5-2018

Math I Benchmark Testing: Friend or Foe?

Shelley Gainey West

Math I Benchmark Testing: Friend or Foe?

By
Shelley Gainey West

A Dissertation Submitted to the
Gardner-Webb University School of Education
in Partial Fulfillment of the Requirements
for the Degree of Doctor of Education

Gardner-Webb University
2018

Approval Page

This dissertation was submitted by Shelley Gainey West under the direction of the persons listed below. It was submitted to the Gardner-Webb University School of Education and approved in partial fulfillment of the requirements for the degree of Doctor of Education at Gardner-Webb University.

Jenny T. Sabin, Ed.D.
Committee Chair

Date

Don Phipps, Ed.D.
Committee Member

Date

Jennifer Putnam, Ed.D.
Committee Member

Date

Jeffrey Rogers, Ph.D.
Dean of the Gayle Bolt Price School
of Graduate Studies

Date

Acknowledgements

I would like to thank God for bringing me to this point and then bringing me through it.

I want to thank my husband, Maris West, for his love, support, and forgiveness when I lost my "sunny disposition."

I would like to thank my mom, Melisia Gainey, for modeling the doctoral process when she traveled the path herself; and my dad, "Duck" Gainey, for always encouraging me to pursue my dreams. Thank you to my brother, Raiford Gainey, for always believing in me; and to my brother, Thomas Gainey, for his support during this process.

This dissertation would not have been possible without the incredible insight of my dissertation chair, Dr. Jenny Sabin. Thank you for always being my advocate. I would also like to thank my committee members, Dr. Jennifer Putnam and Dr. Don Phipps, for your time and commitment to see me through this process. It has truly been a pleasure.

In addition, I am indebted to Zoe Nulty who stuck by me through this entire journey. There are no words to describe what your friendship and unwavering support have meant to me. To the rest of my cohort, Andrea, Elizabeth, Stacey, and Tricia, I cannot imagine this journey with any other group. Just keep swimming…

This dissertation could not have been written without the assistance provided by Andrea Lilley and Ashley Padgett from the East County Public Schools Central Office. I would like to thank them both for their willingness to "dig in" and get what I needed.

Abstract

Math I Benchmark Testing: Friend or Foe? West, Shelley Gainey, 2018: Dissertation, Gardner-Webb University, Mathematics Teachers/End-of-Course/Formative Assessment/Benchmark Assessments/Testing

This quantitative study addresses the ability of locally developed Math I benchmark assessments to predict student performance on the Math I end-of-course (EOC) state assessment for a rural county in North Carolina. Many districts in North Carolina lack funds to purchase commercially developed benchmarks and must rely on district personnel to develop these assessments. Locally developed benchmark assessments are not typically validated or considered psychometrically sound. Signal Detection Theory (SDT) was used to determine cut-off values on the locally developed Math I benchmark assessments for pass/fail grouping of students that will inform instructional interventions prior to the Math I EOC state assessment. The Math I benchmark assessments were determined to be accurate predictors of student performance on the North Carolina Math I EOC assessment, and the sensitivity and specificity of the Math I benchmark assessments were calculated to greater than 70% accuracy for each benchmark based on the identified cut-off scores.

Table of Contents

Chapter 1: Introduction

Introduction

Every year across North Carolina, Math I (also known as Algebra I) students are administered benchmark assessments as a predictor of end-of-course (EOC) summative assessment performance. Administering benchmark assessments allows teachers to assess student learning and allow for instructional adjustments. The practice of administering benchmark assessments is primarily driven by the Every Student Succeeds Act (ESSA, 2015) which, similar to the No Child Left Behind Act of 2001 (NCLB, 2002), reauthorized the Elementary and Secondary Education Act (ESEA, 1965) passed in 1965. ESSA (2015) allowed individual states to determine their accountability goals but stipulated that the goals must address testing proficiency, English language proficiency, and graduation rates. North Carolina READY Accountability Background Brief (2016) components for high school are EOC test scores, American College Testing (ACT) scores, graduation rates, math course rigor, ACT WorkKeys scores, and graduation project participation. Consequently, student performance on the North Carolina Math I EOC assessment has been a contributor to the overall accountability evaluation of the school. The evaluation model considers the level of proficiency as well as student learning growth in the formula for a school's letter grade. The school's letter grade is based on a 15-point scale using a calculation that is 80% achievement and 20% learning growth. The state of North Carolina utilizes the Education Value-Added Assessment System (EVAAS) developed by SAS Institute, Incorporated, as a tool to improve student learning, measure student learning growth, and measure teacher effectiveness (SAS, 2016). In addition to measuring student learning growth, EVAAS provides projections for student performance on EOC summative assessments (SAS, 2016); however, the individual student projections are based on the previous year's cohort of students with similar testing characteristics and not the individual student's testing history exclusively (SAS, 2016). Further, the results of the EVAAS student growth data and projections are not content standard specific and do not inform instruction to improve student learning and achievement. Consequently, common formative assessments, or benchmarks, are given to students at regular intervals to inform teachers and students about the state of student learning as related to the grade or course content standards.

Benchmark testing can provide individualized, real-time, predictive data for student performance on summative EOC assessments. Furthermore, as described by Bailey and Jakicic (2012), benchmark assessments provide periodic information on student progress towards proficiency goals. Benchmark assessments also allow students to become familiar with the format and vernacular of EOC summative assessment questions. This study allows teachers of Math I in East County, North Carolina, to better serve students by allowing the teachers to group students for real time instructional interventions.

Chapter 1 of this quantitative study details the problem statement for the research surrounding the development and use of benchmarks as predictors of EOC assessment performance. In addition, the purpose of the study is presented along with the research questions the study answers. Finally, the terms used in the study are defined in Chapter 1.

Problem Statement

Many school districts in North Carolina do not have the funding to use commercially developed benchmark assessments aligned to North Carolina's defined standards, or the commercially developed benchmarks are not based on North Carolina's standards. Consequently, these school districts have to depend on local input to develop benchmark assessments such as Schoolnet (Pearson, 2017) assessment banks, released assessment items, or teacher developed assessment items. Locally developed benchmarks are not always vetted for validity or reliability as described by American Educational Research Association, American Psychological Association, & National Council of Measurement in Education (2014). In addition, the benchmarks developed in Home Base, North Carolina's digital platform of classroom management and instructional resources, are not statistically analyzed for their ability to predict student performance on the North Carolina Math I EOC assessment (North Carolina Department of Public Instruction [NCDPI], 2013b). Therefore, the validity and reliability of the benchmarks used as predictive indicators are unknown (NCDPI, 2013b).

Commercially developed benchmark assessments are typically validated and are utilized throughout the United States (Olson, 2005). While there does not appear to be a concrete figure for the quantity of school districts using commercially developed benchmarks, the market for benchmark assessments has grown (Olson, 2005). Olson noted the following status regarding the growth of the commercially developed benchmark market:

> Eduventures Inc., a market-research firm based in Boston, identified benchmark assessments as one of two high-growth areas in the assessment industry, alongside state exams, with a compound annual growth rate of greater than 15 percent. The company predicted that by 2006, what it called "the formative-assessment market"—using a term sometimes treated as a synonym for benchmark assessment—would generate $323 million in annual revenues for vendors. (p. 13)

Commercially developed benchmarks are only available to districts with the necessary funds to purchase them. Districts that do not possess these funds may utilize a less expensive option, such as commercially developed test banks that may not be aligned to state standards, to develop benchmark assessments. Consequently, the impact of this research study to determine the predictive nature of the locally developed benchmarks could be significant to this school district. As a replication of the study Thompson (2016) completed in a school district in Georgia, this study could be replicated in other North Carolina school districts.

Further, the data from the benchmark formative assessments in East County were not utilized by teachers to identify students needing instructional remediation. Determining cut-off scores from the benchmark formative assessment data for individual student EOC summative proficiency would allow teachers to identify students needing instructional interventions. Performing benchmark formative assessments is an exercise in futility if the information gained is not utilized.

Background

The literature related to benchmark assessments indicates that commercially developed benchmarks are solid predictors of student success on summative assessments at the elementary level; however, the research on the current state of the predictive validity of locally developed benchmark formative assessments for high school is limited (Brown & Coughlin, 2007; Thompson, 2016). In addition, the research related to the predictive nature of locally developed benchmarks on summative assessments at any grade level is sparse. This study contributes to the current level of information on how to use benchmarks as a key indicator of student achievement and school improvement, and it allows teachers to identify students needing instructional interventions.

According to North Carolina School Report Cards (2016), East County's traditional high schools serve students in Grades 9-12 on a traditional calendar and bell schedule. There are two nontraditional high schools in East County. The first is East County Early College High School which operates on a community college schedule and requires students to apply and be accepted based on defined criteria. Students at the East County Early College High School graduate with both a high school diploma and an associate's degree in 5 years. The second one is East County Ed Tech Center, which is an alternative public school serving students in Grades 6-12 who are assigned for disciplinary reasons by the superintendent or are selected to be part of a high school dropout prevention program. There are five schools in East County serving eighth-grade students, all operating on a traditional calendar and bell schedule (North Carolina School Report Cards, 2016). All students enrolled in North Carolina Math I in the eighth grade or high school take the North Carolina Math I EOC assessment at the end of the course, according to ESSA (2015); however, the performance of eighth graders taking Math I is not a factor in a school's overall performance until those students complete their sophomore year in high school. In this way, the Math I scores of eighth graders are effectively "banked."

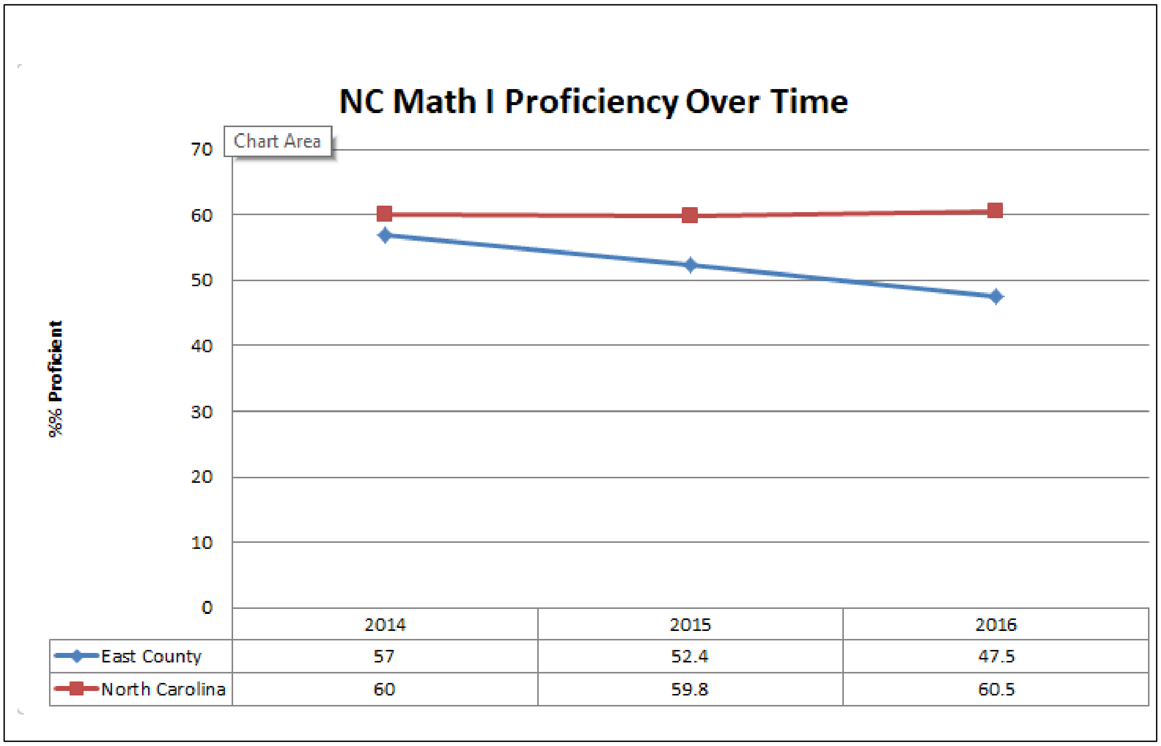
According to North Carolina School Report Cards (2016), 52.5% of East County Schools' Math I students scored Level 1 or Level 2 (below proficient) on the North Carolina Math I EOC summative assessment. Table 1 illustrates the percentage of students who scored at each performance level on the 2015-2016 North Carolina Math I summative assessment for East County and for the State of North Carolina (North Carolina School Report Cards, 2016) and includes the banked eighth-grade scores from 2 years prior to the 2015-2016 school year.

Table 1

2015-2016 North Carolina Math I Student Performance

Achievement Level (Command of Knowledge and Skills)	1 Limited	2 Partial	3 Sufficient	4 Solid	5 Superior
East County	40.4%	12.1%	9.7%	29.4%	8.4%
State	25.0%	14.5%	10.7%	34.3%	15.6%

As illustrated in Table 1, 47.5% of East County student test scores were considered proficient. Although Level 3 does not meet North Carolina Standards for College and Career Readiness, Level 3 is considered grade-level proficient (North Carolina School Report Cards, 2016); however, East County Schools' percentage of students scoring at or above grade-level is well below the state percentage of 60.6%. Further, the history of East County North Carolina Math I EOC assessment student performance demonstrates a consistent deficit in student scores. Figure 1 reflects the history of Math I EOC assessment scores in East County according to North Carolina School Report Cards (2016) and also includes, for each school year, the banked eighth grade Math I scores from 2 years prior to each school year.



	2014	2015	2016
East County	57	52.4	47.5
North Carolina	60	59.8	60.5

Figure 1. East County Math I EOC Assessment Scores Over Time.

As depicted in Figure 1, the Math I EOC assessment student performance has trended down over the past 3 years, while the state percentage of proficient students has remained steady. These data support the need for this research study of Math I formative benchmark assessments as an accurate predictor of North Carolina Math I EOC assessment performance for the purpose of improving student proficiency in Math I.

As a high school math teacher in East County, the researcher was well versed in the administration of the locally developed Math I formative benchmark assessment and the data available from each administration. In addition, the researcher has administered the North Carolina Math I EOC assessment multiple times and was familiar with the data results of this assessment.

Purpose of the Study

The purpose of this study was to acknowledge and analyze two concerns regarding the use of benchmark assessments in East County, North Carolina Math I courses. First, this study investigated how accurately individual student performance on North Carolina Math I benchmarks in East County was predictive of individual student performance on the North Carolina Math I EOC assessment. The second purpose of the study was to utilize Signal Detection Theory (SDT) to determine a cut-off score on the Math I benchmark for indicating pass/fail on the Math I EOC assessment. Identifying the cut-off score for each benchmark that corresponds to the predictive proficiency score enhanced the sensitivity and specificity of the East County Math I benchmark assessment for predicting student performance on North Carolina Math I EOC assessment. This process provided a benchmark score for each benchmark that can be used to group students for remediation as needed.

Theoretical Framework

This study employed SDT as the theoretical framework to determine the performance scores on the benchmark assessments which define the sensitivity and specificity of the prediction model. The sensitivity refers to the percentage of students who passed the North Carolina Math I EOC assessment and were predicted to pass based on their benchmark scores. The specificity refers to the percentage of students who did not pass the Math I EOC assessment and were predicted not to pass based on their benchmark scores. SDT is a decision-making process grounded in uncertainty and based on statistical techniques (Tuzlukov, 2001). Dating back to World War II, SDT was used to refine the process of reading radar signals. It was necessary to determine if a strong signal on radar was an enemy ship or just a big fish. In this study, SDT was used to determine the benchmark cut-off score for pass or fail performance on the Math I EOC assessment.

Research Questions

The research questions for this study were

1. How accurately does student performance on the Math I benchmark assessments for eighth grade, year-long, and semester Math I courses predict student performance on the North Carolina Math I EOC assessment as measured by student achievement?
2. To what degree can the sensitivity and specificity of the Math I benchmark assessments for eighth grade, year-long, and semester Math I courses be determined by analyzing the cut-off scores for the Math I benchmark assessment?

This study focused on individual student performance as related to achievement levels on the North Carolina Math I EOC and did not consider student growth or teacher growth.

Definitions

For the purposes of this study, the following vocabulary were used.

Adequate yearly progress (AYP). Defined by the U.S. Government under NCLB (2002) to determine how schools and school districts are performing on standardized tests.

Benchmark assessments. Locally or commercially developed interim assessments that provide periodic information on student progress towards proficiency levels defined by the state (Bailey & Jakicic, 2012).

Cut-off score. The score predicting pass or fail status on an assessment (Thompson, 2016).

Effect size. The average improvements in student assessment scores divided by the range of assessment scores for similar groups of students on the same tests (Black & Wiliam, 1998a).

EOC assessment. State-level summative assessment measuring student learning in specific courses based on course content standards and goals (NCLB, 2002).

End-of-grade (EOG) test. State-level summative assessment measuring student learning in grades kindergarten through eighth grade based on grade-level standards and goals (NCLB, 2002).

Formative assessment. Formal or informal assessment given during the duration of a topic to inform instruction or learning (Black & Wiliam, 1998a).

Kurtosis. Measures the weight of the tail (heavy tailed or light tailed) with respect to the normal distribution (NIST/SEMATECH, 2013).

Meta-analysis. A subset of similar but related studies quantitatively pooled to determine statistical significance (Black & Wiliam, 1998a).

Reliability. The consistency of assessment scores across replications (American Educational Research Association et al., 2014).

Receiver Operator Characteristic (ROC) curve. The graphical representation of the sensitivity versus the specificity (Green & Swets, 1966).

Schoolnet. An assessment question bank and platform with both commercially developed and teacher developed assessment questions (Pearson, 2017).

SDT. A decision-making process grounded in uncertainty and based on statistical techniques (Green & Swets, 1966).

Skewness. Describes the lack of symmetry of a data distribution on the right or left of the center point (NIST/SEMATECH, 2013).

Student achievement. A measure of student learning based on specific standards and goals (NCLB, 2002).

Summative assessment. Assessment given at the end of a unit of study, end of a course, or end of a grade (Black & Wiliam, 1998a).

