCAT Levels of 1 ($\Delta R^2 = .11$) or 5 ($\Delta R^2 = .13$).

ANCOVA results and discussion. A combination of 2012 FCAT scores and number of interim assessments explained 62% of the variance in 2013 FCAT scores. After controlling for 2012 FCAT scores, the number of interim assessments explained only 2% of the variance in 2013 FCAT scores. The group with the highest mean ($M = 222.63$, $SD = 18.03$) had a low number of IBA scores and a high number of LSA scores during the course of the 2012-2013 school year. Neither number of IBA scores nor number of LSA scores had a statistically significant effect, nor did the interaction between number of IBA and number of LSA scores.

In summary, after statistically controlling for the 2012 Grade 5 Mathematics FCAT 2.0 developmental scale scores, which explained 60% out of the 62% variance in 2013 FCAT scores, there were no statistically significant main or interaction effects for NUM_IBA or NUM_LSA. Although students who had a low number of IBA scores (0 or 1 score) scored higher than students who had a high number of IBA scores (3 scores), they did not score significantly higher. Results showed that those who had a high number of LSA scores (6, 7, or 8 scores) scored higher on the 2013 FCAT than students with a low number of LSA scores (0, 1, 2, or 3); however, again this difference was not statistically significant at the .01 level. Additionally, there was no statistically significant NUM_IBA by NUM_LSA interaction. Despite the lack of statistical significance, scores in Group 3, with a low number of IBA scores and a high number of LSA scores, had the highest overall average 2013 Grade 6 Mathematics FCAT developmental scale score.

Chapter 4 included a report of the results of data analyses. The presentation included a summary of findings for six multiple regression models and ANCOVA. Collinearity was present within the data set, specifically between BLACK and WHITE predictor variables and between IBA and prior year FCAT scores. Interim assessment scores contributed a statistically significant 7% out of 72% total FCAT variance explained. LSAs had a larger capacity to explain variance in FCAT scores than did IBAs. Interim scores contributed more to the predictive models for the middle FCAT Levels than for the extremes. Again, LSAs explained more FCAT score variance than IBAs for every Level except Level 5. The highest performing group was Group 3, with a low number of IBAs and a high number of LSAs. Group 3 also had the largest difference in average FCAT score from 2012 to 2013. Chapter 5 includes a discussion about the study's findings, an overview of possible policy implications, and recommendations for future research studies.

CHAPTER 5

Discussion

The purpose of the present study was to evaluate the predictive ability and frequency-based achievement outcomes of two distinctly different yet simultaneously administered district-developed interim assessment programs. Chapter 5 begins with a review of key aspects of the design, sample, and instruments used regarding two aspects of interim assessment evaluation: predictive utility and achievement outcome differences. A discussion regarding findings for each research question is followed by overall results for each regression predictor variable. Implications for educational policymakers and leaders and recommendations for future research were also included to inspire ideas about how to take next steps toward a greater understanding of interim assessment utility. The chapter concludes with a summary.

Review of Study Design and Methods

The present study examined student and school characteristics as predictors of 2013 Grade 6 Mathematics FCAT 2.0 developmental scale scores. The study included regression modeling to explore the predictive power of two interim assessment programs versus other predictors such as prior year scores, student GPA, student demographics, and school grade. The regression sample was also stratified into prior year achievement levels to explore differences in interim assessment predictive power by level of past academic performance. Finally, an ANCOVA model provided insight into the instructional utility of the two types of interim assessment by comparing achievement score means for equal-sized subsets of the original sample with low or high numbers of each type of interim assessment after controlling for prior year performance. In each model, the dependent variable was 2013 Grade 6 Mathematics FCAT 2.0 developmental scale scores.

Sample and sources of data. The original data set included 9,038 records of students enrolled in either a standard or advanced Grade 6 Mathematics course during the 2012-2013 school year. For the regression models, the set was further refined to remove those with missing average interim scores, as these were predictors in the regression equation (*n* = 5,348). ANCOVA required a balanced cell design, thus 100 cases were randomly selected from each of the four groups of interest (*n* = 400).

Two distinctly different types of interim assessment programs were administered over the course of the same school year in the district of focus: Interim Benchmark Assessment (IBA) and Learning Schedule Assessment (LSA). The first program, IBA, had been in place for nine years (T. Ballentine, personal communication, October 23, 2013) and was constructed as a "dipstick" check of students' mastery to predict performance on the state test. The test consisted of 48 multiple-choice items covering all course benchmarks. Due to budgetary constraints on paper and printing costs, the same form of the test was administered three times in the school year: Fall, Winter, and Spring just before the state test. Reliability for data on IBAs collected from a 2011 cohort of students was moderate (KR-20 = .75); however, this value was low compared to FCAT 2.0 reliability coefficient, .93, for the 2013 cohort of students (FDOE, 2013f). Subject matter experts reviewed items and the test both during development and after each administration year to ensure content validity, or the test's alignment with standards. Uniform administration dates and procedures contributed to construct validity, which was measured in part by using an externally developed ex post facto test and item discrimination analysis. These reports were not available to classroom teachers.

The second program, LSA, was only in its second year of implementation and had an instructional purpose, with each test aligning to one of the district's Learning Schedule modules. The LSA tests were 10-20 multiple-choice items each, and available online or on paper. The LSA program was more flexible than IBA in terms of administration and test security; teachers and school-based educators decided when to give the LSAs and had access to the tests after administration, whereas the IBAs were given in a strict district-wide window of time and were kept locked until the end of the school year. Also, professional development was offered by the district of focus to address all aspects of the LSA program—from item and test development to administration, reporting, and iterative lesson study. Point-biserial values, one measure of reliability, were available for the LSAs to every classroom teacher. However, the test reliability coefficients were not calculated at the time of the present study. Subject matter experts reviewed items and the test during the extended iterative development process to ensure content validity. A curriculum-aligned administration schedule contributed to construct validity.

Research Question Answers and Discussion

Research question one. To what extent can variance in middle school student scores on mathematics high-stakes state tests be explained by scores on district interim assessments after controlling for prior scores, student demographic variables, and teacher-assigned grades?

Results of regression analyses indicated a moderate to high degree of correlation among the variables of interest. Overall, 65% of the variance in 2013 Grade 6 Mathematics

FCAT 2.0 developmental scale scores was explained by the combination of demographic variables, student GPA, and prior year scores. Adding interim scores to the model resulted in an additional 7% of explanatory power, which was statistically significant. Structure coefficients were larger for LSA scores (r_s = .88) than for IBA scores (r_s = .85), implying that LSA scores had a larger capacity to explain variance in FCAT scores than IBA scores. Two questions follow from these results.

Is 7% practically significant? Because 65% of state standardized test scores were explained using historical student data, which require no additional interim testing, some might argue that decisions about district-wide intervention and extension plans should be made based on these factors alone. Test administration usually prohibits instruction for at least one class period each, and assessment programs with strict security policies intended to mimic the high-stakes state testing environment often disrupt entire school days at a time. The IBA assessment program is one of these strict security programs, or what Gong called a *state-test mirror* (2010). In the 2012-2013 school year, students in the district of focus had the opportunity to experience 141 instructional days prior to the start of FCAT. In the case of 6th grade Mathematics, this meant 8 LSA tests plus 3 IBA administrations consumed roughly 11 class periods, or 8% of the available instructional time, to administer tests. The question stands: is 8% of instructional time lost to interim testing worth the added 7% of additional explanatory power?

If the data are used instructionally and formatively at the classroom level and beyond, perhaps the instructional time lost in the short-term is compensated by a clearer direction for future instruction. Assessments such as the LSA tests are better suited for this type of ex post facto analysis because they are open to teachers to use in their classrooms after the tests are administered and the content is limited to one instructional unit.

Teachers will most likely develop their own chapter or unit tests independently if none are provided by the district, and many times, will continue to do so even in addition to district-developed tests. They realize the benefit of assessment for students both individually as they become increasingly metacognitive given sufficient feedback, and collectively as they experience modified instructional lesson plans tailored to their changing needs, either in the form of re-teaching or extension. District-developed interim assessment programs are uniquely beneficial in that results for any one test can be aggregated and compared across teachers, class periods, schools, and geographical groups of schools (Perie et al., 2009). This allows for mid-year programmatic and policy changes, and provides feedback to school and district administrators about the state of instruction within classrooms and schools for follow-up visits and discussions. Additionally, public members and representatives such as school board members want to know how students are doing to hold superintendents or other administrators accountable.

Another common concern is that given the context of high-stakes testing (Nichols et al., 2010), consequential accountability (Hanushek & Raymond, 2005), and teacher performance pay within the U.S., educators place a high value on state standardized test scores (Darling-Hammond, 2004; Goertz, 2007; Hamilton, 2003; Kim, 2010; Linn, 2000; Ravitch, 2010) and, therefore, measures to aide in predicting those scores. The working hypothesis is that if educators have a sufficiently good prediction and adequate time to alter instruction, they can improve the potential future of lower-performing students. Given the competitive and stressful situation that is reality for many who work in the field of education today, it is no surprise that some choose to forgo instructional time in order to measure what students know at a particular point in time and how that relates to how they may perform in the future. For these educators, the 7% additional explanatory power might be worth the opportunity cost of lost instructional time.

Why were the LSA assessments better predictors than the IBA assessments? The IBA assessment program was intended solely to have a predictive purpose. In fact, test security requirements for the IBA program prohibited what is typically seen as instructional use. In other words, the items were not available after the tests were administered to principals, teachers, or students and therefore could not be applied to altering lesson plans or providing specific feedback to students. During development, the IBA tests were designed based on FCAT blueprints and test item specifications. They included the same proportion of items for each group of standards as the FCAT. The formatting was as close as possible to the FCAT, down to the instructions read verbally by the proctor prior to the test. Administrators wanted to mimic the environment of the state standardized test as closely as possible to get the best prediction about how students would perform at the end of the school year. Some also wanted to increase students' testing stamina because test fatigue was a stated concern for some schools in the district of focus. Some educators also viewed the IBA assessment program as a drill for everyone in the school district to identify and work through any complications ahead of the actual FCAT administration in April.

It is surprising, then, that even with all of the protocols ensuring that the test was as much like the FCAT as possible, it did not predict student performance on the FCAT better than the instructionally-purposed LSA assessment program. Some might say that an interim assessment program such as the IBA is superfluous at best, and detrimental to instruction at worst. This might be because after eight years of implementation, teachers and students were desensitized to the IBA tests and the process surrounding them. Perhaps the teachers felt resentment that the payoff was not sufficient to warrant several days of lost instruction per school year (Dee et al., 2013) and repeated FCAT-like drills. Teachers were more in control of the LSA program from the beginning of the item development process. Teachers decided what went into the tests, how they were delivered, and when their students took each test. Teachers became further invested as they analyzed data collected from the LSA tests in groups as part of professional learning communities and other professional development. Overall, teachers were treated as professional educators in the fundamental philosophy of LSA program, and as simply proctors in the IBA program. The reciprocated attitude from teachers had the potential to carry over to student performance on the tests. In other words, if teachers took the test seriously, their students took the test seriously.

