A CAUSAL-COMPARATIVE STUDY OF THE AFFECTS OF BENCHMARK ASSESSMENTS ON MIDDLE GRADES SCIENCE ACHIEVEMENT SCORES

by

Melissa Ritchie Galloway

Liberty University

A Dissertation Presented in Partial Fulfillment

Of the Requirements for the Degree

Doctorate of Education

Liberty University

2016

A CAUSAL-COMPARATIVE STUDY OF THE AFFECTS OF BENCHMARK ASSESSMENTS ON MIDDLE GRADES SCIENCE ACHIEVEMENT SCORES

by Melissa Ritchie Galloway

A Dissertation Presented in Partial Fulfillment

Of the Requirements for the Degree

Doctor of Education

Liberty University, Lynchburg, VA

2016

APPROVED BY

David Nelson, Ph.D., Committee Chair

Michael Schlabra, Ed.D., Committee Member

Diane Vautrot, Ph. D., Committee Member

Scott Watson, Ph.D., Associate Dean, Graduate Programs

ABSTRACT

The purpose of this causal comparative study was to test the theory of assessment that relates benchmark assessments to the Georgia middle grades science Criterion Referenced Competency Test (CRCT) percentages, controlling for schools who do not administer benchmark assessments versus schools who do administer benchmark assessments for all middle school students including those enrolled in Special Education (SPED) and English to Speakers of other Languages (ESOL) programs across the state of Georgia. CRCT pass percentages were collected from fifteen schools that administered benchmark assessments and fifteen schools that did not administer benchmark assessments. The data was collected from The Governor's Office of Achievement website. A *t* test was used to determine if there was a statistically significant difference between eighth grade science CRCT pass percentages of schools who administer benchmark assessments compared to those schools who did not administer benchmark assessments. The *t* test resulted in no statistically significant difference for the whole group and the SPED group; however, there was a statistically significant difference for the ESOL group with the non-benchmark mean being higher than the benchmark mean. Future research was recommended that included determining if a correlation exists between the number of assessments administered each year and standardized test scores, the impact ACCESS (Assessing Comprehension and Communication in English State-to-State) scores have on standardized test scores, the impact of English proficiency on standardized test scores, and determining if teachers who use data gained from benchmarks have a positive impact on standardized test scores compared to those teachers who do not use the data obtained from benchmark assessments to drive their instruction.

Dedication

I would like to dedicate this research first and foremost to my Lord and Savior, Jesus Christ. It is through his constant strength and guidance that I was able to complete this process. I would also like to dedicate this research to my family; without them I surely would have failed and quit before I even got to the research. I would especially like to thank my very loving and patient husband, Stacy. When we met in 2002, I was one year into my bachelor's degree, and he has been with me every step of the way since. He is my rock, and without him I would not be where I am today. He has encouraged me not to quit and pushed me when I needed it. I am also very thankful that he has put up with me through this process. I know I have not been the easiest person to live with during this time, and he is an angel for putting up with me.

I would also like to thank my two children, Paris and Dylan. One thing that I most regret in this life is that I had to spend so much time away from them while they were growing up doing schoolwork. It broke my heart every time I had to tell them we could not do something because I had schoolwork that needed to be done. I am so proud of the young adults that they have become, and without their support and understanding I would have not been able to complete this process.

I would also like to dedicate this research to my parents. At a very young age my parents instilled in me the importance of education. They always pushed me to do my best in everything I attempted. I haven't always been the best daughter to them, but they have always been the best parents to me. Even when my world fell down around me and I felt I had no one to turn to, my parents were there picking the pieces back up putting my world together again.

I would also like to dedicate this research to my Granny Betty. My Granny left this world five short years ago, but she has left a lasting impression on my life. She is the one who

first introduced me to church. Some of my most precious memories are of going to church with her and singing in the choir when I was little. She led me to the Lord, and I will be eternally thankful for that.

Lastly, I would like to dedicate this research to my remaining family, friends, and co-workers. They have all been an integral part of the process, and I am very blessed to have each and every one of them in my life.

Acknowledgement

I would like to acknowledge Dr. David Nelson as my chair on the committee for this research. Dr. Nelson has been an integral part of the process. He has provided essential feedback that has made me a better student. He has been very understanding and cooperative when things were not going as planned. With his excellent leadership, I was able to push through to the end.

I would also like to acknowledge Dr. Scott Watson as my research consultant. Dr. Watson has also been an integral part of the process beginning with my first class, EDUC 919, in the dissertation process. He has guided me and led me to finding the best possible research for me. I appreciate everything he has done for me with this research.

I would also like to acknowledge my other two committee members, Dr. Michael Schlabra and Dr. Diane Vautrot. Dr. Schlabra has also provided insightful feedback and encouragement when needed. Dr. Vautrot has also been essential to me completing the process. She has become a mentor during the process, and without her guidance and encouragement, I would not have finished this process. I have spent more one on one time with her discussing and working on this research than anyone else. She has inspired me to do my very best. She always provided meaningful feedback and provided suggestions on how I could make my research better. I am truly grateful to have been able to work with her during this process.

List of Tables

List of Figures

List of Abbreviations

ACCESS – Assessing Comprehension and Communication in English State to State

AYP – Adequate Yearly Progress

BCPAF – Benchmarks Curricular Planning and Assessment Framework

CCGPS – Common Core Georgia Performance Standards

CCRPI – College and Career Readiness Performance Index

CCS – Common Core Standards

CRCT – Criterion Referenced Competency Test

ELL – English Language Learners

ESEA – Elementary and Secondary Schools Act

ESL – English as a Second Language

ESOL – English to Speakers of Other Languages

GPS – Georgia Performance Standards

IEP – Individualized Education Plan

IRB – Institutional Review Board

NCLB – No Child Left Behind

SIP – School Improvement Plan

SPED – Special Education

Table of Contents

CHAPTER ONE: INTRODUCTION

Background

In 2001, President Bush signed into legislation the No Child Left Behind Act (NCLB), a reauthorization of the Elementary and Secondary Schools Act (ESEA). NCLB was designed to stand "on four basic premises: stronger accountability for schools and teachers; increased flexibility and local control over federal funds; greater schooling options for parents; and a focus on proven, research-based teaching methods" (Rush & Scherff, 2012, p. 91). By far, the most pressing aspect of NCLB to teachers across the nation has been the increase in school accountability. Stronger accountability meant that states had to develop and implement yearly standardized tests in grades 3 through 8 in subjects of mathematics and reading. If schools failed to be proficient on these assessments, the schools were penalized for not meeting the requirements of adequate yearly progress, or AYP.

Not only have schools been punished for not meeting AYP, schools were also required to be 100% proficient in reading and mathematics by 2014. This fact has been a dark looming cloud in education since the authorization of NCLB. However, in 2011, President Obama offered states a break concerning the impending proficiency deadline. The President allowed states to submit a waiver that exempted them from certain demands issued in NCLB. In her report to *The American Prospect*, Abby Rapport (2012) stated:

> To get flexibility from NCLB, states must adopt and have a plan to implement college and career-ready standards. They must also create comprehensive systems of teacher and principal development, evaluation and support that include factors beyond test scores, such as principal observation, peer review, student work, or parent and student feedback. (para. 7)

Beginning the 2012-2013 school year, Georgia was no longer required to meet the demands of AYP. In 2012, Waltz stated:

> In exchange for the exemption, Georgia officials agreed to implement a new school rating system that will judge schools on a wide variety of factors, including test scores, attendance, and college-readiness. It will also reward the highest-performing schools and force low performing schools and schools with wide achievement gaps to accept interventions. (para. 2)

This school rating system was called the College and Career Readiness Performance Index (CCRPI). The CCRPI was designed to measure student gains on standardized tests in all core subjects. Instead of looking for a specific score range like NCLB, CCRPI looks at the gains students make from year to year.

At the beginning of the 2012-2013 school year, Georgia implemented a new set of standards called the Common Core Georgia Performance Standards (CCGPS) in the subjects of mathematics, English language arts, and reading. The CCGPS was a part of the national Common Core Standards (CCS) created by the Governors Association and the Council for Chief State School Officers. The purpose of the Governors Association and the Council for Chief State School Officers was to create a set of standards in language arts, reading, and mathematics that would prepare students for college and career. The CCS are not a national curriculum as each state does not 100% have identical standards. Each state that has joined the Common Core initiative adopted 85% of the CCS with the other 15% being created by each individual state.

One feature of the new CCS for Georgia is the implementation of literacy standards into science, social studies, and technical subjects. These literacy standards focus on building the student's knowledge through their demonstrations of reading, writing, and speaking in textual

context. The main idea is to teach students how to read and write using nonfiction, and to take what the student has learned and be able to speak about the subject in a knowledgeable and educated manner.

Problem Statement

Because of school accountability, districts across the nation have been scrambling to ensure success on end of the year standardized tests. One way educational leaders have tried to ensure success was through various forms of assessment. Wiliam (2010) stated, "Assessment is a key process in education. It is only through assessment that we can find out whether instruction has had its intended effect, because even the best designed instruction cannot be guaranteed to be effective" (p. 107). One type of assessment that was implemented across the nation utilizes benchmark assessments. Benchmark assessments are quarterly tests administered to students that mimic end of the year state standardized assessments. Benchmarks are categorized as summative assessments because benchmarks evaluate learning that has taken place over a specific set of standards. Even though benchmarks are summative assessments, they can be used formatively. Since each benchmark administered is geared toward a specific set of standards, data can be derived from the assessment. The data collected informs the teachers of the strengths and weaknesses of each student. Teachers can then take this information and build upon it in the classroom by providing small group or individualized remediation for weaknesses or enrichments for strengths.

Theoretically, benchmarks help teachers monitor the progress students are making through the curriculum. Benchmark assessments not only monitor student progress through the curriculum, benchmarks also allow teachers to use the data derived from the assessment to drive future instruction. Bulkley, Olah, and Blanc (2010) stated, "the emphasis on testing is spurred

by federal, state, and district accountability policies that have pressed educators to use data to monitor student progress toward well-defined learning goals" (p. 115). Because benchmark assessments have become a tool of choice by districts and administrators alike, the researcher wanted to know if the assessments were truly beneficial; however, the researcher believed that benchmarks could not be truly beneficial until teachers become 100% invested in the process. Teachers have to believe benchmarks make a difference, and teachers have to use the data gained from these assessments to drive future instruction in the classroom. If benchmarks are not used properly, the benchmarks become a burden and get viewed as another hoop that teachers must jump through each year.

Significance of the Study

This study was very significant in the world of education because the study allowed educators and administrators alike to see the impact benchmark assessments have had and could continue to have on end of the year state standardized tests. Although Georgia received an exemption for NCLB and the pressures that go along with NCLB have been somewhat alleviated, school accountability and student achievement are still important. Instead of facing the demands of AYP, teachers and students across the state of Georgia will be measured with the CCRPI. Under NCLB students were only measured on reading and mathematics standardized tests; however, under CCRPI students will be measured in all subjects: reading, English language arts, mathematics, science, and social studies.

This study is specifically designed around science benchmark assessments and the science Criterion Referenced Competency Test (CRCT). With the new evaluation system in place across Georgia, it is important to know if these quarterly assessments are helping students

make gains on the science CRCT, or if a teacher's time would be better spent providing more meaningful formative assessments and enriched learning experiences.

The researcher specifically wanted to find out if benchmark assessments were affecting the Criterion Referenced Competency Test (CRCT) percentage scores of Special Education (SPED) and English to Speakers of Other Languages (ESOL) students. SPED and ESOL students were the two target groups outlined in the School Improvement Plan (SIP) at the researcher's home school. The SIP focuses on science achievement across the school and SPED and ESOL students specifically. Having science achievement in the SIP made this study significant because it told the researcher if the time spent administering benchmark assessments actually made a difference on science CRCT scores.

Purpose Statement

The purpose of the study was to determine whether benchmark assessments affected the percentage scores on the CRCT in science. This study was a causal comparative study, and it looked at CRCT score percentages from thirty different middle schools to determine if schools who gave benchmarks and schools who did not give benchmarks had comparative percentages between general education students, SPED, and ESOL students. Research indicated a need for addressing the concerns that many school districts have about state testing and how student scores can be increased. Gilmer County Schools recently focused attention, time, and resources on benchmark assessments across the curriculum. Some argue benchmark testing is a waste of instructional time while others state that benchmark testing is increasing student improvement on the CRCT. Benchmark assessments are in place to ensure student achievement and knowledge by progress monitoring, so that the students who are not on target can be remediated before the CRCT at the end of the year.

Benchmark tests are created by the teachers in each grade level based on content area. Once a benchmark test is created, the teachers in that grade and content across the county approve or disapprove of the test. Once it is approved by all, it is entered into a program that was purchased by the county called ThinkGate. ThinkGate is a software program that combines curriculum, instruction, assessment, and data utilization. The program runs specific reports by class or student to determine individual strengths and weaknesses based on each Georgia Performance Standard. This data is to be used to drive instruction and group students based on skills that need to be strengthened.

Benchmark assessments provide teachers and administrators an outlet for which to focus the strains from federal and state accountability. Benchmarks can also be used as a determinacy factor for student performance on end of the year state standardized assessments. Benchmarking was designed to ensure no surprises in student achievement at the end of the year occur. If students are successful on benchmark assessments throughout the year, then students should be successful on the end of the year standardized assessment.

Research Questions

RQ1: How does the administration of benchmark assessments affect Georgia Criterion Referenced Competency Test (CRCT) science pass percentage scores among middle grade students?

RQ2: How does the administration of benchmark assessments affect Georgia Criterion Referenced Competency Test (CRCT) science pass percentage scores among middle grade students who are in the English to Speakers of Other Languages (ESOL) program?

RQ3: How does the administration of benchmark assessments affect Georgia Criterion Referenced Competency Test (CRCT) science pass percentage scores among middle grade students in the Special Education (SPED) program?

Hypotheses

$H_0$1: There is no significant difference in Georgia Criterion Referenced Competency Test (CRCT) science pass percentage scores of middle grade students who were administered benchmark assessments compared to Georgia Criterion Referenced Competency Test (CRCT) science pass percentage scores of middle grade students who were not administered benchmark assessments.

$H_0$2: There is no significant difference in Georgia Criterion Referenced Competency Test (CRCT) science pass percentage scores of middle grade students in the English to Speakers of Other Languages (ESOL) program who were administered benchmark assessments compared to Georgia Criterion Referenced Competency Test (CRCT) science pass percentage scores of middle grade students in the English to Speakers of Other Languages (ESOL) program who were not administered benchmark assessments.

$H_0$3: There is no significant difference in Georgia Criterion Referenced Competency Test (CRCT) science pass percentage scores of middle grade students in the Special Education (SPED) program who were administered benchmark assessments compared to Georgia Criterion Referenced Competency Test (CRCT) science pass percentage scores of middle grade students in the Special Education (SPED) program who were not administered benchmark assessments.

Identification of Variables

According to Gall, Gall, and Borg (2007):

> Causal-comparative research is a type of non-experimental investigation in which researchers seek to identify cause-and-effect relationships by forming groups of individual in whom the independent variable is present or absent –or present at certain levels –and then determine whether the groups differ on the dependent variable. (p. 306)

In this study, the independent variable was identified as the presence or absence of a science benchmark assessment. The dependent variable, in which the study determined if the varying groups displayed a difference, was the percentage scores for each school from the science CRCT from the 2011-2012 school year. The CRCT was a mandated assessment set forth by the state of Georgia to determine student growth toward state issued education standards. Since this was a causal comparative study the data was ex post facto; therefore, the independent variable had already occurred and could not be manipulated.