Validity. The level at which the assessment measures what it is designed to measure (American Educational Research Association et al., 2014).

Conclusion

Chapter 1 of this quantitative study details the problem statement for the research surrounding the development and use of North Carolina Math I benchmarks as predictors of North Carolina Math I EOC assessment performance. The purpose of the study and the research questions the study sought to answer are presented. Finally, the terms used in the study are defined. Chapter 2 provides current research on benchmark assessments, benchmark interpretations, and uses. Chapter 3 provides the methodology for this study.

Chapter 2: Literature Review

Introduction

Student achievement on standardized summative assessments is of interest to students, parents, teachers, and administrators. Similar to an autopsy, standardized summative assessments provide information after the fact. Earl (2012) described summative assessments as assessments *of* learning as opposed to assessments *as* learning or assessments *for* learning. Benchmark formative assessments, described as assessments *for* learning (Earl, 2012), have become a key player in preparing students for standardized summative assessments. According to Earl, “The trick is to get the balance right” (p. 30). Earl’s description of formative assessment is supported by Black and Wiliam (1998b) in which they described formative assessment as,

> All those activities undertaken by teachers --and by their students in assessing themselves -- that provide information to be used as feedback to modify the teaching and learning activities. Such assessment becomes formative assessment when the evidence is actually used to adapt the teaching to meet student needs. (p. 2)

This description by Black and Wiliam (1998b) is the cornerstone of benchmark assessments for education. If some type of navigation to establish the current level of proficiency as it relates to the desired level of proficiency is not utilized, it will be very difficult to reach the desired level.

In an exploratory study, Goertz, Nabors Oláh, and Riggan (2009) stated,

> While teachers accessed and analyzed interim assessment data, we found that this information did not substantially change their instructional and assessment practice. Teachers used interim assessment results largely to decide what content

> to re-teach and to whom, but not to make fundamental changes in the way that this content or these students were taught. Teachers' use of classroom-based formative assessment did not necessarily lead to changes in instructional strategies either. (p. 6)

Chapter 2 details the theoretical framework for this study and thoroughly presents the research findings for the history of formative assessment and its uses as a tool for predicting student performance on state standardized assessments. The research findings on policies and procedures governing formative and summative standardized assessments are examined, specifically NCLB (2002). In addition, the research findings surrounding assessment standards and the predictive validity of benchmarks are discussed. Further, Chapter 2 presents the research findings for the applications of SDT and ROC as the theoretical framework of this study. Research on the use of benchmark assessments for Math I in the state of North Carolina is presented. This chapter concludes with a summary of the research findings surrounding the use of benchmarks as predictors of student performance on state standardized assessments.

Theoretical Framework

This research study utilized SDT to determine the scores on the Math I benchmark assessments that correspond to the sensitivity, or true positive rate, and the specificity, or true negative rate, of the prediction model. Figure 2 illustrates the SDT theoretical framework that was utilized to determine the cut-off scores of the Math I benchmark assessments.

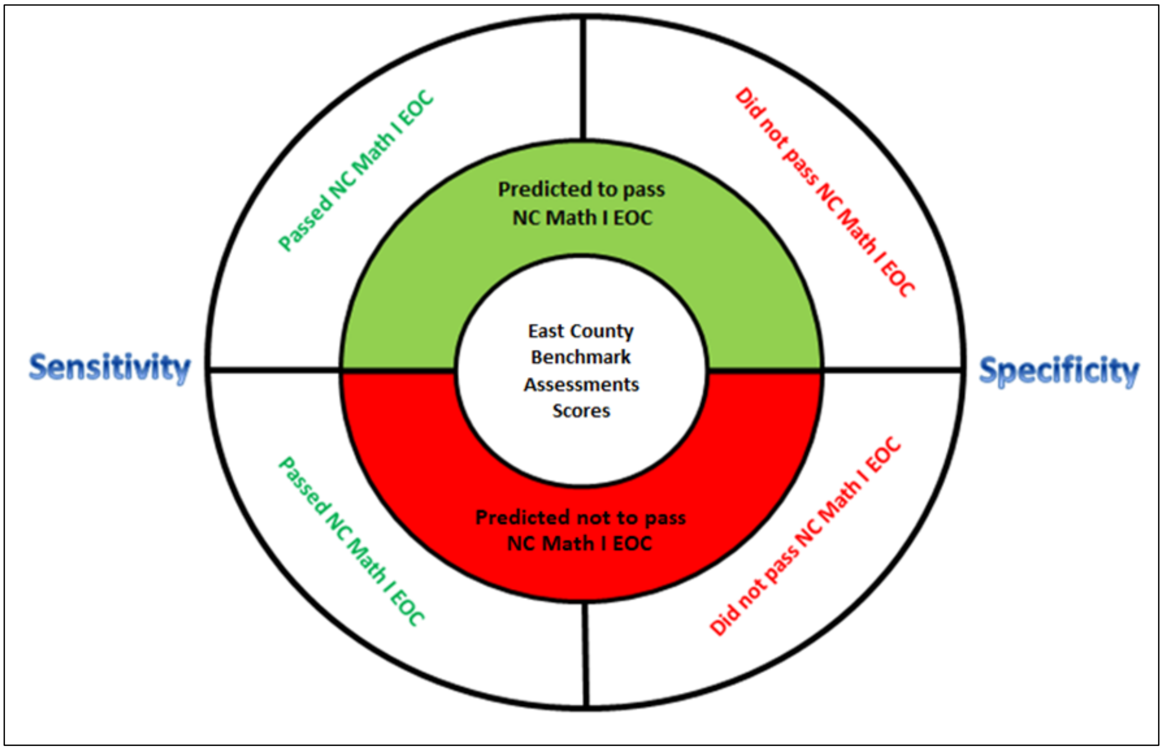

Passed NC Math I EOC
Did not pass NC Math I EOC
Predicted to pass NC Math I EOC
Sensitivity
East County Benchmark Assessments Scores
Specificity
Predicted not to pass NC Math I EOC
Passed NC Math I EOC
Did not pass NC Math I EOC

Figure 2. Theoretical Framework.

As displayed in Figure 2, the sensitivity refers to the percentage of students who passed the North Carolina Math I EOC assessment and were predicted to pass based on the benchmark performance. Likewise, the specificity refers to the percentage of students who did not pass the North Carolina Math I EOC assessment and were predicted not to pass.

SDT was used during World War II to perfect the reading of radar signals that appeared as dots on a circular screen (Tuzlukov, 2001). The radar operator had to determine if the dots on the screen were an enemy aircraft or not. SDT has also been used in psychological settings to evaluate a person’s ability to accurately detect a stimulus (Green & Swets, 1966). For the purposes of this study, SDT was used to determine the cut-off scores of the East County Math I benchmark formative assessment.

This study was a replication of the SDT math benchmark study performed by Thompson (2016) which identified cut-off scores and determined the predictive nature of the math benchmarks used in a school district in Georgia.

History of Formative Assessments

Informal formative assessment, such as questioning and probing students, has been used since the beginning of time. Socrates probably used formative assessment to guide his instruction (Greenstein, 2010); however, the term formative assessment, as compared to summative assessment, was first coined by Scriven (1967). As Scriven sought to acknowledge the work of others, he thoroughly examined the goals and roles of assessment evaluation. “Intellectual progress is possible only because newcomers can stand on the shoulders of giants” (Scriven, 1967, p. 1). Scriven’s work proved to be seminal, in itself, in the history of formative assessment. Scriven laid the foundation for the distinction between summative and formative assessment evaluation. He described the importance of evaluating the effectiveness of instruction while it is still transformable by using the term formative evaluation to differentiate it from evaluating the effectiveness of instruction at the conclusion of the delivery. The evaluation at the conclusion of the instruction was termed summative assessment by Scriven (1967).

Following Scriven’s (1967) ground-breaking work, Bloom, Hastings, and Madaus (1971) were among the first to discuss the actual implementation of formative assessment evaluation as part of their mastery learning initiative. Bloom (1977) continued to explore formative assessment as it related to feedback and instructional adjustments for mastery learning. Mastery learning is based on self-paced, module-type instruction in which students must demonstrate mastery of a unit or module through practice and formative assessment before progressing to the next sequential unit or module.

Acceptance. The idea of formative assessment continued to receive growing attention through the next 2 decades after the 1970s (Greenstein, 2010) culminating in a formative assessment explosion after the publication of Black and Wiliam's (1998b) "Inside the Black Box: Raising Standards Through Classroom Assessment." This publication was a game-changing summary of a study that pooled data from 250 formative assessment research studies for statistical analysis (Black & Wiliam, 1998a). The study strongly supported the case for the implementation of formative assessment evaluations. Black and Wiliam have become highly regarded formative assessment experts as they continue to explore the applications and impacts of formative assessment (Popham, 2008).

The meta-analysis study conducted by Black and Wiliam (1998b) yielded results indicating between a .4 and .7 effect size. The effect size measures the difference between two groups, and results of .4 to .7 indicate an average to above average difference in the two groups considered in the calculations. This effect size translates to a 15-point percentile change, such as 50th percentile to 65th percentile, with the implementation of formative assessment (Black & Wiliam, 1998b).

As the use of formative assessments proliferated, state and federal legislation, specifically NCLB (2002) and ESSA (2015), instituted assessment at specific intervals in public education and state mandates of exit standards and specific promotion criteria (gateway years). Consequently, formative assessment was implemented as much to meet state and federal legislation as it was to improve instruction. Formative assessments have proven to be beneficial in improving student learning in numerous studies, most notably, the meta-analysis study by Black and Wiliam (1998b). According to Alber (2014), formative assessment was not designed to catch students in a "gotcha" quiz. Formative assessment is centered on collecting information for instructional adjustments to improve student learning outcomes. Alber surmised that formative assessment is to inform, not to punish.

The benefits of formative assessments, particularly common formative assessments such as benchmarks, include encouraging teacher efficiency and student equity and providing the data to determine if the prescribed curriculum has been learned (Bailey & Jakicic, 2012). Black, Harrison, Lee, Marshall, and Wiliam (2003) stated, "Evidence of surveys of teacher practice shows that formative assessment is not at present a strong feature of classroom work" (p. 2).

Criticisms. While the work of Black and Wiliam (1998a, 1998b) with regard to the positive impact of formative assessment citing a .4 to .7 effect size seems to be widely accepted, Dunn and Mulvenon (2009) contended the research presented by Black and Wiliam (1998a) for the benefits of formative assessment is lacking empirical evidence.

Dunn and Mulvenon (2009) described the meta-analysis of Black and Wiliam (1998a) and the subsequent summary publication (Black & Wiliam, 1998b) as merely "a perfunctory review of literature on formative assessments for a manuscript on statistical methods" (p. 1). In fact, Dunn and Mulvenon stated the meta-analysis "evolved into a critical analysis of both the operationalization of formative assessment and the methods employed to document the impact of formative assessments" (p. 1). While Dunn and Mulvenon acknowledged that the literature on the impact of formative assessment is limited, they conceded that it would be irresponsible to deny that formative assessment can assist teachers in improving classroom instruction and outcomes. In their critical review of the meta-analysis of Black and Wiliam (1998a), Dunn and Mulvenon found issues with eight of the 250 studies used by Black and Wiliam (1998a).

According to Dunn and Mulvenon (2009), the study most heavily relied on for the conclusions drawn by Black and Wiliam (1998a) was a meta-analysis from Fuch and Fuch (1986) in which 83% of the participants were handicapped (Black & Wiliam, 1998a, 1998b), yet they were part of the statistical pool that was generalized to the population. Dunn and Mulvenon further concluded that the majority, 69%, of studies used in the Fuch and Fuch meta-analysis were of fair quality, as described by Fuch and Fuch.

Among the other issues with the meta-analysis of Black and Wiliam (1998a, 1998b) cited in the critical review by Dunn and Mulvenon (2009) was that one of the eight problematic studies only included math self-assessments for 8 to 14 year olds. Another study utilized only one formative assessment teacher compared to one non-formative assessment teacher, and the delineation between formative assessment effects and teacher effects was not clear (Dunn & Mulvenon, 2009). Another study of Black and Wiliam (1998a, 1998b) compared a novice teacher's use of formative assessment to a veteran teacher's use of formative assessment and also did not account for teacher effect (Dunn & Mulvenon, 2009). Further, Dunn and Mulvenon cited issues with one of the studies that exclusively used low socioeconomic kindergartners who were part of a large program of activities with embedded formative assessment. It was not clear whether the program itself or the formative assessment was the true catalyst for improved student learning (Dunn & Mulvenon, 2009).

The remaining three studies of the eight studies cited by Dunn and Mulvenon (2009) as not creditable in the meta-analysis of Black and Wiliam (1998a, 1998b) had similar inconsistencies. They cited group discussions without evidence the groups were on task, a concentrated fourth-grade study, and a study with a small sample size for the remainder of the eight studies specifically labelled as unacceptable in Black and Wiliam's (1998a, 1998b) meta-analysis study (Dunn & Mulvenon, 2009). Dunn and Mulvenon further argued that the research was limited, and there was not enough empirical evidence to conclusively state that formative assessment has a significant impact on student learning.

Another critical review of the formative assessment meta-analysis of Black and Wiliam (1998a, 1998b) by Bennett (2011) contended the studies used to support the significant impact of formative assessment are too diverse to be used in a meta-analysis. Bennett stated that the research "includes studies related to feedback, student goal orientation, self-perception, peer assessment, self-assessment, teacher choice of assessment task, teacher questioning behavior, teacher use of tests, and mastery learning systems" (p. 11). Based on this variety alone, Bennett concluded that an effect size statistic is not meaningful in the meta-analysis study by Black and Wiliam (1998a, 1998b).

Summary. While the history of formative assessment has numerous contributors, the most notable standouts are Scriven (1967) who is largely credited with defining the term formative assessment; Bloom et al. (1971) credited with early formative assessment application; and Black and Wiliam (1998a, 1998b) credited with groundbreaking evidence of the overall positive impact of formative assessment for student learning. Critics argue the groundbreaking meta-analysis of Black and Wiliam (1998a, 1998b) is not without flaw and should be examined more closely. Whether the degree of positive impact of formative assessment is significant or minimal to student learning, none of the historical research denies that formative assessment does have a positive impact.

Policy and Procedures

In 1981, under the leadership of President Ronald Reagan, the National Commission on Excellence in Education was charged with the task of reviewing and integrating the data and academic literature on the quality of learning and instruction in the nation's schools, colleges, and universities (U.S. Department of Education, 1983). The commission released their findings in a report titled *A Nation at Risk* in April 1983 (U.S. Department of Education, 1983). This report sparked the achievement testing and standards-based education reform era in the United States (Jorgensen & Hoffman, 2003).

Following the recommendations of the commission’s report, the Improving America's Schools Act of 1994 (IASA, 1994) was passed, which actually reinstated the ESEA of 1965. Title I, designed to improve student learning for low socioeconomic students, was the foundation of ESEA (U.S. Department of Education, 1994). Another law passed in 1994, the Goals 2000: Educate America Act (U.S. Department of Education, 1994), provided for a concentration on the needs of all students, not just the low socioeconomic students. IASA mandated that all states would have the following: content and performance standards; assessments aligned with those standards in one grade of each of three spans: 3-5, 6-9, and 10-12; and an accountability system to identify schools that were not helping all students perform as expected on those assessments (i.e., schools whose students could not achieve the standards).

The reward to each state for encouraging higher student learning outcomes was more flexibility to develop and manage their respective federal education funds (U.S. Department of Education, 1994). Meanwhile, at the state level, education funding was being argued in court via the case of *Leandro v. the State of North Carolina* (1997) which culminated in the decision that neither school districts nor counties have any constitutional right to equal funding, but every child has a right to a “sound basic education.” Each year from 1994 to 2000, most states implemented content standards and performance standards and began collecting data as well as utilizing secure test forms annually (Jorgensen & Hoffman, 2003).

NCLB was implemented under the leadership of President George W. Bush in January of 2002. As part of this act and the subsequent ESSA (2015), implemented under the leadership of President Barack Obama, assessment became a key factor in how students and schools were evaluated (Jorgensen & Hoffman 2003). These acts meant schools were faced with the threat of take-over at the state level and reduced federal funding for unfavorable evaluations.

Following NCLB (2002) and prior to the implementation of ESSA (2015), the United States experienced an education reform initiative under the leadership of President Barack Obama for standardizing curriculum in the United States. The Common Core States Standards Initiative, focusing on mathematics and English language arts (ELA), was a collaborative effort among the Gates Foundation, the Bill and Melinda Gates Foundation, the National Governors Association, and the Council of Chief State School Officers, which was published in December 2008 (Bidwell, 2014; Race to the Top [RttT], 2017).

The American Recovery and Reinvestment Act, a stimulus bill, became law in February 2009. The act provided $4.35 billion for the U.S. Department of Education with “no strings attached” (RttT, 2017). The $4.35 billion was used to fund the RttT program which had roots in NCLB (2002) and encouraged states with significant cash needs to compete for monetary grants. Although the original $4.35 billion allocated to the U.S. Department of Education was unconditional in requirements, awards to qualifying states under the RttT program required certain commitments from those states. Among the requirements for receiving grant funds, commitment had to take place within 2 months of the publication of the standards. This rapid turnaround did not allow time for the proper vetting of the standards (RttT, 2017).

By the end of the 1990s, every state in the United States had developed its own educational standards and assessments in compliance with IASA (1994). Consequently, each state had defined proficiency on its assessments according to its established standards. With the adoption of the Common Core State Standards, each state had to revise its curriculum standards and assessments for mathematics and ELA and literacy (Development Process, 2017). Specifically in mathematics, assessment is not limited to evaluating the correct answer but includes the process a student uses for arriving at the answer (Ainsworth, 2016). The purpose of the Common Core Standards was to guarantee all students graduate from high school with the skills and knowledge to be successful in college, career, and life (About the Standards, 2017).