Another possible explanation for the IBA program's lower predictive power is that the students took the same exact IBA test three times in the same school year. Item and/or answer memorization had a higher probability in this case than if each administration consisted of an equated form. Score inflation, a false representation of what the student knows about the subject, might have been one result of this memorization phenomenon (Hamilton, 2003; Koretz & Beguin, 2010; Linn, 2000), leading to an overestimation of FCAT scores and a reduction in the predictive power and validity of the IBA program.

Research question two. To what extent can variance in middle school student scores within achievement levels on mathematics high-stakes state tests be explained by scores on district interim assessments after controlling for prior scores, student demographic variables, and teacher-assigned grades?

In general, interim assessment scores added the most predictive power to models for 2012 Grade 5 Mathematics FCAT 2.0 Achievement Level 3 ($\Delta R^2 = .19$) and Level 4 ($\Delta R^2 = .21$), and added the least to models for the extremes: FCAT Level 1 ($\Delta R^2 = .11$) and Level 5 ($\Delta R^2 = .13$). LSA scores explained more variance in 2013 FCAT scores (ranging from 18 to

37%) than did IBA scores for every 2012 FCAT Level except Level 5 (IBA explained 37% versus LSA explained 26%). Two questions arise from these findings.

Why did interim scores predict the FCAT scores for students in the center better than for students on the extremes? One possible explanation for this phenomenon is that teachers tend to focus more on increasing students who are just under the proficiency cut point. These students are also called *bubble students* and are purposefully targeted by many schools because they are considered the easiest ones to move from non-proficient to proficient (Booher-Jennings, 2005; Jacob, 2005; Moon et al., 2007; Neal & Schanzenbach, 2010). However, if this strategy worked, the students in this central group would have much more variance on the FCAT and the interim scores would not have been better predictors. In this case, a better predictor meant not as much movement, or a greater correlation between the prior year and 2013 FCAT.

Perhaps a better explanation is the phenomenon of *regression to the mean*. Over time, scores from subsequent tests will tend to gravitate toward the average. This is not just the human socialization process at work, it is a statistical occurrence observed across all life forms. Higher performing students from the prior year will regress, and lower performing students will learn more. This higher motility means that it is harder to predict which students will remain in the group, and which students will move.

Another explanation is that the students on the extremes represent volatile student performance. In other words, the scores can be attributed to measurement error or ephemeral qualities such as test fatigue, illness on the day of the test, or testing environment issues.

Why were LSA scores worse at predicting Level 5 performance than IBA scores? This could be attributable to the student, the teacher, or both. Students who scored a Level 5 the prior year may have reacted to the greater length and challenge of the IBA, and not as much to the shorter LSA tests. Another possible explanation is that these students are more likely to recall prior questions and their answers given the same test three times in the school year, and are more likely to have researched the questions after the first administration of the test and are more likely to incorporate what they have learned. These would culminate in a higher average IBA score, regardless of the initial baseline score. However, as was already established, many times students who scored a Level 5 the prior year will fall back to a lower Level. This would mean the prediction is worse.

An alternate explanation was that because the IBA test was modeled after the FCAT, the students who took the test seriously and performed well, despite any negative connotation in the school or classroom, might have been more likely to also perform well on the FCAT. Students who scored poorly due to peer pressure, teacher persuasion, or indifference might have been more likely to perform worse on the FCAT.

Many times, students who scored a Level 5 on the prior year FCAT are attending the same schools or are in the same classes. These classes and schools tend to adhere to a condensed curriculum timeline, placing an even higher price on instructional time. Teachers under added stress of a faster pace may have been more reluctant to test frequently and might have preferred a less frequent test such as the IBA. Perhaps also the teachers of students in these schools or classes did not believe that the LSA tests were better than their own and did not take them as seriously. If the school administrator wanted the teachers to administer the LSA tests even so, the teachers might have done so out of compliance and not because they believed in the benefits of the program.

Yet another possible explanation exists in viewing not what happened with the students who scored a Level 5 the prior year, but with students who scored a Level 1 or 2 the prior year. By law, these students are required to be enrolled in remedial courses and given additional educational support. However, there is a wide variance among schools and even teachers in which safety nets are used, and how they are implemented. The 2012-2013 school year was also the second year of ESE inclusion, where students who had historically been in separate classrooms were integrated into standard classrooms. Many teachers struggled with how to reach the diverse needs in their classrooms. With all of these changes and variance in support, a larger, less frequent test such as the IBA might not capture the progress made over the course of the school year as well as frequent, unit-aligned tests such as the LSA.

Research question three. To what degree does achievement, as measured by mathematics high-stakes state tests, of middle school students who have experienced less frequently administered, predictive interim assessments differ from the achievement of students who have experienced more frequently administered, instructional assessments, after controlling for prior scores?

Research question three addressed the number of interim assessments taken (frequency), whereas research questions one and two addressed the average scores on the interim assessments (performance). A combination of prior year FCAT scores and number of interim assessments explained 62% of the variance in 2013 FCAT scores. After controlling for prior year FCAT scores, the number of interim assessments explained only 2% of the variance in 2013 FCAT scores for the ANCOVA sample (n = 400). The group with the highest mean (M = 222.63, SD = 18.03) had a low number of IBA scores and a high number of LSA scores during the course of the 2012-2013 school year. Neither number of

IBA scores nor number of LSA scores had a statistically significant effect, nor was the interaction between number of IBA and number of LSA scores statistically or practically significant.

There are several studies that point to a connection between an increase in testing frequency and criterion performance, up to a point of diminishing returns where more testing has a negative effect on academic achievement (Bangert-Drowns, Kulik, & Kulik, 1991; Başol & Johanson, 2009; Hausknecht, Halpert, DiPaolo, & Gerard, 2007). The findings about number of IBA scores in the present study do not fit in with these findings, however. Overall (see Figure 15), students who took a lower number of IBAs scored better on average (M = 217.36, SD = 19.73) than those who took a higher number (M = 215.47, SD = 19.13). Although students who had a low number of IBA scores (0 or 1 score) scored higher than students who had a high number of IBA scores (3 scores), they did not score significantly higher.

This contradictory result may stem from the educational context surrounding the test or the combination of both types of interim testing programs occurring in the same school year. The 100 students in Group 3, with a combination of low number of IBA scores and high number of LSA scores, had the highest mean 2013 Grade 6 Mathematics FCAT 2.0 developmental scale score (M = 222.63, SD = 18.07). The lowest mean belonged to group 4, the group with the low number of both types of interim scores (M = 212.09, SD = 20.00). This finding does fit in well with prior studies. With a high number of both IBA and LSA test scores, group 1 had better average 2013 FCAT scores (M = 213.52, SD = 16.74) than the group with a low number of both tests, but worse than groups 1 or 2 with a combination of low and high numbers of interim scores. This is similar to the point of diminishing returns seen in the literature (e.g., Bangert-Drowns et al., 1991).

Findings about the number of LSA scores were much more aligned with expectations based on the literature. Results showed (see Figure 15) that those who had a high number of LSA scores (6, 7, or 8 scores) scored higher on the 2013 FCAT (*M* = 218.08, *SD* = 17.97) than students with a low number of LSA scores (0, 1, 2, or 3; *M* = 214.75, *SD* = 20.71). However, this difference was not statistically significant at the .01 level.

Analysis of Predictors Across Regression Models and Discussion

Independent variables associated with student demographics (ELL, ESE, FREE-RED, GENDER, BLACK, HISP, WHITE, MULTI, AMER_IN) as well as with academics (GPA, SCHOOL_GRADE, FCAT_SS_2012, AVG_IBA, AVG_LSA) were examined to answer research questions pertaining to relationships between the independent variables and the 2013 Grade 6 Mathematics FCAT 2.0 developmental scale scores, FCAT_SS_2013.

Five of these variables were negatively correlated with the dependent variable, FCAT_SS_2013, (ELL, ESE, FREE_RED, BLACK, and HISP). This finding was expected, as the literature repeated negative effects of English language learner status, special education status, poverty, and racial achievement gaps on standardized testing (Diamond & Spillane, 2004; Hanushek & Raymond, 2005; Harris & Herrington, 2006; Watanabe, 2008; Wei, 2012).

English Language Learner (ELL) status. The ELL predictor had a low negative correlation ($r = -.13$) with 2013 FCAT scores; however, it did not contribute much to the predictive models ($-.15 \leq r_S \leq 0$). This is perhaps due to the relatively small size and poor prior performance of the ELL population: 2.7% of the regression sample, where 60.7% of those students scored a Level 1 on the 2012 Grade 5 Mathematics FCAT 2.0 test. Perhaps in a district with a larger and more diverse ELL population, the effect would be larger.

Another possible explanation is that students identified with ELL status are more likely to lack the language advantage of their native English-speaking peers, and as a result tend to underperform on achievement tests by comparison (Fry, 2007; Reardon & Galindo, 2009).

Exceptional Student Education (ESE) status. Students with an ESE code made up 11% of the regression sample. Overall, ESE status had a statistically significant, low negative correlation (r = -.23) to 2013 FCAT scores. Further, for all models except Level 5, ESE status had a negative contribution ($-.40 \leq r_s \leq -.12$) to predictive power. In the study sample, ESE status included Gifted and many types of impairments. There was a large difference in ESE predictive power for the Level 5 model (r_s = .35) and Level 1 model (r_s = -.40). This is most likely because most students who have scored a Level 5 the prior year and have one or more ESE codes were identified as Gifted, and most students who have scored a Level 1 or 2 the prior year and have one or more ESE codes were identified as having an impairment. Also, whereas 21% of students in the regression sample scored a Level 1 on the prior year FCAT (n = 1,107), only 3% of all students scored a Level 5 (n = 162). This could potentially contribute to an inflated positive predictive power for the Level 5 model compared to the larger group who scored a Level 1.

Free and reduced lunch (FREE_RED) enrollment status. A commonly used proxy for socio-economic status, free or reduced-cost lunch is available to students with financial need on a sliding scale based on family income. Although a greater proportion of the student body may have been eligible for the self-elected reduced-cost lunch program, 56% of the regression sample was coded by the school district as receiving either free or reduced-price lunches. Poverty has a well-established link to decreased academic performance (Coleman, 1988; Borman & Dowling, 2006; Kozol, 1991; Sirin, 2005). Although FREE_RED had a low negative correlation (r = -.22) with 2013 FCAT scores overall, the predictor variable did not contribute much to the overall predictive model (-.26 ≤ r_s ≤ -.06). It is possible that that this is due, in part, to collinearity and multicollinearity with a combination of other predictors. Correlations with SCHOOL_GRADE (*r* = -.26), FCAT_SS_2012 (*r* = -.24), and AVG_LSA (*r* = -.23) exceeded the correlation between FREE_RED and 2013 FCAT scores (*r* = -.22).

