A Historical Perspective of Testing and Assessment Including the Impact of Summative and Formative Assessment on Student Achievement

Carole Sanger Brink
University of Southern Mississippi

Follow this and additional works at: https://aquila.usm.edu/dissertations

Part of the Curriculum and Instruction Commons, Educational Administration and Supervision Commons, and the Educational Assessment, Evaluation, and Research Commons

Recommended Citation

This Dissertation is brought to you for free and open access by The Aquila Digital Community. It has been accepted for inclusion in Dissertations by an authorized administrator of The Aquila Digital Community. For more information, please contact aquilastaff@usm.edu.
A HISTORICAL PERSPECTIVE OF TESTING AND ASSESSMENT INCLUDING
THE IMPACT OF SUMMATIVE AND FORMATIVE ASSESSMENT
ON STUDENT ACHIEVEMENT

By

Carole Sanger Brink

Abstract of a Dissertation
Submitted to the Graduate School
of The University of Southern Mississippi
in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy

December 2011
ABSTRACT

A HISTORICAL PERSPECTIVE OF TESTING AND ASSESSMENT INCLUDING THE IMPACT OF SUMMATIVE AND FORMATIVE ASSESSMENT ON STUDENT ACHIEVEMENT

by Carole Sanger Brink

December 2011

In 2007, Georgia developed a comprehensive framework to define what students need to know. One component of this framework emphasizes the use of both formative and summative assessments as part of an integral and specific component of the teachers’ performance evaluation. Georgia administers the Criterion-Referenced Competency Test (CRCT) to every elementary student in Grades 1 though 8. Before 2008, the state tested eighth-grade students on a quality core curriculum. In 2008, the state began testing students on a Georgia Performance Standard curriculum. A direct comparison of the curriculum change should have contained items to test both curriculums. However, this was not done. Therefore, the current study was designed to examine if differences in students achievement occurred because of the curriculum change.

Archival CRCT and ITBS data from 21 middle schools were analyzed to determine if assessment changes affected student achievement. Results of two doubly multivariate, repeated measures ANCOVAs found no statistically significant differences between the two curriculums. However, the lack of significance could be attributed to the small sample size. The increase in scores at the end of the three-year period measuring the quality core curriculum and at the end of the three-year period measuring the Georgia Performance Standard curriculum provided partial support to the hypothesis
of a difference in achievement between eighth-grade students who were taught and then tested under different curriculums.

Recommendations for practice include the provision that educators be engaged in professional development in regards to the use of data. Most principals and district leaders do not have the skills to navigate high-stakes testing results. More importantly, though, are the university systems that should augment an instructional strategy class and add a data leadership class to the current list of courses needed to earn a leadership degree. Another recommendation is to those who develop criterion-referenced tests. Changing the score scales on the criterion-referenced competency tests when the curriculum changes make it very difficult to study data to determine progress over time. In the statistical world, this creates a confounding variable that may be hard to control.
A HISTORICAL PERSPECTIVE OF TESTING AND ASSESSMENT INCLUDING
THE IMPACT OF SUMMATIVE AND FORMATIVE ASSESSMENT
ON STUDENT ACHIEVEMENT

by

Carole Sanger Brink

A Dissertation
Submitted to the The University of Southern Mississippi
Graduate School In Partial Fulfillment of the Requirements
For the Degree of Doctor of Philosophy

Approved:

David E. Lee
Director

Rose McNeese

J.T. Johnson

Ronald Sytron

Susan A. Siltanen
Dean of the Graduate School

December 2011
DEDICATION

With great love and admiration, I dedicate my dissertation to my grandparents Joseph and Lillian Baumgartner who have passed away long ago. They were lifelong educators; teaching was their passion not their job. I thank them for their constant conversations about improving education and how we test students; this has always instilled a sense of the importance of history as a source of forward movement. Most of all, I will be forever grateful to them for their constant reminder that life is fun if you do what you are called to do.

Joseph Auburn Baumgartner was born on November 18, 1905, and died March 28, 1977. He began his career as a chemical engineer receiving his degrees at New River State College, his masters at Marshall University. He became a teacher teaching 4th, 5th, and 6th grade in a two-room school in Ingram Branch, West Virginia. During World War II, he was drafted and worked as a chemist at Union Carbon and Carbide; he was doing what he called secret stuff and never told us what he did. After the war, Grandpa Baumgartner taught science at Collins High School in Oak Hill, West Virginia and ended his career as a principal in the Pax School System West Virginia. He retired in 1974 with thirty years of teaching and administrative service.

Lillian Margaret Moses Baumgartner was born on March 21, 1904 and died March 18, 1977. She received her first teaching degree from Concord State Normal School and then a four-year teaching degree from Concord College. She taught elementary school at Levesy, Page, Summerlee, and Oak Hill Elementary Schools in West Virginia. She also taught at Collins Junior High School, Oak Hill, West Virginia where she finished her career and retired with thirty plus years of teaching.
ACKNOWLEDGMENTS

There are individuals who warrant acknowledgment for their substantial contributions to not only the completion of the dissertation, but to the entire process of coursework involved. Without each of these individuals, simply stated and simply a reality, this accomplishment would not have been possible. I will be forever grateful.

First, I must heartily extend a huge thank you to the outstanding professors at the University of Southern Mississippi who provided mentoring, coaching, and instruction toward the completion of both the dissertation and degree. To my committee chair, Dr. David E. Lee, I sincerely appreciated the needed direction he gave toward the completion of this dissertation. His constant encouragement was most valuable. I also thank Dr. J. T. Johnson for stepping in at the last moment and providing welcomed statistical advice on Chapter IV. Dr. Ronald Styron is also to be commended for stepping in to assist as USM experienced several faculty changes. Finally, Dr. Rose McNeese, thank you for taking on this cohort. Without you, my fellow cohort members and I would have lost significantly. To sincerely thank you, Dr. McNeese, from the bottom of my heart just does not seem adequate.

I want to thank my staff. These wonderful educators were supportive of my goal to achieve a Ph.D. To my parents, Herb and Rita Sanger, who funded this endeavor financially, emotionally, and spiritually, I will forever be grateful. Finally, I could not be more indebted for the encouragement from my husband, James, and my three children, Edward, Claire, and Laura. Sorry for all the weekend events I missed, I know you wanted me there. Your sacrifice made this lifelong goal possible for me.
TABLE OF CONTENTS

ABSTRACT........................................................................................................................................... ii

DEDICATION......................................................................................................................................... iv

ACKNOWLEDGMENTS .............................................................................................................................. v

LIST OF TABLES..................................................................................................................................... viii

LIST OF FIGURES.................................................................................................................................. ix

CHAPTER

I.  INTRODUCTION .............................................................................................................................. 1

Background of the Study
Statement of the Problem
Purpose of the Study
Definition of Terms
Limitations/Delimitations
Assumptions
Justification of the Study
Summary

II.  LITERATURE REVIEW .................................................................................................................... 10

Types of Tests and Their History
History of Assessment in the United States
Curriculum Assessment in the State of Georgia
Summative and Formative Assessment
Trends in Testing/Assessment in the United States
The Reform of Assessment
Types of Assessments
High-Stakes Testing
Next Steps for Assessments in the United States: New History
Summary

III. METHODOLOGY .............................................................................................................................. 40
Research Design
Setting
Data Collected
Procedures
Data Analysis
Limitations
Ethical Considerations
Summary

IV. RESULTS ..............................................................................................................48

Description of the Data
Analysis of Eighth-Grade CRCT Reading Scores
Analysis of Eighth-Grade CRCT Mathematics Scores

V. SUMMARY OF FINDINGS, CONCLUSIONS, AND RECOMMENDATION .............................52

Findings
Conclusion
Recommendations for Practice
Recommendations for Future Research
Concluding Remarks

APPENDIXES ............................................................................................................63

REFERENCES ............................................................................................................65
LIST OF TABLES

Table

1  Demographic Description of School in Study (Percent of Students) .................................. 42
2  Means and Standard Deviations of Variables of Interest ($n = 21$).................................... 47
LIST OF FIGURES

Figure

1  Converted Eighth-Grade CRCT Reading Scores by Year .................................................49

2  Converted Eighth-Grade CRCT Mathematics Scores by Year ........................................51
CHAPTER I
INTRODUCTION

Educators have explored the specialized needs of assessments for decades. Teachers, parents, administrators, board members, and other stakeholders wrestle with statewide accountability testing, benchmark tests, classroom tests, and tests that range from placing gifted students to tests that identify special education students. These stakeholders, especially the teachers, should embrace student assessment as a way to understand students. In 2007, the state of Georgia developed a comprehensive framework to describe and define what students need to know and what schools need to know, understand, and be able to do (Georgia Department of Education, 2007b). This framework is called the School Keys: Unlocking Excellence through the Georgia School Standards. These keys “are the foundation for Georgia’s comprehensive, data-driven system of school improvement and support” (Georgia Department of Education, 2007b, p. 3). One key, the assessments key emphasizes the use of both formative and summative assessments as part of an integral and specific component of teachers’ performance evaluation (Bryant & Bradford, 2010).

Background of the Study

The history of assessment of students began when the doors of schoolhouses were first opened. By the early 19th century, teachers tested their students to see if they had mastered what was taught (U.S. Department of Education, 2008). If students failed, they were held back or retained. In addition, these teachers only administered one type of assessment. This type of assessment was given in a whole class environment and was known as recitation (Giordano, 2005). Current educators administer this type of assessment as well, but refer to this type of test as summative. Only one change has been
made–some students may receive accommodations for the testing environment (Laprairie, Johnson, Rice, Adams, & Higgins, 2010).

By the 20th century, the field of educational psychology was established (Giordano, 2005). This field drastically altered the testing of students. According to Giordano, educational psychologists changed the face of educational testing forever by introducing the standardized test. These psychologists introduced testing that became the norm for measuring student intelligence and objective measurement of knowledge of content. Shepard (2000) called this historical concept scientific measurement. Testing became a science, an area of study for educators. However, these tests were not used for accountability, but for judgment of students’ levels of aptitude, students’ mastery of content, and students’ potential to move to higher levels of education.

As the 20th century moved into the 1950s, this basic practice of testing for student aptitude and mastery of content continued, but the use of these tests expanded to the practice of tracking students. Giordano (2005) reported that the use of program accountability became one of the main purposes of testing in the 1960s. In the 1970s, students were subjected to minimum competency tests centered on student proficiency based on the results of standardized tests. The use of standardized testing was for district accountability in the 1980s. Currently the idea of standards-based accountability is in full stride (Linn, 2000).

No Child Left Behind (NCLB) has a major impact on public education (Jennings & Renter, 2006). Accountability is normal in public schools because of NCLB. Fifteen years of standards-based reform has resulted in local school districts as well as many state educational officials questioning this heavy reliance on student test results as a measure of not only student achievement, but also educational achievement. The provision of the
NCLB law assumes that “external accountability and the imposition of sanctions will force schools to improve and motivate teachers to change their instructional practices, resulting in better school performance” (Orfield & Wald, 2000, p. 39). Therefore, the problem is that students’ sole academic achievement in schools is based on test results. NCLB provides no progress clause for students who show growth. Students must meet standards on a test given once a year. Even if students show significant growth from one year to the next, but still do not meet expectations, this external accountability by NCLB will label them as a student who does not meet expectations.

Another problem with this external accountability placed on schools is the cost. According to Brad Johnson, a chief financial officer for one of the largest school district in the state of Georgia, the state-mandated test costs the district nothing (personal communication, September 2009). However, the costs to the local schools to provide test preparation materials, remediation for students at risk to fail, and teacher training fall to the local principal. Jennings and Renter (2006) reported that the requirements of NCLB have resulted in state and local educational officials expanding the roles of school without any additional federal funding to carry out the mandates of NCLB. The states have no choice but to be highly engaged in public education because the mandates of NCLB affect all public schools. Jennings and Renter reported expanding roles of the state including the creation of testing programs for elementary students in Grades 3 through 8 and for high school. State officials must set minimum testing scores for student to meet.