As standardized summative assessments have become the norm across the United States, it is inappropriate not to acknowledge the concern about the interpretation of them. As described by Popham (2008),

> The chief indicator by which most communities judge a school staff's success is student performance on standardized achievement tests. These days, if a school's standardized test scores are high, people think the school's staff is effective. If a school's standardized test scores are low, they see the school's staff as ineffective. (p. 8)

As described by Madaus (1988), politicians place such an extreme emphasis on standardized assessments that the general public is convinced these assessments are an unequivocal effectiveness index. Ansley (2000) described the relationship between standardized assessments and school accountability:

> These tests, like many other aspects of education, have become pawns in a political chess game. In most such states, these tests are transformed from evaluation devices to high stakes accountability tools. . . . This is a large departure from the purposes for which these tests are constructed. (p. 278)

The implementation of standardized assessments, as required by federal law, has driven the necessity for benchmark assessments. The impact of the high stakes standardized assessments has been described by Abrams and Madaus (2003) through the following seven principles:

> **Principle 1**: The power of tests and examinations to affect individuals, institutions, curriculum and instruction is a perceived phenomenon: if students, teachers, or administrators believe that the results of an examination are important, it matters little whether this is really true or false – the effect is produced by what individuals perceive to be the case.
>
> **Principle 2**: The more any quantitative social indicator is used for social decision making, the more likely it will be to distort and corrupt the social processes it is intended to monitor.
>
> **Principle 3**: If important decisions are presumed to be related to test results, then the teacher will teach to the test.
>
> **Principle 4**: In every setting where a high-stakes test operates, a tradition of past exams develops, which eventually de facto defines the curriculum.
>
> **Principle 5**: Teachers pay particular attention to the form of the questions on a high-stakes test (for example, short answer, essay, and multiple-choice) and adjust

> their instruction accordingly.
>
> **Principle 6**: When test results are the sole or even partial arbiter of future educational or life choices, society tends to treat test results as the major goal of schooling rather than as a useful but fallible indicator of achievement.
>
> **Principle 7**: A high-stakes test transfers control over the curriculum to the agency which sets or controls the exam. (pp. 32-34)

Regardless of how summative standardized assessments are perceived, supporters are unable to identify benefits with empirical evidence to dissuade the notion that teachers are actually teaching to a test (Gruber, 2006).

Assessment Standards

Current practices reveal too much emphasis on student grades rather than formative assessment. In addition, research indicates that implementing classroom formative assessment improves student performance on state and national assessments (Black et al., 2003). Under NCLB (2002), each state is required to establish and implement rigorous academic standards and assessments in reading and math for Grades 3-8 and rigorous academic standards and assessments in math, science, and reading at least once in Grades 9-12. In addition, each state must establish yearly statewide progress objectives (The Bill, 2002). Flexibility is provided for the selection and/or development of each state's assessments, but each assessment must be aligned with the state curriculum standards (NCLB, 2002). With regard to the development of assessments, American Educational Research Association et al. (2014) provided recommendations that are highly regarded in the assessment arena.

According to American Educational Research Association et al. (2014), assessment validity, assessment reliability/precision, and fairness in assessments provide the foundation for high-quality assessments. American Educational Research Association et al. (2003) stated,

> The Standards are not in and of themselves legislation or law; however, they should be considered and implemented, when appropriate, by individuals in the field of testing. It is vital to utilize these standards, since the improper use of tests can cause considerable harm to test takers and other parties affected by test-based decisions. These doctrines provide assessment professionals with guidelines for the evaluation, development, and use of testing instruments. (p. 1)

Validity in assessments describes the degree to which evidence and theory support the interpretations of scores for proposed uses of tests. Assessment validation is a process and involves collecting evidence that provides a solid, scientific foundation for the interpretation of assessment scores. Evidence for validity includes test content, response processes, internal structure, and relations to other variables (American Educational Research Association et al., 2014).

Reliability/precision in testing, as described by American Educational Research Association et al. (2014), referred to the "consistency of the scores across instances of the testing procedure" (p. 33). The necessity for precision grows as the level of importance of the scores grows for student promotion and teacher employment. Reliability/precision is directly influenced by the variability within replications of the assessment. Further, reliability/precision of scores has implications for validity in assessments as well (American Educational Research Association et al., 2014).

Fairness in educational testing is one of the major constructs that needs to be addressed for an assessment to be considered high quality. It is not sufficient that an assessment be fair to individuals with disabilities or diverse backgrounds exclusively; it should be fair to the entire population in the targeted assessment group (American Educational Research Association et al., 2014).

Assessment design and development are critical to high-quality assessments. American Educational Research Association et al. (2014) provides research-based recommendations for assessment design and development. The recommendations for item development and review include documenting all processes, utilizing professionals for item review and scoring, and careful consideration of time constraints. The recommendations from American Educational Research Association et al. (2014) for high-quality assessments support the predictive ability of formative assessment performance to summative assessment performance.

Predictive Validity of Benchmark Assessments

"The ability of one assessment tool to predict future performance either in some activity (success in college, for example) or on another assessment of the same construct" (Brown & Coughlin, 2007, p. 2) is the definition of predictive ability. To determine the predictive validity of a benchmark assessment to the associated summative assessment, there are multiple statistical methods that can be used. Many school districts across the United States lack the funding to purchase commercially developed benchmark assessments. Therefore, the benchmark assessments in these districts are locally developed by classroom teachers, and district personnel and are not typically properly validated. Districts with more funding are able to purchase commercially developed, psychometrically sound benchmarks that are assumed to be properly validated (Brown & Coughlin, 2007).

Research conducted by Brown and Coughlin (2007) indicated the predictive ability of some commercially developed benchmark assessments, specifically being used in the mid-Atlantic region of the United States, were not adequate predictors of summative assessments. In addition, Thompson (2016) noted the research for the predictive ability of benchmark assessments at the high school level is very limited.

Payne (2013) studied the predictive ability of benchmark assessments in an American Indian school district. The benchmark assessments were developed by teachers using a purchased formative and benchmark assessment question bank program. All questions were reported to be properly vetted and validated (Payne, 2013). The research study had sample sizes of 44 students for Algebra I, 68 students for biology, and 52 students for English I. In the study, Payne found that benchmark scores exhibited a positive correlation to state standardized summative assessment (EOC) scores for high school subjects (Algebra I, biology, and English I). Consequently, the conclusion was that benchmark assessment scores were accurate predictors of student achievement on state standardized summative assessments. Payne's summary of this research study stated, "Regardless of whether a district is embarking on a new benchmark assessment program or has one currently in place, this study indicates how important it is for teachers to have a firm grasp of formative assessment" (p. 142). According to Payne, when teachers and administrators are equipped with the confidence that the benchmark assessments are accurate predictors of student performance on the standardized summative assessments, they can use that knowledge to improve and personalize student learning in the classroom.

Stockman (2016) also studied the predictive ability of a benchmark assessment to student performance on a state standardized summative assessment. Using 816 high school students from a school district in Maryland, Stockman sought to determine the predictive ability of the Algebra I benchmark assessment to the Algebra I state summative assessment in Maryland. The study was conducted from a STEM (Science, Technology, Engineering, and Mathematics) perspective and demographic focus. As the desire to increase the variety of students pursuing STEM careers grow, mathematics is recognized as a key component. Stockman (2016) stated,

> In order to remain a global power, the United States needs to develop a strong workforce in the STEM fields. To build this pipeline of STEM employees, high schools need to increase student success in the area of mathematics. Helping improve success in mathematics requires the educational assessment tools to help all students regardless of their race, gender, grade level, or socioeconomic status. (p. 93)

Stockman's (2016) research study explored the predictive ability of a locally developed benchmark assessment for Algebra I. In this study, a non-experimental design using multiple regression analyses to examine the predictive validity of local benchmark assessments on summative standardized state assessment scores was utilized.

Stockman's (2016) findings were detailed for grade level, gender, race, and free/reduced meal status. Overall, the results indicated the benchmark assessment was a strong predictor of student achievement on the state standardized summative assessment (Stockman 2016). Controlling for grade level, Stockman found that there were no significant differences in student achievement for eighth-grade students (using seventh-grade scores as a baseline) taking the state standardized summative assessment. There were significant differences in scores of ninth graders, for which this course is considered grade-level appropriate, and tenth and eleventh graders' scores, taking the same course; however, this was accurately supported by the predictive validity of the benchmark scores. Controlling for race, the predictive validity of the benchmark assessments was also positively correlated and with lower achievement scores for Blacks and other minorities, compared with scores for White students (Stockman, 2016). There was no significant difference in achievement on the state standardized summative assessment based on gender or free/reduced meal status which also correlated with the benchmark assessments scores (Stockman, 2016). Stockman, like Thompson (2016), asserted there is limited research available on the predictive ability of benchmark assessments for state standardized summative assessments, particularly for high school math. Stockman concluded, "Valuable benchmarks can be created when they are aligned to the same priority standards as the state assessment" (p. 97).

Ainsworth (2016) conducted a research study to determine the significance of the predictive nature of Coordinate Algebra common formative assessments on the state standardized summative assessment for Coordinate Algebra with a specific focus on demographics and students with disabilities. The common formative assessments were developed by a team of teachers and instructional experts within the district and were vetted (Ainsworth, 2016). Although the common formative assessments were more numerous than traditional benchmark assessments, they served a similar purpose and were deemed "non-negotiables" by the district (Ainsworth, 2016).

Ainsworth's (2016) study utilized multiple regression analysis for the prediction tool. Findings included that the district common formative assessments were accurate predictors of student performance on the state standardized summative assessment. These findings were echoed in the subgroup of students with disabilities; however, a significant relationship between the common formative assessments and state standardized summative assessments by demographic variable of students with disabilities was not found (Ainsworth, 2016).

Thompson's (2016) study, conducted in a small school district in Maryland, researched the predictive ability of the locally developed Algebra 1 benchmark assessment student scores to student performance on the state Algebra 1 EOC assessment. Using SDT and ROC analysis, Thompson was able to determine the predictive ability of the Algebra 1 benchmark assessment and cut-off scores as pass/fail predictors of the Algebra 1 EOC assessment. This quantitative study utilized a correlational design, regression analysis, and ROC analysis. Thompson acknowledged that the significance of this research study was

> contingent upon the practical application of the results to justify the continued use of benchmark assessments through improved teaching, modifications to curriculum, improved preparation for assessments, and guiding district instructional planning. Aside from the student, the classroom teacher will benefit from gaining knowledge about student content mastery through the interpretation and analysis of benchmark results. . . . When schools are given an early indication of student performance on end-of-year assessments, it allows for interventions to take place which may change the predicted outcome of the assessment. Without knowledge of the predictive ability of benchmarks, students who are on the borderline of passing or failing the state assessment may not be identified and thus, not receive the extra help that could raise the chance of them passing the test. (p. 96)

Thompson (2016) pointed to a powerful truth in this research study. If the information gained from a benchmark assessment is not utilized as an instructional intervention, the benefits of the benchmark process are not realized (Thompson, 2016).

SDT

As part of the predictive validity of benchmark assessments, SDT can be utilized to determine cut-off scores for pass/fail criteria on the associated summative assessment, using the ROC curve analysis. Used in the medical field for clinical assessment and diagnosis (McFall & Treat, 1999), SDT was used in World War II to refine the process of reading radar signals (Green & Swets, 1966; Tuzlukov, 2001). During World War II, it was necessary for the military to be able to tell the difference between an enemy submarine or enemy aircraft (signal) and a school of fish or flock of birds (noise) respectively. While radar would indicate any of these, it was not able to decisively distinguish between them (Tuzlukov, 2001). Green and Swets (1966) are considered pioneers of SDT (Abdi, 2007; Tuzlukov, 2001) based on their work during World War II.

SDT is most often explained using a scenario and a table of the four possible responses to the scenario (Abdi, 2007; Green & Swets, 1966; Tuzlukov, 2001). Table 2 reflects the four possible scenarios using SDT.

Table 2

SDT Response Scenarios

Reality	Receiver Perception Yes	Receiver Perception No
Signal is present	Hit	Miss
Signal is not present	False Alarm	Correct Rejection

As depicted in Table 2, there are four possible outcomes in SDT. These outcomes maximize the sensitivity and specificity. The sensitivity, or the true positive rate, is defined as the number of hits divided by the sum of the hits and misses (Abdi, 2007; Green & Swets, 1966; Tuzlukov, 2001). The specificity, or true negative rate, is defined as the number of correct rejections divided by the sum of the correct rejections and false alarms (Abdi, 2007; Green & Swets, 1966; Tuzlukov, 2001).

The ROC curve is the sensitivity plotted on the y-axis to the corresponding specificity plotted on the x-axis. Each individual point on the graph represents sensitivity/specificity pairing. The smaller the distance between the curve and the top left corner of the graph, the more accurate the test is (Abdi, 2007; Green & Swets, 1966; Tuzlukov, 2001).

Thompson (2016) used SDT to determine cut-off scores on Algebra I benchmark assessments. He defined the sensitivity as the proportion of students who passed the state summative assessment and were predicted to pass based on their Algebra 1 benchmark scores. Likewise, Thompson defined the specificity as the proportion of students who did not pass the state summative assessment for Algebra I and were predicted not to pass based on their Algebra 1 benchmark scores. SDT is a decision-making process grounded in uncertainty and based on statistical techniques (Tuzlukov, 2001).

In addition to Thompson's (2016) study, Anderson, Alonzo, and Tindal (2010) used SDT to determine a maximum cut-off score for Curriculum-Based Measurement (CBM) mathematics formative assessments to predict student achievement on summative state assessments in third through eighth grade. Hintze and Silberglitt (2005) used ROC curve analysis along with two other types of analysis, discriminative and logistic regression, to determine benchmark cut-off scores. They recommended using ROC for districts that need to make multiple decisions such as diagnostic or entitlement decisions. The research findings of Nese, Park, Alonzo, and Tindal (2011) also supported ROC curve analysis as a solid predictor of student performance on summative assessments based on benchmark performances.

North Carolina Math I Benchmarks

The state of North Carolina, in compliance with state and federal legislation, administers EOC assessments for Math I, Biology, and English II at the high school level. In order to gauge student progress, benchmark assessments are administered multiple times during a semester for each of these courses in most school districts. Many districts in North Carolina develop their own benchmark assessments for various reasons including lack of funds to purchase commercially developed benchmarks. North Carolina Public Schools (2011) releases EOC assessment items which are publicly available; however, they issued the following statement regarding the use of these released items:

> The state strongly cautions local school systems against using the forms (released items) to "teach the test." While the EOG and EOC tests summarize student achievement and serve as an important indicator of aggregate student performance for state and federal reporting, they are not the sole measure of student performance. The use of released tests should be part of an overall plan in school systems for implementing the new state testing program that includes formative, benchmark, and summative assessments. Because the released passages and questions will never again be used on the state assessments, the use of these test questions for test preparation would be misleading. (p. 16)

NCDPI (2013a) provided guidelines for developing benchmarks for districts using the Schoolnet question bank house in Home Base aligned to the curriculum standards for each course. Home Base is North Carolina's digital platform of classroom management and instructional resources and was initially developed with RttT funds (NCDPI, n.d.a).

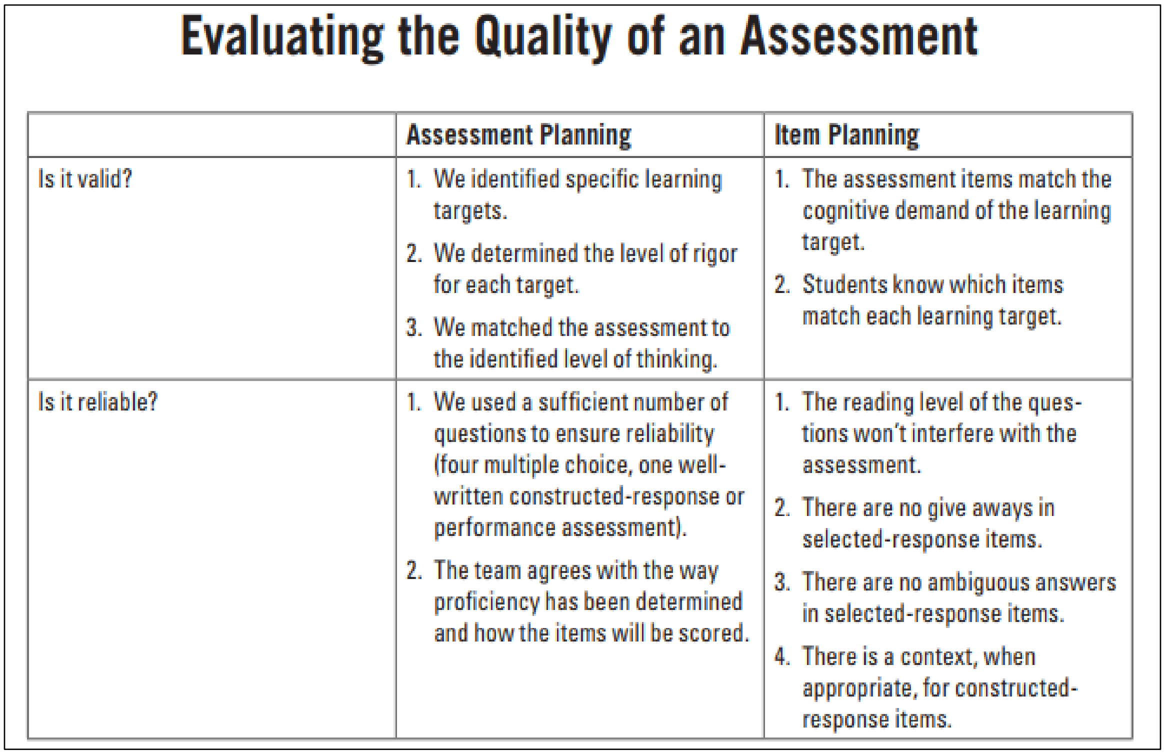
Although parents, students, teachers, and administrators all have access to Home

Base, the district testing coordinator has exclusive access to the Home Base benchmark question bank for the purpose of building the benchmark assessments. Strict access to benchmark assessment items ensures assessment security at all times. Benchmark assessment items are not available for classroom use; however, assessment items can be manually created or edited by the district testing coordinator if using Home Base to build a benchmark assessment (NCDPI, 2013a). Any assessment items entered into Home Base must be associated with a curriculum standard (NCDPI, n.d.b). North Carolina's initial purchase of Home Base included assessment item banks for both classroom and district benchmark use from multiple sources (NCDPI, n.d.b). Although all students across the state of North Carolina take the same Math I EOC summative assessment, the associated formative assessment items used to build the benchmark assessment for Math I vary from district to district, and the final product is typically not validated (NCDPI, 2013b).