Gender (GENDER). In a literature review regarding stereotype threat, Smith and Hung (2008) discuss the negative stereotype against females in mathematics and resulting deficit in academic performance on standardized tests. Others (Cheema & Galluzzo, 2013; Ma, 2008), have reported that gender differences in academic achievement have decreased or reversed. In the present study, structure coefficients for GENDER were trivial in every model (-.04 ≤ r_s ≤ .07), therefore this variable was not a practical predictor of 2013 FCAT Grade 6 Math scores, regardless of 2012 FCAT Level. Moreover, other predictors such as free/reduced lunch status or average IBA scores explained only 5% of the variance in gender (Tolerance = .95). This means that gender did not contribute to multicollinearity in the model, nor was it related to academic achievement performance in the present study.

Black or African-American, Non-Hispanic (BLACK) and White, Non-Hispanic (WHITE). Evidence exists to support no change, or even an increase, in the Black-White achievement gap since NCLB and consequential accountability policies have been implemented (Diamond & Spillane, 2004; Harris and Herrington, 2006; Lee, 2008; Lee & Wong, 2004; Wei, 2012). Even where the achievement gap seemed to narrow, Lee & Reeves (2012) associated the reduction with “long-term statewide instructional capacity and teacher resources rather than short-term NCLB implementation” (p. 209).

The findings in the present study are consistent with those in the current literature. BLACK had a low negative correlation to 2013 FCAT scores (r = -.24), whereas WHITE had a low positive correlation (r = .21). Collinearity did exist in the data set between the WHITE and BLACK predictors, evidenced by a high correlation between the two (r = -.70). The high negative correlation between WHITE (38% of the regression sample, n = 2,030) and BLACK (52% of the regression sample, n = 2,763) was attributed to the fact that these were the largest racial demographic groups, and the variables are mutually exclusive—students are not able to select more than one racial category; multiracial students select "MULTI". Another consideration was that ASIAN, a much smaller proportion of the student sample, was the reference category for variable "dummy" coding. Perhaps if either WHITE or BLACK were used as the reference category, this collinearity would not have been an issue.

Despite collinearity, BLACK ($-.29 \le r_s \le -.04$) and WHITE ($.04 \le r_s \le .25$) were relatively unimportant to the prediction equation. There was a wide variation in structure coefficients for achievement level regression models: where the predictor BLACK explained a trivial amount of the variance in predicted 2013 FCAT scores ($r_s^2 = (-.04)^2 = .00$) for the Level 1 model, the predictor explained 8% of the variance for the Level 5 model ($r_s^2 = (-.28)^2 = .08$). This difference in performance by prior year FCAT achievement level supports what some (Hanushek & Raymond, 2005; Watanabe, 2008) have documented: an increase in the Black-White achievement gap via re-segregation and superficial teaching. This could also be attributed to a historically higher poverty rate and lack of socio-emotional capital among students identified as Black (Kozol, 1991).

Hispanic/Latino (HISP), Multiracial, Non-Hispanic (MULTI), and American Indian or Alaska Native, Non-Hispanic (AMER_IN). The Latino (9%), Multiracial (4%), and American Indian (0.3%) populations in the focus district were relatively small, making it difficult to determine a relationship amidst the larger racial/ethnic populations of BLACK and WHITE. Structure coefficients for HISP ($-.11 \leq r_s \leq .02$), MULTI ($-.18 \leq r_s \leq .05$), and AMER_IN ($-.01 \leq r_s \leq .04$) were close to zero in every model. Consequently, these predictor variables were not practical predictors of 2013 FCAT Grade 6 Math scores, regardless of 2012 FCAT Level.

Asian, Non-Hispanic (ASIAN). This variable was not included in the regression models. As the student group with the highest mean 2013 Grade 6 Mathematics FCAT developmental scale score (M = 240, Potential range = 170-284), ASIAN was used as the reference category for dummy variable coding for the other racial/ethnic variables.

Florida school grade (SCHOOL_GRADE). Coleman and colleagues, in their seminal *Equality of Educational Opportunity* report, argued that school environments were strong predictors of individual student academic achievement: "the social composition of the student body is more highly related to achievement, independent of the student's own social background, than is any other school factor" (Coleman et al., 1966, p. 325). School grade is one measure of academic environment and is composed of school-wide performance on the FCAT, performance by several historically underserved subsets of students, and attendance in accelerated coursework. The majority of students in the regression sample attended an "A" school (34%) or "C" school (32%).

In line with the literature, the present study found that the relationship between school grade and 2013 FCAT scores was positive ($r = .36$). Even though SCHOOL_GRADE had a trivial contribution to the predictive equation, according to beta weights ($.02 \leq \beta \leq .06$), structure coefficients revealed a larger direct effect ($.22 \leq r_s \leq .42$). This discrepancy between beta weights and structure coefficients can be explained by collinearity or multicollinearity (Courville & Thompson, 2001). In other words, SCHOOL_GRADE is useful in the prediction, but the shared predictive power was arbitrarily assigned to another variable (Nathans, Oswald, & Nimon, 2012).

Student grade point average (GPA). Teacher-assigned grades for the first three academic quarters of the 2012-2013 school year were averaged together into a Grade 6 Mathematics grade point average (GPA). The relationship between GPA and 2013 FCAT scores was moderate and positive (r = .53), which was within the expected range of .5 to .6 cited in the literature (Bowers, 2010; Brennan, Kim, Wenz-Gross, & Siperstein, 2001; Linn, 2000; Woodruff & Ziomek, 2004). GPA had lower structure coefficients than prior year FCAT score or interim assessment averages in all of the regression models except Levels 3 and 4. In the Level 3 model, GPA had a larger structure coefficient (r_s = .57) than the prior year FCAT score (r_s = .53). The same was true in the Level 4 model: GPA (r_s = .62) had a larger structure coefficient than the prior year FCAT score (r_s = .51). For these middle level students, the teacher's professional judgment is a better indicator of future success than interim assessments or prior year score. This may be because the teachers are more capable of assessing these students' academic achievement as a result of daily interaction.

2012 Grade 5 Mathematics FCAT 2.0 developmental scale score (FCAT_SS_2012). The prior year test score was expected to be the best predictor, and it was for two of the regression models. Prior year FCAT scores had the highest correlation to 2013 FCAT scores (r = .78) and the largest structure coefficients for the overall model (r_s = .92) and for the Level 1 model (r_s = .73). However, interim assessment scores eclipsed the prior year test scores in the Level 2 (r_s = .51), Level 3 (r_s = .53), Level 4 (r_s = .51), and Level 5 (r_s = .66) models. Additionally, in the Level 3 model, GPA had a larger structure coefficient

(r_s = .57) than prior year FCAT score (r_s = .53). The same was true in the Level 4 model: GPA (r_s = .62) had a larger structure coefficient than prior year FCAT score (r_s = .51).

One explanation for why the prior year FCAT score was a worse predictor than interim averages and/or GPA for Levels 2, 3, 4, and 5 is that the students in the study sample took Grade 5 Mathematics, Reading, and Science FCAT 2.0 tests in the prior year, which is the most state testing they have experienced up to that grade. In fact, Grade 5 Mathematics FCAT 2.0 was administered last out of these three time-intensive state standardized tests during the 2013 school year. It is likely that after so many days of testing, students in the fifth grade experienced testing fatigue. This could mean that the results of the prior year FCAT test were not as valid as would be expected if it were given in isolation. In that case, perhaps mid-year measures such as interim assessments or GPA would be better predictors than the flawed data collected from fatigued students the prior year.

Average interim benchmark assessment score (AVG_IBA). Interim assessments have historically been moderately (r = .6) to highly (r = .8) correlated to state standardized tests (Brown & Coughlin, 2007; Chen, 2011; Kingston et al., 2011; Underwood, 2010; Williams, 2008). The IBA averages in the present study were strongly and positively correlated to the 2013 FCAT scores (r = .72). It is important to note that each test's score reliability coefficient diminishes the correlation between the two. The reliability coefficient for the 2013 Grade 6 Mathematics FCAT 2.0 score was very high (r = .93; FDOE, 2013f) and the IBA had a score reliability coefficient of .745. The upper bound for the correlation coefficient between the IBA and 2013 FCAT is the geometric mean of the two reliability estimates, or .83 (Locke & Spirduso, 2014). Therefore, the .72 correlation in the study sample is very close to the highest the correlation could possibly ever be, .83. One potential explanation for the high correlation among these tests is that both types – interim and state standardized tests – are designed to measure the same content through identified academic standards or benchmarks. Also, the IBA assessment program was designed to mimic the state standardized test in as many ways as possible.

Weighted average learning schedule assessment score (AVG_LSA). Like the IBA averages, LSA weighted averages were strongly and positively correlated to the 2013 FCAT scores (r = .75), again in the expected range of .6 to .8. In the overall regression model, average LSA scores held the second largest structure coefficient (r_s = .88) behind prior year FCAT score (r_s = .92). This was also true for the Level 1 (r_s = .71 vs. prior year FCAT r_s = .73) and Level 5 (r_s = .70 vs. IBA average r_s = .84) models. However, in all of the other models, LSA scores had larger structure coefficients than any other predictor (Level 2 r_s = .84; Level 3 r_s = .86; and Level 4 r_s = .87) models. The LSA program was a better predictor of 2013 FCAT performance for the middle Levels most likely because of the format and frequency of the test. A more frequent test would be better able to capture the growth of the average student over the course of the school year than would prior year FCAT scores or a test that was only given at most three times. Further, the LSA program was easier to use formatively because the items were made available after administration. Teachers were able to review results in professional learning communities and crafted new lesson plans with the help of their peers.

Implications for Policymakers and Educational Leaders

Florida Board of Education and/or Commissioner of Education. Given the considerable and potentially prohibitive cost of interim assessments, both financially to pay for developer salaries and/or materials (Lee, 2008), and in terms of opportunity cost by lost instructional time in classrooms (Dee, Jacob, & Schwartz, 2013; McMurrer, 2008; Ravitch, 2010; Rentner et al., 2006), educational leaders at the state level might reconsider district academic progress reporting requirements. Certainly, flexibility and financial support for test development would help. Many school districts in Florida are relegated to repeated use of the same tests, a practice that leads to score inflation (Hamilton, 2003; Koretz & Beguin, 2010; Linn, 2000). Linn also identified another cost, the loss of validity, when basing important decisions on limited evidence and inflated scores (2000). Perhaps the best use of state public funds would be to pay for assessment-related professional development to support teachers and content specialists at district levels. Darling-Hammond (2004) noted that more of the accountability success stories came from those who focus on "broader notions of accountability, including investments in teacher knowledge and skill" (p. 1047).

Education about statistical processes to evaluate assessments can be prohibitive (Lee, 2008), thus these types of skills are not typically held by local school district personnel; however, it is highly important that interim assessments used for educational policymaking decisions are evaluated (Perie et al., 2009). Psychometricians or other expert staff hired by the state should be available to support school districts attempting to either develop local interim assessments or evaluate externally-developed interim assessments. Or, state administrators could facilitate a partnership between key K-12 school district personnel and the statistics or psychometrics department faculty at the state colleges for these purposes. At the least, it would be beneficial to have a freely available program, perhaps an Excel spreadsheet with macros, a written protocol, and a manual or explanatory paper on how to perform and interpret statistical analyses such as correlations, regression structure coefficients, and ANOVA.