Cox (2006) reported that the state of Georgia has a mission to encourage all students to discover the joy of learning in a positive, culturally diverse, and challenging environment. Cox further reported that this mission has a vision for students: to reach
their full potential as individuals and citizens. However, the key to this vision is offering a curriculum that will identify what students need to know and how teachers can teach these goals. In 1986, public schools in Georgia were directed to follow a curriculum outlined in the Quality Based Education (QBE) Act of 1986. This was a sequenced curriculum known as the Quality Core Curriculum (QCC). This curriculum was a set of standards and expectations for learning. The QCC included content standards for reading, language arts, mathematics, science, social studies, foreign language, fine arts, health, physical education, technology education, career education, and English language learners (Mitzel, 1999).

However, an audit in 2002 concluded that the QCC lacked depth and rigor. Furthermore, the audit found that the curriculum did not meet national standards presented in the NCLB Act, which had just been entered into law in January of 2002. The NCLB reform model was a means for improving student achievement not only in the state of Georgia, but nationwide (Gonzalez, Hamilton, & Stecher, 2003).

Eacker, Dufor, and Burnette (2002) concluded that curriculums should be viable; these curriculums should be clear in defining what students should learn and how schools will plan if students do not meet expectations. Consequently, the state of Georgia revised its curriculum, moving from the QCC to a curriculum that was standards based. The new curriculum was called the Georgia Performance Standards (GPS). This new curriculum, put in place in 2005, provided clear guidance for instruction by defining what level of work a student must produce to meet a standard. The GPS identified the skills students must know and guided teachers on assessment practices (Cox, 2006).
Statement of the Problem

The state of Georgia administers the Criterion-Referenced Competency Test (CRCT) to every elementary student in grades 1 through 8. Before 2008, the state tested eighth-grade students on a quality core curriculum (QCC). In 2008, the state began testing students on a Georgia Performance Standard (GPS) curriculum. In order to make a direct comparison of the curriculum change on the tests scores, the 2008 CRCT should have contained items to test both curriculums (T. Greer, Associate professor, University of Southern Mississippi, personal communication, March 12, 2011). However, this was not done. Therefore, the current study was designed to examine if differences in students achievement occurred because of the curriculum change.

Purpose of the Study

This study sought to determine if differences in student achievement exist between assessments aligned with a performance-based curriculum and assessments that aligned with a quality core curriculum. Archival data were obtained and analyzed to determine if assessment changes affected student achievement. CRCT and ITBS data from 21 middle schools located in the southeastern region of the United States were obtained and used for the study. This study was guided by the following research question:

Is there a significant difference in student achievement between students who are taught and then tested based on a Quality Core Curriculum and those who are taught and tested using the Georgia Performance Standards?

Definition of Terms

The following terms will guide the reader in better understanding the terms used in this study:
**Criterion-Referenced Competency Test (CRCT).** The CRCT measures students’ mastery of the state’s curriculum in Grades 1-8. School improvement teams disaggregate results to identify subgroup patterns and align school improvement plans for the next year. Teachers use data at the beginning of school year to group students by those who do not meet expectations, those who meet expectations, and those who exceed expectations. Teachers review students’ lexile reading levels derived for appropriate selection of books and reading material (Georgia Department of Education, 2010c).

**Iowa Tests of Basic Skills (ITBS).** The ITBS is a nationally normed test that compares one student to another. Teachers and administrators review results to track trends. Teachers and administrators use results for gifted placement, foreign language placement, math placement, and remedial identification (The University of Iowa College of Education, 2010).

**Formative assessment.** Formative assessment is defined as assessment that is integrated into the teaching plan. Ainsworth et al. (2007) wanted educators to think of formative assessments as “assessments for learning that are collaboratively designed, administered, scored, and analyzed by team members” (p. 46). Guskey (2007) promoted the thought that formative assessments can change teaching and learning. Guskey asserted that summative assessment results are not available to teachers until it is too late to help learners. Using formative assessments as a part of the day-to-day instruction serves as meaningful sources of immediate information for teachers to use to re-teach or enrich based on current student information.

**The School Keys.** The School Keys: Unlocking Excellence through the Georgia School Standards is the foundation for Georgia’s comprehensive, data-driven system of school improvement (Georgia Department of Education, 2007b).
Summative assessment. Popham (2008) defined summative assessment as an assessment that tests what students know. Summative assessment can be the standardized tests given at the end of a school year to determine if the students have mastered the curriculum or they can be teacher-made to test the mastery of a unit of study. This form of assessment is an accountability tool for a school, district, and state. A noninclusive list of summative assessments defined and reviewed in this dissertation are (a) state assessments, (b) district or local school benchmarks, (c) end-of-unit tests, and (d) end-of-semester tests. The core point to remember about summative assessment is that it is the final piece of information and is used to determine students’ next steps, such as advancement to a higher class or the next grade level. It is the after of teaching.

Limitations/Delimitations

The following limitations of the study are recognized. Not all principals may choose to participate by sharing the school’s CRCT data. Students’ scores from 21 middle schools were compared. It is not known what test preparations students were subject to prior to the test. Not all students received the same quality of teaching. Although all teachers are required under NCLB to be highly qualified, factors such as experience were not measured. No study of the various teaching methods used were made, nor was the amount of remedial or enrichment instruction implemented in the schools measured. The schools ranged from Title 1 to high achieving, thus possibly limiting the findings of the data. However, it is the expectation of the researcher that this may also be considered positive since a range of schools and performances could provide the study with better validity. The study is delimited to only eighth-grade scores from 21 middle schools in the same school district.
Assumptions

The following assumptions guided this research. It was assumed that the CRCT tests sufficiently measure student performance in reading and math. Eighth-grade scores for reading and mathematics in 21 middle schools were used in the study. Participant selection bias was not a factor. All middle schools had the opportunity to participate. However, the district and local middle school principals had to approve participation. No threat to validity was assumed because the study covered the same testing periods and the same curriculum content for scores at the middle schools.

Justification of the Study

The aim of this study was to examine the specific application or use of a criterion-referenced test for the sole purpose of measuring student achievement. The use of one such test in a state in the southeastern United States is used for holding teachers, schools, and the school districts accountable for student achievement. This criterion-referenced test is also used as a means for promotion and retention in Grades 3, 5, and 8. According to William (2010), criterion-referenced tests are intended to make inferences about the quality of the education provided to students. Therefore, knowing that assessment is a key process in education, the examination of the history of assessment, and the use of formative and summative assessments may determine if this practice is influencing student achievement.

Summary

Porter (1995) helped educators think about the importance of testing and to question the intent of testing. Porter also contended that a fundamental misunderstanding of school reform and school issues exists. An examination of the history of assessment and a study of current assessment practices will guide the reader to question the intent
and uses of tests. It was the intent of this researcher to inform the reader of the history of assessment and to use current archival data to determine if current testing/assessment practices have a positive impact on student achievement.
CHAPTER II
LITERATURE REVIEW

William (2010) reported, “Assessment is a key process in education” (p. 37).

How do educators determine if instruction has had an impact on students? The learning community wants to know if students have learned (William, 2010). This chapter presents a brief history of assessment including types of assessments and assessment reform, examine the practice of using a criterion-referenced test to measure student achievement, study trends of assessments in the United States commonly referred to as high-stakes testing, and a review of literature on the QCC and the GPS. The benefit of knowing where and why assessments were designed is very important, specifically because the uses and degree of accountability has changed.

Types of Tests and Their History

During the early 1900s, the French government asked psychologist Alfred Binet to help decide which students were mostly likely to experience difficulty in schools. The government had passed laws requiring that all French children attend school, so it was important to find a way to identify children who would need specialized assistance (Kamin, 1974). Binet, and his colleague Theodore Simon, began to develop questions that focused on things that had not been taught in school such as attention, memory, and problem-solving skills. Using these questions, Binet determined which ones served as the best predictors of school success. Binet quickly realized that some children were able to answer questions that were more advanced than older children were generally able to answer, while other children of the same age were only able to answer questions that younger children could typically answer. Based on this observation, Binet suggested the
concept of a mental age, or a measure of intelligence based on the average abilities of children of a certain age group (Kamin, 1974).

This first intelligence test, referred to today as the Binet-Simon Scale, became the basis for the intelligence tests still in use today. However, Binet himself did not believe that his psychometric instruments could be used to measure a single, permanent, and inborn level of intelligence (Kamin, 1974). Binet stressed the limitations of the test, suggesting that intelligence is far too broad a concept to quantify with a single number. Instead, he insisted that intelligence is influenced by a number of factors that change over time and can only be compared among children with similar backgrounds (Siegler, 1992).

The Stanford-Binet Intelligence Test

After the development of the Binet-Simon Scale, the test was soon brought to the United States where it generated considerable interest. Stanford University psychologist Lewis Terman took Binet’s original test and standardized it using a sample of American participants. This adapted test, first published in 1916, was called the Stanford-Binet Intelligence Scale and soon became the standard intelligence test used in the United States (Siegler, 1992).

The Stanford-Binet intelligence test used a single number, known as the intelligence quotient (or IQ), to represent an individual’s score on the test. This score was calculated by dividing test takers’ mental age by their chronological age, and then multiplying this number by 100. For example, a child with a mental age of 12 and a chronological age of 10 would have an IQ of 120 (12/10 x 100). The Stanford-Binet remains a popular assessment tool today, despite a number of revisions over the years (Siegler, 1992).
Intelligence Testing During World War I

At the outset of World War I, U.S. Army officials were faced with the monumental task of screening a large number of army recruits. In 1917, the president of the American Psychological Association and chair of the Committee on the Psychological Examination of Recruits, psychologist Robert Yerkes developed two tests known as the Army Alpha and Beta tests. The Army Alpha was designed as a written test, while the Army Beta was administered orally in cases where recruits were unable to read. The tests were administered to over two million soldiers in an effort to help the army determine which men were suited to specific positions and leadership roles (McGuire, 1994).

At the end of WWI, the tests remained in use for a variety of situations outside the military with individuals of all ages, backgrounds, and nationalities. For example, IQ tests were used to screen new immigrants as they entered the United States at Ellis Island. The results of these mental tests were inappropriately used to make sweeping and inaccurate generalizations about entire populations, which led some who considered themselves intelligence experts to exhort Congress to enact immigration restrictions (Kamin, 1974).

The Wechsler Intelligence Scales

The next development in the history of intelligence testing was the creation of a new measurement instrument by American psychologist David Wechsler. Much like Binet, Wechsler believed that intelligence involved a number of different mental abilities, describing intelligence as, “the global capacity of a person to act purposefully, to think rationally, and to deal effectively with his environment” (Boake, 2002, p. 88). Dissatisfied with the limitations of the Stanford-Binet, Wechsler published his new
intelligent test known as the Wechsler Adult Intelligence Scale (WAIS) in 1955 (Boake, 2002).

Wechsler also developed two different tests specifically for use with children—the Wechsler Intelligence Scale for Children (WISC) and the Wechsler Preschool and Primary Scale of Intelligence (WPPSI). The adult version of the test has been revised since its original publication and is now known as the WAIS-III. The WAIS-III contains 14 subtests on two scales and provides three scores: a composite IQ score, a verbal IQ score, and a performance IQ score. Subtest scores on the WAIS-III can be useful in identifying learning disabilities, such as cases where a low score on some areas combined with a high score in other areas may indicate that the individual has a specific learning difficulty (Kaufman, 1990). Rather than score the test based on chronological age and mental age, as was the case with the original Stanford-Binet, the WAIS is scored by comparing the test taker’s score to the scores of others in the same age group. The average score is fixed at 100, with two thirds of scores lying in the normal range between 85 and 115. This scoring method has become the standard technique in intelligence testing and is also used in the modern revision of the Stanford-Binet test.

History of Assessment in the United States

This section of the literature review will focus on the implications of many ideas as they relate to educational assessment and policy. As reported in the earlier part of this chapter, the influence of educational assessments has a long history that found significant increases in standardized testing in post-WWI school systems. Many measurement specialists viewed their quantitative work as scientific, and given that most measurement specialists were trained in psychology they also tended to view themselves as scientists and not as educators (Porter, 1995).
The idea that certain assessments are the basis for students to receive scholarships, honors promotion, and class ranking is part of the history of testing in the United States (William, 2010). William also contended that one of the most distinguishing characteristics of these tests is that the stakes are higher for teachers than students. William believed that the use of high-stakes assessment are “high stakes for teachers but low stakes for students” and are common in the United States, but rare in other countries (William, 2010, p. 109). When Porter (1995) spoke to assessment and the lack of subjectivity as long been viewed as one of the hallmarks of assessment, it also provides a definition of fairness. It is important to embrace the idea there are other definitions of objectivity that have fewer negative connotations. William and Porter agree that assessment is used in ways that measure more than student achievement.