The East County Schools' Central Services mandates benchmark testing dates and testing procedures based on the East County Policy Manual (2016) and creates curriculum pacing guides for all of East County's high schools. The State of North Carolina mandates final exam dates for EOC exams and North Carolina Final Exams (NCFEs) and all curriculum content to be delivered in the classrooms. Recent state legislation, Public Schools/Testing Schedule (2015), mandates a 5-day testing administration window, the last 5 days of the school semester, for all state exams and EOC assessments.

Benchmark Research Needs

Whether the proposed use of benchmark assessments is for instructional purposes, predictive purposes, or evaluative purposes, the data must be accurately interpreted. To accurately interpret benchmark data, teachers and administrators must be properly trained through professional development or in a Professional Learning Community (PLC). In addition to interpreting the data provided by benchmark assessments, it is imperative that the assessment be evaluated for goodness of fit so the users can make connections between the assessment and the content being assessed (Bailey & Jakicic, 2012). Bailey and Jakicic (2012) recommended evaluating the quality of an assessment in a PLC using the reproducible guide in Figure 3.



Evaluating the Quality of an Assessment

	Assessment Planning	Item Planning
Is it valid?	1. We identified specific learning targets. 2. We determined the level of rigor for each target. 3. We matched the assessment to the identified level of thinking.	1. The assessment items match the cognitive demand of the learning target. 2. Students know which items match each learning target.
Is it reliable?	1. We used a sufficient number of questions to ensure reliability (four multiple choice, one well-written constructed-response or performance assessment). 2. The team agrees with the way proficiency has been determined and how the items will be scored.	1. The reading level of the questions won't interfere with the assessment. 2. There are no give aways in selected-response items. 3. There are no ambiguous answers in selected-response items. 4. There is a context, when appropriate, for constructed-response items.

Source: Gareis and Grant, 2008; Stiggins et al., 2004

Figure 3. Assessment Quality Evaluation.

As depicted in Figure 3, Bailey and Jakicic (2012) provided a thorough checklist for assessing specific learning targets when developing an assessment. Including teachers as subject area experts in the building or development of curriculum-based benchmark assessments is a key factor in designing a high-quality assessment (Ainsworth, 2010). Teachers have a vested interest when they are included in the planning and development of instruction and assessment (Ainsworth, 2010; Drago-Severson, 2009). It serves to reason that state standardized summative assessment scores would not increase merely because benchmark assessments are administered. The data from the benchmark must be utilized to impact instruction for summative scores to be affected.

Oswalt (2013) found that the use of formative assessments (not specifically benchmark formative assessments) for detailed feedback occur most often. Oswalt also found that the use of formative assessment varies among classroom teachers in the same school and district. Research findings related to the actual use of benchmark formative assessments by classroom teachers, administrators, and district-level supervisors, particularly at the high school level, is limited (Goertz et al., 2009). Goertz et al. (2009) stated,

> there needs to be research on the quality of data generated by interim assessments. This is a severely neglected area of research, yet poor data precludes effective data use. Claims about the validity of interim assessment results for instructional use need to be investigated as a matter of course. (p. 10)

Goertz et al. (2009) further discussed that there is a need for the association between formative assessments to be examined. Most of the current assessment research primarily explores individual assessments. Examining the scaffolding of information received from various formative assessments and how it impacts instruction requires investigating instruction that is part of the assessment process (Goertz et al., 2009).

Lewis (2010) conducted a correlational study evaluating the degree to which teachers recognize the use of benchmark assessment data as effective, the degree to which the time spent teaching mathematics is related to student mathematics grades, and the degree to which math benchmark assessment data influence teacher instruction. Lewis found, using a Likert scale survey, that teachers do believe continuous use of benchmark formative assessment data is effective. In addition, Lewis found a positive correlation between instructional time and student grades and that teachers rated benchmark data as influential to their instruction.

Conclusion

Formative assessment history has seen notable contributors such as Scriven (1967) who is largely credited with defining the term formative assessment, Bloom et al. (1971) who is credited with early formative assessment application, and Black and Wiliam (1998a, 1998b) credited with groundbreaking evidence of the overall positive impact of formative assessment for student learning. Throughout history, the premise has been that formative assessment is beneficial to student learning when properly utilized; the sticking point has been the degree to which it is beneficial. According to the research, the significance of the beneficial impact weighs heavily on the validity of the formative assessment to inform instruction.

Benchmark formative assessments are an integral part of the assessment process in North Carolina. As state and federal mandates like NCLB (2002) and ESSA (2015) continue to place emphasis on the performance of students on summative assessments, benchmark formative assessments in core content areas like North Carolina's Math I, would remain a tool to influence instructional decisions. Regarding the future of benchmark assessments in North Carolina, Martin (2012) stated,

Future research studies on the use of interim assessments could include a

> comparison of responses from one individual regarding their intended purpose, actual use, and perceived benefits, qualitative research on data analysis methods aligned to specific purposes, case studies showcasing schools in North Carolina using interim assessments for a variety of purposes, and quantitative research studies to determine whether interim assessments can be correlated to improved student achievement. (p. 116)

Whether the benchmark assessments are commercially or locally developed, they can be used for multiple purposes. Research indicates the information provided from benchmark assessments can be utilized for instructional purposes, predictive purposes, or evaluative purposes. Martin (2012) asserted, “it is vital that district and state educational leaders make careful and informed decisions about the purpose and use of interim assessments prior to implementation” (p. 1). Because benchmarks are associated with the high stakes summative assessments mandated by federal law, districts across the United States have used benchmark data to predict student performance on the associated summative assessments; however, research on the predictive ability of the benchmarks used at the high school level is minimal.

The design of benchmark formative assessments is critical to the integrity of the assessment. American Educational Research Association et al. (2014) provided thorough recommendations for assessment design and development which include item development and review. The recommendations stress the importance of validity and reliability in the assessment development process and support the predictive ability of formative assessment performance to summative assessment performance (American Educational Research Association et al., 2014).

Without the funds to purchase commercially developed benchmark assessments assumed to be properly validated, districts like those in North Carolina with locally developed benchmark assessments are not likely to be knowledgeable of the predictive ability of their benchmarks. The majority of the research on predictive ability of benchmarks centers around elementary and middle school commercially developed reading and mathematics benchmarks. Specifically, research studies regarding the predictive validity of locally developed Math I benchmarks are limited. This research study sought to improve the amount of empirical evidence for the predictive validity of locally developed Math I benchmark formative assessments.

To address predictive validity of formative benchmark assessments, research indicates the use of multiple regression analysis techniques. Thompson's (2016) study, in addition to the study by Anderson et al. (2010), used SDT to determine a maximum cut-off score for mathematics formative assessments to predict student achievement on summative state assessments. Hintze and Silberglitt (2005) used ROC curve analysis along with two other types of analysis to determine benchmark cut-off scores. They recommended using ROC for districts that need to make multiple decisions such as diagnostic or entitlement decisions. The research findings of Nese et al. (2011) also support ROC curve analysis as a solid predictor of student performance on summative assessments based on benchmark performances.

In North Carolina, the district testing coordinator has exclusive access to the Home Base benchmark question bank for the purpose of building the benchmark assessments which ensures assessment security at all times. Further, benchmark assessment items are not available for classroom use. North Carolina's initial purchase of Home Base included assessment item banks for both classroom and district benchmark use from multiple sources (NCDPI, n.d.b) and allows the teachers to input items as well.

Although all students across the state of North Carolina take the same Math I EOC summative assessment, each district in North Carolina has the flexibility to develop a common benchmark assessment for that particular district. Consequently, the associated formative assessment items used to build the benchmark assessment for Math I vary from district to district and the final product is typically not validated (NCDPI, 2013b).

Chapter 3: Methodology

Introduction

This research study investigates if student performance on East County Math I benchmarks is an accurate predictor of student performance on the North Carolina Math I EOC state assessment. In addition, it seeks to utilize SDT to determine a cut-off score for each Math I benchmark administration. In this chapter, the participants involved in the study are described as well as the instrumentation utilized. Further, the research design is explained, the procedures that were implemented are presented, and the data collections and analysis are introduced. Finally, a summary of the methodology used for this study is provided.

This study was correlational and sought to determine if the Math I benchmark assessment scores are accurate predictors of student performance on the Math I EOC assessment. Further, the study predicts the cut-off scores for student performance on the North Carolina Math I EOC assessment. A 5-number summary (minimum, first quartile, median, third quartile, and maximum) was evaluated for benchmark assessments and the EOC assessment. Pearson Product Moment Correlation (*r*) was used to determine the strength between the Math I benchmark scores and the Math I EOC assessment scores. In addition, SDT utilizing an ROC curve analysis was used to optimize the predictive cut-off scores on the benchmarks for pass versus not pass on the Math I EOC assessment.

Participants

Data were collected from all Math I students at the middle schools and high schools in the East County school district (n > 500). The participants represent four middle schools and five high schools including an early college high school and an alternative high school. The participants live in a small rural district in which the county seat has approximately 10,000 residents. The schools chosen for the study are all of the public middle and high schools within East County, North Carolina. Each school offers the North Carolina Math I curriculum. The middle schools offer North Carolina Math I as a year-long course, while the high schools offer both a year-long and a semester option for the course. The benchmark assessments are different, depending on whether the course is year-long course or semester course and were, therefore, considered separately.

The middle schools and high schools are located in various residential settings including rural/farm, subsidized housing, inner city, and suburban. The socioeconomic status of the students ranges from below poverty level to upper class. The participants range in age from eighth grade middle school students to tenth grade high school students. Because the students mirror the social, economic, and racial diversity of the East County public school population, the results of this study would be expected to be comparable for future students taking the same Math I benchmarks and North Carolina Math I EOC assessment.

The factors considered in this research study are limited to the student's initial Math I benchmark scores and initial North Carolina Math I EOC assessment scores. Data from students who are repeating the course were eliminated to ensure the validity of the data. Race, demographics, gender, and socioeconomic status were not considered as factors in this study because the study is not examining if there is a performance difference on the Math I benchmarks or the North Carolina Math I EOC assessment related to these factors. The main threat to the internal validity of this study is the student skill level. This study considered students from low- to high-achievement levels as defined by NCDPI (2013b) without distinguishing between them, which allowed for a wide range of student performance data; however, this research study was limited to East

County, North Carolina Math I students because different districts utilize different Math I benchmark assessments.

Procedures

The researcher collected Math I benchmark scores and North Carolina Math I EOC assessment scores from the East County Testing and Accountability data management system with student identifying information deleted. Random student numbers were assigned to the data; and the data were filtered for course repeaters, duplicates, and missing scores. The data from students repeating the course, instances of duplicate data, and data that were incomplete were removed from the analysis. The statistical analysis of the data gathered through these procedures determined how accurately student performance on the local Math I benchmark assessments predicts student performance on the North Carolina Math I EOC state assessment. In addition, these procedures supported the utilization of SDT to optimize cut-off scores for each benchmark assessment.

Research Design

This was a quantitative study that used correlations and regression analysis from Math I benchmark assessments to predict student performance on the Math I EOC assessment. The research questions framing the research design were

1. How accurately does student performance on the Math I benchmark assessments for eighth grade, year-long, and semester Math I courses predict student performance on the North Carolina Math I EOC assessment as measured by student achievement?
    a. The independent variables for this study are the scores for the two administrations of the Math I benchmark assessment.

   b. The dependent variable is the standardized Math I EOC assessment scores for all students enrolled in Math I, in Grades 8-12, in a rural county in North Carolina.

2. To what degree can the sensitivity and specificity of the Math I benchmark assessments for eighth grade, year-long, and semester Math I courses be determined by analyzing the cut-off scores for the Math I benchmark assessment?

   a. The independent variable is the benchmark cut-off score.

   b. The dependent variable is the sum of the sensitivity and specificity values associated with each cut-off score.

This study focused on individual student performance as related to achievement levels on the North Carolina Math I EOC and didi not consider student growth or teacher growth. Table 3 details the benchmark administrations for East County.

Table 3

East County Benchmark Assessments

Course	Benchmarks
8th Grade NC Math I (year long)	2
Year-long NC Math I (9th-10th grade)	2
Semester NC Math I (9th-10th grade)	2
Total	6

As illustrated in Table 3, this study examined six different benchmarks for three different course offerings of NC Math I in East County.

Instruments

The instruments used for this study were the East County Math I benchmark assessments which were administered online via a secure testing platform and the 2016-2017 scores from the North Carolina Math I EOC assessment which was delivered in paper/pencil format. The benchmarks are locally developed from Schoolnet questions and released test items. Further, the benchmark assessments consist of both calculator inactive and calculator active assessment items with the exception of the eighth grade second benchmark administration which was exclusively calculator active. The North Carolina Math I EOC assessment consists of both multiple choice and gridded response items, in addition to calculator inactive and calculator active items. As previously stated, the predictive validity of the benchmarks is the focus of the study. The reliability of the benchmarks was examined by evaluating the benchmark scores over time, and the Math I benchmark assessments are given on the same day across the county to ensure assessment validity.

The Math I benchmark assessments and the North Carolina Math I EOC assessment are secure assessments and are not available to the public to ensure assessment validity; however, testing specifications for the North Carolina Math I EOC assessment are provided in Table 4.

Table 4

North Carolina Math I EOC Assessment Specifications

NC Math I (2016-2017)	Multiple Choice Items	Gridded Response Items
Operational Items	40	10
Field Test Items	8	2
Total	48	12

As seen in Table 4, the North Carolina Math I EOC assessment contains field test items. These items are not discernable on the assessment. The Math I benchmark assessments do not contain field test items or gridded response items. As stated by NCDPI (2016a), “The NC Math 1 assessment included both calculator active and calculator inactive sections. There are 18 items (6 multiple-choice, 12 gridded response/numeric entry) in the calculator inactive section and 42 multiple-choice items in the calculator active section" (p. 2). The Math I benchmarks are required to be administered on the same day across the county; however, the North Carolina Math I EOC assessment is not required to be given on the same day across the state, but it must be administered during the last 5 instructional days of each semester (NCDPI, 2016a). The instruments used in this study provided the information for the data analysis.

Data Analysis

This study was correlational in theme and sought to determine if the Math I benchmark assessments are accurate predictors of student performance on the Math I EOC assessment and what cut-off values predicted the Math I EOC assessment performance. Data dispersion using a 5-number summary was evaluated for each of the eight benchmark assessments and the EOC assessment. Pearson Product Moment Correlation (*r*) was used to determine the strength between the East County Math I benchmark scores and the Math I EOC assessment scores. Normality was examined by evaluating the skewness and kurtosis values, which describe the distribution from the center point of the data (NIST/SEMATECH, 2013) for the East County Math I benchmark scores and the North Carolina Math I EOC assessment scores. In addition, outliers were identified for each assessment. Any outliers beyond three standard deviations from the mean were removed from the dataset, as discussed by Creswell (2013), to adjust the normality of the data. Moreover, since each student score is independent of other student scores, the results of normality tests were reliable.

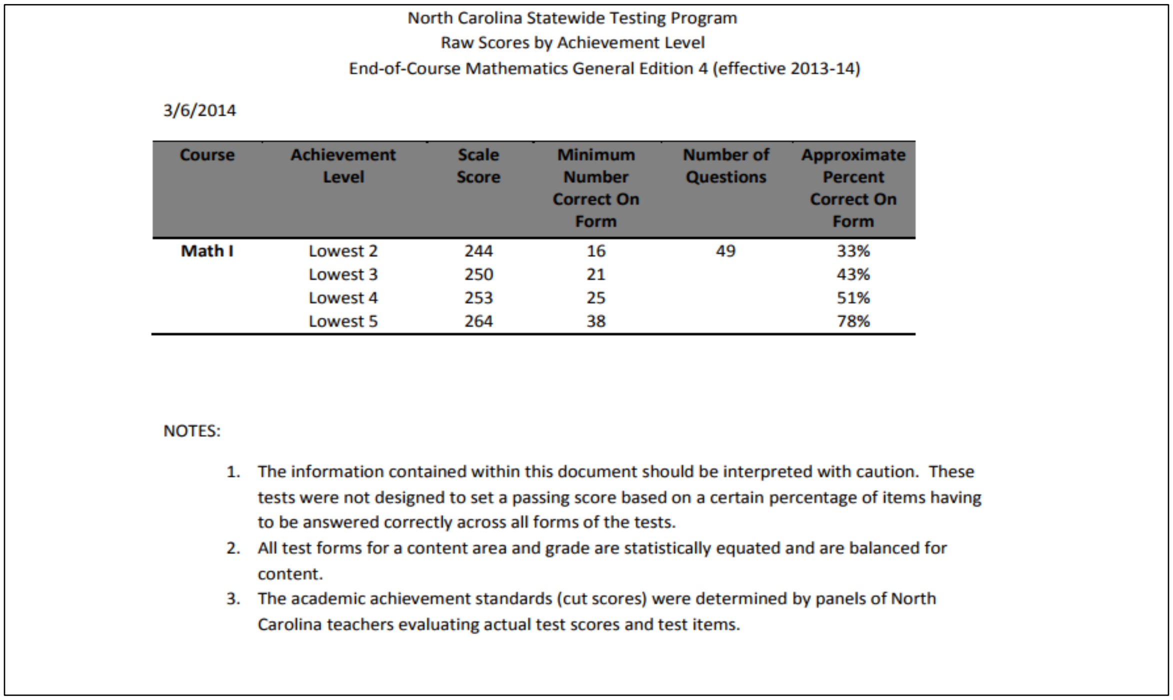
In addition, SDT utilizing an ROC curve analysis was used to optimize the predictive cut-off scores on the benchmarks for pass versus not pass on the Math I EOC assessment and answer the following questions.