Also, the Office of Accountability, Research, & Measurement in the Florida Department of Education (FDOE) should conduct research around which interim assessment methods are most and least effective. A collaborative group including members from the FDOE Office of Assessment, the FDOE Office of Educator Professional Development, local school districts, colleges of education, and perhaps at least one external formative assessment expert, the state should develop a practical, replicable plan to increase teachers' efficacy in formative assessment. This might include hiring a cadre of professional development providers to travel to the school districts in Florida. Finally, policymakers at the state level need to hear limitations regarding high-stakes testing and opportunity costs of assessment prior to making decisions regarding assessment policy.

Local school districts. As part of a continuous improvement philosophy, local school districts need to evaluate whether the current assessment tools being used are effectively achieving the stated outcomes. Metaevaluation of interim assessments using *Student Evaluation Standards* (JCSEE, 2001), *Standards for Educational and Psychological Testing* (JCSEPT, 1999), Perie et al.'s framework, or some other appropriate set of educational assessment standards, should be an ongoing practice by the Assessment and Accountability office in school districts. Where necessary, district administrators or specialists in this office should be given time to seek external training or coursework to support this work. Superintendents and local school board members would benefit from findings of this type of metaevaluative work, which would serve to inform specific assessment practices and broader assessment policies.

In addition, policymakers at the school district level must carefully consider the variety of learning environments and prior abilities within the school district prior to enacting requirements for district-wide testing. The present study found that certain aspects of predictive and instructional utility vary, depending on students' prior performance. For example, average scores on LSAs were able to explain only 14% of the variance in predicted 2013 FCAT scores for students who scored a Level 1 in 2012, but 37% of the variance for students who scored a Level 5 the prior year. Interim scores together added the most predictive power to the achievement level models for Levels 3 (ΔR^2 = .19) and 4 (ΔR^2 = .21), and added the least to models for the extremes: FCAT Levels of 1 (ΔR^2 = .11) or 5 (ΔR^2 = .13).

It is extremely important to identify ahead of time and announce publicly the purpose of interim assessments. The literature is clear that using a test for more than one purpose is not advised (APA, 2013; Black & Wiliam, 2005; Hamilton, 2003; Perie et al., 2009). The current challenge for policy makers and educators is to find alternative accountability frameworks and comprehensive assessment systems that include varying types of assessments intended for improving classroom practice and student achievement, while also avoiding over-testing and some of the negative effects of using any one single low-level test (Volante & Ben Jaafar, 2010).

Not all interim assessments are the same. In the present study, LSA averages predicted FCAT scores better than IBA averages for all students (r_s = .88 for LSA; r_s = .85 for IBA) as well as for students in every FCAT achievement level, except Level 5 (r_s = .70 for LSA; r_s = .84 for IBA). Additionally, the group with the highest average FCAT developmental scale score was the group with the high number of LSAs and the low number of IBAs (*M* = 222.63, *SD* = 18.03) versus the average for the ANCOVA sample (*M* = 216.41, *SD* = 19.43). The instructionally-purposed LSA program included more frequent tests, each aligned to only one instructional unit, available online and on paper, and had minimal security requirements, allowing for teachers and students to formatively use the data after the test administration was complete. Further, professional development wrapped around the LSA process, including how to develop items, administer the test, access reports, and modify instruction based on results.

This is something to consider as decisions are made about the scope, frequency, platform availability (computer vs. only paper-and-pencil), security level, and purpose of the interim testing program. Another consideration should be whether to embed professional development about how to develop quality test items, how to navigate whatever system is used to administer the tests, and how to interpret results and adjust instruction. Teachers and other educators may also gain improved motivation and morale as a result of assessment-based professional development (Hamilton, 2003); it is impossible to tease out how much of an impact the professional development surrounding the LSA program aided the implementation.

Also, more assessment is not always better (see Table 18 & Figure 13). Scholastic, supported by the Bill and Melinda Gates Foundation, surveyed over 10,000 teachers in 2012. The teachers called for "multiple, more frequent measures of teaching and learning" to assess student achievement and teacher performance (Scholastic, 2012, p. 25), similar to the LSA assessment program in the present study, which was more instructionally purposed and lent itself to a more formative usage. However, in the ANCOVA section of the present study, results indicated that the group with the highest level of assessment actually performed worse on average (*M* = 213.52, *SD* = 16.74) than groups with a low number of one type of assessment. This could be the result of over-testing and the resulting student apathy.

Judging how much to test is important and should not be extreme in either direction. The lowest performance belonged to the group that experienced little if any interim assessment (*M* = 222.63, *SD* = 18.03). One possible implication is that formal feedback is a necessary component of instruction. It is important to note that the present study is a correlational study, and any implications of causation are not warranted by these results. In other words, students did not necessarily score better on the FCAT because they took a certain amount of interim tests; however, it is clear that the two are related.

School-based administrators. As advocates for the children and educators in their schools, school-based administrators must carefully weigh the benefits of interim testing, such as predictive power, against the amount of time necessary to administer such tests. Interim assessments, and in particular, shorter curriculum-based interim assessments similar to the LSAs in this study, can offer more predictive power than student demographics or GPA, and in some cases even the prior year's standardized test score. However, as much as it is necessary to predict how students will do on the high-stakes summative test, it is also necessary to teach that which will be tested.

Although students in the group with a low number of IBA scores and a high number of LSA scores outperformed the other groups, the present study found no practically or statistically significant difference overall in the number of interim assessments taken. As such, it is imperative that school-based administrators invest time, money, and other resources to professional development centered on formative classroom testing. Much research has been done on formative assessment and the potential instructional benefits from effective formative assessment cycles, which include such things as clearly defined goals; collaboratively-developed open-ended tasks, items, and tests; rich and timely student feedback; and remediation or extension based on a learning partnership between students and teachers (Black & Wiliam, 1998a, 1998b; Filsecker & Kerres, 2012; Kluger & DeNisi, 1996; Shepard, 2006; Shute, 2008).

Teacher-assigned grades, reported as the course GPA in the present study, did not predict FCAT scores better than interim test averages in any of the models. However, GPA was a better predictor in some cases (Level 3 and 4) than the prior year's FCAT score. This could mean that whereas teacher-assigned grades alone may not be sufficient to produce a quality prediction, GPA used in conjunction with an interim average and prior year test scores would be fairly powerful.

Recommendations for Future Studies

The present study was delimited to one large, urban school district in Northeast Florida; samples from other Florida school districts should be analyzed to increase generalizability of the present findings, particularly where students participated in two or more simultaneous interim assessment programs during the same school year. Grade 6 was chosen because it is the last school grade where the majority of students are still enrolled in the same course. However, other grade levels and other subjects should be explored. Further, where homogeneity of assessments is possible, conducting studies of assessment practices and utility across multiple school districts would lead to greater generalizability of findings.

Only two portions of the predictive and instructional purpose sections from Perie et al.'s interim assessment framework (2009) were addressed in the present study. Other aspects of these sections, as well as aspects of an interim test program with an evaluative purpose, should be addressed for any district-developed or purchased interim assessment programs before, during, and after implementation. Further research is necessary to investigate how interim assessments are used to evaluate programs, professional development practices, and district initiatives.

Although the present study evaluated two types of test design, *state-test mirror* and *non-cumulative instructional mirror* (Gong, 2010), more research needs to be done on other types of interim test design and use. Another extension would be to match types of interim tests with assumptions about curriculum, instruction, and student learning.

Both types of interim assessments were comprised of multiple-choice items, which provide limited information about what students actually know (Kim, 2010; Perie et al., 2009). Most likely, the reliability and validity of data collected by both interim programs would increase as a result of including more open item formats, usually afforded by a computer-based assessment platform. Similar metaevaluative studies could and should be done to assess the predictive and instructional utility of tests with open-ended responses. Another aspect of open-ended performance tasks is scoring. Hopefully as computer-scoring programs gain credibility, research on the resulting scores will be easier to conduct.

Though experimental studies are rare in education, a controlled experiment would be ideal to truly compare learning with and without interim testing. A quasi-experiment comparing two similar districts is another possibility, assuming much thought is given to the educational and assessment context in each district.

Although the quantitative statistical techniques involved in the present study (multiple regression and ANCOVA) are appropriate for answering the research questions, other factors were not included in the analysis or accounted for explicitly in the results. These factors, such as test quality, professional development practices, teacher pedagogical content knowledge, professional collaborative efforts between teachers, student-teacher interactions, school leadership, curricular programs, technology usage, and school culture potentially could impact the results of the present study. These all represent areas for possible future research.

The qualitative paradigm would be an incredibly useful lens in this topic area to uncover how teachers and other educators describe their participation in interim assessment development, administration, scoring, data usage to inform subsequent instruction, and perceived successful practices among their peers. Topics such as students' beliefs about interim assessment, the nature of mathematics or other subjects, purposes for learning, and motivation are extremely important to the field of assessment and might best be served by using qualitative methods. Also interesting would be research around how the results from interim assessments, particularly those with an evaluative purpose, are used at a district level to make adjustments to programs or activities, curricular pacing and content, or policies based on competition or collaboration.

Teacher-assigned grades may include many components other than academic performance (Bowers, 2010; Willingham, Pollack, & Lewis, 2002). A study comparing teachers' surveyed predictions for student performance on high-stakes assessments with district-developed or purchased test scores would be informative. Another aspect of this research might be to determine how well teachers can assess their students during the school year, and what to do about it.

Summary

According to Perie et al. (2009)'s framework for evaluating interim assessments with predictive purposes, the tests "should be significantly more related to the criterion measure [2013 FCAT, in the present study] than other measures (e.g., teachers' grades) that could be used" (p.10). Average scores for both interim assessment types (IBA and LSA) were more related to 2013 FCAT scores than any other predictor including student GPA, except for 2012 FCAT scores. Further, within prior year achievement level groups, the average IBA scores and average LSA scores were better predictors than any other variable, with LSA scores more consistently outperforming IBA scores.

Instructional utility was evaluated in the present study using Perie et al.'s (2009) second criterion for instructional interim assessments: that the assessment program should provide evidence "demonstrating that the assessment system has contributed to improved student learning" (p. 10). Although not statistically significant at the .01 level, groups with a more moderate total number of interim assessments (either Low IBA and High LSA or High IBA and Low LSA) outperformed groups with all or nothing. Again, LSA tests were the favored type as students who took more LSA tests did better than students who took fewer, while the opposite was true for IBA tests.

Overall, the two types of interim assessment programs evaluated in the present study were good predictors of the state high-stakes test, 2012 Grade 6 Mathematics FCAT 2.0. However, more research must be done to identify with certainty whether or not the act of taking the tests and receiving feedback has contributed to improved student learning.