Testing specialists recognize their power and try to understand both the intended and unintended consequences of the political uses of educational assessments (Firestone & Bader, 1992). It may be uncomfortable to consider some of these issues, but it is no longer ethical to simply focus on technical issues and dismiss other policy issues as being outside of one’s purview. The State of Georgia’s Code of Ethics for Educators (Georgia Professional Standards Commission, 2010) added another code to be followed regarding ethical behavior in regards to testing. In part it stated:

(k) Standard 11: Testing - An educator shall administer state-mandated assessments fairly and ethically. Unethical conduct includes but is not limited to:

1. Committing any act that breaches test security; and 2. Compromising the integrity of the assessment. (p. 5)

The educational measurement community can play an essential role by providing expert judgment and advice to help policymakers contextualize the proper uses of
educational assessments. An example of policymakers’ blind trust in numbers is exemplified by the recent re-authorization of the Elementary and Secondary School Act (No Child Left Behind), with its strong support for quantification. One of the four basic principles of NCLB is that more accountability will lead to increases in educational achievement. More accountability translates into significant increases in the amount of testing to meet the federal requirement of adequately yearly progress (AYP) on measurable educational objectives. Ultimately, the goal of NCLB is to have all American students proficient by the year 2014 (Linn, Baker, & Betebenner, 2002). NCLB is the reason school systems have increased the use of standardized testing. Each state is required to have a testing program or federal funding will be withheld (Hogan, 2007).

NCLB is reform that distinguishes the differences between students in terms of each individual’s learning outcomes as measured by a test or tests (William, 2010). The NCLB perspective has implications for standards-based educational reform that requires students to be classified into categories, such as those used by National Assessment of Educational Progress: basic, proficient, and advanced. Many states report academic progress on their own assessment measures of student achievement in reading and math (Lee, 2010). However, Lee reported that many states responded to the NCLB policy mandates and translated them into their own test-driven accountability plan thus leaving each state to their own testing policy. Consequently, states made AYP reporting academic achievement and reported that achievement in many different ways; resulting in students, teachers, school districts, and states being ranked, but ranked using a variety of different assessment results (Lee, 2010). NCLB does allow each state to establish school accountability protocols, but the federal government still independently measures schools and reports these results on a national report card. Comparison of the National
Assessment of Educational Progress assessments and state assessments yield that individual state assessments are not as rigorous as they should be (Lee, 2010).

Lee (2010) suggested need for concern when measuring student and school performance using state tests to categorize teachers and students. Lee concluded that state tests designed to measure student achievement are not only tricky and controversial, but each state’s different tests have produced varied results. The NCLB was intended to assure that 100% of students secure reading and math proficiency (Lee, 2010). Could the low precision of many test scores as reflected in the high standard errors of measurement suggest that there will be a significant number of inaccurate classifications? Lee’s research suggested that the NCLB Act has been an assessment mental quick fix. Lee recommended that federal and state governments coordinate accountability efforts by setting goals that are more realistic and use multiple measure of school success to improve the quality of education for all stakeholders.

Curriculum Assessment in the State of Georgia

The Gwinnett County Public Schools, one of the three largest school districts in the state of Georgia implemented a technical advisory committee in 2010. Gwinnett County serves as a model for how local school districts can implement a sound system of educational assessment (Gwinnett County Public Schools, 2010). Policymakers in Gwinnett County balance their high-stakes testing with other assessment pieces. Every educational decision is made on the basis of multiple sources of information. Gwinnett County school administrators strongly resist the sole use of test scores, and create an environment for assessment that maintains a deep concern for individual students (Lee, 2010). The educational measurement specialists in Gwinnett have personal knowledge of
every student failing the Gwinnett County Gateway tests, and contribute to the
development of detailed plans to help each student succeed.

In Cobb County, another of the three largest school districts in Georgia, teachers,
local building administrators, and district administrators serve on a district-wide
assessment advisory committee (R. Benson, Chief Academic Officer and Assistant
Superintendent, Cobb County Public Schools, Georgia, personal communication, 2010).
This committee also serves as a model for how local districts across the state can
implement a sound testing plan. This committee has developed a standards-based
assessment chart as well as an assessment template that each school must complete and
submit. The committee advises the local board of education on exactly how each school
in this district must administer nationally standardized tests, state assessments, and board
of education mandated assessments (J. Jones, Chief Accountability Office, Cobb County
Public Schools, Georgia, personal communication, May 18, 2011). However, more
important is that this committee and this district allow local schools to implement local
assessments that meet individual school and student needs instead of mandated
benchmarks as in the past (Cobb County Public Schools, 2011).

Quality Core Curriculum

The QBE Act of 1985 required the Georgia Board of Education to develop a
curriculum to be implemented in all public schools in the state. This curriculum required
a state assessment. The Act authorized the Board to create student competencies that
each student should master before the end of the student’s public education (Mitzell,
1999). In compliance with this mandate, the state Board of Education, along with a task
force, created a draft of the basic curriculum content for all Georgia public schools. After
review by the task force, teachers, and local school systems, the QCC curriculum was implemented in August 1988 (Mitzell, 1999).

Each student in each grade level was given a set of concepts to master. The Georgia Criterion-Referenced Competency Test (CRCT) assessed the students’ mastery of the QCC content (Mitzell, 1999). The QBE Act required the Georgia Department of Education to revise the QCC curriculum systematically. The then governor and secretary of the Department of Education created a task force to revise the QCC curriculum (Mitzell, 1999). This group asked for half of the revision writers to be teachers and began working on revisions in 1995.

After a long revision process including stakeholders from all areas of the community and a balance of race and gender, a draft of the QCC went to review (Mitzell, 1999). Reviews were positive. The new QCC curriculum was touted for being clear and specific. Correlations between the Iowa Tests of Basic Skills and QCC were a part of this revision process and were considered positive. After this revision process, the QCC was implemented in all Georgia public schools in 1998 (Mitzell, 1999).

Gandal (1995) reported that the QCC curriculum met the standards of the American Federation of Teachers criteria for core subjects. However, Massell, Kirst, and Hoppe’s (1997) concluded that the QCC curriculum was not comprehensive enough to be used as a statewide assessment. Other researchers indicated that the QCC was simply focused on basic skills and did not extend the learner’s knowledge past basic skills (Firestone & Bader, 1992; Marzano, 1997). Firestone and Bader (1992) found that the QCC was somewhat effective in rural areas of Georgia, but in areas that were more affluent, the QCC curriculum restricted courses that had rigor and challenged learners. Therefore, an audit of the QCC was conducted in 2002. The audit concluded that the
QCC curriculum lacked depth and rigor. The audit also concluded that the curriculum did not meet the national standards presented in the NCLB Act that had just become law in January 2002.

*Georgia Performance Standards*

Eacker et al. (2002) concluded that for curriculums to be viable, these curriculums should be clear in defining what students should learn and how schools will plan if students do not meet expectations. Consequently, the state of Georgia revised its curriculum, moving from the QCC to a curriculum that was standards based. This new curriculum, referred to as the Georgia Performance Standards (GPS) and put in place in 2005, provides clear guidance for instruction by defining what level of work a student must produce to meet those standards. The GPS identifies the skills students must know and guides teachers on assessment practices (Cox, 2006). The GPS was aligned with the CRCT, taking the guesswork out of teaching. The GPS provides guidance on best instructional practices that have been proven to be effective in places such as Michigan and Texas (Davila-Medrano, 2003).

The GPS goes into much greater depth than the QCC. The QCC simply gave the teacher a standard. In contrast, the GPS gives the teacher not only the standard, but a suggested task, sample student work, and teacher commentary (Davila-Medrano, 2003). Additionally, the GPS curriculum provides clear expectations for assessment. These standards define the level of work that demonstrates proficiency levels. Not provided by the QCC, the GPS curriculum provides teachers with skills needed to problem solve reason, communicate, and make connections with other information. Most importantly, the GPS advises teachers how to assess students on how well and to what extent they
know the content or can apply the information to other content or problems (Ravitch, 1996).

Essential to the development of the GPS curriculum was to ensure all students in every school had access to challenging programs (Bransford, Brown, & Cocking, 1999). These content standards provided in the GPS make it possible for teachers to prepare lessons. Testing experts use them as a foundation for student tests to determine if they meet standards or to what extent students meet standards.

Massell et al. (1997) reported that recent research that compares a standards-based curriculum with a curriculum that is not standards based, like the QCC, showed no difference in student achievement. Contrary to this, history reveals that researchers such as Bruner and Dewey stated that a standards-based curriculum could be integrated from general standards to specific standards. In addition, they agreed that the more comprehensive a curriculum is the more effective it is (Giordano, 2005).

Criterion-Referenced Competency Test

The CRCT was developed to test student mastery of the QCC. Currently it measures student mastery of the GPS. The CRCT is used by schools and the state to measure strengths and weaknesses (Cox, 2006). The first use of the CRCT was in spring 2000. Only English/language arts, reading, and math were tested and only in Grades 4, 6, and 8. Science and social studies were added to the CRCT in spring 2002. Scale scores on all reports ranged from 150 to 500. Not only was a total scale score reported but each domain of each content area received a scale score. Scale scores were equivalent across tests from the same content area, such as reading, and across the grade levels. Scale scores 350 or above were identified as exceeding standards. Scale scores 300–349 were identified as meeting standards, and scale scores below 300 were identified as not
meeting standards. Scores below 300 indicated a need for remedial instruction (Cox, 2006).

With the implementation of the GPS, the scoring changed. Although scale scores were still used, 850 or above was identified as exceeding standards, 800–849 was identified as meeting standards, and below 800 was identified as not meeting standards. If a student, school, or system did not meet standards, the state’s level of minimum proficiency was not met and some type of intervention was required under NCLB (Cox, 2006). By Georgia law, the CRCT is used as an indicator of the pass/fail status of a school and district under NCLB. If a school or a district does not meet standards, it is deemed not to have made AYP and is labeled a failing school (Georgia Department of Education, 2011c).

Currently, Georgia administers the CRCT in Grades 1-8 in reading, language arts, math, social studies, and science. Since spring 2004, third graders must meet or exceed standards in the content area of reading in order to be promoted to the next grade. Currently, students in Grades 3, 5, and 8 must meet expectations on reading and math to be considered for promotion to the next grade level (Cox, 2006).

Summative and Formative Assessment

Black and William (1998) reported that, “The assessment process is characterized as a cycle of involving elicitation of evidence, which when interpreted appropriately may lead to action, which in turn, can yield further evidence and so on” (p. 141). Assessments provide information about students and are done for different reasons and are used in different ways. According to Clemson and Clemson (1996), assessment may be criterion referenced, where success is measured against the task itself, or norm referenced, locating work in relation not only to the task but also to the work of others.
Criterion-referenced and norm-referenced tests are considered summative assessments. Popham (2008) defined summative assessment as tests of what students know. A summative assessment can be the standardized tests that are given at the end of a school year to determine if the students have mastered the curriculum or they can be a teacher-made test that measures students’ mastery of a unit of study. This form of assessment is an accountability tool for a school, district, and state. Summative assessments may include (a) state assessments, (b) district or local school benchmarks, (c) end-of-unit tests, and (d) end-of-semester tests.

The core point to remember in regards to summative assessment is that it is the final piece of information and is used to determine students’ next steps such as advancement to a higher-level class or next grade level. It is the after of teaching. Black and William (1998) defined summative assessment as tests that are conducted and nothing happens at the end of the test that changes instruction.