1. How accurately does student performance on the Math I benchmark assessments for eighth grade, year-long, and semester Math I courses predict student performance on the North Carolina Math I EOC assessment as measured by student achievement?
2. To what degree can the sensitivity and specificity of the Math I benchmark assessments for eighth grade, year-long, and semester Math I courses be determined by analyzing the cut-off scores for the Math I benchmark assessment?

The null hypotheses were

1. H_o: Individual Math I benchmark scores are not accurate predictors of student performance on the North Carolina Math I EOC assessment.
2. H_o: Changing Math I benchmark cut-off scores would not change the sensitivity and specificity of the regression model to predict student outcomes on the North Carolina Math I EOC assessment.

The independent variables for this study were the Math I benchmark scores. The dependent variable was the North Carolina Math I assessment score, which is also interval data. Student scores on each Math I benchmark were calculated as the percentage of points earned to the number of possible points. The North Carolina Math I EOC assessment score is a scale score calculated by NCDPI. NCDPI (2013b) provides the information depicted in Figure 4 with regard to scale scores.



North Carolina Statewide Testing Program
Raw Scores by Achievement Level
End-of-Course Mathematics General Edition 4 (effective 2013-14)

3/6/2014

Course	Achievement Level	Scale Score	Minimum Number Correct On Form	Number of Questions	Approximate Percent Correct On Form
Math I	Lowest 2	244	16	49	33%
	Lowest 3	250	21		43%
	Lowest 4	253	25		51%
	Lowest 5	264	38		78%

NOTES:

1. The information contained within this document should be interpreted with caution. These tests were not designed to set a passing score based on a certain percentage of items having to be answered correctly across all forms of the tests.
2. All test forms for a content area and grade are statistically equated and are balanced for content.
3. The academic achievement standards (cut scores) were determined by panels of North Carolina teachers evaluating actual test scores and test items.

Figure 4. North Carolina Math I EOC Scale Scores.

As observed in Figure 4, there is no scale score for achievement Level 1. It is implied that a scale score below 244 results in an achievement Level 1. The actual raw score to scale score conversion for each version of the North Carolina Math I EOC assessment (NCDPI, 2013b), including Level 1 scale scores, is available in the Appendix. The extraneous variables that could impact the results include commitment of the teacher to teach the curriculum in the manner it was developed. Overall, utilizing these data analysis tools, with the instrumentation described, is justifiable because the goal of this study was to determine a way to predict all levels of student performance on the North Carolina Math I EOC assessment based on their benchmark scores.

Limitations of the Study

There were several limitations, or influences the researcher cannot control, of this study. The researcher is a math teacher in the district of the study. The type of questions on the benchmarks and state assessment in North Carolina Math I are objective, reducing the impact of subjective scoring inconsistencies. Assessments containing subjectively scored responses such as English II may not be a viable fit for the methodology utilized in this study. Further, the study was limited to a single district with locally developed benchmark assessments which should not be generalized to other settings.

Delimitations of the Study

One of the delimitations for this study, which are boundaries set by the researcher, is that North Carolina Math I was selected as the primary focus, rather than including all three EOC tested areas (Math I, Biology, and English II) in order to analyze the specific impact of the North Carolina Math I benchmark predictability for North Carolina Math I EOC assessment performance. The results of this particular study are not assumable for other content areas. In addition, demographic data and teacher data were not considered as factors in this study because the study focused on determining the predictability of the North Carolina Math I benchmark regardless of student race or ethnicity or the teacher-student assignment. Furthermore, the study focused on proficiency, not growth, for the data analysis.

Conclusion

This quantitative study evaluated if the locally developed Math I benchmark assessments are accurate predictors of student performance on the Math I EOC assessment. It also optimized cut-off scores on the Math I benchmarks for indicators of pass versus not pass performance on the Math I EOC assessment. For this study, the participants were all students in the East County school district taking Math I during the 2016-2017 school year. The data were collected from the data management system through the East County Testing and Accountability department and analyzed using correlational and regression analysis as well as 5-number summary statistics and ROC curve analysis which are detailed in Chapter 4.

Chapter 4: Results

Introduction

This quantitative research study investigated if student performance on East County Math I benchmarks was an accurate predictor of student performance on the North Carolina Math I EOC state assessment. It utilized correlational analysis, regression analysis, ROC curve analysis, and SDT to determine a cut-off score for each Math I benchmark administration. In this chapter, the results of the study are thoroughly described to investigate the following research questions.

1. How accurately does student performance on the Math I benchmark assessments for eighth grade, year-long, and semester Math I courses predict student performance on the North Carolina Math I EOC assessment as measured by student achievement?
    a. The independent variables for this study were the scores for the two administrations of the Math I benchmark assessment.
    b. The dependent variable was the standardized Math I EOC assessment scores for all students enrolled in Math I, in Grades 8-12, in a rural county in North Carolina.
2. To what degree can the sensitivity and specificity of the Math I benchmark assessments for eighth grade, year-long, and semester Math I courses be determined by analyzing the cut-off scores for the Math I benchmark assessment?
    a. The independent variable is the benchmark cut-off score.
    b. The dependent variable is the sum of the sensitivity and specificity values associated with each cut-off score.

This study focused on individual student performance, defined as achievement levels, on the North Carolina Math I EOC, and did not consider student growth or teacher growth. This chapter details the results of this quantitative research study by explaining the methods and procedures, the data screening, the descriptive statistics analysis, the linear regression models, and the ROC curve analysis.

Methods and Procedures

Descriptive statistics were employed to describe the characteristics of the six East County Math I benchmark assessments as well as the North Carolina Math I EOC. A linear regression model was calculated for each benchmark with the benchmark score as the independent variable and the North Carolina Math 1 EOC score as the dependent variable. The linear regression models were then used to calculate a predicted North Carolina Math I EOC assessment score for each individual benchmark score. From these calculations, SDT using ROC analysis was applied to determine cut-off scores for each benchmark, maximizing sensitivity and specificity. The Math I benchmark cut-off score identified through this statistical process is the indicator by which a student is predicted to be successful or unsuccessful on the North Carolina Math I EOC. A successful indicator on the North Carolina Math I EOC assessment is identified as a 250 or higher scale score corresponding to Level 3: Sufficient Command of Knowledge and Skills, categorized as meeting on-grade-level proficiency standards (NCDPI, 2016b). Although Level 3 does not meet North Carolina Standards for College and Career Readiness, Level 3 is considered grade-level proficient (NCDPI, 2016b); however, a Level 4, corresponding to a minimum scale score of 253, indicates success on the North Carolina Math I EOC assessment under the North Carolina Standards for College Career Readiness (NCDPI, 2013b).

Data Screening

East County Math I benchmark and North Carolina Math I EOC assessment data were provided by the East County Executive Director of Instructional Services. The 618 data records were de-identified and provided electronically in an Excel spreadsheet to the researcher. The records of students who did not take the North Carolina Math I EOC but did take one or more of the Math I benchmarks were excluded from the dataset provided by the Executive Director of Instructional Services. The reason they were not included is because the data were queried by North Carolina Math I EOC assessment score. After screening, the data from 524 student scores were used for analysis.

The elimination of unusable data is described as follows: 58 student records were removed for lack of any benchmark scores; 28 student records were removed because they represented students who repeated the course; and seven student records were removed because the data indicated only one portion of the benchmark was taken, either calculative active or calculator inactive. One student score was removed as an outlier due to a Z-score of -3.68, in Benchmark 5, representing a value of greater than three standard deviations from the mean (Creswell, 2013).

For organization purposes, Table 5 indicates the association of each benchmark in East County.

Table 5

Benchmark Association

Benchmark	Association
Benchmark 1	high school semester Math I, first benchmark
Benchmark 2	high school semester Math I, second benchmark
Benchmark 3	high school year-long Math I, first benchmark
Benchmark 4	high school year-long Math I second benchmark
Benchmark 5	eighth grade year-long Math I benchmark, first administration
Benchmark 6	eighth grade year-long Math I benchmark, second administration

As indicated in Table 5, Benchmark 1 and Benchmark 2 corresponded to the 2016-2017 Math I semester course. Benchmark 3 and Benchmark 4 corresponded to the 2016-2017 year-long Math I course. Benchmark 5 and Benchmark 6 corresponded to the 2016-2017 eighth grade year-long Math I course.

The normality of the data for each individual benchmark was determined by analyzing the skewness and kurtosis of the benchmark scores. Table 6 indicates the skewness and kurtosis for each benchmark administration.

Table 6

Benchmark Data Normality

Benchmark	1	2	3	4	5	6
Skewness	.46	.38	.12	.38	.07	-.25
Kurtosis	.30	.19	-.42	.32	-.44	-.33

As evidenced by the normality presented in Table 6, all skewness and kurtosis data fell between -1 and +1. According to McNeese (2016), skewness and kurtosis values between -1 and +1 indicate acceptable data normality. Skewness describes the lack of symmetry of a data distribution on the right or left of the center point (NIST/SEMATECH, 2013). Kurtosis measures the weight of the tail (heavy tailed or light tailed) with respect to the normal distribution (NIST/SEMATECH, 2013). Initially, the skewness and kurtosis for Benchmark 5 were .38 and .93 respectively. Once the Z-scores were calculated for each data point, an outlier was identified (Z-score of -3.68) and this data point was removed, reflecting an improved skewness and kurtosis of .07 and -.44 respectively.

Z-scores for the minimum and maximum values of each Math I benchmark were calculated and outliers were identified. For data with unacceptable normality, any outliers beyond three standard deviations from the mean can be removed from the data set as discussed by Creswell (2013). Benchmarks 1, 2, and 4 each contained one data point identified as an outlier; however, because the normality of each of these benchmarks was within acceptable range, these outliers were not removed. Since each student score is independent of other student scores, the results of normality tests were reliable.

Z-scores for the minimum and maximum values on the North Carolina Math I EOC assessment were calculated. Two outliers (Z-score = 3.15) were identified; however, the skewness (.51) and the kurtosis (-.29) were both within an acceptable range as previously described. Consequently, neither of these outliers was removed.

Descriptive Statistics Analysis

A 5-number summary (minimum, first quartile, median, third quartile, and maximum) and other summary statistics were evaluated for each Math I benchmark assessment and the North Carolina Math I EOC assessment. Table 7 indicates the summary statistics for each benchmark and the North Carolina Math I EOC.

Table 7

Summary Statistics

Benchmark	1	2	3	4	5	6	EOC
Minimum	.03	.04	.14	.17	.23	.38	231
Quartile 1	.25	.24	.24	.29	.39	.63	240
Median	.31	.29	.31	.37	.45	.71	248
Quartile 3	.41	.36	.38	.43	.55	.85	255
Maximum	.69	.60	.55	.74	.71	1.00	279
Mean	.33	.31	.31	.37	.47	.73	248.52
Standard Error	.01	.01	.01	.01	.01	.02	.42
Mode	.31	.29	.31	.37	.45	.67	251
Standard Deviation	.11	.09	.09	.10	.10	.14	9.70
Sample Variance	.01	.01	.01	.01	.01	.02	94.00
Range	.66	.56	.41	.57	.48	.62	48
Count	423	384	149	151	97	54	524

As indicated in Table 7, the sample size (count) varies with each benchmark assessment; however, because the benchmarks are not being compared to each other, the different sample sizes are not relevant. The minimum scores for Benchmark 1 and Benchmark 2 were extremely low. The maximum score for Benchmark 6 indicated a perfect score of 1.00 or 100%. Also, as evidenced in Table 7, the North Carolina Math I EOC summary statistics are based on scale scores, while the benchmark scores are based on percentages. Further, the standard error for the benchmark data and the North Carolina Math I EOC data is low.

The Pearson Product Moment Correlation (*r*) was used to determine the strength between the Math I benchmark scores and the Math I EOC assessment scores. Table 8 provides the Pearson Product Moment Correlation (*r*) for each benchmark and the North Carolina Math I EOC.

Table 8

Pearson Product Moment Correlation (r)

Benchmark	1	2	3	4	5	6
Pearson Correlation (*r*)	.56*	.53*	.32*	.54*	.63*	.56*
P	.00	.00	.00	.00	.00	.00

*Correlation is significant at the .05 level (2-tailed).

As evidenced in Table 8, Benchmark 1, Benchmark 2, Benchmark 4, Benchmark 5, and Benchmark 6 all show a strong (.5-1.0) correlation to the North Carolina Math I EOC, while Benchmark 3 shows a moderate (.3-.5) correlation to the North Carolina Math I EOC assessment ("Pearson Product-Moment Correlation," 2013). The correlation between each benchmark and the North Carolina Math I EOC assessment was calculated to be significant at the 0.05 level for a two-tailed distribution. The *p* values were all less than .00. Consequently, the significance of each correlation justifies the calculations of linear regression models.

Research Question 1. The results in Table 8 answer the first research question: How accurately does student performance on the Math I benchmark assessments for eighth grade, year-long, and semester Math I courses predict student performance on the North Carolina Math I EOC assessment as measured by student achievement? The null hypothesis (H_o) for the first research question, individual Math I benchmark scores are not accurate predictors of student performance on the North Carolina Math I EOC assessment, is rejected based on the strength and significance of the correlation coefficients between each benchmark and the North Carolina Math I EOC. A moderate to strong predictive relationship exists between the East County Math I benchmark assessments and the North Carolina Math I EOC.

Linear Regression

The first research question was examined through the predictive model generated for each benchmark using linear regression from scatterplots. According to Urdan (2010), linear regression "is a statistical technique that is very closely related to correlation" (p. 146). Urdan explained that regression analysis produces an equation for predicting the value of one variable, given the value of the second variable. Figure 5 displays the scatterplot data that were used to calculate the linear regression equation for Benchmark 1.

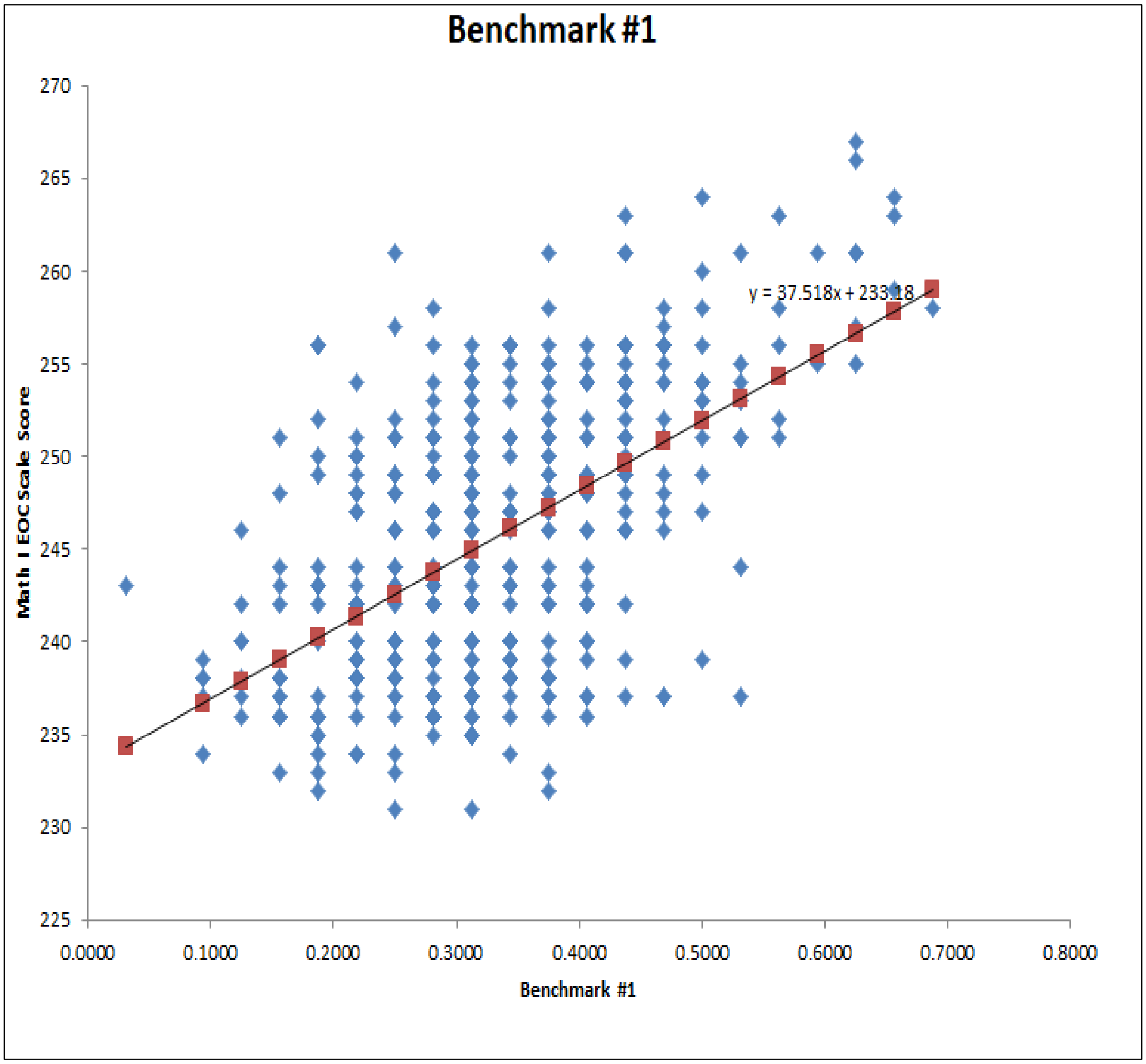


Figure 5. Benchmark 1 Scatterplot.

Figure 5 illustrates the positive linear regression characteristic evident between East County Math I Benchmark 1 and the North Carolina Math I EOC assessment. Benchmark 1 was the first benchmark administered to the semester Math I students. The linear regression line approximates the trend of the data by passing through the center of the scatterplot data points. The positive trend indicates that as Math I Benchmark 1 assessment scores increase, North Carolina Math I EOC assessment scores increase.

Similarly, Figure 6 indicates the scatterplot data for Benchmark 2, the second benchmark administered to the students taking semester Math I.