Appendix A – University Instructional Review Board Waiver

From: "Champaigne, Kayla"
Subject: RE: Waiver for IRB review
Date: December 16, 2013 at 9:35:19 AM EST
To: Tavy
Cc: "Daniel, Larry G" "Kasten, Katherine"
"O'Connor, Dawn"

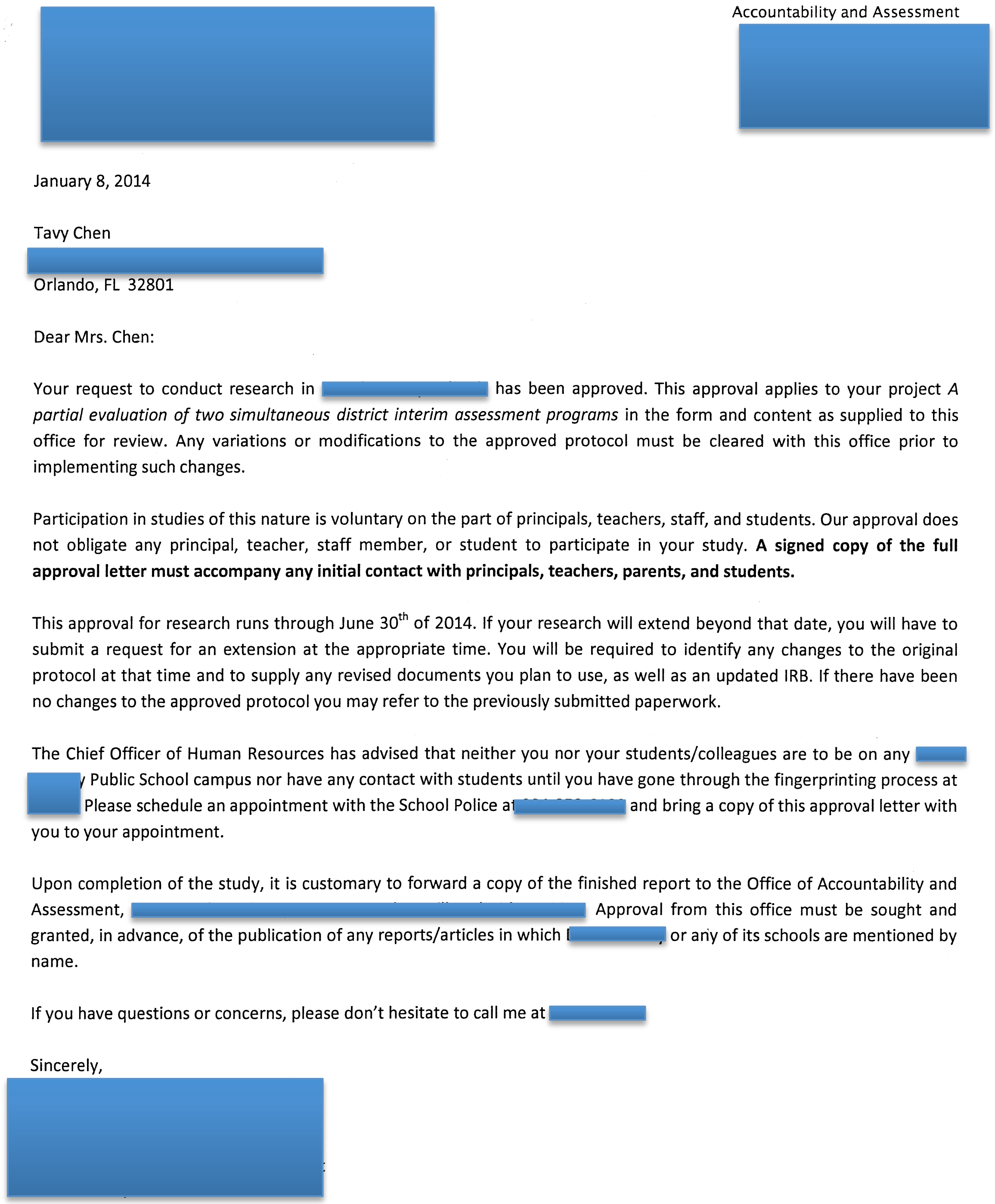
Hi Tavy,

Thank you for confirming that there will be no interaction with individuals and that the recorded data are not individually identifiable. Based on the information you submitted, the IRB member I consulted confirmed that your project is not research involving human subjects and therefore does not need to be reviewed by the UNF IRB prior to initiation. This determination was made based on the understanding that you will have no way to re-identify the data or associate identities with the information. Please keep a copy of this email which will serve as the waiver for your project. Thank you so much for being conscientious and taking the time to contact the UNF IRB about your project. We appreciate that you understand the value of IRB review of projects that may involve human subject research. Feel free to let us know if you have any questions or concerns. Have a great week and good luck with this project!

Sincerely,

Kayla Champaigne, CIP
Research Integrity Coordinator
Office of Research and Sponsored Programs
University of North Florida

Appendix B – District Instructional Review Board Approval



[redacted]

Accountability and Assessment

[redacted]

January 8, 2014

Tavy Chen
[redacted]
Orlando, FL 32801

Dear Mrs. Chen:

Your request to conduct research in [redacted] has been approved. This approval applies to your project *A partial evaluation of two simultaneous district interim assessment programs* in the form and content as supplied to this office for review. Any variations or modifications to the approved protocol must be cleared with this office prior to implementing such changes.

Participation in studies of this nature is voluntary on the part of principals, teachers, staff, and students. Our approval does not obligate any principal, teacher, staff member, or student to participate in your study. **A signed copy of the full approval letter must accompany any initial contact with principals, teachers, parents, and students.**

This approval for research runs through June 30th of 2014. If your research will extend beyond that date, you will have to submit a request for an extension at the appropriate time. You will be required to identify any changes to the original protocol at that time and to supply any revised documents you plan to use, as well as an updated IRB. If there have been no changes to the approved protocol you may refer to the previously submitted paperwork.

The Chief Officer of Human Resources has advised that neither you nor your students/colleagues are to be on any [redacted] [redacted] Public School campus nor have any contact with students until you have gone through the fingerprinting process at Please schedule an appointment with the School Police at [redacted] and bring a copy of this approval letter with you to your appointment.

Upon completion of the study, it is customary to forward a copy of the finished report to the Office of Accountability and Assessment, [redacted] Approval from this office must be sought and granted, in advance, of the publication of any reports/articles in which [redacted] or any of its schools are mentioned by name.

If you have questions or concerns, please don't hesitate to call me at [redacted]

Sincerely,

[redacted]

Appendix C – Correlation Coefficients for RQ2

2012 Grade 5 Mathematics FCAT 2.0 - Level 1

Variable	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1. FCAT_SS_2013 (dependent)	--														
2. ELL	-.08*	--													
3. ESE	-.25*	-.10*	--												
4. FREE_RED	-.07	.03	.03	--											
5. GENDER	.03	-.01	-.16*	-.01	--										
6. BLACK	-.02	-.34*	-.06*	.14*	.01	--									
7. HISP	-.01	.45*	-.02	-.03	.03	-.42*	--								
8. WHITE	.04	-.04	.07	-.12*	.01	-.69*	-.17*	--							
9. MULTI	.03	-.04	.09*	-.04	-.04	-.24*	-.06	-.10*	--						
10. AMER_IN	.03	-.01	-.02	.02	.03	-.04	-.01	-.02	-.01	--					
11. SCHOOL_GRADE	.17*	-.06	-.02	-.11*	-.03	-.13*	.01	.14*	.03	.03	--				
12. GPA	.19*	.21*	-.10*	-.06	.15*	-.09*	.07	-.02	.05	-.04	-.07	--			
13. FCAT_SS_2012	.45*	-.15*	-.19*	-.06	.08*	.03	-.02	.03	.02	.02	.11*	.14*	--		
14. AVG_IBA	.37*	-.08*	-.15*	-.03	-.04	-.05	-.03	.08*	.02	.06	.08*	.06	.26*	--	
15. AVG_LSA	.43*	-.04	-.14*	-.05	.03	.02	.01	-.02	.02	-.02	.26*	.35*	.26*	.23*	--

* $p < .01$.

2012 Grade 5 Mathematics FCAT 2.0 - Level 2

Variable	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1. FCAT_SS_2013 (dependent)	--														
2. ELL	.01	--													
3. ESE	-.15*	-.04	--												
4. FREE_RED	-.04	.02	-.01	--											
5. GENDER	.02	0	-.13*	-.02	--										
6. BLACK	-.07*	-.11*	-.07*	.24*	0	--									
7. HISP	.01	.18*	-.01	-.02	0	-.35*	--								
8. WHITE	.06*	-.03	.08*	-.24*	-.01	**-.72***	-.19*	--							
9. MULTI	-.02	-.03	.02	0	.01	-.23*	-.06	-.12*	--						
10. AMER_IN	-.01	-.01	-.01	-.05	.01	-.05	-.01	-.03	-.01	--					
11. SCHOOL_GRADE	.15*	-.02	.04	-.21*	-.04	-.24*	.09*	.19*	.01	-.01	--				
12. GPA	.29*	.10*	-.05	-.07*	.15*	-.06	.02	.01	.02	.02	.09*	--			
13. FCAT_SS_2012	.30*	-.02	-.07*	-.05	-.03	-.08*	0	.09*	-.01	-.04	.07*	.16*	--		
14. AVG_IBA	.40*	-.04	-.14*	-.03	-.04	-.11*	.02	.11*	-.02	.01	.08*	.14*	.26*	--	
15. AVG_LSA	.49*	.04	-.09*	-.07*	.07*	-.11*	.04	.07*	.01	.02	.29*	.50*	.23*	.30*	--

* *p* < .01.

2012 Grade 5 Mathematics FCAT 2.0 - Level 3

Variable	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1. FCAT_SS_2013 (dependent)	--														
2. ELL	-.05	--													
3. ESE	-.14*	.01	--												
4. FREE_RED	-.13*	.02	-.03	--											
5. GENDER	.06*	0	-.10*	0	--										
6. BLACK	-.14*	-.08*	-.08*	.25	.03	--									
7. HISP	-.03	.12*	.03	0	-.01	-.29	--								
8. WHITE	.11*	-.04	.04	-.21	-.01	-.67	-.27	--							
9. MULTI	0	0	.06	-.03	-.01	-.19	-.08	-.18	--						
10. AMER_IN	.02	-.01	.04	-.06	-.01	-.06	-.02	-.06	-.02	--					
11. SCHOOL_GRADE	.22*	-.05	.04	-.22	0	-.25	-.01	.23	.01	.05	--				
12. GPA	.39*	.04	-.03	-.13	.14	-.14	0	.06	.03	.04	.11*	--			
13. FCAT_SS_2012	.37*	0	-.04	-.07	-.06	-.12	-.01	.11	.01	-.04	.12*	.19*	--		
14. AVG_IBA	.52*	-.08*	-.07*	-.09*	-.02	-.12*	-.05	.12*	.01	0	.08*	.25*	.35*	--	
15. AVG_LSA	.59*	0	-.09*	-.13*	.10*	-.15*	-.03	.10*	.04	.01	.31*	.55*	.30*	.41*	--

* *p* < .01.