Formative assessment is defined as assessment that is integrated into the teaching plan. Ainsworth et al. (2007) wanted educators to think of formative assessments as assessments that measure learning. These assessments are developed, scored, and analyzed by a team of teachers. Guskey (2007) promoted the thought that formative assessments can change teaching and learning. Guskey asserted that summative assessment results are not available to teachers until it is too late to help learners. Using formative assessments as a part of the day-to-day instruction serves as meaningful sources of immediate information for teachers to reteach or enrich based on current student information. Formative assessment is not about the after, but about the now of teaching. Formative assessment should be considered as the art of teaching, not a thing to do to students after a lesson. Formative assessments may include (a) student criteria
and goal setting, (b) observations, (c) self- and peer assessments, and (d) student recordkeeping.

Reeves (2007) edited a compilation of researchers’ work in the area of assessment. Ainsworth et al. (2007) agreed that both summative and formative assessments are important and can be used to influence student achievement. Most evidence reported that concepts embedded in reform models such as NCLB are summative and are not effective in increasing student achievement (William, 2007). However, Reeves and Shepard (2000) asserted that formative assessment informs, affects, and even maps future improvement in instructional practices.

Formative assessments are constructed to monitor student progress during the ongoing learning process (Stiggins, 2008). These concepts fall into two categories: assessment OF learning (summative) and assessment FOR learning (formative). Assessment for learning (AfL) helps enhance teaching while assessment of learning (AOL) is of limited value. While the concept of AFL is not new, the concept has advanced greatly since the work of Black and William (1998). Black and William posed the question, “Is there evidence that improving formative assessment raises standards?” and concluded that the answer was “an unequivocal yes” (p. 3). Their research has been credited with showing that developing AfL is one of the most powerful ways of improving learning and teaching and raising standards. Formative assessment is the process of seeking and interpreting evidence for use by learners and their teachers to decide where the learners are in their learning, where they need to go, and how best to get there (Stiggins, 2002).

Central to the idea of AfL is the idea of the role of students; the role of children as active learners, involved in their own learning is paramount. One way to begin this
process is by sharing the learning objectives and success criteria of each lesson with students (Stiggins, 2002). The process of learning has to be in the minds of both learner and teacher when assessment is planned and when the evidence is interpreted. Learners should become as aware of the how of their learning as they are of the what.

The value of sharing the learning objectives is especially apparent when students are given time to reflect on and discuss what they have learned and inform the teacher where they have experienced problems. Students are able to begin the process of self-assessment by comparing what they have learned with the objectives of the lesson and are able to evaluate their performance by checking their work against the success criteria. Sadler (1989, as cited in Black & William, 1998, p. 9) described this as “central to learning” as “pupils can only achieve a learning goal if they understand that goal and can assess what they need to reach it.” After children have become experienced in self-assessment then peer assessment, where one child gives feedback to another may follow. Black and William described this method as valuable because “students may accept from one another criticisms of their work, which they would not take seriously if made by their teacher” (p. 10). Peer feedback will often be verbal in the children’s own level of language, which may benefit some children.

Feedback itself is also central to the idea of AfL. As Kyriacou (1997) stated, “It is not true that ‘practice makes perfect.’ Rather it is practice plus feedback that makes perfect” (p. 93). Black and William (1998) stated, “It is the nature rather than the amount that is critical when giving feedback on both oral and written work” (p. 8). McTighe and O’Connor (2005) concluded that pupils make a greater gain in progress when they receive commentary feedback that is not associated with a grade. McTighe and O’Connor contended that the use of both oral and written feedback is one of the seven
most important practices that enhance student learning. Consequently, positive marking schemes have been implemented by many schools. Georgia’s implementation of the School Keys introduced a rubric to help teachers provide commentary when an aspect of the work has been done well or needs improvement (Georgia Department of Education, 2007a). It is based on the premise that it is important to communicate specifically on student work that meets or does not meet expectations. Students need to be able to understand specific learning outcomes. Students need to be told why their learning meets or does not meet expectations if they are to build on their success.

Formative assessment may include verbal feedback that may also be more appropriate during a lesson to correct misconceptions as a lesson progresses. Whatever form the feedback takes, it is vital that it is implemented in a method that helps the student identify and apply evaluation criteria and monitor learning in an ongoing fashion. Preserving motivation is key; “Assessment for learning should be sensitive and constructive because any assessment has an emotional impact” (Stiggins, 2002, p. 760). Comments should focus on the work produced rather than on the individual if they are to be helpful for both learning and motivation. Black and William (1998) succinctly stated, “Comments should identify what has been done well and what still needs improvement and should give guidance on how to make that improvement” (p. 9).

Questioning is also at the heart of AfL as, by nature, it provides a tool for gathering information quickly and on the spot. Teachers use information collected in this way formatively during every lesson, using responses to inform the next questions. If used effectively this technique can elicit existing ideas from children in order to provide a base for new ideas to be scaffolded upon (following a constructivist view of learning). However, some teachers do not teach lessons that include discussions in a manner that
may affect student learning (Black & William, 1998). One criticism is that not enough
time is given for student responses; that the teaching design around questions are those
that can be answered quickly without thought (William, 2007). A good question should
be open and require emphasis on higher-order thinking skills, creative thinking, and
analysis. A teacher should allow time for children to answer. If a thought out and
reasoned answer is desired, students will need time to articulate their responses, but must
begin with a well-designed question.

Summative assessment is now largely associated with NCLB. These tests are
used by teachers and districts to place children on a scale of achievement. There is a
danger (whether the level given is decided by teacher assessment or a formal test) that
children may become labeled at an early age as underachieving and that this may
perpetuate a cycle of constant low achievement. However, if handled correctly, these
level descriptors should help teachers identify how best to move children to the next level
and to achieve their individual potential. Achieving individual potential is important,
especially in the field of formative learning where the importance of individual learning
styles is stressed.

The Iowa Tests of Basic Skills (ITBS) is one example of how national data from
summative assessments are now being used formatively to help individual pupils.
Schools are able to analyze pupil level performance data against the results of pupils
nationally. This use of summative data for formative purposes is an area where
educational standards may be raised (William, 2010). The role of summative tests should
be to drive students’ course of study (Joughin, 2010), but is often only used to measure
student success and make future educational decisions based on the results.
The ITBS is administered to students in Grades 3, 5, and 7 in the state of Georgia. The scores are used to determine gifted eligibility, to place students in tiered mathematics classes, and to determine a rank for students, schools, and districts compared to others across the nation. The ITBS can provide information about each student’s strengths. According to The University of Iowa College of Education (2010), the specific purposes of the ITBS are:

1. To help determine the extent to which individual students have the background and skills needed to deal successfully with the academic aspects of an instructional program or a planned instructional sequence;
2. To estimate the general developmental level of students so that materials and instructional procedures may be adapted to meet individual needs;
3. To identify the areas of greatest and least development to use in planning individual instruction for early intervention;
4. To establish a baseline of achievement information so that the monitoring of year-to-year developmental changes may begin;
5. To provide information for making administrative programming decisions that will accommodate developmental differences;
6. To identify areas of relative strength and weakness in the performances of groups (e.g., classes), which may have implications for curriculum change—either in content or emphasis—as well as for change in instructional procedures;
7. To provide a basis for reports to parents that will enable home and school to work together in the students’ best interests.
The widespread belief that schools are not helping all students achieve has spurred efforts to reform our schools (McChesney & Hertling, 2009). Concerns have been raised that the ways students are taught and assessed do not lead students to acquire needed knowledge or skills, nor help them apply and use their knowledge and skills appropriately. Content standards containing the types of knowledge, skills, and behaviors now believed needed for all students to achieve at high levels are being developed at the national and state levels. Starting with such efforts as the National Council of Teachers of Mathematics (1989) Curriculum and Evaluation Standards for School Mathematics, content standards are being developed in the arts, civics, economics, English, foreign languages, geography, health education, history, physical education, science, and social studies.

School reform is also motivated by the belief that there are competencies needed for graduates to enter the workforce successfully. The Secretary’s Commission on Achieving Necessary Skills developed generic competencies and foundation skills that all workers will need in the future (U.S. Department of Labor, 1991). The foundation skills include flexible problem solving, respecting the desires of the customer, working well on teams, taking responsibility for one’s own performance, and continuous learning. These skills were developed to guide the efforts of educational reform in the direction of helping more students make the transition to work successfully. Collectively, these standards represent substantial challenges for American schools. They imply that all students will need to achieve at much higher levels. New strategies for assessment are also implied by these content standards.
Georgia, in an effort to reform public schools, developed a comprehensive school reform model called the *School Keys* (Georgia Department of Education, 2007a). These keys are the foundation for Georgia’s comprehensive, data-driven system of school improvement. This initiative is associated with several respected research frameworks, such as Marzano’s backwards design. These keys define what schools need to know, understand, and be able to do. They are intertwined models both couched in the notion of improved student achievement. The Assessment Key is very comprehensive and detailed. The key contains an overall definition of assessment, three standards that contain at least one substandard, a rubric to rate each substandard, and elements or operation descriptors for the standard. The standards include:

1. A cohesive and comprehensive system is in place to ensure that administrators and instructional personnel use assessment data to design and adjust instruction.

2. A variety of effective and balanced assessment is routinely and systematically implemented by all instructional personnel as part of a comprehensive school-based assessment and evaluation system.

3. Assessment and evaluation data are analyzed to plan for continuous improvement for each student, subgroup of students, and the school as a whole (Georgia Department of Education, 2007b).

School districts in Georgia had to develop a systematic and comprehensive training and redelivery regime to educate teachers quickly and effectively on the *School Keys*. Teachers were and are expected to use the data from summative and formative assessment to improve student achievement (R. Benson, Chief Academic Officer and
Assistant Superintendent, Cobb County Public Schools, Georgia, personal communication, 2010).

The importance of a balanced assessment plan, like the *Schools Keys*, is not the kind of assessments educators administer, but rather how an assessment is used as a part of instruction to support learning (Shepard, 2000). Popham (2004) suggested that teachers may implement both formative and summative assessments, but do not know what to do with the assessment results. Popham recommended that everyone who is working with students to close achievement gaps become assessment literate. These educators must understand each different assessment and what that assessment might reveal. Chappuis, Chappuis, and Stiggins (2009) contended that educators will better attend to students’ academic needs if balanced assessment systems are built. However, they add that these multiple and different measures of student achievement must be accompanied with educators who are assessment literate.

**The Reform of Assessment**

Student assessment is at the top of the reform list for policymakers at the national and state levels (Linn, 2000). Assessment is important because it is widely believed that what gets assessed is what gets taught, and that the format of assessment influences the format of learning and teaching (O’Day & Smith, 1993). Assessment is viewed as the way to set more appropriate targets for students, focus staff development efforts for the nation’s teachers, encourage curriculum reform, and improve instruction and instructional materials in a variety of subject matters and disciplines (Darling-Hammond & Wise, 1985). The hope of policymakers is that changes in assessment will not only bring about the needed changes in students, but also in ways schools are organized (Linn, 2000).
Interest in performance assessment has also been justified on the basis that using such measures will promote educational equity (Linn, 2000).

However, outside pressure external on testing programs can be ignored or resisted by local educators (Shepard, 2000). There is also ample evidence of the distortions in teaching that external testing programs can create (Smith & Shepard, 1989). Rather than encourage reform of teaching, inappropriate teaching to the test may occur (as opposed to teaching to the domain covered by the test). Rather than creating opportunities for all students to learn to high levels, even new forms of assessment may lead to tracking and limiting opportunities for some students (Darling-Hammond, 1994).

Assessment reform should occur along with professional development, instructional development, and other strategies designed to assure that all of the changes are mutually supported (Georgia Professional Learning Standards, 2011). Coordination of assessment reform at the national and state levels with assessments at the local level is also important, so that each will present a coherent view of student performance, not simply be stuck together. In addition, as mandated by the QBE Act of 1985 in the state of Georgia, a curriculum must be developed and maintained. This curriculum must specify what students are expected to know in each subject at each grade level.

Another consideration to apply to the study of assessment reform is the idea of progress monitoring. Progress monitoring is a set of assessment tools that actually monitor the curriculum (National Research Center, 2006). Slope in assessment scores occur sometimes after a curriculum has been implemented. Slope may be applied to states who are implementing a new curriculum and consequently testing the curriculum over a period of time. To effectively reform assessment practice, what is being taught must be monitored. The National Center for on Student Progress Monitoring (2011)
reports not only is progress monitoring important, it is scientifically based. Then reform models may monitor testing results looking for slope which is exactly what it implies, a steepness or a rise between two tests scores measured from one curriculum to another.