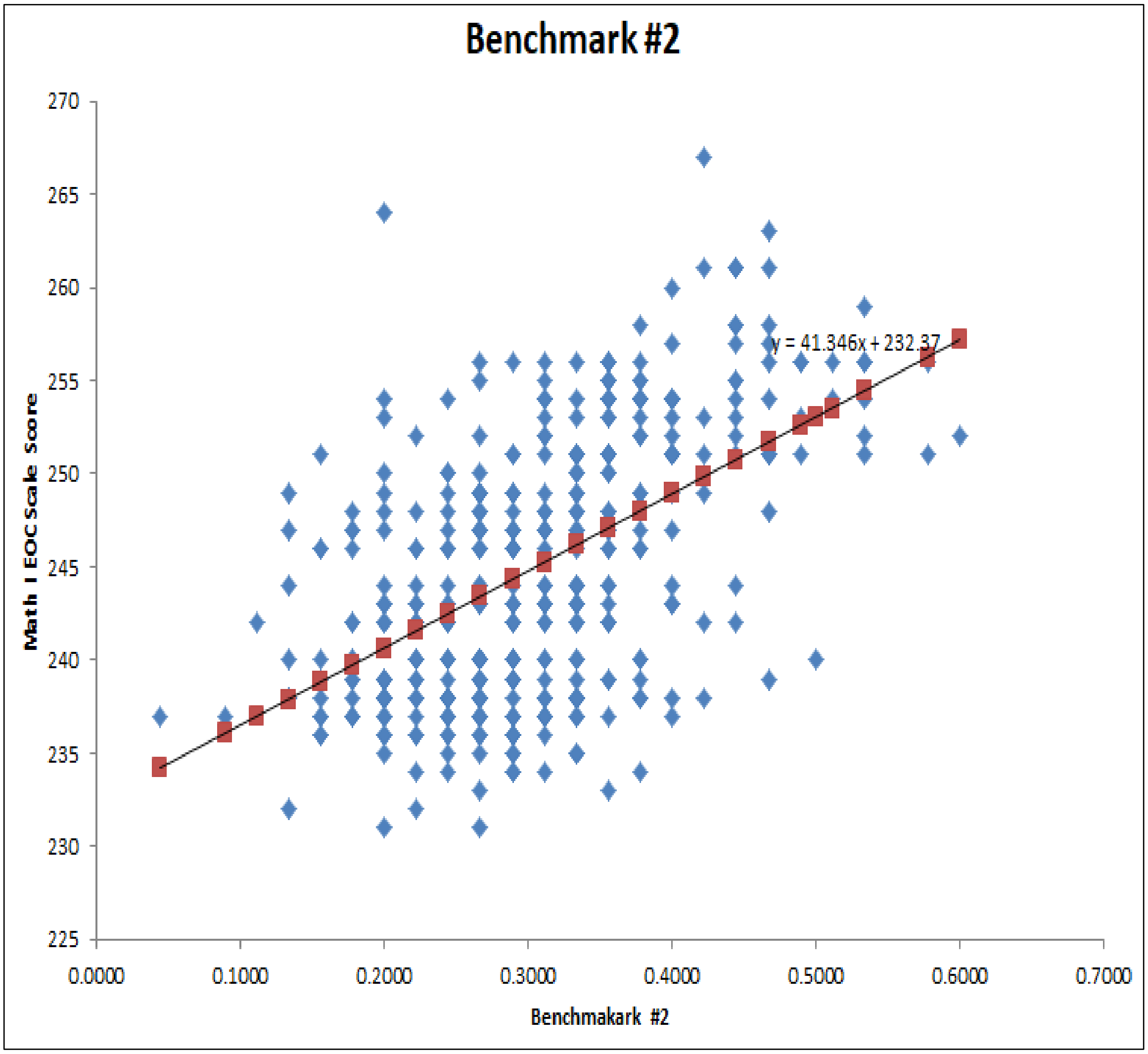


Figure 6. Benchmark 2 Scatterplot.

Figure 6 illustrates the positive linear regression characteristic evident between East County Math I Benchmark 2 and the North Carolina Math I EOC assessment. Figure 6 contains 39 fewer data points than Figure 5 because of the data screening process previously described; however, the linear regression line in Figure 6 indicates a similar trend by passing through the center of the scatterplot data points. The positive trend indicates that as Math I Benchmark 2 assessment scores increased, North Carolina Math I EOC assessment scores increased. The scatterplot for the first benchmark administered to the year-long high school Math I students is visible in Figure 7.

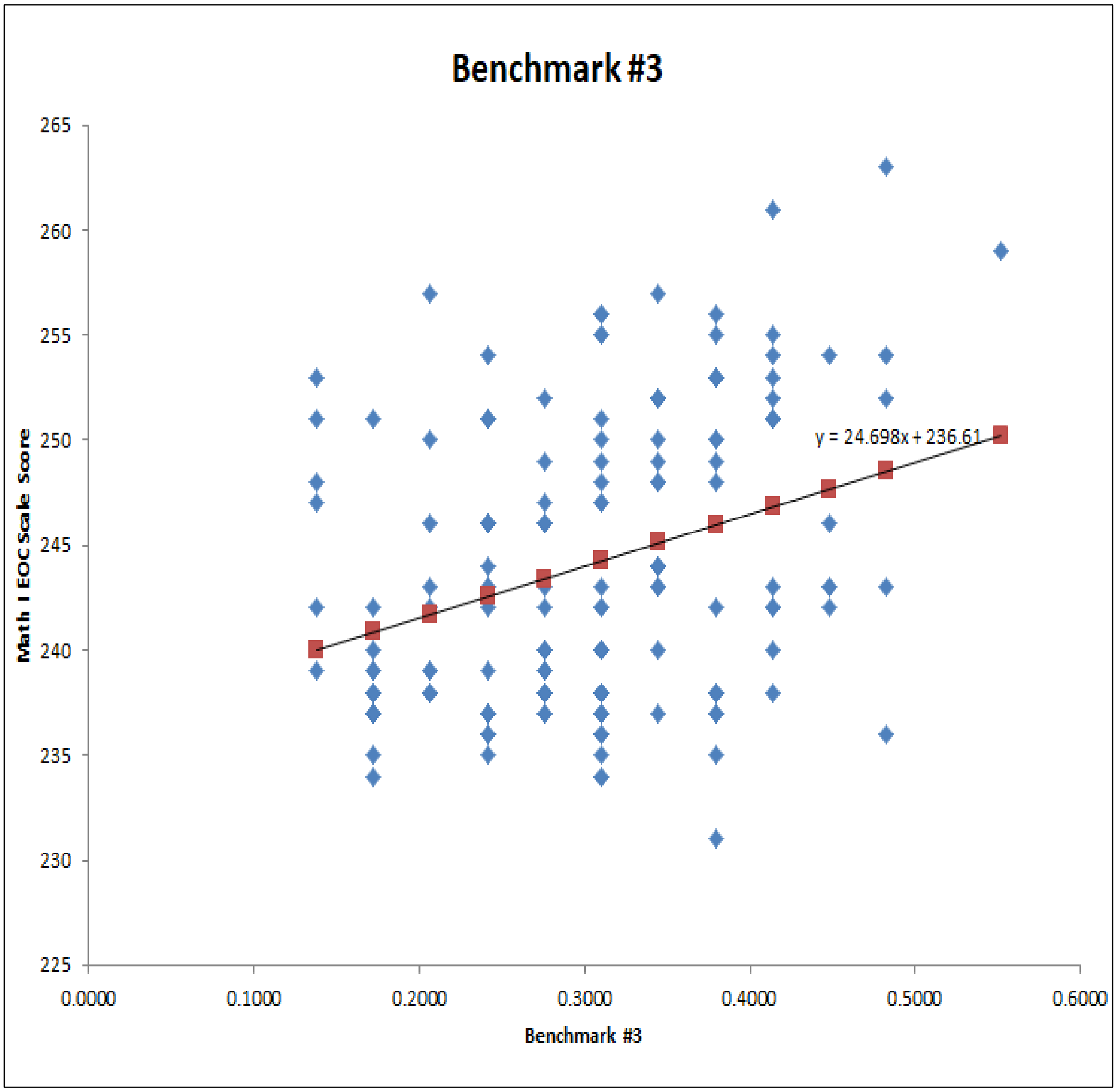


Figure 7. Benchmark 3 Scatterplot.

Figure 7 illustrates the positive linear regression characteristic evident between East County Math I Benchmark 3 and the North Carolina Math I EOC assessment.

Figure 7 contains 149 data points. The linear regression line in Figure 7 indicates a positive trend by passing through the center of the scatterplot data points. Further, the positive trend indicates that as Math I Benchmark 3 assessment scores increased, North Carolina Math I EOC assessment scores increased. The scatterplot for the second benchmark administered to the year-long high school Math I students is visible in Figure 8 as Benchmark 4.

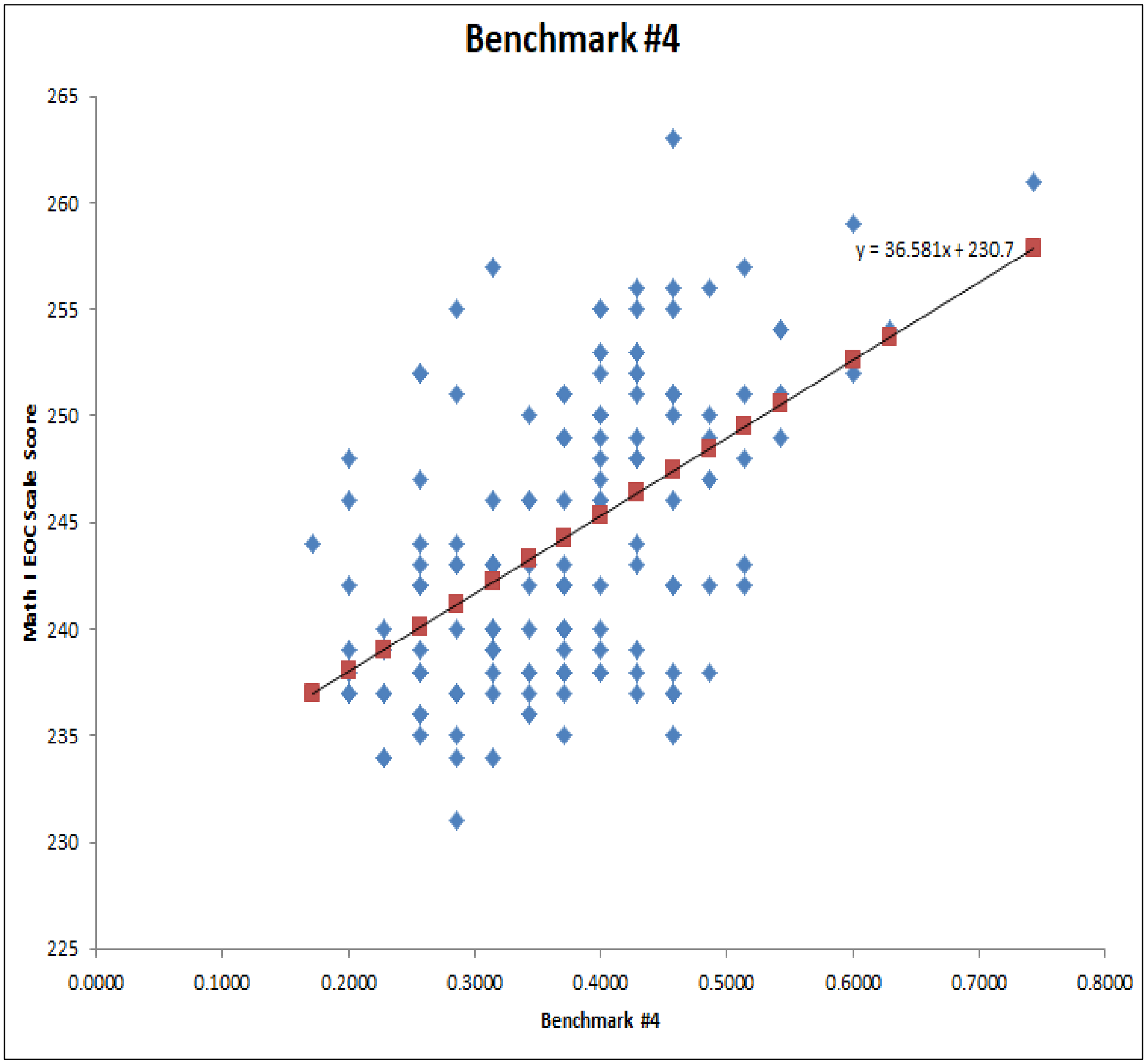


Figure 8. Benchmark 4 Scatterplot.

Figure 8 illustrates the positive linear regression characteristic evident between

East County Math I Benchmark 4 and the North Carolina Math I EOC assessment. Figure 8 contains two more data points than Figure 7 due to the data screening process previously described; however, the linear regression line in Figure 8 indicates a similar trend by passing through the center of the scatterplot data points. The positive trend indicates that as Math I Benchmark 4 assessment scores increased, North Carolina Math I EOC assessment scores increased. The scatterplot for the first benchmark administered to the eighth grade year-long Math I students is visible in Figure 9.

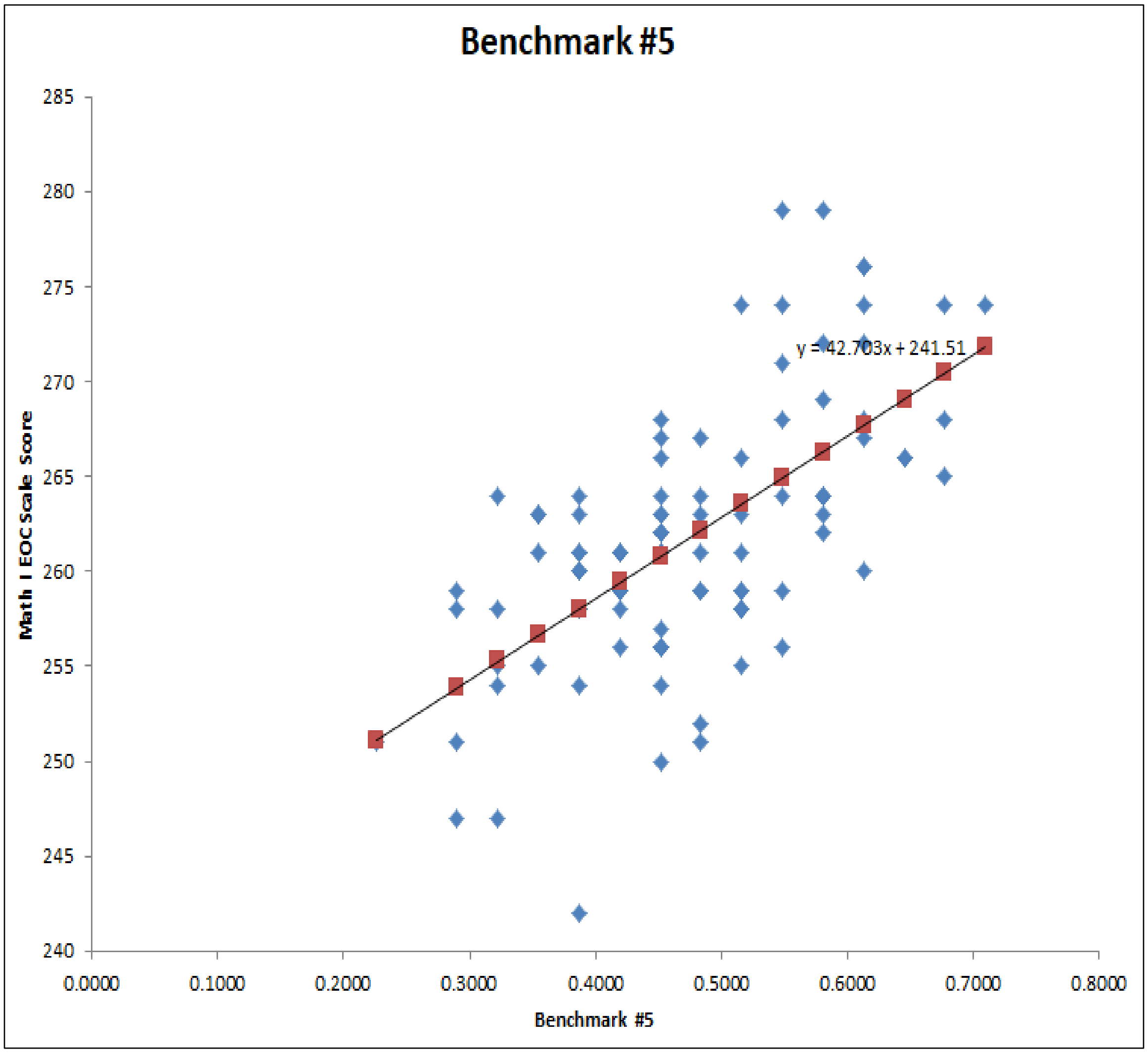


Figure 9. Benchmark 5 Scatterplot.

Figure 9 illustrates the positive linear regression characteristic evident between East County Math I Benchmark 5 and the North Carolina Math I EOC assessment. Figure 9 contains 97 data points. The linear regression line in Figure 9 indicates a positive trend by passing through the center of the scatterplot data points. Further, the positive trend indicates that as Math I Benchmark 5 assessment scores increased, North Carolina Math I EOC assessment scores increased. The scatterplot for the second benchmark administered to the eighth grade year-long Math I students is visible in Figure 10.