2012 Grade 5 Mathematics FCAT 2.0 - Level 4

Variable	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1. FCAT_SS_2013 (dependent)	--														
2. ELL	-.06	--													
3. ESE	-.08	-.02	--												
4. FREE_RED	-.08	.02	-.01	--											
5. GENDER	.07	-.03	-.09	-.06	--										
6. BLACK	-.09	-.02	.05	.30*	-.05	--									
7. HISP	-.01	.09	-.05	.05	.04	-.19*	--								
8. WHITE	.03	-.04	.03	-.26*	-.03	-.66*	-.31*	--							
9. MULTI	-.03	-.02	.01	-.02	-.02	-.13*	-.06	-.22*	--						
10. AMER_IN	.02	0	-.01	.05	.04	-.02	-.01	-.04	-.01	--					
11. SCHOOL_GRADE	.18*	-.07	-.08	-.24*	.05	-.21*	-.06	.21*	-.05	-.08	--				
12. GPA	.43*	.03	-.05	-.18*	.23*	-.12*	.03	.04	-.03	.03	.06	--			
13. FCAT_SS_2012	.35*	-.03	-.04	-.06	-.02	-.04	-.07	.04	.05	.02	.10	.25*	--		
14. AVG_IBA	.57*	-.06	-.02	-.05	0	-.10*	.01	.05	.01	-.03	.10*	.33*	.35*	--	
15. AVG_LSA	.60*	.03	-.05	-.12*	.08	-.08*	0	.02	-.03	.04	.22*	.54*	.32*	.52*	--

* p < .01.

2012 Grade 5 Mathematics FCAT 2.0 - Level 5

Variable	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1. FCAT_SS_2013 (dependent)	--														
2. ELL	-.01	--													
3. ESE	.26*	-.04	--												
4. FREE_RED	-.16	-.04	-.11	--											
5. GENDER	-.03	.09	-.13	.03	--										
6. BLACK	-.20*	-.03	-.19	.36*	.17*	--									
7. HISP	-.08	-.01	-.08	.07	-.02	-.07	--								
8. WHITE	.14	-.12	.25*	-.25*	-.14	-.60*	-.26*	--							
9. MULTI	-.13	-.01	-.08	-.02	-.02	-.07	-.03	-.26*	--						
10. AMER_IN	[a]	[a]	[a]	[a]	[a]	[a]	[a]	[a]	[a]	[a]					
11. SCHOOL_GRADE	.16	-.15	.16	-.18	-.03	-.06	0	.07	-.13	[a]	--				
12. GPA	.44*	.05	.08	-.18	.20*	-.12	-.04	.02	-.06	[a]	.11	--			
13. FCAT_SS_2012	.48*	-.06	.27*	0	-.10	-.18	-.09	.16	-.09	[a]	.11	.13	--		
14. AVG_IBA	.61*	-.04	.26*	-.17	-.15	-.16	-.08	.16	-.20	[a]	.12	.17	.49*	--	
15. AVG_LSA	.51*	0	.16	-.16	.14	-.21*	-.15	.23*	-.17	[a]	.23*	.43*	.41*	.51*	--

* *p* < .01.

[a] There were no American Indian/Native American students who scored a Level 5 on the 2012 FCAT.

References

Abrams, L. M., Pedulla, J. J., & Madaus, G. F. (2003). Views from the classroom: Teachers' opinions of statewide testing programs. *Theory Into Practice, 42*(1), 18-29.

American Psychological Association. (2013). *Appropriate use of high-stakes testing in our nation's schools*. Retrieved from http://www.apa.org/pubs/info/brochures/testing.aspx

Amrein-Beardsley, A. (2008). Methodological concerns about the education value-added assessment system, *Educational Researcher, 37*(2), 65–75.

Amrein, A., & Berliner, D. (2002). High-stakes testing and student learning. *Education Policy Analysis Archives, 10*(18).

Au, W. (2011). Teaching under the new Taylorism: High-stakes testing and the standardization of the 21st century curriculum. *Journal of Curriculum Studies, 43*(1), 25-45.

Bangert-Drowns, R. L., Kulik, J. A., & Kulik, C. C. (1991). Effects of frequent classroom testing. *Journal of Educational Research, 85*(2), 89-99.

Başol, G. & Johanson, G. (2009). Effectiveness of frequent testing over achievement: A meta analysis study. *International Journal of Human Sciences, 6*(2), 99-121.

Bennett, R. E. (2011). Formative assessment: A critical review. *Assessment in Education, 18*(1), 5-25.

Black, P. J., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy, and Practice, 5*(1), 7-74.

Black, P. J., & Wiliam, D. (2005). Lessons from around the world: How policies, politics and cultures constrain and afford assessment practices. *The Curriculum Journal, 16*(2), 249-261.

Blanc, S., Christman, J. B., Liu, R., Mitchell, C., Travers, E., & Bulkley, K. E. (2010). Learning to learn from data: Benchmarks and instructional communities. *Peabody Journal of Education, 85*(2), 205-225. doi: 10.1080/01619561003685379

Booher-Jennings, J. (2005). Below the bubble: "Educational triage" and the Texas accountability system. *American Educational Research Journal, 42*, 231-268.

Borman, G. D., & Dowling, N. M. (2006). Longitudinal achievement effects of multiyear summer school: Evidence from the Teach Baltimore randomized field trial. *Educational Evaluation and Policy Analysis, 28*(1), 25-48.

Bowers, A. (2010). What's in a grade? The multidimensional nature of what teacher-assigned grades assess in high school. *Educational Research and Evaluation: An International Journal on Theory and Practice, 17*(3), 141-159.

Brennan, R. T., Kim, J., Wenz-Gross, M., & Siperstein, G. N. (2001). The relative equitability of high-stakes testing versus teacher-assigned grades: An analysis of the Massachusetts Comprehensive Assessment System (MCAS). *Harvard Educational Review, 71*, 173–215.

Brookhart, S. M. (2011). Educational assessment knowledge and skills for teachers. *Educational Measurement: Issues and Practice, 30*(1), 3-12.

Brown, R. S., & Coughlin, E. (2007). *The predictive validity of selected benchmark assessments used in the Mid-Atlantic region* (Issues & Answers Report, REL 2007-No. 017). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Mid-Atlantic. Retrieved from http://ies.ed.gov/ncee/edlabs

Bulkley, K. E., Olah, L., & Blanc, S. (2010). Introduction to the special issue on benchmarks for success: Interim assessments as a strategy for educational improvement. *Peabody Journal of Education, 85*(2), 115-124.

Bulkley, K. E., Christman, J. B., Goertz, M. E., & Lawrence, N. R. (2010). Building with benchmarks: The role of the district in Philadelphia's benchmark assessment system. *Peabody Journal of Education, 85*(2), 186-204. Doi: 10.1080/01619561003685346

Burch, P. (2010). The bigger picture: Institutional perspectives on interim assessment technologies. *Peabody Journal of Education, 85*(2), 147-162.

Campbell, D. T. (1976). *Assessing the impact of planned social change.* Hanover, NH: Dartmouth College.

Carnoy, M., & Loeb, S. (2002). Does external accountability affect student outcomes? A cross-state analysis. *Educational Evaluation and Policy Analysis*, *24*(4), 305–331. doi:10.3102/01623737024004305

Chappius, S. (2005, August 10). Is formative assessment losing its meaning? *Education Week, 24*(44), 38.

Cheema, J. R., & Galluzzo, G. (2013, November). Analyzing the gender gap in math achievement: Evidence from a large-scale US sample. *Research in Education, 90*, 98-112.

Chen, T. W. (2012, February). *Understanding problems with collinearity of predictors in prediction of achievement test scores*. Paper presented at the annual meeting of the Eastern Educational Research Association, Hilton Head, SC.

Chudowsky, N., Chudowsky, V., & Kober, N. (2007). *Answering the question that matters the most: Has student achievement increased since 2002? State test score trends through 2006–07.* Washington, DC: Center on Education Policy.

Clune, W. H. (1993). The best path to systemic educational policy: Standard/centralized or differentiated/decentralized? *Educational Evaluation & Policy Analysis, 15*(3), 233.

Clune, W. H., & White, P. (2008). *Policy effectiveness of interim assessments in Providence public schools* (WCER Working Paper No. 2008-10). Madison, WI: University of Wisconsin-Madison, Wisconsin Center for Education Research.

Coleman, J. (1988). Social capital in the creation of human capital. *American Journal of Sociology, 94*(1), 1–25.

Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Mood, A. M., Weinfeld, F. D., & York, R. L. (1966). *Equality of educational opportunity.* Washington, DC: U.S. Department of Health, Education, and Welfare.

Courville, T., & Thompson, B. (2001). Use of structure coefficients in published multiple regression articles: β is not enough. *Educational and Psychological Measurement, 61*, 229-248.

Crane, E. (2008). *Interim assessment practices and avenues for state involvement*. Washington, DC: Council of Chief State School Officers.

Darling-Hammond, L. (2010). *The flat world and education: How America's commitment to equity will determine our future*. New York, NY: Teachers College Press.

Dee, T. S., & Jacob, B. (2011). The impact of No Child Left Behind on student achievement. *Journal of Policy Analysis and Management, 30*(3), 418-446.

Dee, T. S., Jacob, B., & Schwartz, N. L. (2013). The effects of NCLB on school resources and practices. *Educational Evaluation and Policy Analysis, 35*(2), 252-279.

Desimone, L. (2013). Teacher and administrator responses to standards-based reform. *Teachers College Record, 115*(8).

Diamond, J. B., & Spillane, J. P. (2004). High-stakes accountability in urban elementary schools: Challenging or reproducing inequality? *Teachers College Record, 106*(6), 1145-1176.

Downey, C., Steffy, B., Poston, W., & English, F. (2009). *50 ways to close the achievement gap.* Thousand Oaks, CA: Corwin Press.

Dunn, K. E., & Mulvenon, S. W. (2009). A critical review of research on formative assessments: The limited scientific evidence of the impact of formative assessments in education. *Practical Assessment, Research & Evaluation*, *14*(7). Retrieved from http://pareonline.net/getvn.asp?v=14&n=7

Daniel, L. G., & Onwuegbuzie, A. J. (2001, February). *Multiple regression: A leisurely primer.* Paper presented at the annual meeting of the Southwest Educational Research Association, New Orleans, LA.

Dweck, C. S. (1999). *Self-theories: Their role in motivation, personality, and development*. Philadelphia: Taylor & Francis.

Feuer, M. J. (2010). A call for innovation: Challenges raised by 21st-century test-based accountability systems. *Measurement: Interdisciplinary Research and Perspectives, 8*(2–3), 59–69.

Figlio, D. N., & Ladd, H. (2008). School accountability and student achievement. In H. Ladd & E. Fiske (Eds.), *Handbook of research in education finance and policy* (pp. 166–182). New York, NY: Routledge.