Types of Assessments

New content standards may require different assessment methods. In addition to multiple-choice exercises, assessment techniques now being considered are short-answer, open-ended; extended-response, open-ended; individual interviews; performance events; performance tasks in which students have extended time; projects; portfolios; observations; and anecdotal records. A broader repertoire of techniques is increasingly being used.

School Improvement Strategies

The information about student achievement needed at various levels of the educational system is different. Parents have different needs than do teachers, who in turn, have different needs than do school principals. District administrators need broader, systemwide information, while at the state level, there is concern about equity across districts and identification of state priorities. Nationally, policymakers are concerned about differences between states and how competitive American students are with their peers in other countries (DiMartino & Miles, 2004).

School reform occurs at all levels of the educational system. Teachers work with individual students in a classroom. Sometimes teachers make data-driven instruction changes by revamping classroom instruction based on an assessment (Ainsworth et al., 2007). At the school level, educators use school information to set long- and short-range objectives and decide how to accomplish these. At the district level, educators target particular areas of the curriculum for attention. At the state level, incentives for
improving instructional programs may be most important. Improving student achievement can take place at each of these levels.

*Useful Assessment Designs*

Student achievement is usually measured with available student test data, often using information from district or state testing programs. Information collected less formally in classrooms is not typically included in school improvement plans, even though such information could provide valuable insights into student learning. However, current districts may indicate a change in this practice (R. Benson, Chief Academic Officer and Assistant Superintendent, Cobb County Public Schools, Georgia, personal communication, 2010). The nature of information needs should form the basis for an assessment design. In a top-down model, policymakers develop an assessment design that meets their needs, hoping the data may be useful by persons at lower levels. An alternative is to build the assessment system needed at the local level, aggregating the information upwards to the district, state, and national levels.

Another model, based on the assumption that multiple approaches will allow different users’ needs to be met, is to develop a comprehensive assessment system using different assessment formats. Various assessment strategies can be implemented together at the different levels to provide for the different information needs in a coordinated, coherent manner (Darling-Hammond, 1994). For example, local districts can adopt a portfolio system for improving instruction, while the state carries out matrix-sampling across important standards. The information collected by the state can become part of the student’s portfolio, thereby strengthening the portfolio’s quality. The state could also provide opportunities for teachers to learn to score the open-ended written and performance assessments, thereby enhancing teachers’ capabilities of observing and
rating student performances in their classrooms (Thomas et al., 2005). In this case, the elements of the system at the different levels build on and support the elements at other levels. It is also anticipated that information collected at the different levels can be reported in a more understandable manner, since the same standards apply in different ways. This assessment model enhances the reforms of schools that so many desire (Thomas et al., 2005).

High-Stakes Testing

The advent of the No Child Left Behind federal legislation has plunged educators into an unprecedented era of high-stakes testing. With this high-stake testing, comes the notion that testing and more testing will drive improvement in instruction and student achievement (Stiggins, 2008). Yet there are many flaws to this approach, an approach that, under a masque of academic excellence, threatens to undermine the tenets of exemplary instructional practice, and leave behind the very students that the legislation and testing movement purport to be helping (Lee, 2010). In a society with issues that include race and income, it is of paramount importance that standards be created to determine what all students should know and be able to do upon graduation from middle school and from high school (Stiggins, 2008).

The absence of standards almost guarantees unfair distribution of resources and access to knowledge, based upon income, color of skin, and the community and neighborhood in which one lives. Two Massachusetts Department of Education studies found statistically significant differences in algebra and U.S. history courses in suburban, rural, and urban districts. Suburban students were given more homework, had longer class periods, had greater access to educational materials, and received a more rigorous curriculum than urban and rural students enrolled in courses with the same title
(Massachusetts Department of Education, 1986). The Massachusetts Department of Education concluded that minority students were not provided the same curriculum as White students were. According to Newmann, Bryk, and Nagoaka (2001), standards form a tool to promote high quality curriculum and instruction for all.

Standards that promote equity are broad and streamlined in nature. They are designed to ensure that every student graduates with the skills and knowledge to be a contributing member of a democratic society, prepared for a future productive life (DiMartino & Miles, 2004). Broad, streamlined middle grades standards promote relevant and rigorous instruction and curriculum that builds upon the principles of early adolescent development. These standards allow for locally determined assessments that require students to demonstrate mastery over essential knowledge through varied means, including portfolios, demonstrations, and exhibitions (DiMartino & Miles, 2004; Thomas et al., 2005).

However, standards can also reinforce and amplify current inequities that pervade public education (Bransford et al., 1999). The National Forum to Accelerate Middle Grades Reform reinforced its belief in standards and assessments, quality instruction, and higher levels of learning for every student (Ames, 2008). The National Forum reported that no single test should ever be the sole determinant of a student’s academic future, whether it be promotion to the next grade, special placement, or transition from middle grades to high school (Ames, 2008). Rothstein (2000) questioned the validity of assessing a student’s knowledge at one point in time. Studies have documented that how students fare on standardized tests can be greatly influenced by many external factors, including stress over taking the test, amount of sleep, distractions at the testing site, time of day, and emotional state (Kamin, 1974; Sacks, 2000). Revisit the origins of testing in
the United States—resulting in tests such as the Stanford-Binet—to understand that the roots of high-stakes testing lie in sorting students (Sacks, 2000).

It is important to take a look at the impact of state-level high-stakes testing on the standards movement, and the resulting impact on curriculum and instruction. The goal of this dissertation was to examine the history of assessment and the effects on student achievement. Educators need to examine the industry claims of validity and reliability of high-stakes standardized tests, and question the very reasons our government and the education industry have embraced these tests (French, 2003)

Standardized tests are also poor predictors of how well students can apply the knowledge that they do demonstrate on these tests (French, 2003). A recent study examined data from 18 states that have implemented high-stakes testing programs to assess whether students gained any knowledge that they could apply elsewhere, other than learning the necessary facts for performing on a state’s high-stakes test (Amrein & Berliner, 2002). Amrein and Berliner concluded that,

Analyses of these data reveal that if the intended goal of high-stakes testing policy is to increase student learning, then that policy is not working. While a state’s high-stakes test may show increased scores, there is little support in these data that such increases are anything but the result of test preparation and/or the exclusion of students from the testing process. (p. 2)

Kane and Staiger (2001) found “between 50 percent and 80 percent of the improvement in a school’s average test scores from one year to the next was temporary and was caused by fluctuations that had nothing to do with long-term changes in learning or productivity” (p.37). According to Kane and Staiger, these findings are caused by the variations in student population in a given grade from year to year, variations in test
conditions, and the presence of rewards and punishments from the state or district. Other studies have concluded that students lost from the testing pool—through grade retention, dropping out, or being excluded from taking the test because of having special education or bilingual status—create the appearance of an improving school when in fact the opposite may be true (Test scores unreliable means of assessing school quality, 2001).

Next Steps for Assessments in the United States: New History

A commission comprised of governors and heads of state departments of education from all the states of the union formed the Common Core State Standards Initiative in spring 2009. The Initiative was developed to create a set of national standards that will hold all students across the United States to the same academic expectations (Kendall, 2011). In June 2011, the common core standards for English language arts and mathematics, and literacy in history/social studies, science, and technical subjects were published (Kendall, 2011).

Resnick and Berger (2010) stressed that our current assessment practices is simple; it is a test-based accountability system, not a standards-based accountability system. Resnick and Berger made a case that this current system does not support teaching and learning because teachers often resort to teaching the test by using practice materials that practice the format of the tests, especially when students are at-risk learners. Kendall (2011) stated that schools must have the flexibility to incorporate common core standards in ways that engage students in learning and that learning is applied to choices that can be made after graduation.

For the first time in the history of education and assessment in the United States, the common core standards is an agreed upon set of instructional standards that will be implemented in 47 states across the United States (Georgia Department of Education,
Implementation of these standards will help teachers better prepare students for success beyond graduation (Georgia Department of Education, 2011b). There are three reasons for adopting the common core standards. The common core standards integrate some of the current GPS, but offer a more rigorous curriculum that 47 states are ready to implement. Secondly, the common core standards will not only allow but will facilitate comparison of students’ achievement scores from state to state. Finally, the implementation of the common core standards will allow states to save money. With a common curriculum, textbook companies will be able to focus on the common core standards, not standards from state to state to state. With this focus, these instructional resource vendors will be able to reduce prices.

Students will be assessed on these common core standards for the first time in 2014 (Kendall, 2011). This new assessment will influence instruction by asking teachers to focus on creating rigorous lesson plans (Kendall, 2011). Educators must make sure to focus on the tenants of the standards. The standards only describe what to teach, not how to teach, so teachers are still in charge of creating meaningful lessons (Georgia Department of Education, 2011b). The common core standards have captured the best of the past 20 years of standards-based education research and build upon lessons learned from history (Kendall, 2011).

Summary

Education reform has been a topic for educational policymakers since the inception of the public school system, and educators have explored the specialized needs of assessments for decades. The state of Georgia developed a comprehensive system to provide a framework to describe and define what students need to know and what schools need to know, understand, and be able to do. This chapter, although not comprehensive,
is one view of education in regards to the history of testing. In addition, it explored the concepts of standards-based assessment so that the questions of effectiveness can be addressed. The ultimate goal of educators is to help students become educated. To do this, teachers and all educators must help them learn to improve the quality of their work and to see their own capabilities (Ainsworth et al., 2007). However, educators must also assess how they test to be reflective and visionary in their efforts to improve student achievement.
CHAPTER III

METHODOLOGY

The purpose of this study was to examine the difference in student achievement when students are taught and then tested under different curriculums. The study examined the history of testing in the United States and Georgia, and examined the question, what difference in student achievement exists when students are taught and then tested based on a Quality Core Curriculum versus the Georgia Performance Standards?

Research Design

This research was a descriptive study designed to obtain information concerning the status of a phenomena (Ary, Jacobs, & Razavieh, 1990). The aim of the study was to determine what difference exists when comparing the achievement scores of students taught using two types of curriculums. This study was guided by the following research question:

Is there a significant difference in student achievement between students who are taught and then tested based on a Quality Core Curriculum and those who are taught and tested using the Georgia Performance Standards?

Setting

The school district consists of 24 middle schools. One principal did not agree to participate in the study and two of the schools were not open in 2003. Therefore, data were collected from 21 middle schools. The enrollment of each of the middle schools ranged from 814 students to 1367 students. Table 1 contains a demographic description of the 21 middle schools in the study. These middle schools are very diverse in that many have a stable population while others experience mid to high transiency rates. The district community is characterized as diverse with ethnic populations differing greatly
from middle school to middle school. Even though no demographic comparisons were conducted in this study, it is important to visualize the demographic differences to understand and generalize the purpose of the study.