Definitions

1. *Accountability* - Wiliam (2010), defined accountability as "held to account" (p. 108). Wiliam's definition "suggests there is an expectation that when a person, organization, or entity is accountable, they can be expected or required to render an account of their actions or inactions" (p. 108).
2. *Adequate Yearly Progress (AYP)* - An accountability measurement tool set forth in Title I of the No Child Left Behind Act of 2001. It is defined as "a measure of year-to-year student achievement on statewide assessments" by the Georgia Department of Education (2013, para. 1).
3. *Assessment* - Cowie, Jones, and Otrel-Cass (2011), defined assessment as "one of the defining elements through which young people form and perform their identity for school purposes" (p. 354). Assessment is a demonstration of knowledge.

4. *Benchmark Assessments* - Benchmark assessments are defined by Olson (2005a), as a tool used "throughout the year to measure students' progress and provide teachers with data about how to adjust instruction" (p. 13).
5. *Common Core Georgia Performance Standards (CCGPS)* - A new set of standards adopted by Georgia in Reading, English Language Arts, and Mathematics to replace the Georgia Performance Standards. Literacy in Science, Social Studies, and Technical subjects is also a part of CCGPS. The CCGPS are a part of a national set of common core standards. The CCGPS is defined by the Georgia Department of Education (2013) as "a consistent framework to prepare students for success in college and/or the 21st century workplace".
6. *College and Career Readiness Performance Index (CCRPI)* - The Georgia Department of Education (2013) defined the CCRPI as "a comprehensive school improvement, accountability, and communication platform for all educational stakeholders that will promote college and career readiness for all Georgia public school students" (para. 1). The CCRPI is an index system based on points by which schools will be measured. Schools earn points based on various criteria including, but not limited to, standardized test pass percentages, attendance, professional development, and achievement gains in the subgroups of SPED and ESOL. It is set to take the place of AYP.
7. *Criterion Referenced Competency Test* (CRCT) - A form of assessment used by the Georgia Department of Education to determine the amount of skill and knowledge obtained by each student based on the Georgia Performance Standards (GPS). According to the Georgia Department of Education (2013), the CRCT is "designed to measure how well students acquire the skills and knowledge described in the state mandated content

standards in reading, English/language arts, mathematics, science, and social studies" (para. 2).

8. *Formative Assessment* – Formative assessment is an informal or formal assessment used to inform the teacher of students' progress in the learning objectives. Britton (2011) stated, "Formative assessment involves gathering data from students on their progress and comprehension so that instruction can be adjusted to needs their learning needs" (p. 17); while, Bulkley et al. (2010) indicated formative assessments are "built into classroom instructional activities and provide teachers and students with ongoing information about what students are learning" (p. 117).
9. *Summative Assessment* – Summative assessment is a formal assessment used to determine if students have met all learning objectives. According to Assessment and Reporting Unit, Learning Policies Branch, and Office of Learning and Teaching (2005), summative assessments "occur when teachers use evidence of student learning to make judgments on student achievement against goals and standards" (p. 9).

Research Summary

The purpose of the study was to determine whether benchmark assessments affect CRCT percentages in science. This study was a causal comparative study, and it looked at CRCT percentage scores from thirty different middle schools to determine if schools who gave benchmarks and schools who did not give benchmarks had comparative percentages among the whole group of students, SPED, and ESOL students. Research indicated a need for addressing the concerns that many school districts had about state testing and how student scores could be increased.

Causal-comparative designs are a form of ex post facto research. Gall et al. (2007) defined ex post facto research as "designs that rely on observation of relationships between naturally occurring variations in the presumed independent and dependent variables" (p. 306). The causal-comparative design was chosen for the study because it allowed the researcher to look at data that was already available and determine the cause and effect relationship between the data based on a manipulated variable.

The researcher investigated how the independent variable of benchmark assessments affected the dependent variable of CRCT percentage scores. The treatment group consisted of 15 schools in Georgia that administered benchmark assessments throughout the year to monitor the students' progression toward mastery of the GPS in science. The control group consisted of 15 schools that did not use benchmark assessments as a progress monitoring tool. Besides investigating middle school students as a whole group, the researcher also examined the subgroups of students enrolled in the ESOL and SPED programs at the control and treatment level.

CHAPTER TWO: REVIEW OF THE LITERATURE

Introduction

Making the learning experience beneficial for students is an important part of being a teacher. During the learning experience, students must participate in various forms and types of assessments to determine the progress students are making toward specific learning goals. The use of benchmarking provides one type of assessment that teachers use in the modern classroom. Generally, benchmark assessments are administered periodically throughout the year to determine if students are mastering state standards and to assess their progress toward the end of the year state assessment; however, benchmarks take away valuable classroom time. Each benchmark administered takes at least two days: one day to review, and one day to administer the test. If teachers actually go over the test and clear up any confusion the students may have had, it could take an additional third day. This becomes time that the average classroom teacher does not have to give. Because of the recent economic downturn, schools across the nation have been slashing budgets and cutting precious schools days out of the calendar. The time that teachers are spending on benchmarks could be better spent teaching standards and enriching the curriculum if benchmarks do not truly make a difference on end of the year assessments.

This study was designed to determine whether benchmark testing had any effect on student percentage scores of the CRCT in Science. The study looked specifically at middle level students in all grades, students receiving special education services (SPED), and English Language Learners (ELL). A vast amount of research was found that presented a definite need for addressing the concerns that many school districts have about state testing and how student scores can be increased. A review of literature uncovered a wide variety of research that examined the relationships between local and commercial made assessments that would provide

data enabling examiners to evaluate such programs. Research also indicated a need for this study due to many school improvement plans and the need to increase scores in science.

Theoretical Framework

Meaningful assessments allow students to demonstrate a deep understanding of the curricula being taught. According to "Current Perspectives" (Assessment and Reporting Unit et al., 2005), "assessments should be able to reveal the quality of students' understanding and thinking as well as specific content or processes" (p. 1). To employ significant assessment strategies teachers must first understand how students accept and interpret information. Students must understand that everything being taught has meaning. "Meaningful learning occurs when learners are actively involved and have the opportunity to take control of their own learning" (p. 2). Teachers should provide students the opportunities to construct personal knowledge through discovery and higher order learning activities. Knowing how a student accepts knowledge and how students construct personal knowledge directly ties to Piaget's Cognitive Stage Learning Theory and the Information-Processing Theory.

"The core theoretical assumption of Piaget's theory is that children are active thinkers, constantly trying to construct more advanced understandings of the world" (Siegler & Ellis, 1996, p. 211). Piaget's formal operational period covers ages eleven to fifteen, the same ages as the students in the sample of this study. Students identified within this age are able to think abstractly and connect current content to their prior knowledge. "Knowledge creation is viewed as occurring through a complex interplay between preexisting knowledge and new information gathered through interactions with the external world" (Siegler & Ellis, 1996, p. 212). Students in the formal operational period are better prepared for the demands of higher level abstract thinking questions that are presented in benchmark assessments; however, it is the teacher's

responsibility to make these benchmark assessment questions meaningful to the students. The teacher has to reinforce and re-teach the content presented on benchmark assessments to have a positive impact on future state standardized assessments.

In order to understand assessments, educators must first grasp how students receive and accept knowledge. The information-processing theory can help teachers and administrators alike understand this phenomenon. The information-processing theory is based on memory, mental representations, and problem solving. The information-processing theory can help explain the "mental process children apply to the information and, as a result, how they transform, manipulate, and use that information" (Miller, 2011, p. 267).

The information-processing theory works on an input/output system. As a student reads a question on a benchmark assessment, the information is inputted into the brain. Students then compare this information with previously learned/stored information through a process called encoding. Next, the information becomes an output and allows the student to then answer the question.

In a perfect world, this process would work beautifully; however, neither students nor the world is perfect. Some students can successfully process information in this manner thus making the students successful on standardized summative assessments, but the majority of the student population has not fully reached this developmental level. These students are more successful with hands-on learning and formative assessments where they perform various tasks to demonstrate new learning. This is why the researcher believed that benchmark assessments do not positively impact state standardized test scores. The majority of students are not mentally or emotionally ready to make the connection between benchmark assessments and state standardized assessments. That is why it is important for teachers to make the connections for

the students by using the data gained from benchmarks to enhance the classroom learning environment.

Today, the education world is no longer geared toward the general population. In a modern classroom, educators are teaching various types of students. During one class, a teacher could have students from any or all of the following categories: SPED, ESOL, gifted, and general education. Within these four categories there are also two subgroups: high functioning and low functioning. Because there are so many different types of students and no two students learn in the exact same way, there no longer exists a cookie-cutter method to appropriate learning theory, instruction, and assessment.

Swanson (1987) argued that information-processing theory approaches do not adequately apply to students with learning disabilities:

> Their lack of success in the classroom, whether in academic tasks or in social interactions, has been demonstrated by their inability to shift from one strategy to another, to abandon inappropriate strategies, to process information with one strategy and then select another, or even to consider several processing approaches in rapid succession in order to arrive at a solution to a problem. (p. 3)

Because special education students lack the skills to process informational efficiently, benchmark assessments are especially difficult for them. These difficulties make students frustrated and oftentimes the students end up guessing instead of trying to work through the problem which ultimately leads to lower test scores. ELLs have a similar problem of lower test scores. ELLs are struggling with the language of the test because the English language needs to be understood properly and the content language as well. Oftentimes this content language can

be like learning a third language for the ESOL student. Thus, they become frustrated which leads to guessing and lower test scores.

Figure 1 below depicts how Piaget's formal operational period and information-processing theory tie to benchmark assessments and state standardized assessments.

School Accountability

In order to fully understand the literature behind benchmark assessments, the researcher first looked at what led to this growing phenomenon in the world of education. In "Failing Our Children: No Child Left Behind Undermines Quality and Equity in Education," authors Guisbond and Neill (2004), discussed two main injustices that NCLB imposed on the American education system. The first injustice concerns "boosting standardized test scores [to] be the primary goal of schools" (Guisbond & Neill, 2004, p. 12). This belief presents a major flaw in NCLB because it creates a cookie-cutter educational system, and it takes away the high-quality education that all of students deserve. The second injustice states, "because poor teaching is the primary cause of unsatisfactory student performance, schools can be improved by threats and sanctions" (Guisbond & Neill, 2004, p. 12). Because of these "threats and sanctions" teachers channel all of their efforts on boosting test scores (p. 13). "However, these punitive actions fail to address underlying problems such as family poverty and inadequate school funding, which are major reasons that many students start off behind and never catch up" (Guisbond & Neill, 2004, p. 12).

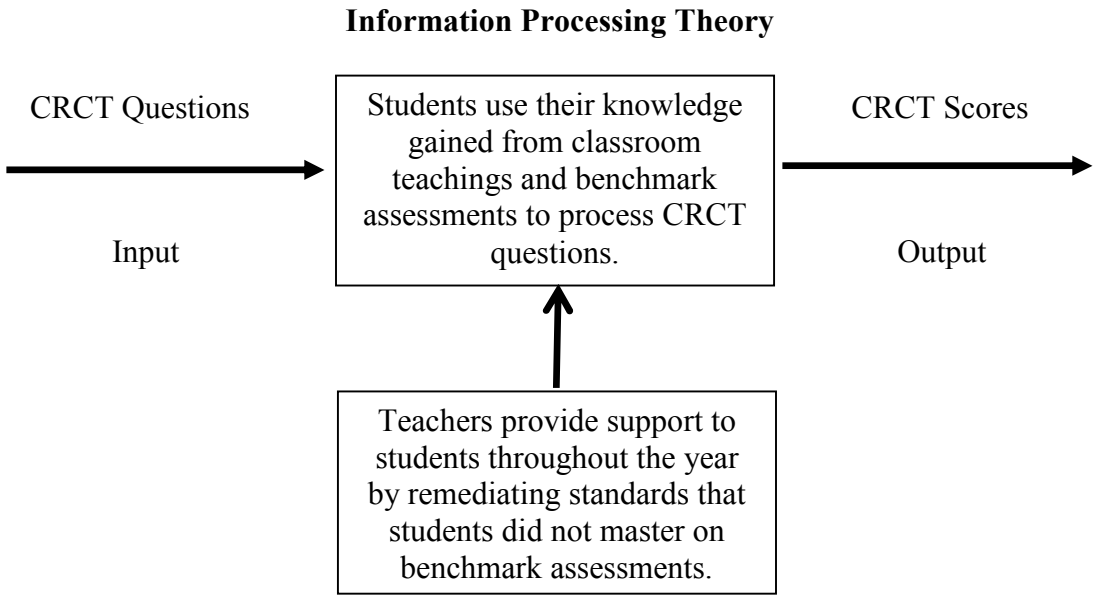

Figure 1. Information processing theory.

The largest punitive measuring tool that schools are facing concerns AYP. The problem with AYP as a measuring tool centers on the assumption that all schools are on a level playing field. In fact, the truth is most schools do not have the adequate resources available to have 100% proficient students. Instead, NCLB and AYP state that if teachers and administrators try a little harder, schools will achieve designated goals. "This reasoning ignores real factors that impede improvements in teaching and learning, such as large class sizes, inadequate books, and outmoded technology, as well as nonschool factors like poverty and high student mobility" (Guisbond & Neill, 2004, p. 13). Because of these factors, Guisbond and Neill "believe that rather than threatening educators with sanctions based on test results, a more effective approach focuses on gathering multiple forms of evidence about many aspects of schooling and using them to support school improvements" (p. 12). One such way is to demonstrate student growth through various forms of assessment.

Summative Versus Formative Assessments

In the article "Outcomes-Based Assessment in Practice: Some Examples and Emerging Insights," author Geoff Brindley (2001) discussed the growing debate between summative and formative assessments. Educational systems worldwide have been under the microscope in recent years to produce a more effective way of reporting student performance. One way systems have met this heightened demand is through the "adoption of systems that use pre-specified descriptions of learning outcomes – known, amongst other terms, as standards, benchmarks, competencies, and attainment targets- as a basis for assessing and reporting learners' progress and achievement" (Brindley, 2001, p. 393). These outcome-based systems have a number of advantages and disadvantages. One advantage is the alignment between the teacher and the curriculum. Having tests ensures that teachers have set goals that must be accomplished and met by a certain time. Another advantage this focused on was the individual needs of the students. Often tests can be broken down into subsets of the curriculum, and having this breakdown allows the teachers to see the areas in which the students need more work. One disadvantage of the outcome-based systems was the "doubts surrounding the validity of outcome statements and the reliability of the assessment tools that are used to elicit student performance" (Brindley, 2001, p. 394). There were questions whether or not these "benchmarks" were a true measure of students' performance. Also, teachers question if a test should be the only measure of student learning.

Creating Formal Assessments

Buck and Trauth-Nare (2009) developed a study that analyzed the process of creating formative science assessments. The importance of creating these formative assessments was to construct scaffolding in the classroom. Buck and Trauth-Nare stated:

> Assessments and the interpretations of their outcomes have direct and lasting impact on teachers, learners, and classroom activities. Assessments should serve to enrich students' understanding of science and not simply measure attainment of content knowledge. To this end, formative assessment should be an essential feature of classroom practice since the development and implementation of formative assessment serves to support science standards and promote learning. (pp. 475-476)

Buck and Trauth-Nare's assessment beliefs directly impacted the purpose of this study by explaining the importance behind implementing effective assessments in the science classroom.

In "Tough Choices in Designing a Formative Assessment System" Sharkey and Murnane (2006) discussed the problems that teachers and administrators face when choosing and implementing a formative assessment system. Sharkey and Murnane outlined two main reasons for investing in formative assessment systems. The first developed as a result of NCLB and the pressures of school accountability that came along with it. "Most state accountability systems also put pressure on educators to improve student performance on standardized tests. A growing number of districts see investments in formative assessments as part of their strategy for improving student achievement" (Sharkey & Murnane, 2006, p. 574). The second reason for investing in a formative assessment system dealt directly with technology. Assessment groups market the increasing technological advances to teachers through these formative assessment systems. Assessment groups boast of capabilities to grade the assessment and then generate reports and tables with pages of student data which can be used to enhance the student learning experience.