Filsecker, M., & Kerres, M. (2012). Repositioning formative assessment from an educational assessment perspective: A response to Dunn & Mulvenon (2009). *Practical Assessment Research & Evaluation*, *17*(16). Retrieved from http://pareonline.net/getvn.asp?v=17&n=16

Florida Department of Education (FDOE). (2007). *2007 Mathematics Next Generation Sunshine State Standards.* Retrieved from http://www.cpalms.org/Public/PreviewStandard/Preview/605

Florida Department of Education (FDOE). (2013a). *2013 guide to calculating school grades: Technical assistance paper.* Retrieved from http://schoolgrades.fldoe.org/pdf/1213/SchoolGradesTAP2013.pdf

Florida Department of Education (FDOE). (2013b). *Assessment and accountability workshop.* Retrieved from http://www.fldoe.org/arm/

Florida Department of Education (FDOE). (2013c). *Race to the Top assessments.* Retrieved from http://www.fldoe.org/arra/racetothetop/assessments

Florida Department of Education (FDOE). (2013d). *Student Performance Results: District Math Demographic Report.* Retrieved from http://app1.fldoe.org/fcatdemographics

Florida Department of Education (FDOE). (2013e). *Understanding FCAT 2.0 reports.* Retrieved from http://fcat.fldoe.org/mediapacket/2013/pdf/2013UFR.pdf

Florida Department of Education (FDOE). (2013f). *Florida Statewide Assessments 2013 Yearbook.* Retrieved from http://www.fldoe.org/fcat2/

Florida Statutes Annotated § 1008.25 2013.

Florida Statutes Annotated § 1008.35 2013.

Florida Statutes Annotated § 1012.34 2013.

Fry, R. (2007). *How far behind in math and reading are English language learners?* Washington, DC: Pew Hispanic Center. Retrieved from http://pewhispanic.org/files/reports/76.pdf

Fuller, B., Wright, J., Gesicki, K., & Kang, E. (2007). Gauging growth: How to judge No Child Left Behind? *Educational Researcher 36*(5), 268-178.

Goertz, M. E. (2007, June). Standards-based reform: Lessons from the past, directions for the future. In *Clio at the table: A conference on the uses of history to inform and improve education policy, Brown University*. Providence, RI: Brown University.

Goertz, M. E., Olah, L., & Riggan, M. (2009). *Can interim assessments be used for instructional change?* (CPRE Policy Briefs. RB-51). Philadelphia, PA: Consortium for Policy Research in Education.

Goertz, M. E., Olah, L., & Riggan, M. (2010). *From testing to teaching: The use of interim assessments in classroom instruction* (CPRE Research Report No. RR-65). Philadelphia, PA: Consortium for Policy Research in Education.

Gong, B. (2010, October). Some implications of the design of balanced assessment systems for the evaluation of the technical quality of assessments. Paper prepared for the 2010 Reidy Interactive Lecture Series sponsored by the Center for Assessment and WestEd, Cambridge, MA.

Gong, B. (2012, September). *Success for all: A personal view of educational challenges and research possibilities from an assessment perspective*. Paper presented at the annual meeting of the Northern Rocky Mountain Educational Research Association (NRMERA), Park City, UT.

Goren, P. (2010). Interim assessments as a strategy for improvement: Easier said than done. *Peabody Journal of Education, 85*(2), 125-129. doi: 10.1080/01619561003688688

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). *Multivariate data analysis* (7th ed.). Upper Saddle River, NJ: Prentice Hall.

Hamilton, L. (2003). Assessment as a policy tool. *Review of Research in Education, 27*, 25–68. Retrieved from http://www.jstor.org/stable/10.2307/3568127

Hausknecht, J. P., Halpert, J. A., Di Paolo, N. T., & Moriarty Gerrard, M. O. (2007). Retesting in selection: A meta-analysis of practice effects for tests of cognitive ability. *Journal of Applied Psychology, 92*, 373-385.

Hanushek, E. A., & Raymond, M. E. (2005). Does school accountability lead to improved student performance? *Journal of Policy Analysis and Management*, *24*(2), 297–327. doi:10.1002/pam.20091

Harris, D., & Herrington, C. (2006). Accountability, standards, and the growing achievement gap: Lessons from the past half-century. *American Journal of Education*, *112*(February), 209–238. doi:10.1086/498995

Henderson, S., Petrosino, A., Guckenburg, S., & Hamilton, S. (2008). *A second follow-up year for "Measuring how benchmark assessments affect student achievement"* (REL Technical Brief, REL Northeast and Islands 2007-No. 002). Washington, DC: U.S.

Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Northeast and Islands.

Henson R. K. (2002). *The logic and interpretation of structure coefficients in multivariate general linear model analyses.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Herman, J., & Baker, E. (2005). Making benchmark testing work. *Educational Leadership, 63*(3), 48-54.

Herrington, C. D., & MacDonald, V. M. (2000). Accountability as a school reform strategy: A thirty-year perspective on Florida. In C. D. Herrington & K. Kasten (Eds.), *Florida 2001: Educational policy alternative* (pp. 7-34). Jacksonville, FL: University of North Florida.

Heritage, M. (2008). *Learning progressions: Supporting instruction and formative assessment.* Washington, DC: Council of Chief State School Officers.

Heritage, M. (2010). *Formative assessment and next generation assessment systems: Are we losing an opportunity?* Washington, DC: Council of Chief State School Officers.

Hill, H. C., Kapitula, L., & Umland, K. (2010). A validity argument approach to evaluating teacher value-added scores. *American Educational Research Journal, 48*(3), 794-831.

Hinkle, D. E., Wiersma, W., & Jurs, S. G. (2003). *Applied statistics for the behavioral sciences* (5th ed.). Boston, MA: Houghton Mifflin.

Jacob, B. (2005). Accountability, incentives and behavior: The impact of high-stakes testing in the Chicago Public Schools. *Journal of Public Economics, 89*, 761-796.

Jacob, B. A., & Levitt, S. (2003). Rotten apples: An investigation of the prevalence and predictors of teacher cheating. *Quarterly Journal of Economics 118*(*3*), 843–877.

Jennings, J. (2012). *Reflections on a half-century of school reform: Why have we fallen short and where do we go from here?* Washington, DC: Center on Education Policy.

Joint Committee on Standards for Educational and Psychological Testing (JCSEPT). (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Joint Committee on Standards for Educational Evaluation (JCSEE). (2001). *Student evaluation standards.* Retrieved from http://www.jcsee.org/ses

Kerr, K., Marsh, J., Ikemoto, G., Darilek, H., & Barney, H. (2006). Strategies to promote data use for instructional improvement: Actions, outcomes, and lessons from three urban districts. *American Journal of Education, 112*(4), 496-520.

Kim, Y. (2010). The Procrustes ' bed and standardization in education. *Journal of Thought, Fall-Winter,* 9-20.

Kingston, N., & Nash, B. (2011). Formative assessment: A meta-analysis and a call for research. *Educational Measurement: Issues and Practice, 30*(4), 28-37.

Kingston, N., Wang, W., Broaddus, A., & Kramer, L. (2011, August). *Kansas interim assessment validity evidence based on the relationship between interim and summative assessment scores.* Paper presented at the annual meeting of the American Psychological Association, Washington, DC.

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Pscychological Bulletin, 119*(2), 254-284.

Koretz, D. M., & Béguin, A. (2010). Self-auditing assessments. *Measurement: Interdisciplinary Research and Perspectives, 8(*2–3), 92–109.

Kozol, J. (1991). *Savage inequalities: Children in America's schools.* New York, NY: Crown Publishers.

Kress, S., Zechmann, S., & Schmitten, J. M. (2011). When performance matters: The past, present, and future of consequential accountability in public education. *Harvard Journal on Legislation*, *48*(1), 185-234.

Lee, J. (2008). Is test-driven external accountability effective? Synthesizing the evidence from cross-state causal-comparative and correlational studies. *Review of Educational Research*, *78*(3), 608–644. doi:10.3102/0034654308324427

Lee, J., & Reeves, T. (2012). Revisiting the impact of NCLB high-stakes school accountability, capacity, and resources: State NAEP 1990-2009 reading and math achievement gaps and trends. *Educational Evaluation and Policy Analysis, 34*(2), 209–231. doi:10.3102/0162373711431604

Li, Y., Marion, S., Perie, M., Gong, B. (2010). An approach for evaluating the technical quality of interim assessments. *Peabody Journal of Education, 85*(2), 163-185. doi: 10.1080/01619561003685304

Lile, B. C. (2012). *Perceptions of Kentucky educators concerning the Kentucky State Assessment System as an accurate reflection of student learning.* (Doctoral Dissertation, Western Kentucky University). Retrieved from http://digitalcommons.wku.edu/cgi/viewcontent.cgi?article=1023&context=diss

Lindquist, E. F. (Ed.). (1951). *Educational measurement.* Washington, DC: American Council of Education.

Linn, R. L. (2000). Assessments and accountability. *Educational Researcher, 29*(2), 4–16.

Linn, R. L., Baker, E. L., & Betebenner, D. W. (2002). Accountability systems: Implications of requirements of the No Child Left Behind Act of 2001. *Educational Researcher, 31*(6), 3-16.

Locke, L., Spirduso, W. W., & Silveman, S. J. (2014). *Proposals that work: A guide for planning dissertations and grant proposals.* Los Angeles, CA: Sage.

Ma, X. (2008). Within-school gender gaps in reading, mathematics, and science literacy, *Comparative Education Review, 52*(3), 437–60.

McMillan, J. H., Venable, J. C., & Varier, D. (2013). Studies of the effect of formative assessment on student achievement: So much more is needed. *Practical Assessment Research & Evaluation*, *18*(2). Retrieved from http://pareonline.net/getvn.asp?v=18&n=2

McMurrer, J. (2008). Instructional time in elementary schools: A closer look at changes for specific subjects. In *From the capital to the classroom: Year 5 of the No Child Left Behind Act.* Washington, DC: Center on Education Policy.

McNeil, L. M., Coppola, E., Radigan, J., & Heilig, J, V. (2008). Avoidable losses: High-stakes accountability and the dropout crisis. *Education Policy Archives, 16*. Retrieved from http://epaa.asu.edu/ojs/article/view/28/154

Mills, J. I. (2008). A legislative overview of No Child Left Behind. In T. Berry & R. M. Eddy (Eds.), *Consequences of No Child Left Behind for Educational Evaluation. New Directions for Evaluation 117*, 9-20.

Moon, T., R; Brighton, C., M.; Jarvis, J., M., & Hall, C. J. (2007). *State standardized testing programs: Their effects on teachers and students.* Retrieved from http://www.gifted.uconn.edu/nrcgt.html

National Center for Education Statistics (NCES). (2012). *NAEP: Measuring student progress since 1964.* Retrieved from http://nces.ed.gov/nationsreportcard/about/naephistory.aspx

National Council on Teacher Quality (NCTQ). (2012, January). *2011 State Teacher Policy Yearbook.* Washington, D.C. Retrieved from www.nctq.org

National Commission on Excellence in Education (NCEE). (1983). *A nation at risk.* Washington, DC.

National Research Council (NRC). (1999). *High stakes: Testing for tracking, promotion, and graduation.* Washington, DC: National Academy Press.

Neal, D., & Schanzenbach, D. W. (2010). Left behind by design: Proficiency counts and test-based accountability. *The Review of Economics and Statistics*, *92*(2), 263-283.

National Research Council. (2001). *Knowing what students know: The science and design of educational assessment.* Committee on the Foundations of Assessment, J. W.

Pellegrino, N. Chudowsky, & R. Glaser (Eds.), Board on Testing and Assessment, Center for Education, Division of Behavioral and Social Sciences and Education. Washington, DC: National Academy Press.