Data Collected

Only eighth-grade CRCT and ITBS scores for reading and math were used. The rationale for this decision was that these scores are used to determine adequate yearly progress. Data came from the collection of the official school summary report from the CRCT and ITBS tests for the subject areas of reading and mathematics at the eighth-grade level. The following summary reports were collected:

- Eighth-grade CRCT summary reports for math from spring 2003 to spring 2010
- Eighth-grade CRCT summary reports for reading from spring 2003, to spring 2010
- Eighth-grade ITBS summary reports from 2003 to 2010

In addition, the state’s performance summary sheets for each year between 2003 and 2010 were obtained to gather the state’s mean score and standard deviation for all eighth graders tested each year.
Table 1

Demographic Description of School in Study (Percent of Students)

<table>
<thead>
<tr>
<th>ID</th>
<th>Asian</th>
<th>Black</th>
<th>Hisp</th>
<th>AmIn</th>
<th>White</th>
<th>MR</th>
<th>SWD</th>
<th>ELL</th>
<th>ED</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3.0</td>
<td>29.2</td>
<td>10.1</td>
<td>1.0</td>
<td>51.3</td>
<td>5.4</td>
<td>8.1</td>
<td>1.3</td>
<td>33.2</td>
</tr>
<tr>
<td>2</td>
<td>1.8</td>
<td>53.9</td>
<td>31.8</td>
<td>0.8</td>
<td>8.7</td>
<td>2.9</td>
<td>9.7</td>
<td>11.1</td>
<td>78.4</td>
</tr>
<tr>
<td>3</td>
<td>1.0</td>
<td>77.2</td>
<td>10.2</td>
<td>0.0</td>
<td>11.6</td>
<td>0.0</td>
<td>14.2</td>
<td>4.0</td>
<td>65.3</td>
</tr>
<tr>
<td>4</td>
<td>10.3</td>
<td>32.7</td>
<td>14.7</td>
<td>0.6</td>
<td>41.7</td>
<td>0.0</td>
<td>14.1</td>
<td>3.9</td>
<td>47.1</td>
</tr>
<tr>
<td>5</td>
<td>14.0</td>
<td>8.6</td>
<td>2.7</td>
<td>0.2</td>
<td>71.5</td>
<td>3.0</td>
<td>10.6</td>
<td>2.7</td>
<td>5.9</td>
</tr>
<tr>
<td>6</td>
<td>16.2</td>
<td>3.2</td>
<td>3.5</td>
<td>0.3</td>
<td>73.3</td>
<td>3.5</td>
<td>10.5</td>
<td>3.2</td>
<td>6.5</td>
</tr>
<tr>
<td>7</td>
<td>2.2</td>
<td>9.4</td>
<td>2.7</td>
<td>0.3</td>
<td>84.4</td>
<td>1.1</td>
<td>8.9</td>
<td>.8</td>
<td>10.0</td>
</tr>
<tr>
<td>8</td>
<td>3.5</td>
<td>47.1</td>
<td>21.4</td>
<td>0.5</td>
<td>23.4</td>
<td>4.0</td>
<td>9.3</td>
<td>6.3</td>
<td>58.2</td>
</tr>
<tr>
<td>9</td>
<td>4.3</td>
<td>43.0</td>
<td>32.4</td>
<td>0.8</td>
<td>17.6</td>
<td>2.0</td>
<td>12.1</td>
<td>7.4</td>
<td>77.7</td>
</tr>
<tr>
<td>10</td>
<td>1.5</td>
<td>54.3</td>
<td>23.8</td>
<td>0.0</td>
<td>16.4</td>
<td>4.1</td>
<td>9.7</td>
<td>6.7</td>
<td>80.7</td>
</tr>
<tr>
<td>11</td>
<td>1.8</td>
<td>37.2</td>
<td>44.5</td>
<td>0.0</td>
<td>15.7</td>
<td>7.7</td>
<td>9.5</td>
<td>18.6</td>
<td>80.7</td>
</tr>
<tr>
<td>12</td>
<td>1.0</td>
<td>6.4</td>
<td>4.4</td>
<td>0.0</td>
<td>81.5</td>
<td>2.2</td>
<td>12.5</td>
<td>0.9</td>
<td>13.5</td>
</tr>
<tr>
<td>13</td>
<td>0.3</td>
<td>62.1</td>
<td>29.7</td>
<td>0.0</td>
<td>4.3</td>
<td>3.6</td>
<td>10.7</td>
<td>13.3</td>
<td>85.7</td>
</tr>
<tr>
<td>14</td>
<td>1.7</td>
<td>14.4</td>
<td>2.9</td>
<td>0.0</td>
<td>79.3</td>
<td>1.7</td>
<td>9.8</td>
<td>0.6</td>
<td>8.6</td>
</tr>
<tr>
<td>15</td>
<td>7.3</td>
<td>11.2</td>
<td>3.6</td>
<td>0.0</td>
<td>75.9</td>
<td>2.0</td>
<td>11.9</td>
<td>2.6</td>
<td>17.2</td>
</tr>
<tr>
<td>16</td>
<td>5.5</td>
<td>16.8</td>
<td>12.1</td>
<td>0.4</td>
<td>61.3</td>
<td>3.9</td>
<td>16.4</td>
<td>1.2</td>
<td>40.2</td>
</tr>
<tr>
<td>17</td>
<td>4.5</td>
<td>23.7</td>
<td>12.4</td>
<td>0.3</td>
<td>57.3</td>
<td>1.7</td>
<td>9.3</td>
<td>5.4</td>
<td>38.1</td>
</tr>
<tr>
<td>18</td>
<td>2.4</td>
<td>29.7</td>
<td>20.7</td>
<td>0.0</td>
<td>46.3</td>
<td>.8</td>
<td>11.8</td>
<td>6.1</td>
<td>41.9</td>
</tr>
<tr>
<td>19</td>
<td>11.0</td>
<td>12.6</td>
<td>8.4</td>
<td>0.0</td>
<td>65.7</td>
<td>2.3</td>
<td>14.6</td>
<td>0.3</td>
<td>15.5</td>
</tr>
<tr>
<td>20</td>
<td>2.3</td>
<td>44.0</td>
<td>43.2</td>
<td>0.8</td>
<td>8.2</td>
<td>1.6</td>
<td>12.8</td>
<td>14.0</td>
<td>82.1</td>
</tr>
<tr>
<td>21</td>
<td>2.3</td>
<td>59.9</td>
<td>11.1</td>
<td>0.4</td>
<td>22.9</td>
<td>3.4</td>
<td>11.1</td>
<td>1.9</td>
<td>61.5</td>
</tr>
</tbody>
</table>

1Hispanic
2American Indian
3Multiracial
4Students with Disabilities
5English Language Learners
6Economically Disadvantaged

Procedures

The research design of this study relied on accurate acquisition of CRCT and ITBS data. The first step in the process was to secure permission from the school district. The district requires all studies conducted using district data of any kind to seek and gain permission to proceed with such a study (see Appendix A for application to conduct the study). Second, the Institutional Review Board process was completed through the
This process enabled the collection of data to commence. After the approval of the school district and Institutional Review Board permission was gained, data were collected through communication with local school principals.

Principals who agreed to participate signed a consent form (see Appendix C). Each principal was asked to provide eighth-grade CRCT and ITBS data for reading and math for the years 2003–2010. However, 19 of the principals were unable to obtain the data from their school records and gave permission for the researcher to find the data for their respective schools in the school district’s archives. District administrators responsible for the archives were very responsive and provided data that several principals could not locate in the local schools’ testing files.

Data Analysis

The scale scores of the CRCT tests for reading from 2003–2005 ranged from 150 to 500. This same scale was used for the mathematics CRCT from 2003–2006. The CRCTs during these periods were administered to measure success on the QCC. However, with implementation of the GPS, the scale score was changed. Beginning in 2006 for reading and in 2007 in mathematics, the scale scores now range from 650 to 900. In order to compare the achievement of the eighth graders across time (2003–2010), the scores were transformed into $z$ scores. A $z$ score has a mean of 0 and a standard deviation of 1. To transform the CRCT scores, the means and standard deviations of the state eighth graders during each of the years of the study were obtained for mathematics and reading. For example, in 2003, the state eighth-grade mean in reading was 342 and the standard deviation was 50. A school’s CRCT reading average of 322 was transformed into a $z$ score of -.40, indicating that the school performed .40 of a standard deviation below the state average.
deviation below the state mean. Another school’s CRCT reading average of 379 was transformed into a \( z \) score of .74, indicating that the school performed .74 of a standard deviation above the state mean.

This same procedure was used to transform the scale scores of the tests when the GPS curriculum was measured. For example, in 2008, the state eighth-grade mean in reading was 829 and the standard deviation was 23. A school’s CRCT reading average of 818 was transformed into a \( z \) score of -.48, indicating that the school performed .48 of a standard deviation below the state mean. Another school’s CRCT reading average of 851 was transformed into a \( z \) score of .96, indicating that the school performed .96 of a standard deviation above the state mean. In this way, a comparison could be made between scores obtained when the QCC was measured and scores obtained when the GPS was measured.

The data were analyzed using SPSS. Four sets of data were entered for each of the 21 schools: (a) 2003–2010 CRCT mathematics results, (b) 2003–2010 CRCT reading results, (c) 2003–2010 ITBS mathematics results, and (d) 2003–2010 ITBS 4 schools using. The ITBS scale scores were used as covariates. The ITBS was measured using the same scale for all years of the study. Two doubly multivariate, repeated measures ANCOVAs were used to examine eighth-grade reading and mathematics scores. The tests within the type of curriculum were measured, not the effect of time. In addition, the average \( z \) scores were graphed for reading and mathematics. The results of the analyses are presented in Chapter IV.

Limitations

There were limitations to this study. The schools ranged from Title 1 (low income) to high-income schools. However, these limitations could also be considered
positive attributes in this study since a vast range of schools and performances may give this study a wider range of generalized information. The CRCT is designed to measure student mastery of the state’s curriculum (Georgia Department of Education, 2010b). The Georgia Department of Education’s 2010 validity and reliability report detailed the process of the CRCT development that leads to the report’s conclusion that the test is in fact valid and reliable (Georgia Department of Education, 2010b).

Ethical Considerations

Student achievement data were collected as school averages and did not include individual test scores for any student. Therefore, no protection of individual participants was necessary. There were no anticipated risks for the principals who agreed to participate in the study. The middle school principals were not required to take part in the study and time required to participate for those who did agree was limited to the retrieval and sharing of the data. The identification of each school was confidential and no individual information about each school or any individual student was shared in this study.

Summary

The purpose of this study was to examine the difference in student achievement when students are taught and then tested under different curriculums. The data used in this study was archival CRCT and ITBS data from 21 middle schools located in the southeastern region of the United States. Math and reading scores were analyzed using two doubly multivariate, repeated measures ANCOVAs.
CHAPTER IV

RESULTS

The state of Georgia administers the Criterion-Referenced Competency Test (CRCT) to every elementary student in Grades 1 though 8. During 2005, 2006, and 2007, the state tested eighth-grade students on a quality core curriculum (QCC). In 2008, the state tested the students on a Georgia Performance Standard (GPS) curriculum. This study was designed to examine if differences in students achievement occurred because of the curriculum change. This study was guided by the following research question:

Is there a significant difference in student achievement between students who are taught and then tested based on a Quality Core Curriculum and those who are taught and tested using the Georgia Performance Standards?

Description of the Data

The scale scores of the CRCT tests for reading from 2003–2005 ranged from 150 to 500. This same scale was used for the mathematics CRCT from 2003–2006. The CRCTs during these periods were administered to measure success on the QCC. However, with implementation of the GPS, the scale score was changed. Beginning in 2006 for reading and in 2007 in mathematics, the scale scores now range from 650 to 900. In order to compare the achievement of the eighth graders across the two curriculums (2003–2010), the scores were transformed into $z$ scores. A $z$ score has a mean of 0 and a standard deviation of 1. ITBS achievement was measured the same from 2003 to 2010. No conversion of these scores was required to use them as covariates in the repeated measures ANCOVAs. Table 2 contains a description of the variables used to analyze the research question.
Table 2