Sharkey and Murnane focused their research on the River City school district. River City has a high population of low socioeconomic students. The school district was consistently

experiencing low achievement scores in mathematics at the elementary and middle school levels. However, district leaders knew that a formative assessment system would not be enough to demonstrate student improvement; therefore, the leaders focused on improving the overall quality of instruction.

> Elements of the strategy include a new kindergarten through fifth-grade math curriculum named "Investigations," instructional support specialists for English language arts and math in every school, a teachers' contract that provides a significant amount of time for professional development, and a district-wide commitment to examining student work. (Sharkey & Murnane, 2006, p. 575)

The bulk of Sharkey and Murnane's research focused on the steps River City school district took in order to implement a formative assessment system. "The superintendent believed that providing teachers with timely information from these student assessments would help them to zero in on the delivery of instruction and individual student learning" (Sharkey & Murnane, 2006, p. 577). During the process, the district faced many challenges including deciding to buy into a company with ready-made tests or create district made assessments, should the test be paper or computer-based, and should the test be compulsory or voluntary. The purpose of the research was not to tell the tale of River City but to inform an audience about the challenges and questions that a district could face upon deciding to invest in a formative assessment system.

Mandated Assessments

Towndrow, Tan, Yung, and Cohen (2010) discussed the practices and professional development that occur during the implementation of science assessments. They specifically focused their research on teachers in Hong Kong and Singapore. Towndrow et al. looked to find out "how in-service science teachers identify with and change their teaching practices in

response to externally mandated assessment policy and practice reforms" (p. 119). Mandated reform can have various effects on how a teacher performs his or her duties. Towndrow et al. found that teachers behaved in three ways in response to mandated assessment: (a) drew on their own personal experiences, (b) focused on the requirements of the reform, and (c) implemented changes that focused on what was best for the students.

Towndrow et al. also focused on how teachers were presented with reform that leads to mandated science assessments through professional development:

> There would seem to be two broad and largely opposing approaches relating to teacher professional development that could be adopted in this respect: one follows a short-term, training based agenda that is usually conducted offsite by an external agent, the other involves the adoption of a more continuous situated and learning based approach. (p. 119)

Offsite professional development is a more popular choice; however, offsite professional development creates disconnect between teachers and the classroom, and oftentimes the reform becomes distorted because of this disconnection. Schools choose this method of professional development because it allows more teachers to be reached simultaneously.

The second approach to professional development provides more teacher-centered instruction. Teachers work directly with an outside person to make changes in the schools and classrooms. It allows the teacher to implement change based on experiences both in and outside of the classroom. Just as with the first approach, the second approach also has flaws. Teachers can become burned out quickly with the new changes, and teachers can even drop the changes once the outside source is no longer present.

Towndrow et al. (2010) stated:

> Despite the abundant professional development opportunities and programs presented to teachers, those initiatives that overly focus on measurable learning outcomes and short term gains might fail to account for teacher participation and teacher learning that is situated within a complex web of context-specific variables including politics, pedagogy, and innovation. (p. 120)

Mandated benchmark assessments are one such piece of reform that teachers are struggling to fit into the classroom. Towndrow et al. stated that advocates of assessment reform "often promote assessment practices that are squarely focused on learning as opposed to teachers' instruction" (p. 120). Furthermore, Towndrow et al.'s research focused on how practical science assessments were implemented into various schools in Hong Kong and Singapore after receiving professional development on assessment reform. While the schools in both settings took initiative on the assessment reform, each school approached it in a different manner; however, each school fell into one of the three previously discussed behaviors. Teachers in Singapore took a more critical stance and focused solely on the implementation of the reform from their training. In Hong Kong, teachers took a more personal approach and drew from their own personal experiences to implement the reform.

Although the research that Towndrow et al. conducted took place in a different country, it directly applies to the implementation of benchmark assessment in the United States. Teachers across the nation attend professional development seminars to receive training on new assessment reform as an answer to mounting pressures of student achievement set forth in the NCLB Act. "One strategy for compelling schools to examine and improve students' performance outcomes is to require more strenuous and intermittent testing in line with state standards for curriculum and instruction" (Bancroft, 2010, p. 53).

Successfully Implementing Benchmark Assessments

Means, Padilla, DeBarger, and Bakia of the U.S. Department of Education, along with a team of researchers and educators, published an article in 2009 that discussed ways to successfully implement benchmark testing programs. According to Means' et al. article, "Implementing Data-Informed Decision Making in Schools—Teacher Access, Supports, and Use," one of the most common obstacles of a successful testing program was training teachers how to interpret data and how to translate data into changes in instructional practice. Professional development was a key factor to teach teachers how to use the data. Another important step was to "build teachers' mutual trust to a point where teachers are comfortable working with colleagues to examine data that reflected on their teaching performance" (Means et al., 2009, p. 48).

Means' et al. article concluded by stating that many schools are making a real effort to make benchmark testing programs successful, but schools still have a long way to go before benchmarks are considered a completely effective method for instruction. Not one school or district in the study showed a totally integrated process, but most schools were making slow, steady, and good faith efforts.

In an article published in 2011, Britton discussed using formative and summative assessment to support learning in the science classroom. Britton used a variety of assessments that she described in detail. Particularly important to this study is Britton's use of pretests. "Pretests are formative assessments that locate misconceptions, misunderstandings, and prior knowledge so that an effective and efficient instructional plan is designed" (p. 21). The use of pretests by Britton is the manner in which benchmarks are used in the school of this study. Benchmarks are assessments administered by classroom teachers to determine what instructional

activities and changes need to be made in the curriculum to prepare students for end of the year state achievement assessments.

In the 2010 article, " School Formative Feedback Systems," author Richard Halverson "presents the concept of a formative feedback system to identify the capacities that many school are developing in the quest to meet the demands of high-stakes accountability policies" (p. 130). The main premise behind a formative feedback system was to use data to drive instruction. This type of system was used to give timely feedback to administrators, teachers, and students alike so that all parties involved knew where the student fell in an educational expectations timeline. If instructors were unable to identify how a student was performing, then the instructors could not determine what further interventions the student might need.

Halverson discussed the formative feedback system in terms of three functions: intervention, assessment, and actuation. These three functions are discussed in terms of the information processing theory. Halverson related interventions to signals, assessments to sensors, and actuation to translations. The brain receives a signal and then uses a sensor to translate the signal. The student receives an intervention for learning and then uses the assessment to actualize the intervention for learning.

Interventions can be described on a whole group level or an individual level. Whole group interventions are things the teachers does for the entire group that helps to guide the learning for that group. A personal intervention may be something that a teacher has done for a single student, such as providing extra time on an assignment as laid out by the student's Individualized Education Plan (IEP). "The learning that results from an intervention is analogous to the signal in classic information processing systems theory" (Halverson, 2010, p. 132). Assessments serve as the sensor to the signal. "Assessments provide the information to

help teachers determine the degree to which signal received (estimates of student learning) correspond with the learning goals built into the interventions" (Halverson, 2010, p. 132). The actuation function of the system concerns the process through which the student assessment results are translated by the teacher into usable data that drive future instruction.

Halverson specifically discussed the use of this type of system with the implementation of benchmark assessments. Halverson stated:

> Benchmark assessments aim to provide timely and appropriate data to guide schools in making effective decisions about teaching and learning. The systems typically involve output processes to deliver the assessment information in student-level or learning-standard-level reports that make sense for guiding teaching and learning. (p. 140)

The test can be designed in many different ways from computer adaptive tests to teacher generated that align with local or state standards. Either way, a large amount of data is produced that helps teachers to design future instruction and interventions.

One main issue that has arisen from the enactment of NCLB concerns teaching state standards without standardization. In the article, "Benchmarks Curricular Planning and Assessment Framework: Utilizing Standards without Introducing Standardization," Erika Feldman (2010) addressed this issue. Standardization uses a standard way to teach across the board. Multiple learning and teaching techniques are not offered to students in this process. Each student and class receives the exact same experience, a standard education. By doing this, teachers try to push round pegs into square holes. However, Feldman offered ideas and plans on how to curricular plan and develop assessment that avoids this downfall.

One way to avoid such standardization is to steer clear of prescribed curriculums when implementing standards and assessments. Instead, the curriculum should be based on

observations and needs of the students. This approach is valued because it allows the teacher to focus on the students instead of the standardized tests. Teachers should also embrace diversity within their own classrooms. Diversity is not just limited to the different cultures that are present, but it can also include different teaching and learning styles and assessments and student learning backgrounds such as public, private, or home-school. Finally, administrators and teachers alike should avoid assessment programs that track and test single students. Feldman (2010) stated, "Individualized assessments usually focus on a single domain of development and do not provide information about activity across different developmental domains" (p. 234). Overall, teachers can avoid the downfall of standardization if the teacher addresses the learning and assessments needs of individual students rather than looking at the groups and classes as a whole unit.

Feldman also discussed the use of a program called Benchmarks Curricular Planning and Assessment Framework (BCPAF). In her article, benchmarks were not simply tests that students took periodically, but benchmarks were a set of standards and goals that teachers developed during curriculum planning. Students were then assessed periodically to determine if students met a particular benchmark with a predisposed time frame. Feldman stated that there were three main steps to the assessment process:

> The first part of the assessment process entails writing down field notes during observations of children in early learning environments. The second part involves breaking up the observation into small episodes and writing descriptive interpretations of those episodes. The third part involves assigning benchmarks based on criteria, decision rules, and examples of behaviors that do and do not meet the criteria. (p. 236)

Benchmark Assessments

In "Benchmark Assessments Offer Regular Checkups on Student Achievement," Olson (2005a) stressed concern that the quality of benchmark tests were lacking due to vendors rushing into the benchmark/formative assessment market:

> The reason that there is a boom in benchmark assessments is that most states and school systems are providing nothing more than autopsy reports right now…They tell you why the patient died at the end of the year, and then marveled that the patient didn't get any better. (p. 13)

Schools and districts want to participate in more preventative strategies to diagnose problems earlier rather than being surprised at the end when it is too late to remediate and improve student learning. Therefore, the benchmark assessment vendors market continues to boom. Vendors are attractive to districts and schools because they offer "extensive reporting systems that break down test results by the same student categories required under federal No Child Left Behind Act" (Olson, 2005a, p. 13). Critics argued that currently no proof exists to support the claim that benchmark testing improves student learning. Olson reported:

> Dylan William, a senior researcher at the Educational Testing Service, wrote an influential review that found that improving the formative assessments teachers used dramatically boosted student achievement and motivation. Now that same evidence, he fears, is being used to support claim about long-term benefits of benchmark assessments that have yet to be proven…"We just don't know if this stuff works." (p.13)

In 2007, Brown and Coughlin completed a study entitled, "The Predictive Validity of Benchmark Assessments." The research supported school districts' concerns about administering periodic assessments "to provide information to guide instruction (formative

assessment), monitor student learning, evaluate teachers, predict scores on future state tests, and identify students who are likely to score below proficient on state tests" (p. 2). The study addressed the key question of whether there was evidence that benchmark assessments could predict state assessments such as the CRCT and the relationship between the two.

The schools used by Brown and Coughlin were mostly concerned with students who did not perform well on state tests and to evaluate how teachers prepare students for state examinations. The results of the tests were then analyzed and evaluated much like the data contained in this study. The results and implications of the Brown and Coughlin study determined that more evidence was needed before a final conclusion could be made and recommended further research on the subject. Brown and Coughlin concluded the study by stating,

> Such evidence is crucial for school districts to make informed decisions about which benchmark assessments correspond to state assessment outcomes so that instructional decisions meant to improve student learning, as measured by state tests, have a reasonable chance of success. (p. 12)

Using Data to Enhance Learning

While the Brown and Coughlin (2007) study discussed the question of the relationship between benchmark testing and state tests, Christman et al.'s 2009 study entitled, "Making the Most of Interim Assessment Data," focused on the question of formative versus interim assessment. In this study, teachers analyzed benchmark data and developed instructional responses to be implemented during the course of study which usually lasted approximately nine weeks. The research team did agree, however, that

> School reformers have embraced data-driven decision-making as a central strategy for improving much of what is wrong with public education. The appeal of making education decisions based on hard data – rather than tradition, intuition, or guesswork – stems partly from the idea that data can make the source of a problem clearer and more specific. (p. 1)

In 2005, Herman and Baker conducted a study called "Making Benchmark: Six Criteria can Help Educators Use Benchmark Tests to Judge Student Skills and to Target Areas for Improvement." The first criterion concerned alignment, which meant aligning test questions with state standards and assessments. The second criterion focused on diagnostic value, which measured how much useful feedback the test provides for instructional planning. The third criterion for effective benchmark testing focused on fairness. This takes into account how many, if any, questions are considered to be biased to the many diverse subgroups. The fourth criterion, technical quality, addresses the quality and accuracy of the information received from benchmark tests. The fifth criterion centered on utility, which was the extent to which intended users find the test results meaningful and are able to use them to improve teaching and learning. Lastly, the sixth criterion of feasibility focused on making benchmark testing worth the time and money. Together, these criteria have been found to make benchmark testing worthwhile (Herman & Baker, 2005).

In their study, "Learning to Learn from Benchmark Assessment Data: How Teachers Analyze Results," Olah, Lawrence, and Riggan (2010) took the study of benchmark assessments a step further by looking at what the teachers actually did with the data that was derived from administering benchmark assessments. The researchers conducted a series of interviews with 25

teachers in a Philadelphia school. Each teacher was visited and interviewed at three different points throughout the school year.

> Visits to the schools were scheduled to coincide with the district's reteaching week, the time in each 6-week curricular cycle in which teachers were allotted five days to revisit, reteach, practice, and enrich content that had been covered in the previous five weeks. (p. 228)

The purpose of these interviews was to determine how teachers analyzed the data and what the teachers did with the data once the analysis was complete. The study focused on the teachers' interpretation of the data. If the teachers discovered patterns in the data, then the researchers probed further to determine what the teachers did with this information. The teachers also looked at what happened to the high achieving students during the remediation week. The teachers wanted to know what types of enrichment these students received during the five days of re-teaching. Overall, the researchers found that teachers used the data derived from benchmark assessment to further student learning inside the classroom. "Although teachers may not always be using them in the way the district intends them to be used, the fact remains that they are consulting, analyzing, and acting on interim assessment results" (p. 244).

In 2007, Matthews, Trimble, and Gay also discussed how to use data from benchmark testing to drive instruction in the article, "But What Do You Do with the Data?" The article specifically focused on the Camden County School System in Camden County, Georgia. In order for the test to be successful, the data must be used and "teachers must accept the data, know what numbers indicate, and be ready to change their instruction" (p. 53). Camden County School System followed a three step plan to get the most out of their data: "schedule intensive data sessions, prepare data for teachers to examine, and lead teachers in data analysis" (p. 53).

Data sessions were scheduled to discuss, with the teachers, the results of the benchmark test. The sessions were scheduled and carried out as quickly as possible so teachers could adjust classroom instruction to meet the needs of the students. "By examining the data reports of the whole school, then looking at the reports that disaggregate the data by grade, by teacher, and by individual student," (p. 54) Camden County Schools gained the best insight possible into the data.

In the article "Benchmark Assessments," Coffey (2009) stated,

> benchmark testing couples student performance with extensive reporting systems in order to break down test results by the same student categories required under the federal No Child Left Behind Act (i.e. race, income, disability, and English proficiency) in addition to providing individual progress reports at the district, school, classroom, and student level. (para. 2)

Benchmark tests can include performance tasks, but they are often seen in the same form as standardized tests. These benchmark tests are used to determine the level of student achievement, and the results are compared to other classrooms of the same content.