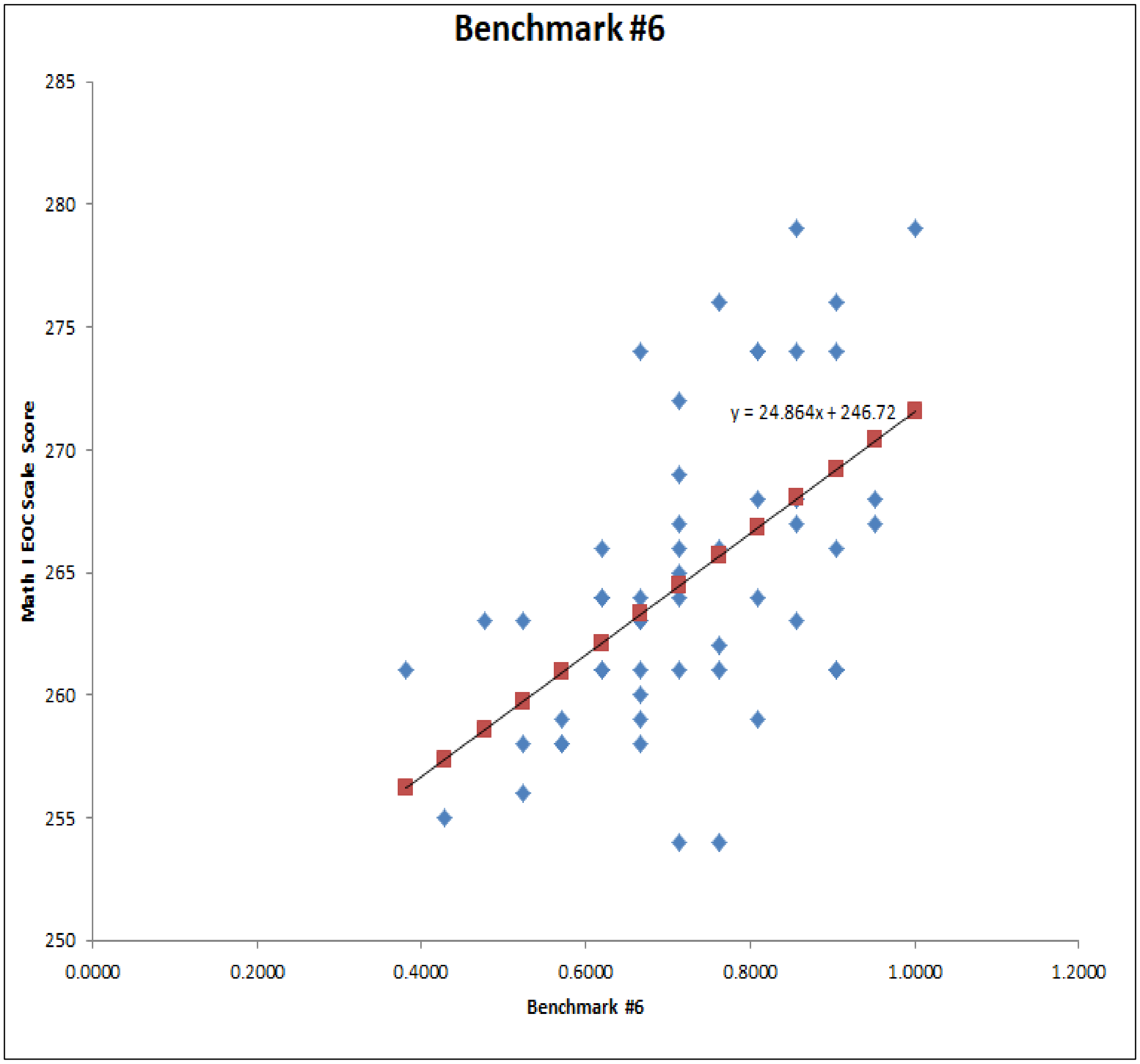


Figure 10. Benchmark 6 Scatterplot.

Figure 10 illustrates the positive linear regression characteristic evident between East County Math I Benchmark 6 and the North Carolina Math I EOC assessment. Figure 10 contains 54 data, 43 less data points than Figure 9, because of the data screening process previously described; however, the linear regression line in Figure 10 indicates a similar trend by passing through the center of the scatterplot data points. The positive trend indicates that as Math I Benchmark 6 assessment scores increase, North Carolina Math I EOC assessment scores increase.

The scatterplots for each of the East County Math I benchmark assessments yielded a predictive linear regression equation. As previously discussed, linear regression equations allow for the calculation of one variable based on the input value of a known second variable (Urdan, 2010). For the purposes of this study, the known input variable was the benchmark score. The benchmark score was used to predict the North Carolina Math I EOC assessment score which was compared to the actual North Carolina Math I EOC assessment score. The linear regression equations for each benchmark are presented in Table 9.

Table 9

Linear Regression Equations

Benchmark	Linear Regression Equations
Benchmark 1	y = 37.52x + 233.18
Benchmark 2	y = 41.35x + 232.38
Benchmark 3	y = 24.70x + 236.61
Benchmark 4	y = 36.58x + 230.70
Benchmark 5	y = 42.70x + 241.51
Benchmark 6	y = 24.86x + 246.72

The linear regression equations in Table 9 use x (independent variable) to represent the Math I benchmark score and y (dependent variable) to represent the North Carolina Math I EOC assessment score. Given a score for a specific benchmark, the associated linear regression equation is used to predict the North Carolina Math I EOC assessment score. The predicted North Carolina Math I EOC assessment score and the actual North Carolina Math I EOC assessment score are necessary to determine the sensitivity and specificity for the ROC analysis.

ROC

The second research question explores the degree to which the sensitivity and specificity can be optimized through SDT and ROC analysis to determine cut-off scores for each Math I benchmark. For the purposes of this study, SDT was used to categorize student benchmark scores and North Carolina Math I EOC assessment scores in one of the following categories: predicted to pass the EOC assessment and actually passed, predicted to pass the EOC assessment and actually failed, predicted to fail the EOC assessment and actually failed, or predicted to fail the EOC assessment and actually passed. ROC curve analysis uses the sensitivity and specificity of these four categories to improve the relationship between the predicted performance on the North Carolina Math I EOC assessment and the actual performance. Figure 11 illustrates the use of SDT for Math I Benchmark 1 based on the linear regression model for this benchmark.

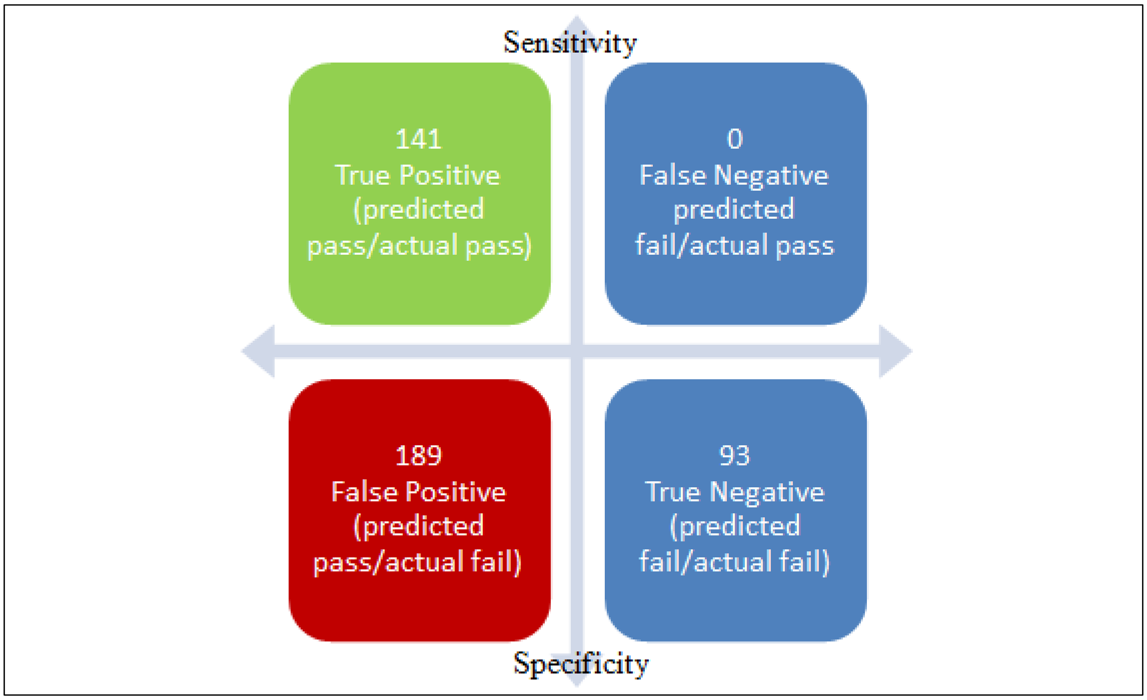


Figure 11. Benchmark 1 SDT.

As evidenced in Figure 11, the linear regression predictive model for Math I Benchmark 1 yielded 189 false positives in which students were predicted to pass the

North Carolina Math I EOC assessment with a scale score of 250 (Level 3) or greater. This prediction was based on a benchmark performance of .45 or greater and yielded an accuracy of 55%. Sensitivity and specificity of 70-80% accuracy is considered high (Statistics How To, 2018). To improve the sensitivity and specificity, the ROC curve for optimizing the sensitivity and specificity by reducing the occurrences of false positives for the benchmark indicator was generated. The ROC curve analysis for Benchmark 1 is visible in Figure 12.

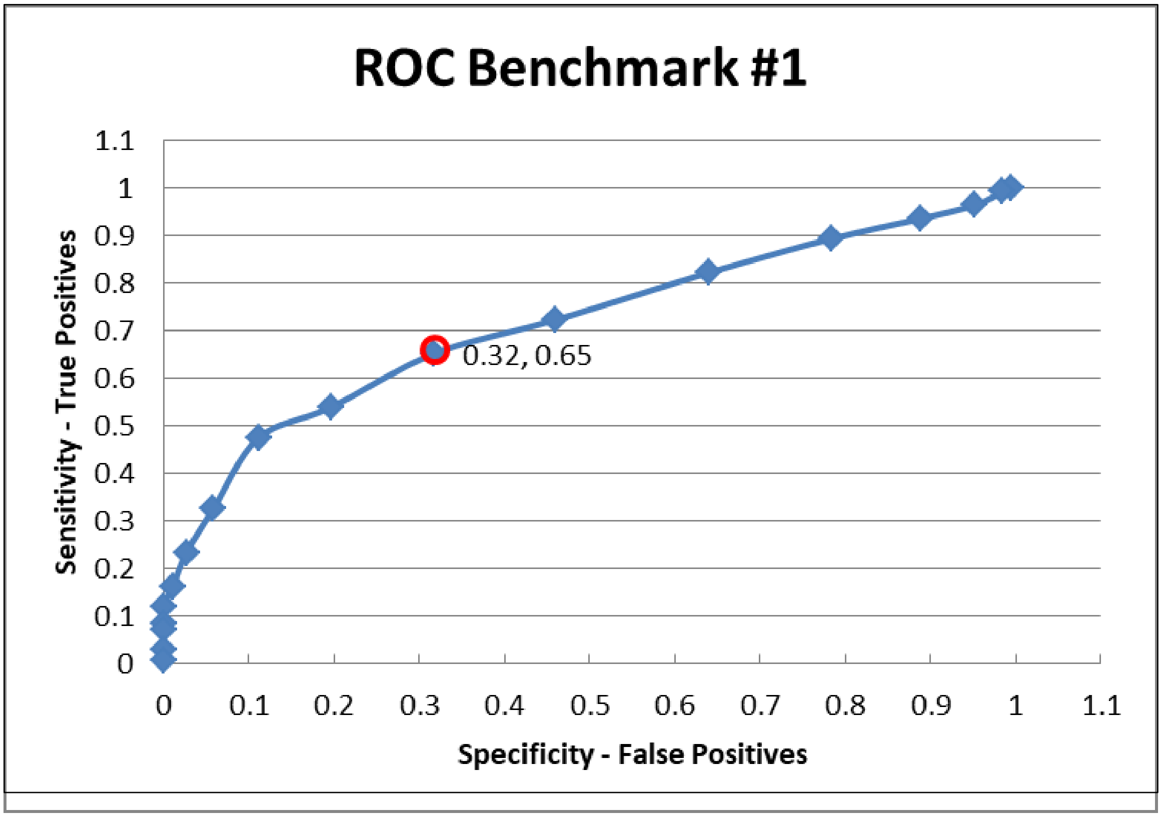


Figure 12. Benchmark 1 ROC Curve.

As illustrated in Figure 12, optimization of the sensitivity and specificity for Benchmark 1 occurs at (0.32, 0.65) corresponding to a benchmark indicator cut-off of 0.38 for passing the North Carolina Math I EOC. Upon this adjustment to the benchmark cut-off using the ROC curve analysis, the sensitivity and specificity accuracy improved to 74% from 55% for this benchmark. This analysis process was repeated for East County Math I Benchmark 2. Figure 13 illustrates the use of SDT for Math I Benchmark 2, based on the linear regression model for this benchmark.

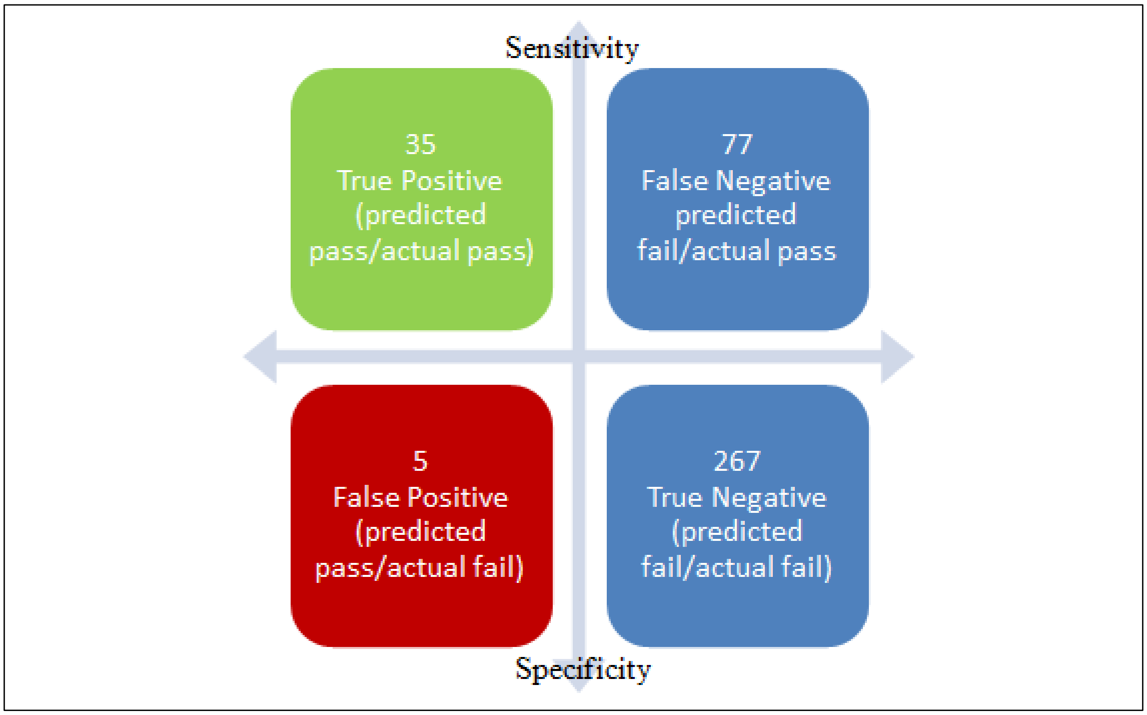
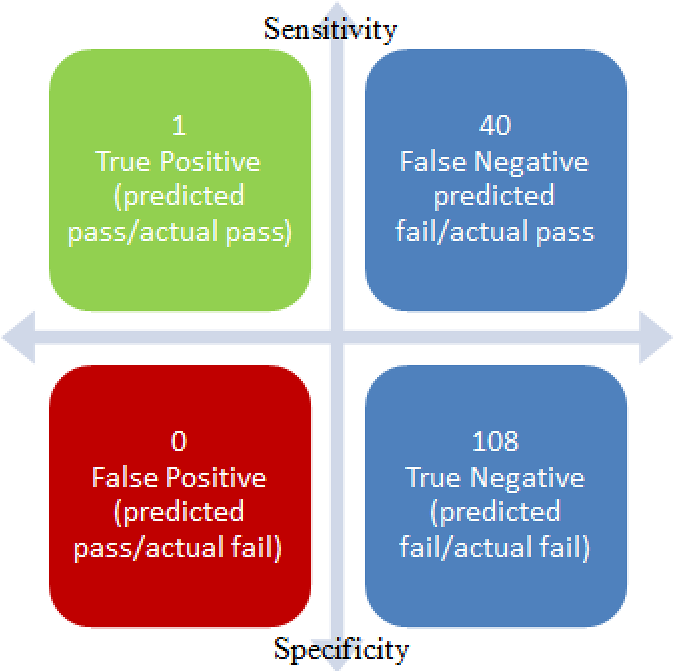


Figure 13. Benchmark 2 SDT.

As evidenced in Figure 13, five false positives occur in the linear regression predictive model for East County Math I Benchmark 2. The 35 true positive and 267 true negative values indicate an accuracy of 79% with a benchmark cut-off value of 0.43. Consequently, the optimization of the threshold for this benchmark is accepted without further adjustment using an ROC curve. This same process was repeated for East County Math I Benchmark 3. Figure 14 illustrates the use of SDT for Math I Benchmark 3 based on the linear regression model for this benchmark.



Sensitivity

1
True Positive
(predicted
pass/actual pass)

40
False Negative
predicted
fail/actual pass

0
False Positive
(predicted
pass/actual fail)

108
True Negative
(predicted
fail/actual fail)

Specificity

Figure 14. Benchmark 3 SDT.

As evidenced in Figure 14, no false positives occur in the predictive linear regression model for East County Math I Benchmark 3. The one true positive and 108 true negative values indicate an accuracy of 73% with a benchmark cut-off value of 0.54. Consequently, the optimization of the threshold for this benchmark is accepted without further adjustment using an ROC curve. This same process was repeated for East County Math I Benchmark 4. Figure 15 illustrates the use of SDT for Math I Benchmark 4 based on the linear regression model for this benchmark.