*Statistical Description of the Variables of Interest (n = 21)*

<table>
<thead>
<tr>
<th></th>
<th>Minimum (Min)</th>
<th>Maximum (Max)</th>
<th>Mean (M)</th>
<th>Standard Deviation (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reading</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ZCRCT 2003</td>
<td>-.4</td>
<td>.8</td>
<td>.20</td>
<td>.41</td>
</tr>
<tr>
<td>ZCRCT 2004</td>
<td>-.3</td>
<td>.9</td>
<td>.24</td>
<td>.38</td>
</tr>
<tr>
<td>ZCRCT 2005</td>
<td>-.4</td>
<td>.8</td>
<td>.21</td>
<td>.36</td>
</tr>
<tr>
<td>ZCRCT 2006</td>
<td>-.4</td>
<td>.9</td>
<td>.28</td>
<td>.41</td>
</tr>
<tr>
<td>ZCRCT 2007</td>
<td>-.5</td>
<td>1.0</td>
<td>.22</td>
<td>.43</td>
</tr>
<tr>
<td>ZCRCT 2008</td>
<td>-.5</td>
<td>1.0</td>
<td>.25</td>
<td>.41</td>
</tr>
<tr>
<td>ZCRCT 2009</td>
<td>-.4</td>
<td>.9</td>
<td>.22</td>
<td>.42</td>
</tr>
<tr>
<td>ZCRCT 2010</td>
<td>-.8</td>
<td>1.0</td>
<td>.24</td>
<td>.49</td>
</tr>
<tr>
<td>ITBS 2003</td>
<td>230</td>
<td>276</td>
<td>252.9</td>
<td>14.6</td>
</tr>
<tr>
<td>ITBS 2004</td>
<td>226</td>
<td>274</td>
<td>250.1</td>
<td>15.3</td>
</tr>
<tr>
<td>ITBS 2005</td>
<td>221</td>
<td>271</td>
<td>246.5</td>
<td>14.5</td>
</tr>
<tr>
<td>ITBS 2006</td>
<td>220</td>
<td>272</td>
<td>245.9</td>
<td>15.1</td>
</tr>
<tr>
<td>ITBS 2007</td>
<td>217</td>
<td>272</td>
<td>246.6</td>
<td>15.3</td>
</tr>
<tr>
<td>ITBS 2008</td>
<td>223</td>
<td>273</td>
<td>246.4</td>
<td>15.5</td>
</tr>
<tr>
<td>ITBS 2009</td>
<td>216</td>
<td>261</td>
<td>235.8</td>
<td>14.5</td>
</tr>
<tr>
<td>ITBS 2010</td>
<td>215</td>
<td>260</td>
<td>236.3</td>
<td>14.8</td>
</tr>
</tbody>
</table>

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Math</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ZCRCT 2003</td>
<td>-.5</td>
<td>1.1</td>
<td>.24</td>
<td>.48</td>
</tr>
<tr>
<td>ZCRCT 2004</td>
<td>-.4</td>
<td>1.5</td>
<td>.26</td>
<td>.50</td>
</tr>
<tr>
<td>ZCRCT 2005</td>
<td>-.5</td>
<td>1.7</td>
<td>.25</td>
<td>.54</td>
</tr>
<tr>
<td>ZCRCT 2006</td>
<td>-.5</td>
<td>1.2</td>
<td>.26</td>
<td>.49</td>
</tr>
<tr>
<td>ZCRCT 2007</td>
<td>-.5</td>
<td>1.2</td>
<td>.25</td>
<td>.52</td>
</tr>
<tr>
<td>ZCRCT 2008</td>
<td>-.6</td>
<td>1.5</td>
<td>.29</td>
<td>.57</td>
</tr>
<tr>
<td>ZCRCT 2009</td>
<td>-.5</td>
<td>1.4</td>
<td>.26</td>
<td>.55</td>
</tr>
<tr>
<td>ZCRCT 2010</td>
<td>-.5</td>
<td>1.3</td>
<td>.31</td>
<td>.55</td>
</tr>
<tr>
<td>ITBS 2003</td>
<td>233</td>
<td>284</td>
<td>252.7</td>
<td>16.6</td>
</tr>
<tr>
<td>ITBS 2004</td>
<td>206</td>
<td>283</td>
<td>250.9</td>
<td>18.9</td>
</tr>
<tr>
<td>ITBS 2005</td>
<td>227</td>
<td>285</td>
<td>251.8</td>
<td>16.1</td>
</tr>
<tr>
<td>ITBS 2006</td>
<td>227</td>
<td>285</td>
<td>251.0</td>
<td>16.4</td>
</tr>
<tr>
<td>ITBS 2007</td>
<td>220</td>
<td>281</td>
<td>249.4</td>
<td>17.1</td>
</tr>
<tr>
<td>ITBS 2008</td>
<td>223</td>
<td>284</td>
<td>248.5</td>
<td>17.6</td>
</tr>
<tr>
<td>ITBS 2009</td>
<td>218</td>
<td>267</td>
<td>238.0</td>
<td>15.4</td>
</tr>
<tr>
<td>ITBS 2010</td>
<td>218</td>
<td>266</td>
<td>239.4</td>
<td>14.6</td>
</tr>
</tbody>
</table>
Analysis of Eighth-Grade CRCT Reading Scores

A doubly multivariate, repeated measures ANCOVA was conducted to determine if the eighth-grade reading scores were significantly different between the 3 years from 2003 to 2005 that the CRCT measured the QCC and the following 3 years from 2006 to 2008 that the CRCT measured the GPS. The tests within the type of curriculum were measured, not the effect of time. The ITBS reading scale scores from 2003 to 2008 were used as covariates. Figure 1 illustrates the eighth-grade CRCT reading scores from 2003 to 2008. The line down the middle of the graph sets the demarcation between the two curriculums.

![Graph of average z scores by year](image)

Figure 1. Converted Eighth-Grade CRCT Reading Scores by Year.

The multivariate effect of test, comparing the first 3-year period (measuring the QCC between 2003 and 2005) to the second 3-year period (measuring the GPS curriculum between 2006 and 2008) was not significant \(F(1,14) = .81, p = .38\).
However, the lack of statistical significance could be attributed to the small sample size. A post hoc pairwise comparison between CRCT reading scores measuring the QCC (from 2003 to 2005) and CRCT reading scores measuring the GPS (from 2006 to 2008) was not statistically significant ($p = .07$), indicating that no difference between the two reading curriculums was found using this small sample of schools.

The GPS reading curriculum was implemented in fall 2005 and tested in spring 2006 (see Figure 1). An increase in scores occurred from the spring 2005 CRCT test of the QCC to the first year implementation of the GPS curriculum and subsequent spring 2006 CRCT test results. However, a drop in scores occurred the next year (spring 2007), after two years of testing the GPS curriculum. At this time, the eighth-grade reading scores were at the same level of achievement before implementation of the new GPS curriculum. By the end of the three-year period after the GPS curriculum was implemented (spring 2008), eighth-grade reading achievement was higher than it was at the end of the three-year period used to measure the QCC curriculum (spring 2005). Although statistically significant differences were not found between the two curriculums, the increase in scores between 2005 and 2006 and between 2005 and 2008 provided partial support to the hypothesis of a difference in reading achievement between eighth-grade students who were taught and then tested under different curriculums.

Analysis of Eighth-Grade CRCT Mathematics Scores

A doubly multivariate, repeated measures ANCOVA was conducted to determine if the eighth-grade mathematics scores were significantly different between the three years from 2005 to 2007 that the CRCT measured the QCC and the following three years from 2008 to 2010 that the CRCT measured the GPS. The tests within the type of curriculum were measured, not the effect of time. The ITBS mathematics scale scores
from 2005 to 2010 were used as covariates. Figure 2 illustrates the eighth-grade CRCT mathematics scores from 2005 to 2010. The line down the middle of the graph sets the demarcation between the two curriculums.

![Graph showing average z scores from 2005 to 2010](image)

**Figure 2.** Converted Eighth-Grade CRCT Mathematics Scores by Year.

The multivariate effect of test, comparing the first three-year period (measuring the QCC between 2005 and 2007) to the second three-year period (measuring the GPS curriculum between 2008 and 2010) was not significant \(F(1,14) = .46, p = .51\). However, the lack of statistical significance could be attributed to the small sample size. A post hoc pairwise comparison between CRCT mathematics scores measuring the QCC (from 2005 to 2007) and CRCT mathematics scores measuring the GPS (from 2008 to 2010) were not statistically significant \(p = .20\), indicating that no difference between the two mathematics curriculums was found using this small sample of schools.
The GPS mathematics curriculum was implemented in fall 2007 and tested in spring 2008 (see Figure 2). An increase in scores occurred from the spring 2007 CRCT test of the QCC to the first year implementation of the GPS curriculum and subsequent spring 2008 CRCT test results. However, a drop in scores occurred the next year (spring 2009), after two years of testing the GPS curriculum. By the end of the three-year period after the GPS curriculum was implemented (spring 2010), eighth-grade mathematics achievement was higher than it was at the end of the three-year period used to measure the QCC curriculum (spring 2007). Although statistically significant differences were not found between the two curriculums, the increase in scores between 2007 and 2008 and between 2007 and 2010 provided partial support to the hypothesis of a difference in mathematics achievement between eighth-grade students who were taught and then tested under different curriculums.

Summary

Tests scores from 21 middle schools from 2003 to 2010 were analyzed to determine if there was a significant difference in student achievement between students who were taught and then tested on a Quality Core Curriculum and those who were taught and tested using the Georgia Performance Standards. Two doubly multivariate, repeated measures ANCOVAs found no statistically significant differences between the two curriculums. A discussion of these findings is presented in Chapter V.
CHAPTER V

SUMMARY OF FINDINGS, CONCLUSIONS, AND RECOMMENDATIONS

The ultimate goal of educators is to help students become educated. To do this, teachers and all educators must help them learn to improve the quality of their work and to see their own capabilities (Reeves, 2007). However, educators must also assess how they test to be reflective and visionary in their efforts to improve student achievement. The primary purpose of this study was to examine the history of assessment and to determine if differences in student achievement exist between assessments aligned with a performance-based curriculum and assessments aligned with a quality core curriculum. Archival data were obtained and analyzed to determine if assessment changes have affected student achievement. CRCT and ITBS data from 21 middle schools located in the southeastern region of the United States were obtained and used for the study. This chapter contains a summary of the study, a discussion of how the findings related to the literature, a final reflective conclusion, and recommendations for future research.

The state of Georgia administers the Criterion-Referenced Competency Test (CRCT) to every elementary student in Grades 1 through 8. During 2005, 2006, and 2007, the state tested eighth-grade students on a quality core curriculum (QCC). In 2008, the state tested the students on a Georgia Performance Standard (GPS) curriculum. Not only did the state change the curriculum in several subject areas, the implementation from the QCC curriculum to the GPS curriculum was staggered. The CRCT continued to be used as the accountability piece to meet the requirements of No Child Left Behind. However, the QCC curriculum tested was measured on a scale score of 150–500 while the rollout of the GPS curriculum tested was measured on a scale score of 650–900.
Consequently, this study was designed to examine if differences in student achievement occurred because of the curriculum change.

The NCLB Act required the implementation of performance standards that would result in better instructional delivery and balanced assessment (DuFour, 2004). Accountability is important and probably not going away any time soon. Ravitch (1996) reported that standards guided teachers to select better assessment practices to not only measure student success but also students’ level of achievement. Mederano (2003) reported that the GPS curriculum was a standards-based curriculum and provided a very clear set of guidelines for improved instructional delivery.

To guide educators in assessing students’ level of achievement, a framework called *School Keys* was provided as the foundation for Georgia’s comprehensive, data-driven system of school improvement (Georgia Department of Education, 2007b). Within the *School Keys* was an assessment key. This comprehensive key required teachers to include both summative and formative assessments in the day-to-day instructional plan. The assessment key contains an (a) overall definition of assessment, (b) three standards that contained at least one substandard, and (c) a rubric to rate each substandard and elements or operation descriptors for the standard.

Findings

This study asked if differences in student achievement exist between assessments aligned with a performance-based curriculum and assessments that aligned with a quality core curriculum. For this question to be answered, the average CRCT scores in reading and mathematics for eighth graders in 21 middle schools were obtained. The QCC standards were used to measure achievement for mathematics from 2005–2007 and the GPS mathematics curriculum was assessed by the CRCT from years 2008–2010. The
QCC standards were used to measure achievement for reading from 2003–2005 and the GPS reading curriculum was assessed by the CRCT from years 2006–2008. The scales used to measure success on the two tests were not similar; therefore, a z transformation was obtained for each score. Two doubly multivariate, repeated measures ANCOVAs were used to examine if differences in achievement occurred because of the changes in the curriculums.

*Analysis of Reading Achievement*

The multivariate effect of test, comparing the first three-year period (measuring the QCC between 2003 and 2005) to the second three-year period (measuring the GPS curriculum between 2006 and 2008), was not significant. The lack of statistical significance could be attributed to the small sample size. The reading scores fluctuated after the implementation of the GPS curriculum, going up the first year, but dropping in the second year of the new curriculum. However, by the end of the three-year period after the GPS curriculum was implemented, eighth-grade reading achievement was higher than it was at the end of the three-year period used to measure the QCC curriculum.