In the article "EQUIPping Teachers" by Marshall, Horton, and White (2009), the power of benchmark assessments was again emphasized. It stated,

> One way to improve our teaching practice is to use a benchmark assessment to obtain a solid point of reference that honestly reflects what we do in the classroom, and then to design a developmental plan to raise the level of performance. (p. 46)

This article focused on inquiry-based instruction and gave a variety of guides to assist teachers in teaching using inquiry while still meeting their standards as measured by benchmark assessments. Most of the guides focused on higher order thinking skills, and used questioning

techniques that challenged students to form a personal understanding about concepts. The EQUIP program helped to bridge the gaps between inquiry teaching and standardized testing. On the other hand, many teachers are apprehensive about benchmark tests and how the use of these assessments will impact the classroom environment.

In the article "Monthly Checkup: A New Principal Works with Teachers to Get the Most out of a District's Benchmark Assessment," author Mary Ann Zehr (2006) discussed Charlotte Kreuder, a principal who helped her teachers use data derived from benchmarks to drive instruction. "That effort, she believes, helped the school narrow a large gap in test performance between African-American and white students and significantly raise the state-test scores of all its students" (p. 36). Kreuder moved to a new school and was hopeful to attain the same success with a low achieving group of ELLs. Kreuder implemented a benchmark program that assessed students from second grade through fifth grade once monthly in math and reading. The assessment was computer-based and aligned with the state achievement test. "The DeKalb district requires teachers to meet in grade-level teams for another two hours each month to look at students' results in various reports and decide how to teach the standards that students are weak on" (p. 36).

Advantages of Benchmark Assessments

The article "Tip of Their Fingers," by Vaishali Honawar (2006) addressed that benchmarks assessments give teachers "an array of student-level information from the benchmark assessment data" (p. 38). This article illustrated the advantage of having a data system to help analyze and breakdown data from benchmark assessments. It also showed how this tool helps equip teachers with what they need to meet the needs of their students' weaknesses individually or as a whole group. Mary Rooney Thorp is a teacher who utilized the

data to figure out where she needed to modify her instruction. She gave a specific example of how she immediately was able to view the results of a benchmark assessment and see that all her students missed the same three questions. At that point she realized that there was a problem and was able to go back to the drawing board and focus on that skill that the class had missed. This example shows how valuable benchmark assessments can be with the right data analysis system. One teacher in the article stated, "The system has made me a more comprehensive teacher and helped me pay attention to what my students are learning" (p. 39).

In the article titled, "Using Literacy Assessment Results to Improve Teaching for English-Language Learners," Helman (2005) discussed the benefits of benchmark assessments when in regards to progress monitoring the success and progress of ELLs. The article stated that "Ongoing assessment of early literacy progress is essential for giving teacher the information they need to measure student progress, identify students who may require additional or individualize assistance, and guide instructional practice" (p. 668).

The study reported on the data that showed ELLs benefited most from a literacy class designed specifically to address their specific needs such as: developing vocabulary, comprehension, background knowledge, and sound-symbol commonalities and differences between home and new language. Another advantage addressed in this article included literacy benchmark assessments. These literacy benchmark assessments "give teachers an opportunity to review student performance along a continuum of development, and in addition to identifying individual needs, early literacy assessments provide teachers with the information they need to organize instructional groupings in the classroom" (p. 674).

On a larger scale, literacy assessments give schools, districts, states, funded projects, and the nation important information about how well various groups of learners are doing at

> meeting agreed-upon expectations and call on us to respond when we see inequities. (p. 675)

Having literacy benchmark assessments for ELLs helped the teachers to determine further language developmental needs for each student.

In the 2009 informative article, "Using Student Achievement Data to Support Instructional Decision Making," author Hamilton et al., supported the idea of benchmark assessments. They explained that student data was readily available in most schools, but the question was what to do with it. "Using data systematically to ask questions and obtain insight about student progress is a logical way to monitor continuous improvement and tailor instruction to the needs of each student," she said of her research (p. 5). Hamilton et al., further explained that having access to this data allows teachers to prioritize instructional time, target struggling students who are having difficulty with particular topics, identify students' weaknesses, and refining instructional methods.

Although the article by Hamilton et al., appears to fully endorse benchmark testing and assessment, they did warn that these tests vary in reliability and level of detail. "No single assessment can tell educators all they need to know to make well-informed instructional decisions," they stressed (p. 6). One unique aspect of this article suggested teaching students to examine their own data and set their own learning goals. This would require teaching students how to interpret this data, setting clear expectations, and providing tools to help them learn from the feedback.

Henderson, Petrosino, Gukenburg, and Hamilton (2007) seemed to agree with Hamilton et al., (2009) in the sense that data acquired from benchmark testing be used to its fullest advantage. She also stressed this point in her research team's study entitled, "Measuring How

Benchmark Assessments Affect Student Achievement." Henderson et al. reported that many schools that implemented a benchmark testing program did not see any substantial differences after the first year of assessments. "The finding might be because of limitations in the data rather than the ineffectiveness of benchmark assessments" (p. 7). She also explained that since this type of testing was relatively new, it was difficult to compare the results to other schools which may have been implementing their own versions of assessments. These uncontrolled conditions are very difficult to interpret because there are no set guidelines for schools to follow.

Henderson and her team believed that it was still too early to observe any real impact from interventions in schools that performed benchmark testing. Henderson and her team recommended continuing to track achievement data and allow teacher collaboration and more data analysis based on individual student performance in order to get a real picture of the effectiveness of this program.

Disadvantages of Benchmark Assessments

The comparison of classrooms and teachers of the same content can be very stressful for teachers. Lynn Olson (2005b) wrote an article, "Not All Teachers Keen on Periodic Tests," that discussed the added stress benchmark assessments can have on teachers and students. This article discusses that benchmark assessments were viewed by the students as "totally meaningless and very intrusive, because it was another interruption, in addition to all the other testing" (Olson, 2005b, p. 13). The benchmark tests are usually given monthly, and the teachers are required to form data teams to digest the results of the benchmark assessments and reform classroom teaching and instruction accordingly.

In the article John W. Hutcheson, states that he left public school and went to private school mainly because of these tests. He says "the benchmark assessments are a reflection of the

standardized tests, which are only a small piece of learning" (Olson, 2005b, p. 13). On the other hand, the article points out that a Norfolk, VA school district has found great success and gains in reading and math and credit data-driven instruction as one key to the district's success. The school districts in this article left analyzing and utilizing the data up to the teachers which could be viewed as a huge burden added on the teachers' work load. In the article, the teachers had a good attitude about benchmark testing because school districts had a data analysis system that broke the results of the assessments down for the teachers by standard and/or skill. These teachers were able to take the data and run with it, while the other teachers were required to plow through it on their own and form an understanding.

In 2010, Bancroft investigated how benchmark assessments worked as an indicator for high achievement scores. Her research focused on English language arts in a low-income high school over the course of three years. "The administration sought to assess students' master of individual standards using three standardized benchmark tests over the school year before giving the state test in late spring" (p. 54). Through interviews with teachers and administrators, Bancroft found four disadvantages within the benchmark testing program:

1. Students who were reading below grade level at the beginning of the three years did not make academic improvements with the administration of benchmark assessments.
2. Teachers became unorganized in their teaching methods during the benchmark process.
3. Teachers felt that they were wasting valuable teaching time to the benchmark process.
4. Self-reflection after the benchmarking process was unrealistic because of a lack of resources within the school.

Bancroft's study showed that a low-income school with a lack of resources and a student population with low academic skills cannot effectively implement a benchmark testing regime.

> Given that teachers and administrators struggled with the classroom realties to carry out this process over three years, in the end the benchmark tests were abandoned as a failed method for understanding and improving teaching and learning in the schools English classes. (p. 54)

In 2002, Leigh Hall conducted research that focused on the issues and purposes of implemented standards and benchmark assessments in the middle school setting. She discussed five main issues in terms of standards and benchmarks:

1. The content that should be taught;
2. The skills that should be taught in addition to content;
3. The importance of specific content knowledge on state exams;
4. Specific skill outside of content that will be tested; and
5. Skills or content missing from the documents that are found in the state exams. (pp. 213-214)

Her focus showed that standards and benchmarks were not for the benefit of the students, rather the standards and benchmarks represented a vehicle for exams to drive instructions.

States expect school districts to align the curriculum being taught with state-mandated standards and to meet standards on end of year tests. However, states did not make curriculum documents interpretable. The standards are very vague and broad, and cover a wide range of material that may or may not be on assessments. It becomes a guessing game for classroom teachers concerning what materials teachers cover. Teachers must decide what they think might be on the state standardized tests. Therefore, teachers make assumptions and generate

benchmarks to gage student progress based on these assumptions, and the content may or may not be on state-mandated end of year assessments. In regards to aligning standards and benchmarks, Hall (2002) stated the following:

> It could be inferred that by having the documents aligned, students would be more likely to be exposed to the content and skills necessary to do well on the exam. How that would increase student achievement and what is meant by that term are still unclear. If student achievement is being measured in terms of test scores, then that is only further evidence that the exams dictate what is important for students to know and be able to do. (p. 216)

Focus on Science

With an increase on school and student accountability, many changes begin to take place in school communities. One of those changes is requiring higher achievement on standardized assessments in science. Liu, Lee, and Linn (2010) studied the different techniques used in science assessment. One such technique was inquiry-based learning. "Many science education researchers have implemented inquiry science teaching programs to improve the current situation of science learning and teaching by placing more emphasis on fostering students' deep scientific understanding and less emphasis on memorizing science facts" (p. 70). This study exposed important information because the focus was on how students learn in science, and how students take gained knowledge and apply it to assessments. This was especially important to this study of benchmark assessments because a student's success on these assessments was due to the knowledge that each student gains in the classroom.

Maerten-Rivera, Myers, Lee, and Penfield (2010) examined the predictors set forth by high stakes tests in science, and it was created in response to an increase in accountability for science. Maerten-Rivera et al. looked specifically at how these predictors had an effect on ELLs.

> This emerging role of science as part of accountability systems is unique relative to reading and mathematics education that dominate schooling based on the tradition of educational policies and practices. The policy change in science is particularly relevant to ELL students who have traditionally not been part of -large-scale testing in science. (p. 938)

ELLs have been exempt from testing in science because of the students' need to be immersed in literacy and mathematics for basic survival in school. However, all students, along with the schools, are now held accountable for all subject areas. The researchers saw the need to determine what effect a student's ESOL status had on their science achievement. The researchers found that there was an achievement gap in science when comparing ESOL students to the entire school population.

In 2010, Cowie et al. studied assessments for learning in science. They believed that providing the correct assessments and proper teaching methods to students prepared them for a future in society. "Student participation and achievement in science is a social justice and equity matter because of the role science and its technological applications play in defining may of the key issues and opportunities facing society today" (p. 347).

Cowie et al. specifically looked at science education in New Zealand and ways to enhance student participation, engagement, and achievement. Cowie et al. found, "A focus on assessment for learning directs attention to the value of students being able to access multiple sources of knowledge and feedback as part of their active engagement with ideas and participation in classroom activities" (p. 363). This fact is the fundamental aspect behind formative assessments that prepare students for end of the year summative assessments. Students have to see value in the content that is being taught. The value students see comes from

teacher feedback and the student continually using the knowledge they gain to construct new and different knowledge.

Overall, the educational community is entering the informational/technological world of reports and analysis. Everyone on all levels feels the stress of meeting expectations because the age of accountability is present. Now educators must figure out, do benchmark assessments help in the early detection of problem areas, or are these assessments just another hoop for students and teachers to jump through? The previous articles demonstrate how data driven instruction from benchmark assessments affects student improvement. When a school has a system in place that analyzes data from benchmark assessments, the teachers must take the data and focus on problem areas while improving academic instruction. This is one way to ensure benchmarks are actually effective in the classroom setting. Simply administering the benchmarks and walking away will not ensure success on end of the year standardized assessments. It all comes down to what the teachers are willing to do with the data they gain from the benchmarks.

CHAPTER THREE: METHODOLOGY

The literature previously reviewed in this study suggests that benchmark assessments should be used to monitor student performance. The literature suggests that teachers should use data gained from benchmarks to drive future instruction in the classroom. However, there is a gap in the literature because no one researcher is addressing the question that educators should be asking. Are benchmarks impacting end of year state standardized achievement tests? Approving or discarding this gap in the literature will help educators decide the true value of benchmark assessments. The purpose of this study was to examine the effects of benchmark assessments on Georgia CRCT percentage scores in middle grades science.

This methodology section will contain a research design, research questions and hypotheses, a description of the participants and setting, instrumentation, procedures, and data analysis.

Design

The researcher will use the causal-comparative design for the methodology of this study. Causal-comparative designs are a form of ex post facto research. Gall et al. (2007) defined ex post facto research as "designs that rely on observation of relationships between naturally occurring variations in the presumed independent and dependent variables" (p. 306). The causal-comparative design was chosen for the study because it allowed the researcher to look at data that was already available and determine the cause and effect relationship between the data based on a manipulated variable.

The researcher investigated how the independent variable of benchmark assessments affected the dependent variable of science CRCT percentages. The treatment group consisted of 15 schools in Georgia that administered benchmark assessments throughout the year to monitor

the students' progression toward mastery of the GPS in science. The control group consisted of 15 schools that did not use benchmark assessments as a progress monitoring tool. Besides investigating middle grades students as a whole group, the researcher also examined the effects of benchmark assessments on science CRCT percentages within the subgroups of students enrolled in the ESOL and SPED programs at the control and treatment.

Participants

The participants in this study came from 30 different middle schools across the state of Georgia. Each school had approximately 600 hundred students. Gall et al. (2007) stated, "In causal-comparative research, there should be at least 15 participants in each group to be compared" (p. 176). Fraenkel et al (2015) also suggested a minimum sample of 30 for causal-comparative research. A sample of 30 schools gave the researcher approximately 18,000 data points in the entire sample. Since the researcher was looking at percentage scores for middle grades science students at each school, there was only 30 samples collected to which the 18,000 data points contributed. The home school of the researcher was one of 15 schools in the experimental group. The home school had the following demographics. The school was located in northwest Georgia in the Appalachian Mountains. The school had been operating for six years. The school consisted of grades seven and eight and had a total school population of 710 students. Seventh grade had 357 students, and eighth grade had 353 students. The school was not widely diverse with 79% of the population being white students. Of the remaining 21% of the population, 0.5% was Asian, 0.5% was African American, 19% was Hispanic, and 1% was multi-racial. The school also had 9% of its population in the SPED program, 5% in the ESOL program, and 67% in the free and reduced lunch program.

The researcher used purpose sampling to gain subjects for this study. Purpose sampling was the best procedure for the study because this sampling method allowed the researcher to choose subjects that fit a specific purpose. Purpose sampling is used in research when a research has prior information on a group that the researcher believes will provide the needed data (Fraenkel et al., 2015, p. 101). The researcher used data available on the Georgia Department of Education website and found schools that were comparative to the researcher's home school free and reduced lunch percentages. A free and reduced lunch percentage told the researcher what percentage of the population was considered to have a low socioeconomic status. Once the researcher compiled a list of schools in the state of Georgia that were similar to the home school based on free and reduced lunch percentages. The researcher then began to contact schools or school districts to determine if the schools administered benchmark assessments to science students in the 2011-2012 school year. The researcher added schools to the treatment and control groups until the sample was large enough to adequately run data analysis. Using this information, the researcher then pulled CRCT data compiled by the state Department of Education to determine the percentage scores for the whole population and the subgroups of SPED students and ESOL students.

Site

The site of the research was a small community in northern Georgia. The community contained approximately 28,000 people, and the largest employer of the area was the local school board. The community was comprised of six schools: one primary school (pre-kindergarten through first grade) one elementary (second through fourth grades), one elementary school (kindergarten through fourth grade), one middle school (fifth and sixth grades), one middle school (seventh and eighth grades), and one high school (ninth through twelfth grade). The

research was conducted to the benefit of the middle school containing seventh and eighth grades. The school's improvement plan had specific goals designed for the improvement of CRCT scores. These goals specifically addressed the importance of raising the scores of the ESOL and SPED populations. One way in particular that the school intended to measure the progress toward these goals was through benchmark assessments.