Nichols, S. L., Glass, G. V., & Berliner, D. C. (2006). High-stakes testing and student achievement: Does accountability pressure increase student learning? *Education Policy Analysis Archives, 14*(1). Retrieved from http://epaa.asu.edu/epaa/v14n1/

Nichols, S. L., Glass, G. V., Berliner, D. C. (2012). High-stakes testing and student achievement: Updated analyses with NAEP data. *Education Policy Analysis Archives, 20*(20), 1-30. Retrieved from http://epaa.asu.edu/ojs/article/view/1048

Onwuegbuzie, A. J., & Daniel, L. G. (2003, February 19). Typology of analytical and interpretational errors in quantitative and qualitative educational research. *Current Issues in Education*, *6*(2). Retrieved from http://cie.asu.edu/volume6/number2/index.html

Pedulla, J. J., Abrams, L. M., Madaus, G. F., Russell, M. K., Ramos, M. A., & Miao, J. (2003). *Perceived effects of state-mandated testing programs on teaching and learning: Findings from a national survey of teachers*. Chestnut Hill, MA: National Board on Educational Testing and Public Policy.

Pellegrino, J. (2002). Knowing what students know. *Issues in Science and Technology, 19*(2), 48-52.

Perie, M., Marion, S., & Gong, B. (2007). *A framework for considering interim assessments.* Retrieved from http://www.nciea.org/publication_PDFs/ConsideringInterimAssess_MAP07.pdf

Perie, M., Marion, S., & Gong, B. (2009). Moving toward a comprehensive assessment system: A framework for considering interim assessments. *Educational Measurement: Issues and Practice, 28(*3), 5-13.

Pilotin, M. (2013). Finding a common yardstick: Implementing a national student assessment and school accountability plan through state-federal collaboration. *California Law Review, 98*, 545-574.

Popham, W. J. (2006, October). *Defining and enhancing formative assessment.* Paper presented at the Annual Large-Scale Assessment Conference, Council of Chief State School Officers, San Francisco, CA.

Popham, W. J. (2008). *Transformative assessment.* Alexandria, VA: ASCD.

Pullin, D. (2013). Legal issues in the use of student test scores and value-added models (VAM) to determine educational quality. *Education Policy Analysis Archives, 21*(6), 1-27.

Ravitch, D. (2010). *The death and life of the great American school system: How testing and choice are undermining education.* New York, NY: Basic Books.

Ready, D. D. (2013). Associations between student achievement and student learning: Implications for value-added school accountability models. *Educational Policy, 27*(1), 92-120.

Reardon, S. E., & Galindo, C. (2009). The Hispanic-White achievement gap in math and reading in the elementary grades. *American Educational Research Journal, 46*, 853–891.

Resnick, L. B. (1987). The 1987 presidential address: Learning in school and out. *Educational researcher*, *16*(9), 13-54.

Rentner, D. S., Scott, C., Kober, N., Chudowsky, N., Chudowsky, V., Joftus, S., & Zabala, D. (2006). *From the capital to the classroom: Year 4 of the No Child Left Behind Act*. Washington, DC: Center for Education Policy.

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science, 18,* 199-144.

Sahlberg, P. (2008). Rethinking accountability in a knowledge society. *Journal of Educational Change*, *11*, 45-61. doi:10.1007/s10833-008-9098-2

Sahlberg, P. (2011). *Finnish lessons: What can the world learn from educational change in Finland?* New York, NY: Teachers College Press.

Samuels, C. A. (2011, August 4). Cheating scandals intensify focus on test pressures. *Education Week, 30*(37).

Scholastic. (2012). *Primary sources: America's teachers on the teaching profession.* Retrieved from Scholastic's website: http://www.scholastic.com/primarysources/pdfs/Gates2012_full.pdf

Scriven, M. S. (1967). The methodology of evaluation. In R. W. Tyler, R. M. Gagne, & M. Scriven (Eds.), *Perspectives of curriculum evaluation. Volume I* (pp. 39-83). Chicago, IL: Rand McNally.

Scriven, M. S. (1974). Evaluation perspectives and procedures. In W. J. Popham (Ed.), *Evaluation in education: Current applications* (pp. 3–33). Berkeley, CA: McCutcheon.

Shepard, L. (2005, October). *Formative assessment: Caveat emptor.* Paper presented at the Educational Testing Service Invitational Conference, New York, NY.

Shepard, L. A. (2008). Formative assessment: Caveat emptor. In C. A. Dwyer (Ed.), *The future of assessment: Shaping teaching and learning* (pp. 279–303). New York, NY: Erlbaum.

Shepard, L. A. (2010). What the marketplace has brought us: Item-by-item teaching with little instructional insight. *Peabody Journal of Education, 85(*2), 246–257.

Shepard, L. A., Davidson, K. L., & Bowman, R. (2011). *How middle-school mathematics teachers use interim and benchmark assessment data.* (CRESST Report 807). Los

Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research*, *78*(1), 153–189.

Simpson, R. L., LaCava, P., & Graner, P. (2004). The No Child Left Behind Act: Challenges and implications for educators. *Intervention in School and Clinic,* 40, 67-75.

Sirin, S. R. (2005). Socioeconomic status and academic achievement: A meta-analytic review of research. *Review of Educational Research, 75*(3), 417-453.

Spring, J. H. (2005). *The American school, 1642-2004*. Boston, MA: McGraw-Hill.

Spring, J. (2011). *The politics of American education*. New York, NY: Taylor & Francis US.

Stedman, L. C. (2011). A preliminary analysis of Atlanta's performance on the National Assessment of Educational Progress. *Critical Education, 2*(9*)*. Retrieved from http://m1.cust.educ.ubc.ca/journal/index.php/criticaled/article/view/175

Smith, C. S., & Hung, L. (2008). Stereotype threat: Effects on education. *Social Psychology of Education, 11*, 243-257. doi: 10.1007/s11218-008-9053-3

Stullich, S., Eisner, E., McCrary, J., & Roney, C. (2006). *National assessment of Title I interim report to Congress, Vol. 1: Implementation of Title I.* Washington, DC: U.S. Department of Education, Institute of Education Sciences.

Thompson, B. (2006). *Foundations of behavioral statistics: An insight-based approach.* New York: Guilford Press.

Thompson, B., & Borrello, G. M. (1985). The importance of structure coefficients in regression research. *Educational and Psychological Measurement, 45,* 203-209.

Underwood, M. K. (2010). *The relationship of 10th-grade district progress monitoring assessment scores to Florida Comprehensive Assessment Test scores in reading and mathematics for 2008-2009.* (Doctoral dissertation, University of Central Florida) Retrieved from http://etd.fcla.edu/CF/CFE0003214/Underwood_Marilyn_K_201008_EdD.pdf

USA Today. (2013, April 14). *School cheating scandal shakes up Atlanta.* Larry Copeland.

U.S. Census Bureau. (2012, March 26). *Growth in urban population outpaces rest of nation, Census Bureau reports.* Retrieved from http://www.census.gov/newsroom/releases/archives/2010_census/cb12-50.html

U.S. Department of Education. (2009). *Race to the Top executive summary.* Washington, DC. Retrieved from http://www2.ed.gov/programs/racetothetop/executive-summary.pdf

U.S. Department of Education. (2010). *A blueprint for reform: The reauthorization of the Elementary and Secondary Education Act.* Washington, DC.

Volante, L., & Ben Jaafar, S. (2010). Assessment reform and the case for learning-focused accountability. *Journal of Educational Thought, 44*(2), 167-188.

Watanabe, M. (2008). Tracking in the era of high-stakes state accountability reform: Case studies of classroom instruction in North Carolina. *Teachers College Record, 110*(3), 489–534.

Wiliam, D. (2004, June). *Keeping learning on track: Integrating assessment with instruction.* Invited address presented at the 30th annual conference of the International Association for Educational Assessment, Philadelphia, PA.

Wiliam, D., & Thompson, M. (2008) Integrating assessment with instruction: What will it take to make it work? In C.A. Dwyer (Ed.), *The future of assessment: Shaping teaching and learning* (pp. 53-82). Mahwah, NJ: Lawrence Erlbaum

Williams, M. N., Grajales, C. A. G., & Kurkiewicz, D. (2013). Assumptions of multiple regression: Correcting two misconceptions. *Practical Assessment, Research & Evaluation, 18*(11), 1-14.

Willingham, W. W., Pollack, J. M., & Lewis, C. (2002). Grades and test scores: Accounting for observed differences. *Journal of Educational Measurement, 39*, 1–37.

Wilson, M., & Draney, K. (2004). Some links between large-scale and classroom assessments: The case of the BEAR Assessment System. In M. Wilson (Ed.), *Toward coherence between classroom assessment and accountability* (103rd Yearbook of the National Society for the Study of Education, Part II, pp. 132–154). Chicago, IL: University of Chicago Press.

Woodruff, D. J., & Ziomek, R. L. (2004). *High school grade inflation from 1991 to 2003* (Research report series 2004–05). Iowa City, IA: ACT.

Wong, K. K. (2013). Politics and governance: Evolving systems of school accountability. *Educational Policy, 27*(2), 410–421. doi:10.1177/0895904813479089

TAVY WELLS CHEN

Curriculum Vitae

http://tinyurl.com/tavychen

EDUCATION

2014	Ed.D. Educational Leadership, University of North Florida
2007	M.Ed. Secondary Mathematics Education, University of North Florida
2000	B.S. Mathematics, University of Florida

RELATED EXPERIENCE

2013	Director, Data & Assessment Duval County Public Schools, Jacksonville, FL
2010-2013	Supervisor, Data & Assessment Duval County Public Schools, Jacksonville, FL
2008-2010	Specialist, Secondary Mathematics PROMiSE grant/Duval County Public Schools, Jacksonville, FL
2004-2008	Teacher, Secondary Mathematics Duval and Hillsborough School Districts, Jacksonville and Tampa, FL

ACADEMIC PRESENTATIONS

Chen, T. W. (2012, February). *Understanding problems with collinearity of predictors in prediction of achievement test scores*. Paper presented at the annual meeting of the Eastern Educational Research Association, Hilton Head, SC.

Schackow, J., Chen, T. W., Ewing, V., Sharpe, C., & Sotolongo, R. (2010, January). *Florida PROMiSE: Summer Institute Follow-Up Days.* Presentation at the semiannual meeting of the Florida Association of Mathematics Supervisors, Orlando, FL.

CERTIFICATIONS

2010-2015	Educational Leadership Certificate, Florida State Department of Education
2008-2018	National Board Certificate, Adolescence and Young Adulthood Mathematics, National Board for Professional Teaching Standards
2005-2015	Professional Teaching Certificate, Mathematics Grades 6-12, Florida State Department of Education

EXAMINATIONS

2009	Graduate Record Examination (GRE) Verbal: 580 Quantitative: 800 Analytical: 800
2008	Florida Educational Leadership Exam (FELE)

PROFESSIONAL MEMBERSHIPS

American Educational Research Association (AERA)
National Council on Measurement in Education (NCME)