*Analysis of Mathematics Achievement*

The multivariate effect of test, comparing the first 3-year period (measuring the QCC between 2005 and 2007) to the second three-year period (measuring the GPS curriculum between 2008 and 2010) was not statistically significant. Again, the lack of significance could be attributed to the small sample size. The mathematics scores fluctuated after the implementation of the GPS curriculum, going up the first year, but dropping in the second year of the new curriculum. However, by the end of the three-year period after the GPS curriculum was implemented, eighth-grade mathematics
achievement was higher than it was at the end of the three-year period used to measure the QCC curriculum.

Summary of the Findings

It is possible that teachers were unsure how to implement a standards-based curriculum. It is also possible that teachers were not trained in formative assessment strategies; how to write them, administer them and effectively use them to make instructional decisions based on the results of the assessment tool used. Furthermore, it is also possible that each middle school implemented or even scheduled reading and mathematics classes differently from school to school. In other words, how reading and mathematics were delivered could have affected the short history of these seesawing grades.

Conclusion

Educators can conclude that implementation of the School Keys stressing formative and summative assessments had a positive impact on student achievement. These changes could be attributed to the Schools Keys being implemented in tandem with the GPS curriculum. The School Keys provide teachers with a direction on not only the importance of both formative and summative assessments, but how to use them in a balanced manner to help guide teachers to make better instructional decisions. Because the state is seeing gains in scores, the use of data to adjust instruction and design lessons that are more appropriate will increase student achievement. Educators are providing not only summative grades to students, but formative feedback that helps students self-monitor their own learning. These keys seem to be most effective when used not only as summative assessments such as the CRCT and ITBS, but when used as a guide to better instructional practices.
Recommendations for Practice

The information from this study is intended to assist teachers, administrators, districts, and states gain a better understanding of testing and history. Continual examination of a single test used to promote students from one grade level to the next is a necessity. However, if educators do not know the basic history of testing or how it has evolved, no reflection can occur when making future decisions. Even Binet stressed the limitations of intelligence tests by stating that they are too broad and cannot be used to determine someone’s intelligence with a single number (Siegler, 1992). In the implementation of this study and in the analysis of the data, a number of issues arose that must be addressed. The following are recommendations to resolve these issues.

Recommendation for a Data Retrieval Bank

Districts should institute some type of data retrieval bank. Only two principals in this study had immediate access to the data required for this study. Nineteen of the school principals did not have immediate access to their data. Although some of the principals were new to the building and some had only been at their building a few years, their ability to obtain test summary sheets for eight years of data should have been easily retrievable. These data are used to determine AYP status. The state’s website does have some of the data, but going back more than three years, the data is reported differently, difficult to find, and does not contain necessary means to measure or analyze data on a statistical level.

Therefore, I recommend that data going back at least 10 years be housed in a data system and at the minimum include means and standard deviations of the scores, averages reported in percentages, and number of students tested. Kudos to the study district for having such data archived at a warehouse and allowing access to the
researcher. However, if we really are making instructional decisions based on data, we need to look at each school using an historical timeline, and that data must be easily accessible. This will enable leaders to mark trends and make better future instructional decisions.

*Professional Development in the Use of Data*

Educators should be engaged in tiered professional development in regards to the storage and use of data. Tier 1 would be teachers. Classroom teachers look at last year’s data and see that a domain, for example, eighth-grade math, statistics, and probability is weak, but now they are teaching a new group of students. How do they ascertain if this current group has the same weakness? This is where the idea of formative assessment presented in Chapter II of this study would benefit them greatly.

Tier 2 is for principals and district administration. While gathering the data for this study, many principals and district leaders did not know the difference between a mean score, a percentile, and a percentage. Most principals and district leaders are outstanding educators, but do they have the skills to navigate high-stakes testing results to make not just adjustments to instruction, but make changes that would actually change their role from manager of a building to a true instructional leader in the building.

The third, and perhaps the most important, tier in this recommendation are the university systems. Systematically, universities could delete or augment an instructional strategy class and add a data leadership class to the current list of courses needed to earn a leadership degree. This is by no means a criticism of our current schools or university. I do believe that true educators are constantly looking for ways to encourage and support each other to do better. Perhaps this one attribute is needed to round out and even change how we look at current testing and educator preparation.
Test Development Issues

Assuming that testing and accountability are not going away, a final and perhaps most important recommendation is to those who write and administer these tests. They should consider that when a curriculum change occurs, students should be administered tests that phase in the new curriculum. In addition, changing the score scales on the criterion-referenced competency tests on such as the CRCT when the curriculum changes make it very difficult to study data to determine progress over time. In the statistical world, this creates a confounding variable that may be hard to control.

Recommendations for Future Research

A limitation of the current study was the small number of middle schools from which the scores were collected. The study was conducted in one school district and analyzed data from only one grade. Thus, the sample size was small and no statistically significant differences were found in the data. In addition, the number of years used in the analysis was limited to three years before and three years after the implementation of the GPS. Large fluctuations in the data were found, furthering limiting the ability to find any significant differences between the two curriculums. If indeed the teachers were unsure how to implement a standards-based curriculum, they had not been trained adequately in how to use formative assessment strategies, and the scheduling of the classes may have been nonstandardized, three years may not be enough time to measure the impact of the new curriculum on student achievement.

Therefore, a recommendation for future research is for a study across a larger number of schools (and therefore, more school districts in the state), more grades, and more years before and after the implementation of the GPS. By expanding the scope of the study, its findings can be more easily generalized across the state. A more
comprehensive analysis of the impact of the curriculum could then be available for policymakers to consider when making future decisions about assessment of Georgia’s students.

A final recommendation would be to consider President Obama’s recent offer for states to waive out of the federal accountability mandates of NCLB. With Georgia waving out of these federal regulations prescribed by NCLB, what will the impact be upon student achievement and how we assess/test students? Will Georgia do a better job of managing performance of its schools and learners than the federal government did? The waiver is not an opt out of accountability, but rather requires states who do waiver to prescribe a set of standards that prepare learners for college and monitor these standards. By studying the results of the waiver, this study could become more encompassing and global in its application.

Concluding Remarks

It is necessary for all educators and policymakers to think about the purpose of testing. All need to question the reason for every test administered. Porter (1995) reported that a fundamental misunderstanding of school reform and school issues exists. Porter helped educators think about the importance of testing and to question the intent of testing. In this study, an examination of the history of assessment and a study of current assessment practices should help the reader to question the intent and uses of tests and to question the results of each test.

Educators must consider making sure tests not only test what is taught, but also that one test should not be the sole determinate of student achievement. Educators must assist teachers with multiple ways to measure and monitor students’ achievement as the School Keys have done. As educators, we should be bold and challenge history to make
changes that best serve students. Curriculums should be viable; these curriculums should be clear in defining what students should learn and how schools will plan if students do not meet expectations. By studying the curriculums and the tests given to measure the achievement of these curriculums, this can be accomplished. With so much dependent on students’ abilities to pass these tests, whether it is to be promoted to the next grade level, prove proficiency on a high school curriculum, or to be accepted into a university, educators must continually revisit testing practices.
APPENDIX A

SCHOOL DISTRICT APPROVAL TO CONDUCT RESEARCH

June 23, 2011

Dear Ms. Brink:

Your research project has been approved. Listed below are the schools where approval to conduct
the research is complete. Please work with the school administrator to schedule administration of
instruments or conduct interviews.

Middle School
Middle School
Middle School
Middle School
Middle School
Middle School
Middle School
Middle School
Middle School
Middle School
Middle School
Middle School
Middle School
Middle School
Middle School
Middle School
Middle School
Middle School
Middle School
Middle School

Should modifications or changes in research procedures become necessary during the research
project, changes must be submitted in writing to the Office of Accountability and Research prior to
implementation. At the conclusion of your research project, you are expected to submit a copy of your
results to this office. Results cannot reference the School District or any District schools
or departments.

BOARD OF EDUCATION

SUPERINTENDENT
Research files are not considered complete until results are received. If you have any questions regarding the process, contact our office at 770-426-3407.

Sincerely,

[Signature]

Dr. Judith A. Jones
Chief Accountability and Research Officer
APPENDIX B
UNIVERSITY IRB APPROVAL TO CONDUCT RESEARCH

THE UNIVERSITY OF SOUTHERN MISSISSIPPI
Institutional Review Board
118 College Drive #5147
Hattiesburg, MS 35406-0001
Tel: 601.266.6820
Fax: 601.266.3509
www.usm.edu/irb

HUMAN SUBJECTS PROTECTION REVIEW COMMITTEE
NOTICE OF COMMITTEE ACTION

The project has been reviewed by The University of Southern Mississippi Human Subjects Protection Review Committee in accordance with Federal Drug Administration regulations (21 CFR 26, 111), Department of Health and Human Services (45 CFR Part 46), and university guidelines to ensure adherence to the following criteria:

- The risks to subjects are minimized.
- The risks to subjects are reasonable in relation to the anticipated benefits.
- The selection of subjects is equitable.
- Informed consent is adequate and appropriately documented.
- Where appropriate, the research plan makes adequate provisions for monitoring the data collected to ensure the safety of the subjects.
- Where appropriate, there are adequate provisions to protect the privacy of subjects and to maintain the confidentiality of all data.
- Appropriate additional safeguards have been included to protect vulnerable subjects.
- Any unanticipated, serious, or continuing problems encountered regarding risks to subjects must be reported immediately, but not later than 10 days following the event. This should be reported to the IRB Office via the "Adverse Effect Report Form".
- If approved, the maximum period of approval is limited to twelve months. Projects that exceed this period must submit an application for renewal or continuation.

PROTOCOL NUMBER: 11071403
PROJECT TITLE: A Historical Perspective of Testing and Assessment Including the Impact of Summative and Formative Assessment on Student Achievement
PROPOSED PROJECT DATES: 07/25/2011 to 10/12/2011
PROJECT TYPE: Dissertation
PRINCIPAL INVESTIGATORS: Carole Sanger Brink
COLLEGE/DIVISION: College of Education & Psychology
DEPARTMENT: Educational Leadership & School Counseling
FUNDING AGENCY: N/A
HSRPC COMMITTEE ACTION: Exempt Approval
PERIOD OF APPROVAL: 07/21/2011 to 07/20/2012

Lawrence A. Hosman, Ph.D.  7-21-2011
HSRPC Chair  Date
APPENDIX C

PARTICIPATING PRINCIPAL CONSENT FORM

District Research Procedures

Participating Principal Consent Form

My signature below indicates that I have read the information provided and have decided to allow my school to participate in the study titled: A Historical Perspective on the History of Testing and Assessment and The Impact of the Practice of Summative and Formative Assessment on Student Achievement to be conducted independently dates of April 2012 and December 2011. I understand that my signature indicates I have agreed to participate in this research project.

I understand the purpose of the research project will be to study CEC 8th grade reading and math results over a period of time and that I will participate in the following manner:


Potential benefits of the study are: Since this study will obtain and analyze archival test data and seek to understand if assessment changes have impacted student achievement, the results may yield a better understanding of the current assessment practices in our district. It is always beneficial to observe the results of practices such as formative and summative assessment practices and perhaps take those results and improve assessment practices in future use.

I agree to the following conditions with the understanding that I can withdraw my school from the study at any time should I choose to discontinue participation.

- The identity of participants will be protected as only 8th grade total scale scores will be used and school will be identified by a marker such as Middle School A, Middle School B, and so forth.
- Information gathered during the course of the project will become part of the data analysis and may contribute to published research reports and presentations.
- There are no foreseeable inconveniences or risks involved to any schools participating in the study.
- Participation in the study is voluntary. If I decide to withdraw permission after the study begins, I will notify the school of my decision.

If further information is needed regarding the research study, I can contact Carole Brink (cell or (name)).

Signature

Principal

Date

Revised May 2009
REFERENCES


Cox, K. (2006). What Georgia educators need to know about Georgia’s testing program. Atlanta, GA: Georgia Department of Education.


e57e7942b6d677375da861&type=d


Georgia Department of Education. (2010b). *What Georgia educators need to know about the Georgia testing program*. Retrieved from www.doe.k12.ga.us/testing


Cambridge, MA.