Teachers used benchmark assessments to determine how their students were progressing through the curriculum. Once a benchmark was administered, teachers then analyzed the data to determine strengths and weaknesses of the student population. A plan was then developed to address the areas of weakness and to enrich the areas of strength.

Instrumentation

The instrument used in the research and from which data was collected is the Criterion Referenced Competency Test (CRCT). The CRCT was the Georgia standardized test given to all students in grades three through eight in the subjects of reading, language arts, mathematics, science, and social studies. In order for students to meet standards and pass the test a score of 800 had be earned. Students who scored below an 800 in reading and/or mathematics in grades three, five, or eight were required to take the test a second time. If they were still not successful, students may not have been promoted to the next grade. Any score received below 800 was categorized as "does not meet standards." Any score received between 800 and 849 was categorized as "meets standards." Any score 850 and above was categorized as "exceeds standards." Maximum scores among the five content areas varied. The CRCT had a Cronbach's alpha score of reliability of 0.858 - 0.932. Cronbach's alpha score tests the "internal consistency of an instrument" (Fraenkel, Wallen, & Hyun, 2015, p. 158). The alpha score varies among the five content areas.

CRCT scores can be used in a variety of ways. The scores can be used to assess individual student achievement in all five subject areas. Each subject area is broken down into content weights. When an individual school receives CRCT scores from the state, each student's score is broken down by the content weights. Using this information, a teacher could determine the area(s) of weakness each student may have had for a particular subject. Each content weight category accounts for a certain percentage of the overall score. Since this study focused on middle grades science achievement scores, the content weights for middle grades science are discussed. The sixth grade science CRCT is divided into three subcategory content weights: geology (40%), hydrology and meteorology (40%), and astronomy (30%). The seventh grade science CRCT is divided into three subcategory content weights: cells and genetics (35%), interdependence of life (50%), and evolution (15%). The eighth grade science CRCT is divided into three subcategory content weights: structure of matter (30%), forces and motion (30%), and energy and its transformations (40%). Once scores are released, students receive a detailed account of how many questions they answered correctly or incorrectly for each category.

CRCT scores can also be used as a measuring tool for teachers, specific classes, schools as a whole, and entire school systems. How well students perform on the CRCT is directly tied to the school's and the system's score on the CCRPI.

Procedures

After successfully completing the requirements of EDUC 919 and receiving a passing grade on the comprehensive exams, the research candidate was enrolled in EDUC 980 where a research prospectus was developed. During this class, the candidate secured a committee for the dissertation process. Upon the completion of EDUC 980, the candidate enrolled in EDUC 989 twice. In this course, a dissertation proposal was developed to be defended. Once the proposal

had been defended and the candidate had received permission from the committee and the Internal Review Board (IRB), the candidate began to contact schools and local school boards for permission to be included in the study. After the candidate had received permission from the participants, data collection began. The researcher then began to compile all data obtained from the state Department of Education website into a spreadsheet. After all the data had been collected, the analysis of the data began. The candidate continued to develop the dissertation through correspondences with the committee chair. Finally, the committee chair worked with the candidate to schedule the dissertation defense.

Data Collection

Data was collected from the sample one time only. The researcher collected middle grades science CRCT pass percentage scores from each school participating in the study from the 2011-2012 school year. The researcher collected CRCT science pass percentage for the school as a whole and for the separate groups of SPED and ESOL. For each school the researcher also collected the number of students tested as whole, SPED, and ESOL. After the data had been collected by the researcher, all schools were alphabetized in a spreadsheet and assigned a numeric code ranging from one to thirty. Each school had three subgroups: whole group, SPED, and ESOL. The control and treatment both had the same subgroups. Once the subgroups were created, the researcher began to input data obtained from the state Department of Education website into the spreadsheet. This kept all the data organized and kept the general education, SPED, and ESOL scores separated. All the data entered in a spreadsheet program could then be transferred to SPSS where the appropriate data analysis tests could be performed.

Data Analysis

The researcher performed a *t* test to determine if the null hypotheses was rejected or accepted. The *t* test for a single mean tests "whether a sample mean differs significantly from a specified population mean" (Gall et al., 2007, p. 317). If problems arose in the score distributions, the researcher would also use the nonparametric Mann-Whitney signed rank test. "Nonparametric statistics are tests of statistical significance that do not rely on any assumptions about the shape or variance of population scores" (Gall et al., 2007, p. 325).

CHAPTER FOUR: FINDINGS

Introduction

The purpose of this causal-comparative study was to determine if the administration of benchmark assessments among middle school science students resulted in a statistically significant difference in student CRCT pass percentages compared to CRCT pass percentages of middle school science students who were not administered benchmark assessments.

This chapter presents the results of the statistical analysis of the comparison of pass percentages on the 2012 middle grade science CRCT between schools that administered benchmark assessments and schools that did not administer benchmark assessments by whole group, by students in the ESOL program, and by students in the SPED program.

Research Questions

A causal-comparative research design and a *t* test were used to address the following research questions and hypotheses:

RQ1: How does the administration of benchmark assessments affect Georgia Criterion Referenced Competency Test (CRCT) science pass percentage scores among middle grade students?

RQ2: How does the administration of benchmark assessments affect Georgia Criterion Referenced Competency Test (CRCT) science pass percentage scores among middle grade students who are in the English to Speakers of Other Languages (ESOL) program?

RQ3: How does the administration of benchmark assessments affect Georgia Criterion Referenced Competency Test (CRCT) science pass percentage scores among middle grade students in the Special Education (SPED) program?

Hypotheses

$H_0$1: There is no significant difference in Georgia Criterion Referenced Competency Test (CRCT) science pass percentage scores of middle grade students who were administered benchmark assessments compared to Georgia Criterion Referenced Competency Test (CRCT) science pass percentage scores of middle grade students who were not administered benchmark assessments.

$H_0$2: There is no significant difference in Georgia Criterion Referenced Competency Test (CRCT) science pass percentage scores of middle grade students in the English to Speakers of Other Languages (ESOL) program who were administered benchmark assessments compared to Georgia Criterion Referenced Competency Test (CRCT) science pass percentage scores of middle grade students in the English to Speakers of Other Languages (ESOL) program who were not administered benchmark assessments.

$H_0$3: There is no significant difference in Georgia Criterion Referenced Competency Test (CRCT) science pass percentage scores of middle grade students in the Special Education (SPED) program who were administered benchmark assessments compared to Georgia Criterion Referenced Competency Test (CRCT) science pass percentage scores of middle grade students in the Special Education (SPED) program who were not administered benchmark assessments.

Descriptive Statistics

This quantitative study evaluated the possible effects that benchmark assessments could have on middle grades science CRCT pass percentages by whole group, ESOL enrollment, and SPED enrollment. Descriptive statistics were calculated to summarize the sample, based upon the experimental and control groups.

The data in this study was collected from The Governor's Office of Student Achievement. The data consisted of CRCT pass percentages for 15 middle schools that administered benchmark assessments during the 2011-2012 school year and 15 middle schools that did not administer benchmark assessments during the 2011-2012 school year. The number of students tested was also collected to determine how many data points contributed to the sample size. After data collection, the researcher found the number of students contributing to the sample size to be quite larger than originally anticipated, because the researcher had to stretch beyond rural areas and gather data from schools located in the urban areas.

The experimental group consisted of 15 middle schools that administered benchmarks and had 15,320 data points contributing to the sample size. The control group consisted of 15 middle schools that did not administer benchmark assessments and had 11,557 data points contributing to the sample size. The total sample size for the entire study was 30 middle schools which consisted of 26,877 data points. Data from two subgroups, ESOL and SPED, was also collected. The ESOL experimental group had 1,512 data points contributing to the sample, and the ESOL control group had 651 data points contributing to the sample. The SPED experimental group consisted of 1,321 data points, and the SPED control group consisted of 1,246 data points.

In Figure 2, the mean for the experimental whole group consisted of 1021.33 students as compared to the control whole group of 770.45 students. The mean for the experimental SPED group consisted of 88.06 students as compared to the control SPED group of 83.06 students. The mean for the experimental ESOL group consisted of 100.8 students as compared to the control ESOL group of 43.4 students.

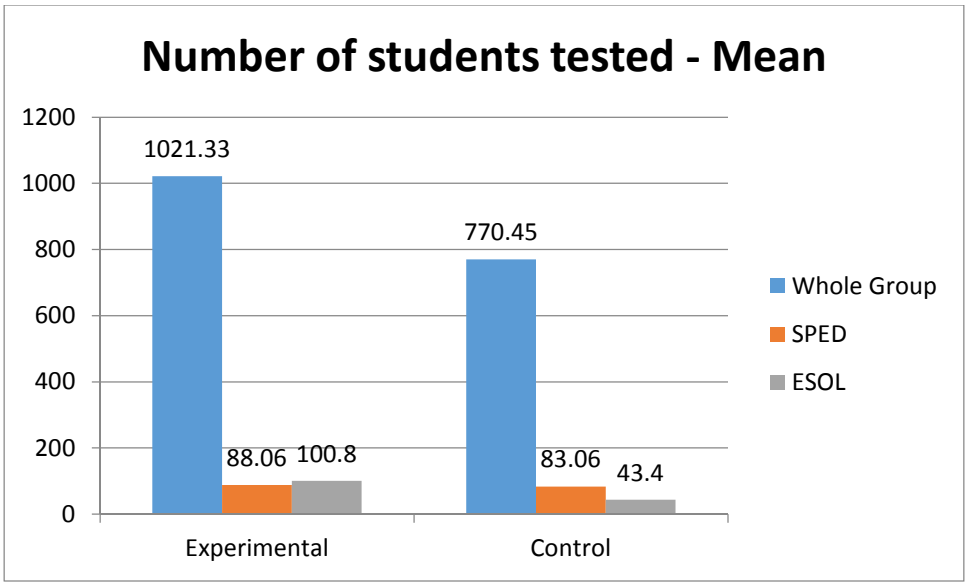

Figure 2. Number of students tested.

The large number of data points contributing to the sample size increased the reliability of the study because it allowed for more student scores to be contributed to the mean. Specific numbers of data points for each of the schools in the sample size can be found in Appendix B. Figure 2 shows the mean number of students tested for each group.

The sample consisted of eighth grade science CRCT pass percentage mean scores. Table 1 shows the mean, median, and standard deviation for the experimental whole, ESOL, and SPED groups. The mean score for the experimental whole group was 64.35 (SD = 12.64). The mean score for the experimental ESOL group was 41.67 (SD = 7.63). The mean score for the experimental SPED group was 30.29 (SD = 12.71).

Table 1 also shows the mean, median, and standard deviation for the control whole, ESOL, and SPED groups. The mean score for the control whole group was 72.21 (SD = 10.61). The mean score for the control ESOL group was 54.15 (SD = 16.46). The mean score for the control SPED group was 37.49 (SD = 9.88).

Table 1

Descriptive Statistics for Whole Group, ESOL, and SPED

	Experimental			Control		
Variable	*M*	*Median*	*SD*	*M*	*Median*	*SD*
Whole	64.35	63	12.64	72.21	75.5	10.61
ESOL	41.67	41.6	7.63	54.15	54	16.46
SPED	30.29	27	12.71	37.49	38.9	9.88

Results

To determine if there was a significant difference between CRCT pass percentages of middle schools that administered benchmark assessments and CRCT pass percentages of middle schools that did not administer benchmark assessments, data was imported from an Excel spreadsheet into SPSS 22.0 for data analysis. This result section summarizes the data analysis results for each research question and determine if each null hypothesis was rejected or accepted.

Research Question One and Null Hypothesis One

An independent sample *t* test was conducted to assess if there were differences in the whole group (benchmark vs. non-benchmark). Prior to analysis, the assumption of normality was assessed using a Shapiro-Wilk test. The result of the test was not significant, $p = .085$, validating the assumption of normality. The assumption of equality of variance was assessed using Levene's test. The result of the test was not significant, $p = .260$, indicating the assumption of equality of variance was met (Green & Salkind, 2011).

The results of the independent sample *t* test were not significant, $t(28) = -1.85$, $p = .076$, suggesting that there was not a difference in the whole group. Based on the statistical results, null hypothesis one was accepted. Results of the independent sample *t* test are presented in Table 2.

Table 2

Independent Sample t test for Whole Group

				Benchmark		Non Benchmark	
Variable	*t*(28)	*p*	Cohen's *d*	*M*	*SD*	*M*	*SD*
Meets	-1.85	.076	0.67	64.35	12.64	72.21	10.61

Figure 3 shows the averages of the whole group. The mean score for benchmark whole group was 64.35. The mean score for the non-benchmark whole group was 72.21. The non-benchmark whole group had a mean score 7.86 percentage points higher than the benchmark group.

Research Question Two and Null Hypothesis Two

An independent sample *t* test was conducted to assess if there were differences in the ESOL group (benchmark vs. non-benchmark). Prior to analysis, the assumption of normality was assessed using a Shapiro-Wilk test. The result of the test was not significant, $p = .059$, validating the assumption of normality. The assumption of equality of variance was assessed using Levene's test. The result of the test was significant, $p = .018$, violating the assumption of equality of variance; therefore, the Welch *t*-statistic, which does not assume equality of variance was used (Stevens, 1999).

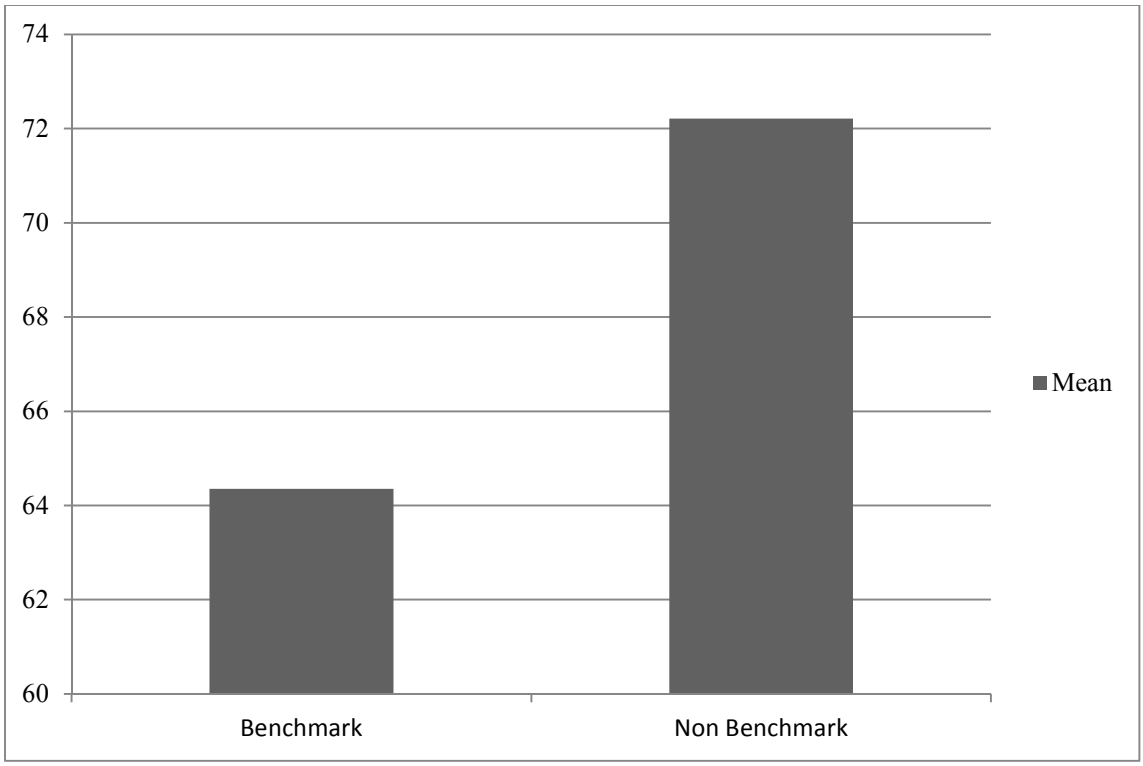

Figure 3. Whole group mean.

The results of the independent sample *t* test were significant, $t(20) = -2.66$, $p = .015$, suggesting that there was a difference in the ESOL group. Schools that administered benchmarks had a significantly lower mean than schools that did not administer benchmarks. According to Cohen (1988), the difference between the two groups was a large effect size. Based on the statistical results, null hypothesis two was rejected. Results of the independent sample *t* test are presented in Table 3.

Table 3

Independent Sample t test for ESOL Group

				Benchmark		Non Benchmark	
Variable	*t*(20)	*p*	Cohen's *d*	*M*	*SD*	*M*	*SD*
Meets	-2.66	.013	0.97	41.67	7.63	54.15	16.46

Figure 4 shows the averages of the ESOL group. The mean score for the benchmark ESOL group was 41.67. The mean score for the non-benchmark ESOL group was 54.15. The non-benchmark ESOL group had a mean score 12.48 percentage points higher than the benchmark group.

Research Question Three and Null Hypothesis Three

An independent sample *t* test was conducted to assess if there were differences in the SPED group (benchmark vs. non-benchmark). Prior to analysis, the assumption of normality was assessed using a Shapiro-Wilk test. The result of the test was not significant, $p = .136$, validating the assumption of normality. The assumption of equality of variance was assessed using Levene's test. The result of the test was not significant, $p = .084$, indicating the assumption of equality of variance was met (Green & Salkind, 2011).

The results of the independent sample *t* test were not significant, $t(28) = -1.73$, $p = .094$, suggesting that there was not a difference in the SPED group. Based on statistical results, null hypothesis three was accepted. Results of the independent sample *t* test are presented in Table 4.

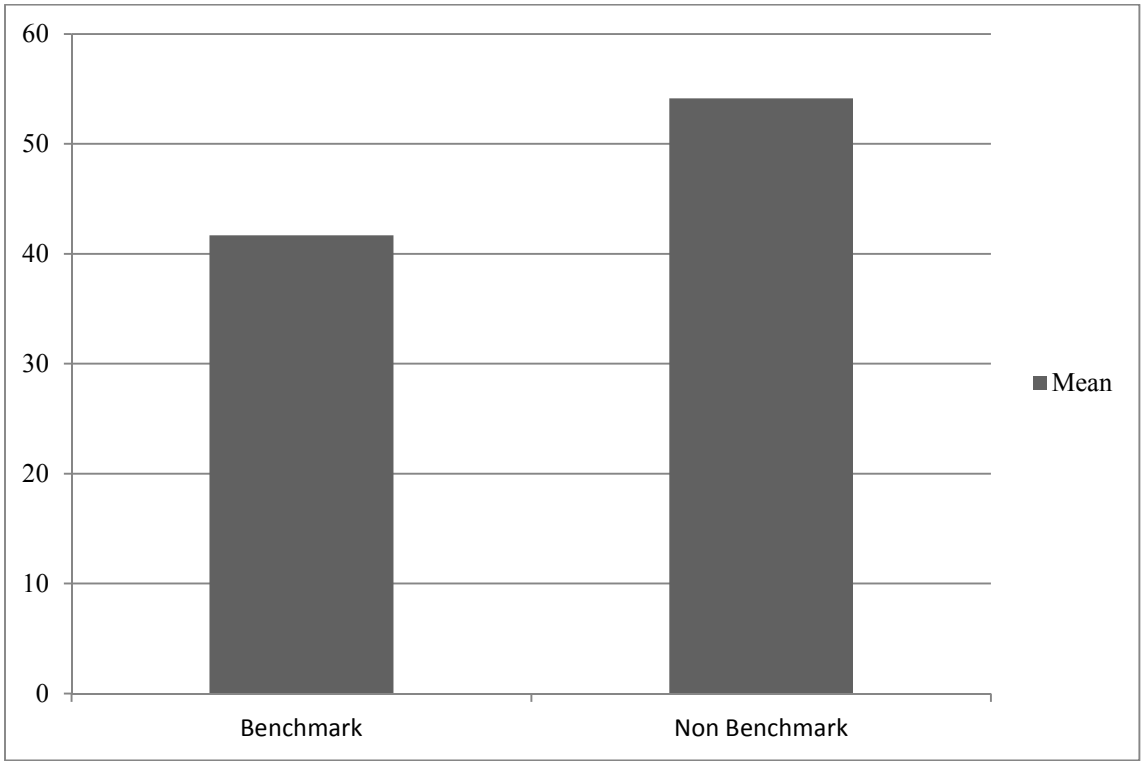

Figure 4. ESOL Group Mean.

Figure 5 shows the averages of the SPED group. The mean score for the benchmark SPED group was 30.29. The mean score for the non-benchmark SPED group was 37.49. The non-benchmark SPED group had a mean score 7.20 percentage points higher than the benchmark group.

Based upon the statistical analysis the researcher rejected the null hypothesis for research question two, as there was a statistically significant difference in middle grades ESOL science CRCT pass percentages when comparing schools that administered benchmark assessments and schools that did not administer benchmark assessments. The statistical results showed the mean of ESOL CRCT pass percentage among middle schools that did not administer benchmark assessment was significantly higher at 54.15 compared to schools that did administer benchmark assessments. The mean for the schools that did administer benchmark assessments was 41.67.

Table 4

Independent Sample t test for SPED Group

				Benchmark		Non Benchmark	
Variable	*t*(28)	*p*	Cohen's *d*	*M*	*SD*	*M*	*SD*
Percent	-1.73	.094	0.63	30.29	12.71	37.49	9.88

Descriptive statistics were computed for the sample for each group and subgroup. A statistical analysis of the data collected was conducted using a *t* test to test for statistically significant difference between 2012 middle grades science CRCT pass percentages of schools that administered benchmark assessments and schools that did not administer benchmark assessments (Fraenkel, Wallen, & Hyun, 2015).

Summary

Based on the *t* test of research question one, the null hypotheses for the whole group was not rejected. The statistical analysis of research question one showed $p = 0.076$. The statistical analysis indicated that there was not a statistically significant difference in the whole group between middle schools that administered benchmark assessments and middle schools that did not administer benchmark assessments (Fraenkel et al., 2015).

Based on the *t* test of research question two, the null hypothesis for the ESOL group was rejected. The statistical analysis indicated a statistically significant difference having a p value $< .05$ (Fraenkel et al., 2015). The statistical analysis of research question two showed $p = .013$. This significant statistical difference was a result of a large difference between the means of CRCT pass percentages among ESOL students between middle schools that administered benchmark assessments and middle schools that did not administer benchmark assessments. The means of the ESOL group can be found in Figure 4.

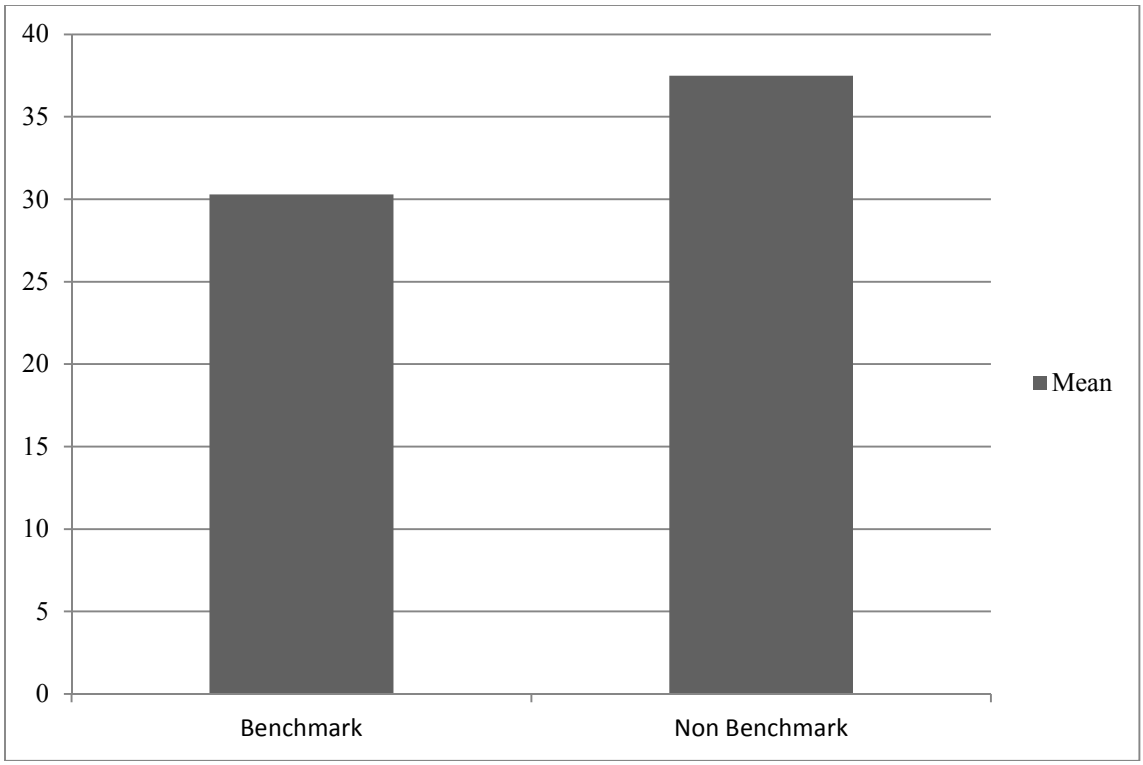

Figure 5. SPED group.

Based on the *t* test of research question three, the null hypotheses for the SPED group was not rejected. The statistical analysis of research question one showed $p = 0.094$. The statistical analysis indicated that there was not a statistically significant difference in the SPED group between middle schools that administered benchmark assessments and middle schools that did not administer benchmark assessments (Fraenkel et al., 2015).

CHAPTER FIVE: DISCUSSION, CONCLUSIONS, AND RECOMMENDATIONS

Discussion

In the last decade, teachers across America have seen a shift in education. Holding teachers and schools accountable for student performance on standardized testing has now become the norm. Because of these higher stakes in accountability, schools scrambled to develop new ideas and methods to predict how students were going to score on these standardized tests. One such method was the use of benchmark assessments.

The purpose of this causal-comparative study was to determine if benchmark assessments administered to middle grades science students had any impact on middle grades science Georgia CRCT pass percentages. To determine this impact, middle grades science CRCT pass percentages of middle schools that administered benchmark assessments were compared to middle grades science CRCT pass percentages of middle schools that did not administer benchmark assessments among the whole group, ESOL students, and SPED students.

Chapter Five provides a summary and discussion of the research findings, the implications of the study in terms of relevant literature and methodology, the study's limitations, and recommendations for future research.

This quantitative study evaluated the possible effects benchmark assessments had on middle grades science CRCT pass percentages by specifically comparing middle schools that administered benchmark assessments and middle schools that did not administer benchmark assessments among the whole group, ESOL students, and SPED students. A *t*-test was used to determine if there was a statistically significant difference between middle grades science CRCT pass percentages of schools who administered benchmark assessments compared to those schools who did not administer benchmark assessments.

Research Hypothesis One

$H_0$1: There is no significant difference in Georgia Criterion Referenced Competency Test (CRCT) science pass percentage scores of middle grade students who were administered benchmark assessments compared to Georgia Criterion Referenced Competency Test (CRCT) science pass percentage scores of middle grade students who were not administered benchmark assessments.

The *t*-test revealed that there was not a statistically significant difference in Georgia CRCT science pass percentage scores of middle grades students who were administered benchmark assessments compared to Georgia CRCT science pass percentage scores of middle grades students who were not administered benchmark assessments, $t(28) = -1.85$, $p = .076$ (Green & Salkind, 2011). Thus, the null hypothesis was not rejected. This means that on average students who were taking benchmark assessments were not scoring any higher on standardized tests than students who were not taking benchmark assessments.

Even though the results of hypothesis one are not statistically significant, means of both groups being compared can be viewed. The experimental group (benchmark) had a mean of 64.35, while the control group (non-benchmark) had a mean of 72.21. From this data, it was determined that schools not administering benchmark assessment had a higher pass percentage mean than schools that did administer benchmark assessments. These differences are not significant, but it can be discussed why there is a discrepancy between the two. These results can be seen in Figure 6 below.

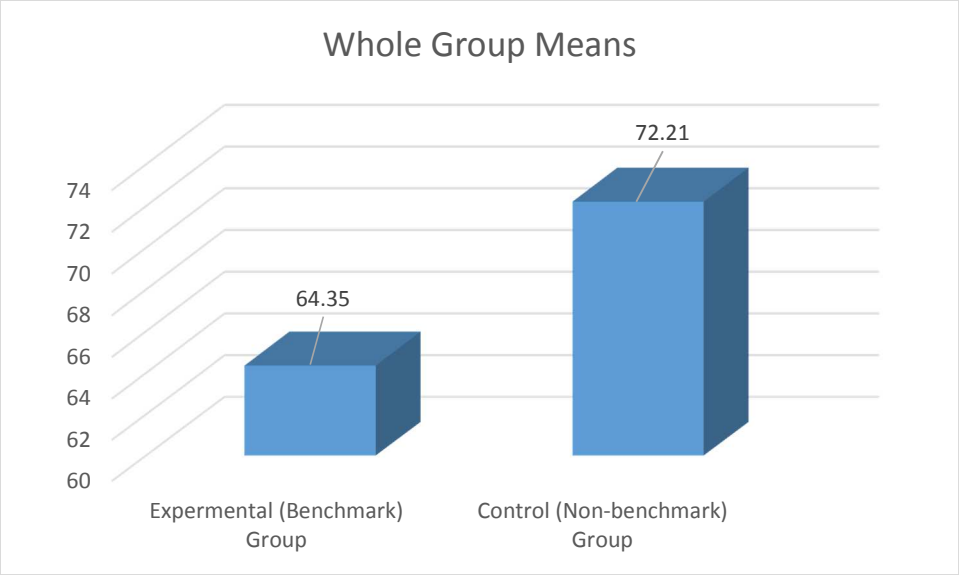

Figure 6. Whole group (Experimental vs. Control) means.

Mertler (2009) stated, "Assessing student performance is one of the most critical aspects of the job of a classroom teacher" (p. 101). Teachers are constantly assessing students to prepare for end of the year state standardized assessments. Benchmarks are assessments that students take on top of chapter and unit tests administered by the teacher throughout the year. At what point do educators decide that students are being assessed too much? Data that teachers gain from administering chapter and unit tests can easily take the place of benchmark data.

One factor which may have contributed to a higher mean in the non-benchmark group is the amount of instructional time taken away to administer these benchmark assessments. In Bancroft's (2010) study, teachers felt too much of their instructional time was spent on benchmark assessments. If teachers spent a day reviewing for each benchmark and a day administering each benchmark, teachers would have used eight instructional days during the year for benchmark assessments. That number could have been more depending on how much time the teachers dedicated to remediation once the data was analyzed.

Brindley (2001) expressed "doubts surrounding the validity of outcome statements and the reliability of the assessment tools that are used to elicit student performance" (p. 394). Brindley questioned whether or not these assessments were true measures of student performance. Based on the results of this study, the researcher believed that benchmark assessments are not an effective tool for monitoring whole group student progression through content standards and for predicting outcomes on state standardized assessments. According to the data, students who took benchmarks throughout the year had lower mean pass percentages compared to students who did not take benchmark assessments.

Research Hypothesis Two

$H_0$2: There is no significant difference in Georgia Criterion Referenced Competency Test (CRCT) science pass percentage scores of middle grade students in the English to Speakers of Other Languages (ESOL) program who were administered benchmark assessments compared to Georgia Criterion Referenced Competency Test (CRCT) science pass percentage scores of middle grade students in the English to Speakers of Other Languages (ESOL) program who were not administered benchmark assessments.

The *t*-test revealed a statistically significant difference in Georgia CRCT science pass percentage scores of middle grade students in the ESOL program who were administered benchmark assessments compared to Georgia CRCT science pass percentage scores of middle grades students in the ESOL program who were not administered benchmark assessments was found, $t(20) = -2.66$, $p = .015$ (Green & Salkind, 2011). Thus, the null hypothesis was rejected.

There are many factors contributing to the rejection of null hypothesis two. According to the information-processing theory, students should have been able to use knowledge from benchmark assessments to aid in answering CRCT questions. However, with the ESOL group

this theory did not work. One explanation for the lack of success can be explained by examining the language skills of the ESOL students (Maerten-Rivera et al., 2010). ESOL students are learning in environments where their native language is not being spoken. Plus, science vocabulary is often difficult for native English speaking students, so students who have a native language other than English are going to struggle as much or more than native English speakers. Allen and Park (2011) stated, "The language used by many ESL students is conversational English, while science classes require fluency in academic English" (p. 29). Thus, making success on science achievement tests very difficult for ESL students.

Maerten-Rivera et al. (2010) examined the emerging role ELLs were playing in standardized testing. Maerten-Rivera et al. found a gap in student achievement existed when comparing ESOL students to a whole group population. Maerten-Rivera et al. stated, "ELL students have traditionally been excluded from content area instruction, including science, due to the perceived urgency of developing basic literacy and numeracy" (p. 938). This gap in ESOL achievement could be attributed to the fact that ESOL students have not received as much science instruction as native English speaking students throughout their school career.

Figure 7 shows the differences between the means of the ESOL group and the whole group. The mean score for the experimental (benchmark) whole group was 64.35 compared to the experimental ESOL group mean score of 41.67 with a 22.68 percentage point gap. The mean score for the control (non-benchmark) whole group was 72.21 compared to the control ESOL group mean score of 54.15 with an 18.06 percentage point gap.

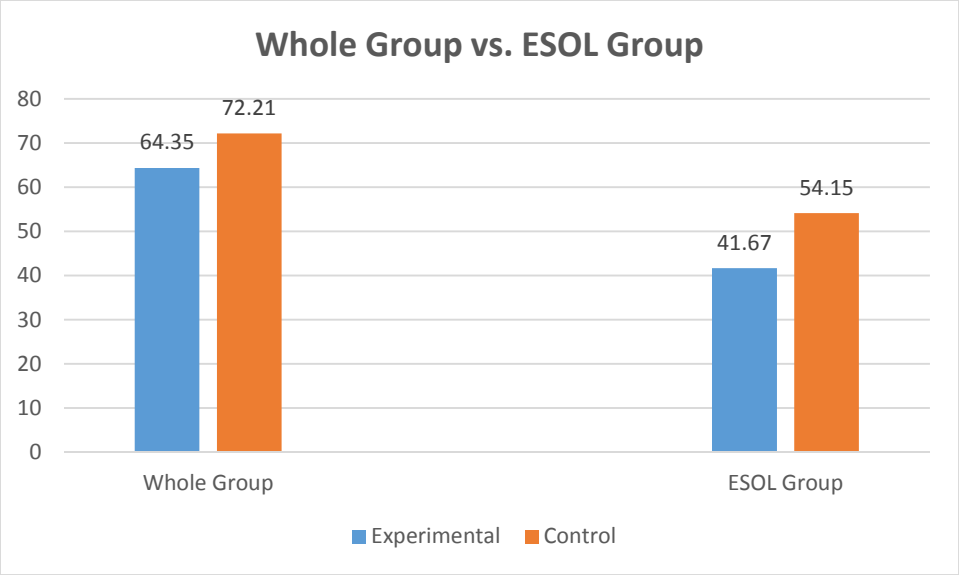

Figure 7. Whole group vs. ESOL group.

Language acquisition and proficiency of the ESOL student could help explain why there is such a large discrepancy between the two groups. Bunch, Shaw, and Geaney (2010) stated the following:

> A deeper understanding of the language demands facing students from language minority backgrounds can inform educators and researchers in their efforts to envision the supports needed for students to learn and demonstrate what they have learned and to capitalize on potential language development opportunities in content-area classrooms. (p. 185)

In Figure 4, there is also a discrepancy between the experimental ESOL group and control ESOL group. The control ESOL group is 12.48 percentage points higher than the experimental ESOL group. According to this data, ESOL students who were not administered benchmark assessments had a higher pass percentage on the eighth grade science CRCT when compared to ESOL students who were administered benchmark assessments.

Research Hypothesis Three

$H_0$3: There is no significant difference in Georgia Criterion Referenced Competency Test (CRCT) science pass percentage scores of middle grade students in the Special Education (SPED) program who were administered benchmark assessments compared to Georgia Criterion Referenced Competency Test (CRCT) science pass percentage scores of middle grade students in the Special Education (SPED) program who were not administered benchmark assessments.

The *t* test revealed that there is no statistically significant difference in Georgia CRCT science pass percentage scores of middle grades students in the SPED program who were administered benchmark assessments compared to Georgia CRCT science pass percentage scores of middle grade students in the SPED program who were not administered benchmark assessments was found, $t(28) = -1.73$, $p = .094$ (Green & Salkind, 2011). Thus, the null hypothesis was not rejected.

Even though the results of hypothesis three are not statistically significant, the means of both groups can be compared. The experimental group (benchmark) had a mean of 30.29, while the control group (non-benchmark) had a mean of 37.49. From this data, it was determined that schools not administering benchmark assessments had a higher pass percentage mean than schools that did administer benchmark assessments. These difference are not significant, but the discrepancy between the two can be discussed. These results can be seen in Figure 8 below. Swanson (1987) examined the role that information-process theory played the in the education of a learning disabled child. Swanson indicated that, "Learning disabled children may be viewed as failing to assemble, adapt, alternate, assess, and abandon certain cognitive programs in the process of performing a task relatively simple to nondisabled children" (p. 4). This explains why

the SPED group did not have a significant difference and why the non-benchmark group was higher than the benchmark group.

Students with learning disabilities may not be cognitively developed enough to process the question on the CRCT while accessing and interpreting their prior knowledge to develop an answer to a question on the CRCT. This is because students with disabilities face challenges that include processing deficits in the areas of response inhibition, sustained attention, metacognition, auditory and visual processing, long term memory, short term memory, working memory, verbal reasoning, nonverbal reasoning and abstract reasoning. Any single one of the areas could significantly impact a student's ability to process the questions on the CRCT and determine the correct response (Fayette County Schools, 2012).

Wiliam (2010) stated, "It is only through assessment that we can find out whether instruction has had its intended effect, because even the best-designed instruction cannot be guaranteed to be effective" (p. 107). In this study, the researcher found that when each research question was tested, the control group for each question had a higher CRCT pass percentage mean. This means that in terms of the CRCT, instruction was more effective in classrooms where benchmark assessments were not administered. Figure 9 shows the CRCT pass percentage means for the entire study.

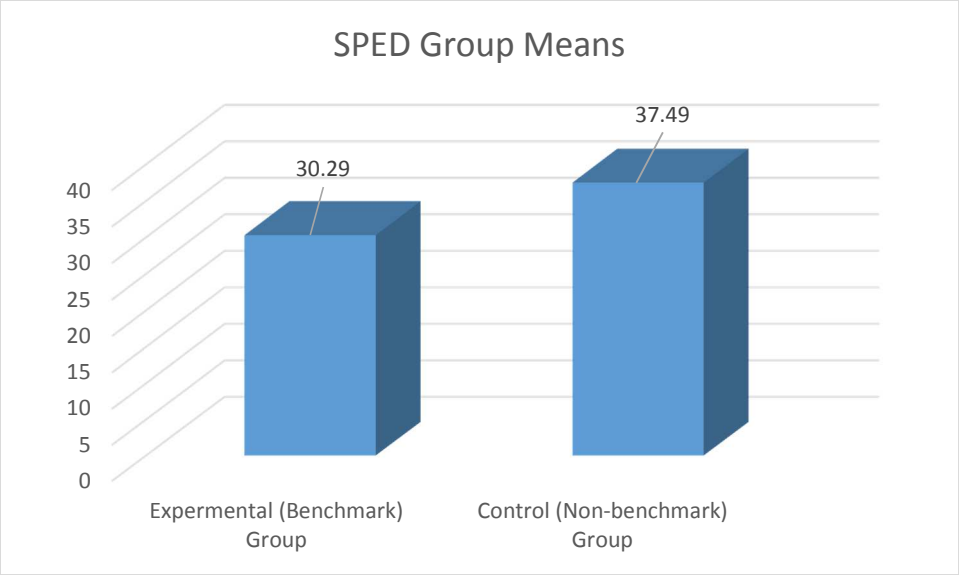

Figure 8. SPED group means.

Conclusions

This study was important to the education community as a whole because the study fills the gap that is lacking in available research concerning the effects of benchmark assessments. In the review of literature, the bulk of the studies reviewed focused on advantages (Hamilton et al., 2009; Helman, 2005; Henderson et al., 2007; Honawar, 2006) or disadvantages (Bancroft, 2010; Hall, 2002; Olson, 2005a), the importance of implementing benchmark assessments properly (Britton, 2010; Buck & Trauth-Nare, 2009; Feldman, 2010; Halverson, 2010; Means et al., 2009; Sharkey & Murnane, 2006), or how the data from benchmarks was being used (Christman et al., 2009; Coffey, 2009; Matthew, Trimble, and Gay, 2007; Herman & Baker, 2005; Marshall et al., 2009; Olah et al., 2010; Zehr, 2006).

Sharkey and Murnane (2006) discussed how one district faced challenges in deciding to implement assessments using a purchased assessment system. Means et al. (2009) discussed how to implement benchmarks assessments successfully. Means et al. suggested that data obtained should drive classroom instruction. Towndrow et al. (2008) discussed the problems

faced with mandated assessment. What all of this literature failed to discuss is whether or not these implemented, data-driven, mandated assessments are actually working.

Educators need to know if the time they are spending administering benchmark assessments is actually valued time. One way to determine this is to find out how teachers are using data they have gained from benchmark assessments. Are the teachers administering these assessments remediating students based on student need determined by the benchmark? In order for a benchmark to serve its purpose, students need to be remediated based on their results.

The literature reviewed in this study also suggested that proper feedback and the appropriate use of assessment systems will lead to improvements in test scores. Halverson (2010) discussed the use of a feedback system to drive instruction:

> Because school staff cannot rely on standardized test results to directly inform changes in their classroom-level practices, schools must also engage in instructional system redesign – first to link everyday classroom practices with school wide outcomes, and second to develop data-driven practices that give teachers local, ongoing information to benchmark student learning process. (p. 130)

Giving students appropriate feedback on benchmarks is important and detrimental to their success because it allows students to learn from their mistakes and their victories; however, what the study does not explain is if using appropriate feedback systems translates to any improvements on standardized test scores. This research is important because it fills the gap of determining if benchmarks are making a difference on standardized tests.

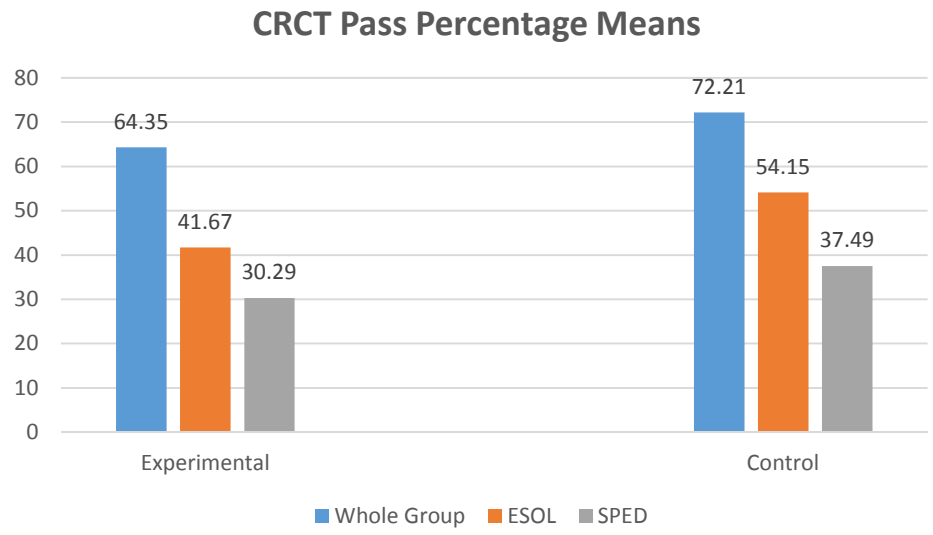

Figure 9. CRCT pass percentage means.

Olson (2005a) discussed benchmarks as a "bandwagon" that districts were jumping on to address concerns of higher stakes and accountability. She also argued that there is data lacking to validate the effectiveness of benchmark assessments. This research answers that question of validation. Based on the data analysis comparing middle grades science CRCT percentage scores of schools who administered benchmark assessments to middle grades CRCT percentage scores of schools who did not administer benchmark assessments, benchmark assessments did not have any bearing on increased standardized test scores. In fact, the schools that did not administer benchmark assessments had a higher percentage mean across the whole group and the two subgroups of ESOL and SPED.

Only one study was found to show what effect benchmarks actually have on standardized tests. Bancroft (2010) studied how benchmark assessments impacted reading achievement scores among high school students. Bancroft's study took place over three years in a low income high school. Bancroft found that students who read below grade level did show improvements after the administration of the benchmarks. Teachers in Bancroft's study also expressed

concerns about wasting too much instructional time on the benchmark process. The results of this study corroborate Bancroft's finding because the schools that did administer benchmark assessments had lower percentage means on the middle grades science CRCT compared to the schools that did not administer benchmark assessments.

Implications

In order for students to transfer and apply knowledge from benchmark assessments to state standardized tests, the teacher plays an active role in helping the student make the connections. According to Piaget's formal operational period, middle school age students are able to think abstractly allowing them to connect current content to their prior knowledge (Siegler & Ellis, 1996). However, students still need assistance from the teacher to make connections within the content being taught. It is essential for teachers to take the information/content from the benchmarks to remediate and help students master the standards. If the connections are not made from the benchmark back to the content, the benchmark assessments become a meaningless assessment to students.

Teachers are obligated to afford students every opportunity to learn. "Meaningful learning occurs when learners are actively involved and have the opportunity to take control of their own learning" (Assessment and Reporting Unit et al., 2005, p. 2). Not allowing students to take ownership of their learning by using higher level thinking skills and relying solely on the benchmark could possibly explain why the benchmark groups had a lower mean.

The results of this study can also be directly tied back to the information-processing theory (Miller, 2011). The information-processing theory explains how knowledge is synthesized by the brain. It relies on students being mature enough to read a question and be able to access different levels of prior knowledge to determine an answer for the questions.

Middle school students lack the maturity level that allows them to fully make connections between prior knowledge and assessment questions. Thus, when taking the science CRCT, students are not making the appropriate connections back to content on benchmark assessments.

The results of this study indicated the CRCT percentage means for schools where benchmarks were not administered was higher than CRCT percentage means for schools where benchmarks were administered. One such explanation for this is time management. When administering a benchmark assessment, teachers take at least two days away from instruction every time, which is approximately eight days of instruction per year. This approximation can be higher if teachers use the data properly to reteach key concepts missed on the benchmark.

Brindley (2001) suggested that benchmark assessments were not a true measure of student performance. This study took Brindley's research a step further by determining if the benchmark assessments did or did not impact student performance on state mandated assessments. The researcher predicted that benchmark assessments would not positively impact middle school science CRCT pass percentages, and according to the data collected and tested in this study, the researcher's predictions were true.

Limitations

This study examined whether or not benchmark assessments had an impact on middle grades science CRCT pass percentages. In this section, the researcher will discuss the limitations that were present in the study.

The data in this study was collected from The Governor's Office of Student Achievement. The researcher only collected numerical data relating to sample size and CRCT pass percentages for whole schools and the subgroups of ESOL and SPED students. One limitation for the study concerns what schools did or did not do with data gained from

benchmark assessments. The researcher only knew which schools administered benchmark assessments. It is not known what was done with data after the benchmarks were administered. The research did not have any knowledge of what teachers were doing with the data obtained from benchmark assessments. In order for benchmark assessments to beneficial teachers must take the data and drive future instruction. Remediation should occur for weak standards and enrichment should occur for strong standards. Also, the researcher had to trust when asking if specific schools administered benchmarks that administrators were telling the truth.

Recommendations for Future Research

Buck and Trauth-Nare (2009) stated, "Assessments should serve to enrich students' understanding of science and not simply measure attainment of content knowledge" (p. 475). Since schools that administered benchmark assessments had lower CRCT pass percentage means, the researcher concluded that benchmark assessments are not truly assessing the knowledge of students. This researcher recommends that future research concerning how benchmark assessments are specifically aligned to content objectives and standards and determining what can be done to make benchmark assessment more effective so that the assessments fulfill the intended purpose of positively impacting state standardized test scores would be beneficial. If benchmark assessments are not impacting CRCT scores, then educators must take a step back and examine the constructs of the benchmark.

The researcher also recommends future research to determine if teachers who are using data from benchmarks to drive instruction have increased benchmark scores compared to teachers who only administer benchmark assessments and do no use the data for future instruction.

It is also recommended that research be done to determine if a correlation exists between the total numbers of assessments a student takes in a year and scores on state standardized assessments.

The researcher also recommends that research be done to determine if a correlation exists between the types of benchmark assessments administered and state standardized test scores. Teacher created tests may be less likely to show improvements on state standardized assessments because teachers truly do not know what is on the state standardized assessments. Whereas, state created benchmark assessments would be better correlated to the actual end of the year assessment.

Another recommendation for future research would be to determine what factors are the most beneficial for ESOL student success on state standardized assessments. A future researcher could look at whether or not an ESOL student's Assessing Comprehension and Communication in English State to State (ACCESS) level score affects CRCT scores. Research could also be done to determine if an ESOL student's level of English proficiency impacts their CRCT scores. This could be expanded by specifically looking at each individual ACCESS level and determining if benchmark assessments benefit any of the ACCESS levels.

To extend this study, qualitative research could also be conducted to determine if administrators, teachers, students, and parents felt there was any validity in benchmark assessments. This research could be conducted through surveys and interviews with various groups of people.

The results of this study revealed that benchmark assessments do not necessarily have a positive effect on CRCT scores. In fact, analysis of data revealed that CRCT pass percentage mean scores for each group were actually higher among schools that did not administer

benchmark assessments. These results are very significant to the science education community because the results show that spending less time assessing and more time teaching yields better standardized test scores. This study fills in the gaps left by the literature that is available for benchmark testing by demonstrating that the administration of local benchmark assessments at the study sites do not improve test scores on state mandated assessments.

REFERENCES

Allen, H., & Park, S. (2011). Science education and ESL students. *Science Scope*, *35*(3), 29.

Assessment and Reporting Unit, Learning Policies Branch, & Office of Learning and Teaching. (2005). *Current perspectives on assessment*. Retrieved from https://www.eduweb.vic.gov.au/edulibrary/public/teachlearn/student/assessment_current_per.pdf

Bancroft, K. (2010). Implementing the mandate: The limitations of benchmark tests. *Educational Assessment Evaluation & Accountability*, *22*, 53-72.

Brindley, G. (2001). Outcomes-based assessment in practice: Some examples and emerging insights. *Language Testing*, *18*, 393-407.

Britton, T. (2011). Using formative and alternative assessments to support instruction and measure student learning. *Science Scope*, *34*(5), 16-21.

Brown, R., & Coughlin, E. (2007). The predictive validity of selected benchmark testing. *Regional Education Laboratory*, (17). Retrieved from http://www.ctb.com/media/articles/pdfs/resources/PredictiveValidity.pdf

Buck, G. A., & Trauth-Nare, A. E. (2009). Preparing teachers to make the formative assessment process integral to science teaching and learning. *Science Teacher Education*, *20*, 475-494.

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

Bunch, G., Shaw, J., & Geaney, E. (2010). Documenting the language demands of mainstream content-area assessment for English learners: Participant structures, communicative

modes and genre in science performance assessments. *Language and Education, 24*(3), 185-214.

Christman, J. B., Neild, R. C., Bulkley, K., Blanc, S., Liu, R., Mitchell, C., & Travers, E. (2009). Making the most of interim assessment data. *Research for Action*. Retrieved from http://www.researchforaction.org/publication/details/558

Coffey, H. (2009). *Benchmark assessments*. Retrieved from http://www.learnnc.org/lp/pages5317

Cohen, J. (1988). *Statistical power analysis for the behavior sciences* (2nd ed.). St. Paul, MN: West Publishing Company.

Cowie, B., Jones, A., & Otrel-Cass, K. (2011). Re-engaging students in science: Issues of assessment, funds of knowledge, and sites for learning. *International Journal of Science and Mathematics Education*, *9*, 347-366.

Fayette County Schools. (2012). Processing deficits. *Fayette County School System Department of Exceptional Children's Services and School Psychological Services, 1-14.*

Feldman, E. (2010). Benchmarks curricular planning and assessment framework: Utilizing standards without introducing standardization. *Early Childhood Education Journal*, *38*, 233-242.

Fraenkel, J., Wallen, N., & Hyun, H. (2015). *How to design and evaluate research in education* (9th ed.). New York: McGraw-Hill.

Gall, M., Gall, J., & Borg, W. (2007). *Educational research: An introduction* (8th ed.). New York: Logman.

Georgia Department of Education, Office of Accountability. (2013). *Adequate yearly progress (AYP)*. Retrieved from http://www.gadoe.org/ayp/Pages/default.aspx

Georgia Department of Education, Office of Assessment. (2013). *Criterion-referenced competency test (CRCT)*. Retrieved from http://www.gadoe.org/Curriculum-Instruction-and-Assessment/Assessment/Pages/CRCT.aspx

Georgia Department of Education, Office of Curriculum and Instruction. (2013). *Common core Georgia performance standards (CCGPS)*. Retrieved from http://www.gadoe.org/Curriculum-Instruction-and-Assessment/Curriculum-and-Instruction/Pages/CCGPS.aspx

Georgia Department of Education, Office of Curriculum, Instruction, Assessment, & Accountability. (2013). *College and career ready performance index (CCRPI)*. Retrieved from http://www.gadoe.org/CCRPI/Pages/default.aspx

Governor's Office of Student Achievement: K-12 Public Schools Report Card. (n.d.). Retrieved from https://gaawards.gosa.ga.gov/analyitcs/saw.dll?dashboard

Green, S., & Salkind, N. (2011). *Using SPSS for Windows and Macintosh: Analyzing and understanding data* (6th ed.). Upper Saddle River, NJ: Pearson/Prentice Hall.

Guisbond, L., & Neill, M. (2004). Failing our children: No Child Left Behind undermines quality and equity in education. *The Clearing House*, *78*(1), 12-16.

Hall, L. (2002). Social studies standards, benchmarks, and assessments: An analysis of an eighth-grade exam. *The Social Studies*, 213-217.

Halverson, R. (2010). School formative feedback systems. *Peabody Journal of Education*, *85*, 130-146.

Hamilton, L., Halverson, R., Jackson, S., Mandinach, E., Supovitz, J., & Wayman, J. (2009). Using student achievement data to support instructional decision making. *National*

Center for Education Evaluation. Retrieved from http://ies.ed.gov/ncee/wwc/pdf/practiceguides/dddm_pg_092909.pdf

Helman, L. A. (2005). Using literacy assessment results to improve teaching for English-language learners. *The Reading Teacher*, *58*(7).

Henderson, S., Petrosino, A., Gukenburg, S., & Hamilton, S. (2007). Measuring how benchmark assessments affect student achievement. *National Center for Education Evaluation*. Retrieved from http://ies.ed.gov/ncee/edlabs/regions/northeast/pdf/REL_2007039

Herman, J. L., & Baker, E. L. (2005). Making benchmark: Six criteria can help educators use benchmark tests to judge student skills and to target areas for improvement. *Educational Leadership*.

Honawar, V. (2006). Tip of their fingers. *Education Week*, *25*(35).

Liu, O., Lee, H., & Linn, M. (2010). Multifaceted assessment of inquiry based science learning. *Educational Assessment*, *15*, 69-86.

Maerten-Rivera, J., Myers, N., Lee, O., & Penfield, R. (2010). Student and school predictors of high stakes assessment in science. *Science Education*, *94*(6), 937-962.

Marshall, J., Horton, R., & White, C. (2009). Equipping teachers. *The Science Teacher.*

Matthews, J., Trimble S., & Gay, A. (2007). But what do you do with the data? *Education Digest: Essential Reading Condensed for Quick Review, 73*(3), 53-56.

Means, B., Padilla, C., DeBarger, A., & Bakia, M. (2009). Implementing data-informed decision making in schools--teacher access, supports, and use. *U.S. Department of Education*, Retrieved from http://www.eric.ed.gov/ERICDocs/data/ericdocs2sql/content_storage_01/0000019b/80/43/48/3e.pdf

Mertler, C. (2009). Teachers' assessment knowledge and their perceptions of the impact of classroom assessment professional development. *SAGE Publications, 12*(2), 101-113.

Miller, P. (2011). *Theories of developmental psychology* (5th ed.). New York, NY: Worth Publishers.

Olah, L., Lawrence, N., & Riggan, M. (2010). Learning to learn from benchmark assessment data: How teachers analyze results. *Peabody Journal of Education*, *85*, 226-245.

Olson, L. (2005a). Benchmark assessments offer regular achievement. *Education Week*, 25(13).

Olson, L. (2005b). Not all teachers keen on periodic tests. *Education Week*, *25*(13).

Rapport, A. (2012, February 10). The coming battle over NCLB exemptions. *The American Prospect*, *23*(1).

Reichrath, M., Georgia Department of Education, Office of Curriculum, Instruction, Assessment, & Accountability. (2013). *College and career ready performance index (CCRPI)*. Retrieved from http://www.gadoe.org/CCRPI/Pages/default.aspx

Rush, L. S., & Sherff, L. (2012). NCLB 10 years later. *English Education*, *44*(2), 91-101.

Sharkey, N., & Murnane, J. (2006). Tough choices in designing a formative assessment system. *American Journal of Education*, *112*(4), 572-588.

Siegler, R., & Ellis, S. (1996). Piaget on childhood. *Psychological Science*, *7*(4), 211-215.

Statistics Solutions. (2014). Statistics Solutions Pro (Version v1.15.02.16) [Online computer software]. Retrieved from http://ssp.statisticssolutions.com/

Stevens, J. (1999). *Intermediate statistics* (2nd ed.). Mahwah, NJ: Routledge Academic.

Swanson, H. (1987). Information processing theory and learning disabilities: An overview. *Journal of Learning Disabilities*, *20*(1), 3-7.

Towndrow, P., Tan, A., Yung, B., & Cohen, L. (2010). Science teachers' professional development and changes in science practical assessment practices: What are the issues? *Research in Science Education*, *40*, 117-132.

Walz, M. (2012, February 09). Georgia receives NCLB waiver. *GPB News*. Retrieved from http://www.gpb.org/news/2012/02/09/georgia-receives-nclb-waiver

Wiliam, D. (2010). Standardized testing and school accountability. *Educational Psychologist*, *45*(2), 107-122.

Zehr, M. A. (2006). Monthly checkup: A new principal works with teachers to get the most out of a district's benchmark assessments. *Education Week*, *25*(35), 36-37.

APPENDIX A: IRB Letter

LIBERTY UNIVERSITY
INSTITUTIONAL REVIEW BOARD

November 18, 2014

[redacted]

IRB Application 1834: A Causal-Comparative Study of the Effects of Benchmark Assessments on Eighth-Grade Science Achievement Scores

[redacted]

The Liberty University Institutional Review Board has reviewed your application in accordance with the Office for Human Research Protections (OHRP) and Food and Drug Administration (FDA) regulations and finds your study does not classify as human subjects research. This means you may begin your research with the data safeguarding methods mentioned in your approved application.

Your study does not classify as human subjects research because it does not involve obtaining private information about individuals.

Please note that this decision only applies to your current research application, and that any changes to your protocol must be reported to the Liberty IRB for verification of continued non-human subjects research status. You may report these changes by submitting a new application to the IRB and referencing the above IRB Application number.

If you have any questions about this determination, or need assistance in identifying whether possible changes to your protocol would change your application's status, please email us at irb@liberty.edu.

Sincerely,

[redacted]

Fernando Garzon, Psy.D.
Professor, IRB Chair
Counseling

(434) 592-4054
LIBERTY
UNIVERSITY.

APPENDIX B: Data Points for Groups

Table 5

Data Points for Groups

	# of Data Points					
	Control			Experimental		
School Code	Whole	ESOL	SPED	Whole	ESOL	SPED
1	736	14	103	-	-	-
2	673	53	80	-	-	-
3	1027	57	98	-	-	-
4	887	41	109	-	-	-
5	-	-	-	941	58	54
6	-	-	-	681	35	62
7	-	-	-	969	127	128
8	611	135	62	-	-	-
9	-	-	-	994	335	81
10	-	-	-	1419	132	114
11	1008	32	106	-	-	-
12	-	-	-	1124	20	89
13	-	-	-	958	20	105
14	-	-	-	944	11	79
15	555	59	21	-	-	-
16	842	62	95	-	-	-
17	-	-	-	1290	134	94
18	-	-	-	884	12	85
19	-	-	-	752	17	93
20	970	17	90	-	-	-
21	-	-	-	916	391	70
22	888	48	86	-	-	-
23	651	28	82	-	-	-
24	457	39	67	-	-	-
25	-	-	-	1046	11	74
26	-	-	-	1080	67	100
27	-	-	-	1322	142	93
28	462	44	39	-	-	-
29	937	12	116	-	-	-
30	853	10	92	-	-	-

APPENDIX C: CRCT Meets Percentages for Groups

Table 6

CRCT Meets Percentages for Groups

	Meets %					
	Control			Experimental		
School Code	Whole	ESOL	SPED	Whole	ESOL	SPED
1	78	57.1	48.5	-	-	-
2	85.4	81.1	50.1	-	-	-
3	78.5	71.9	36.7	-	-	-
4	66.9	36.5	31.2	-	-	-
5	-	-	-	83	46.6	46.3
6	-	-	-	82	40	37.1
7	-	-	-	67.1	44.8	35.1
8	69.9	54	27.4	-	-	-
9	-	-	-	51.9	25.1	22.2
10	-	-	-	79.3	43.2	42.1
11	65	46.9	26.5	-	-	-
12	-	-	-	59.5	40	21.3
13	-	-	-	47.3	45	15.2
14	-	-	-	55.3	45.5	19
15	81.3	57.6	47.7	-	-	-
16	76.5	58.1	38.9	-	-	-
17	-	-	-	80.6	35	53.2
18	-	-	-	57.5	41.6	14.1
19	-	-	-	44	58.8	21.5
20	61	47.1	36.1	-	-	-
21	-	-	-	68.3	49.3	38.5
22	77.6	68.7	40.7	-	-	-
23	72.2	53.6	45.1	-	-	-
24	72.8	35.9	41.8	-	-	-
25	-	-	-	63	36.4	17.6
26	-	-	-	58.3	35.8	27
27	-	-	-	68.2	38	44.1
28	75.5	38.7	33.3	-	-	-
29	42.1	25	13.8	-	-	-
30	80.5	80	44.5	-	-	-