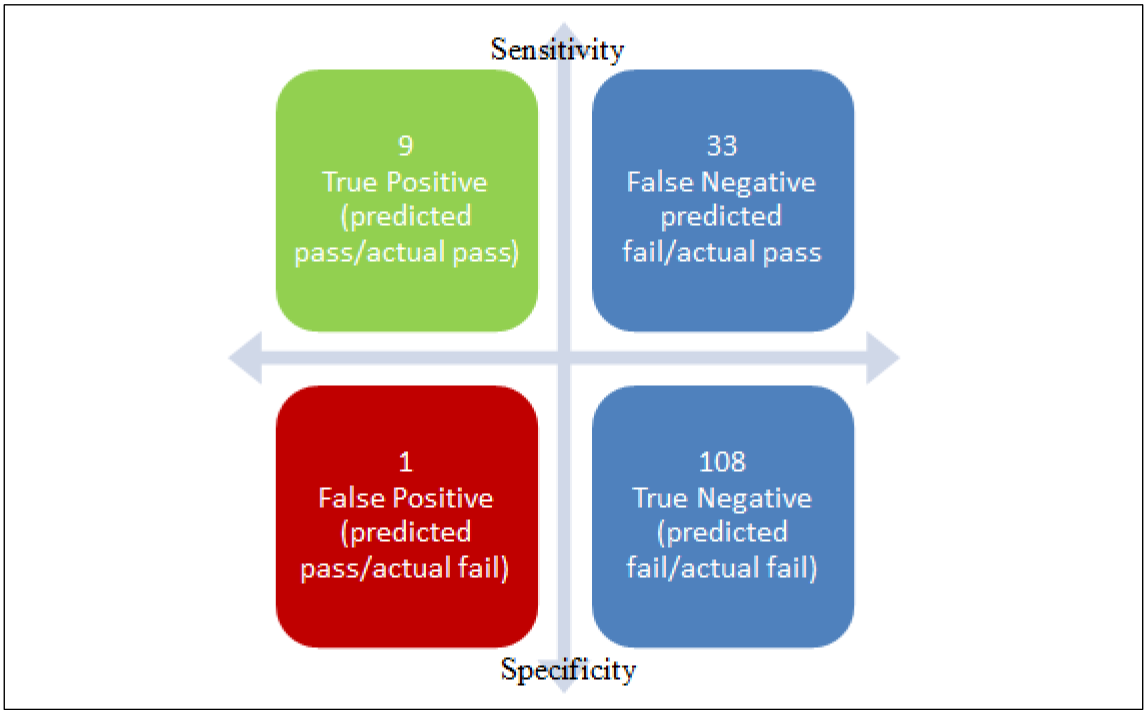


Figure 15. Benchmark 4 SDT.

As evidenced in Figure 15, one false positive occurs in the predictive linear regression model for East County Math I Benchmark 4. The nine true positive and 108 true negative values indicate an accuracy of 78% with a benchmark cut-off value of 0.54. Consequently, the optimization of the threshold for this benchmark is accepted without further adjustment using an ROC curve. This same process was repeated for East County Math I Benchmark 5. Figure 16 illustrates the use of SDT for Math I Benchmark 5 based on the linear regression model for this benchmark.

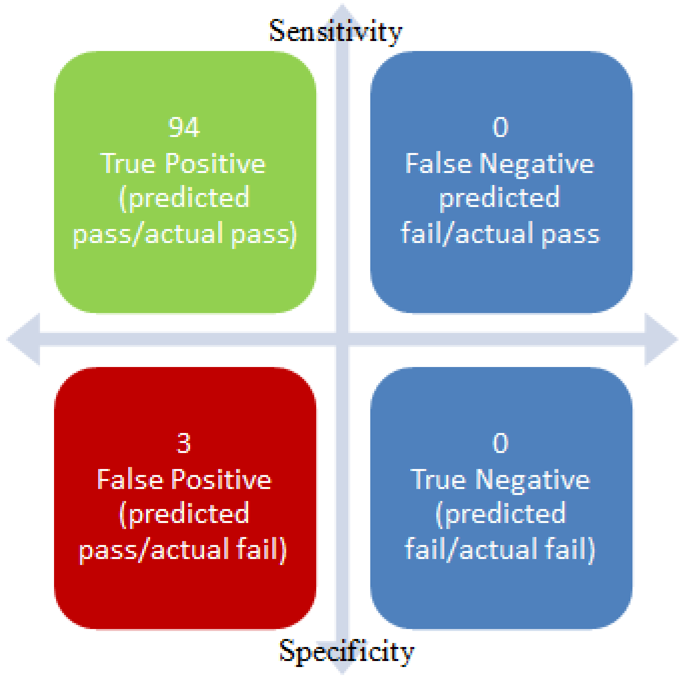
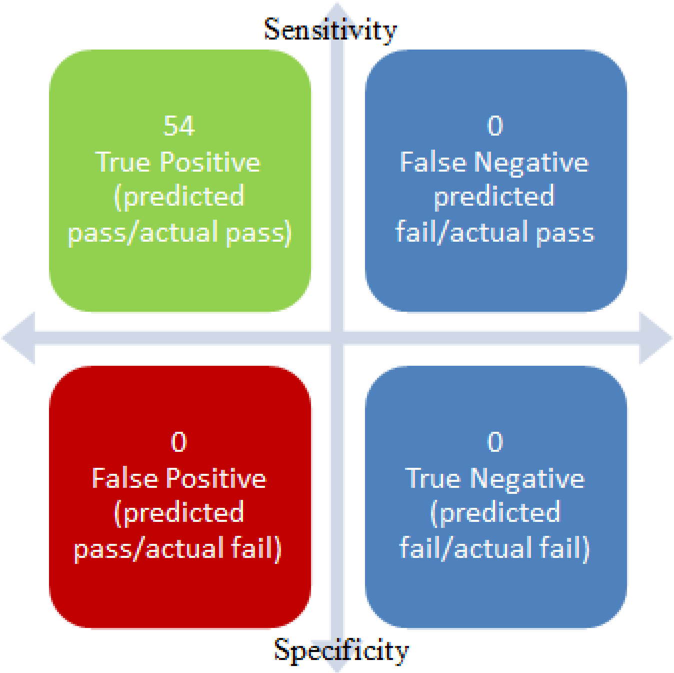


Figure 16. Benchmark 5 SDT.

As evidenced in Figure 16, three false positives occur in the predictive linear regression model for East County Math I Benchmark 5. The 94 true positive and zero true negative values indicate an accuracy of 97% with a benchmark cut-off value of 0.20. Consequently, the optimization of the threshold for this benchmark is accepted without further adjustment using an ROC curve. This same process was repeated for East County Math I Benchmark 6. Figure 17 illustrates the use of SDT for Math I Benchmark 6 based on the linear regression model for this benchmark.


Sensitivity

54
True Positive
(predicted
pass/actual pass)

0
False Negative
predicted
fail/actual pass

0
False Positive
(predicted
pass/actual fail)

0
True Negative
(predicted
fail/actual fail)

Specificity

Figure 17. Benchmark 6 SDT.

As evidenced in Figure 17, zero false positives occur in the predictive linear regression model for East County Math I Benchmark 6. The 54 true positive and zero true negative values indicate an accuracy of 100% with a benchmark cut-off value of 0.13. Consequently, the optimization of the threshold for this benchmark is accepted without further adjustment using an ROC curve.

Research Question 2. The SDT analysis and the ROC curve analysis provide the answer to Research Question 2. The sensitivity and specificity of the Math I benchmark assessments for eighth grade, year-long, and semester Math I courses can be determined to a degree of greater than 70% accuracy for each benchmark. In fact, the range of accuracy is 73% on Benchmark 3 to 100% on Benchmark 6 using cut-off scores for sensitivity and specificity.

Conclusion

Chapter 4 reported the results of this research study on the six East County Math I Benchmarks. The results included a description of the methods and procedures, the data screening, the descriptive statistics analysis, the linear regression models, and the ROC curve analysis. Each benchmark demonstrated a strong correlation to the North Carolina Math I EOC assessment with the exception of Benchmark 3. The sensitivity and specificity for eighth grade, year-long, and semester Math I courses was determined by analyzing the cut-off scores for each East County Math I benchmark assessment and calculated to be greater than 70% accurate for each benchmark. The discussion of these findings is detailed in Chapter 5.

Chapter 5: Discussion

Introduction

The purpose of this study was to acknowledge and analyze two concerns regarding the use of benchmark assessments in East County North Carolina Math I courses. First, this study investigated how accurately individual student performance on North Carolina Math I benchmarks in East County was predictive of individual student performance on the North Carolina Math I EOC assessment. In addition, this study utilized SDT to determine a cut-off score on the Math I benchmarks for indicating pass/fail on the Math I EOC assessment. This study was correlational in design. A 5-number summary (minimum, first quartile, median, third quartile, and maximum) and summary statistics were evaluated for benchmark assessments and the EOC assessment. Pearson Product Moment Correlation (*r*) was used to determine the strength between the Math I benchmark scores and the Math I EOC assessment scores. In addition, SDT utilizing an ROC curve analysis was used to optimize the predictive cut-off scores on the benchmarks for pass versus fail on the Math I EOC assessment. As previously described, SDT is a decision-making process grounded in uncertainty and based on statistical techniques (Green & Swets, 1966). ROC is the graphical representation of the sensitivity versus the specificity (Green & Swets, 1966).

Identifying the cut-off score for each benchmark that corresponded to a predictive proficiency score enhanced the sensitivity and specificity of the East County Math I benchmark assessments for predicting student performance on the North Carolina Math I EOC assessment. This process provided a score for each benchmark that can be used to group students for remediation as needed. As described by Black et al. (2003), benchmark performance as a formative assessment can be used to identify students who require timely instructional interventions.

Summary of the Results

The results of this research study are summarized by the research questions.

Research Question 1: How accurately does student performance on the Math I benchmark assessments for eighth grade, year-long, and semester Math I courses predict student performance on the North Carolina Math I EOC assessment as measured by student achievement? The data indicated that student performance on the Math I benchmarks does accurately predict student performance on the North Carolina Math I EOC assessment. This conclusion is based on the strength and significance of the correlations. Table 10 displays the benchmark and North Carolina Math I EOC assessment correlations and the interpretations of strength at the 95% confidence level.

Table 10

Benchmark Correlations and Strength Interpretations

Benchmark	Correlation	Strength
Benchmark 1	.56*	strong
Benchmark 2	.53*	strong
Benchmark 3	.32*	moderate
Benchmark 4	.54*	strong
Benchmark 5	.63*	strong
Benchmark 6	.56*	strong

*Correlations were calculated using a 95% confidence interval.

As illustrated in Table 10, five of the six benchmarks and the North Carolina Math I EOC assessment demonstrated strong correlations ranging from r =.53 to r = .63 with one correlation considered moderate (Benchmark 3, r = .32). In addition, each correlation was found to be significant at the 95% confidence interval which supported developing predictive linear regression models for each benchmark and the North Carolina Math I EOC. The linear regression models worked well as a predictive tool for implementing SDT.

Research Question 2: To what degree can the sensitivity and specificity of the Math I benchmark assessments for eighth grade, year-long, and semester Math I courses be determined by analyzing the cut-off scores for the Math I benchmark assessment? Using SDT and ROC curve analysis, this study demonstrated that the sensitivity and specificity of the Math I benchmark assessments predictive ability can be optimized to greater than 70% accuracy which is considered high accuracy (Statistics How To, 2018). Using a scale score of 250 on the North Carolina Math I EOC assessment as the baseline for passing, a corresponding cut-off value for each Math I benchmark was determined. Benchmark 1 required the development of an ROC curve analysis because the original indications using the predictive linear regression model for the SDT yielded only 55% accuracy. An accuracy of 55% is not considered "high" (Statistics How To, 2018), which justified an adjustment. Upon the adjustment to the benchmark cut-off using the ROC curve analysis, the accuracy improved to 74%. Table 11 summarizes the passing benchmark cut-off values for each benchmark and the corresponding accuracy of using each benchmark score.

Table 11

Benchmark Cut-Off Values and Accuracy

Benchmark	Cut-off Value	Accuracy
Benchmark 1	.38	74%
Benchmark 2	.43	79%
Benchmark 3	.54	73%
Benchmark 4	.54	78%
Benchmark 5	.20	97%
Benchmark 6	.13	100%

Table 11 indicates the benchmark cut-off values calculated in this study for passing the North Carolina Math I EOC assessment and the corresponding accuracies for each cut-off value. Each benchmark is unique based on the course (Math I semester, Math I year-long, or eighth grade Math I) and the cut-off values are relative to each respective benchmark exclusively.

Findings

The findings from this quantitative research study support the following conclusion for the first research question: East County Math I benchmark assessments are accurate predictors of student performance on the North Carolina Math I EOC assessment. Consequently, the East County Math I benchmarks can allow educators to make informed decisions about positively impacting the instruction and learning of the North Carolina Math I curriculum prior to the EOC assessment.

The findings from this quantitative research study support the following conclusion for the second research question: The sensitivity and specificity of the East County Math I benchmark assessments can be determined with greater than 70% accuracy for each benchmark based on the identified cut-off scores.

Connections to Existing Research on Benchmark Testing

The results associated with Research Question 1 parallel the findings of previous studies analyzing the relationship between benchmark assessments and EOC assessments. The results showed that a moderate to strong correlation exists between East County Math I benchmarks and the North Carolina Math I EOC assessment, supporting similar results from Brown and Coughlin (2007). Further, this study replicates methods used in Thompson's (2016) study and the study by Anderson et al. (2010) utilizing SDT to determine a maximum cut-off score for mathematics formative assessments to predict student achievement on summative state assessments. This study, like the study by

Hintze and Silberglitt (2005), incorporated ROC curve analysis to determine benchmark cut-off scores. Hintze and Silberglitt recommended using ROC for districts that need to make multiple decisions, such as diagnostic or entitlement decisions. The research findings of Nese et al. (2011) also supported ROC curve analysis as a solid predictor of student performance on summative assessments based on benchmark performances. Thompson stated,

> The versatility of ROC analysis to identify benchmark cut-off scores is that it is computationally simple, easy to implement, and allows educators to emphasize a particular outcome. A district could use this method to identify the students who are most likely to have a specific outcome on the state assessment, with the intention of getting help to those students who need it or to accelerate those students who have mastered the assessed content. Similarly, if the district wishes to focus on the students who are most likely on the bubble between passing and not passing the state assessment, ROC analysis could be used to identify the borderline students who are most likely to affect a change in the schools' performance on the assessment. (p. 90)

The ROC analysis employed in this research study possesses beneficial characteristics for educators to optimize cut-off scores for reducing false positive results as Hintz and Silberglitt (2005) and Thompson (2016) indicated.

Implications for Student Assessment

According to Black and Wiliam (1998a), "Teachers need to know about their pupils′ progress and difficulties with learning so that they can adapt their own work to meet pupils′ needs -- needs that are often unpredictable and that vary from one pupil to another" (p. 2). North Carolina's Math I EOC assessment scores are available in an acceptable timeframe, sometimes the same day the exam is taken; however, because it is an EOC assessment, no opportunity exists for effective instructional adaptations. Benchmark assessments, however, provide timely results. Their value lies in the fact that the information provided can be immediately utilized to adapt or adjust instruction (Black & Wiliam, 1998a).

The implications for student assessment from this study quantify the value of the East County Math I benchmarks, assuming their continued implementation. Although the data utilized in this study were summative, the practical use of the information the benchmark assessment data provide has the greatest opportunity for improving instruction and student learning in the future. Black and Wiliam (1998b) recognized the implications for student assessment:

> Learning is driven by what teachers and pupils do in classrooms. Teachers have to manage complicated and demanding situations, channeling the personal, emotional, and social pressures of a group of 30 or more youngsters in order to help them learn immediately and become better learners in the future. Standards can be raised only if teachers can tackle this task more effectively. What is missing from the efforts alluded to above is any direct help with this task. (p. 1)

With the heavy emphasis on standardized testing in the United States, undeniable justification exists for classroom formative assessment in a suitable time interval for which instructional adjustments are still an option (Black & Wiliam, 1998a). In fact, Bailey and Jakicic (2012) stated, “While they offer many benefits, the primary goal of common formative assessments is to provide information about student learning and to identify which students are in need of additional time and support” (p. 73). Blankstein, Houston, and Cole (2010) summarized this point with the following:

> The job of the teacher, always, is to establish each student's current location on the continuum of learning and to help students progress to a higher level as quickly as possible, but at a pace consistent with achieving mastery and deep learning. It is also the duty of the teacher (and of the school) to give extra time and support to those at risk of not achieving that level of proficiency necessary to cope with the requirements of the next level of schooling, in order to reduce achievement gaps. (p. 35)

According to Payne (2013), when teachers and administrators are equipped with the confidence that the benchmark assessments are accurate predictors of student performance on the standardized summative assessments, they can use that knowledge to improve and personalize student learning in the classroom. Graham and Ferriter (2010) explained this initiative through the lens of PLCs:

> Collaborative data analysis is not easy. On one hand, it requires a pretty sophisticated look at numbers – a skill that does not come naturally to a lot of people. On the other hand, it means exposing professional successes and perceived shortcomings to ones colleagues and that is a difficult endeavor in even the safest environments. (p. 157)

However, as Black et al. (2003) described, implementing sound formative assessment practices and data analysis requires significant changes for most teachers. Further, Graham and Ferriter (2010) explained, "Any non-trivial change in classroom teaching involves the teacher both in taking risks and, at least during the process of change, in extra work" (p. 2).

This quantitative study has even more merit considering North Carolina's consolidated state plan for ESSA (2015). Under North Carolina's ESSA plan (NCDPI, 2017), Math 3 will become an EOC tested curriculum beginning in the 2018-2019 school year; however, Math I will continue to be an EOC tested curriculum as well. With these testing stakes increasing, the need for accurate, predictive benchmark assessments also increases.

The design and development of benchmark formative assessments is crucial to the predictive ability of the assessment. American Educational Research Association et al. (2014) provides guidelines for assessment design and development that include rigorous item development and review to ensure the validity and reliability of an assessment. These guidelines stress the importance of validity and reliability in the assessment development process and support the predictive ability of formative assessment performance to summative assessment performance (American Educational Research Association et al., 2014). Further, American Educational Research Association et al. (2014) stated the following regarding the validity of assessments:

> Not all tests are well-developed, nor are all testing practices wise or beneficial, but there is extensive evidence documenting the usefulness of well-constructed, well-interpreted tests. Well-constructed tests that are valid for their intended purposes have the potential to provide substantial benefits for test takers and test users. Their proper use can result in better decisions about individuals and programs than would result without their use and can also provide a route to broader and more equitable access to education and employment. (p. 1)

It follows from this statement that the importance of analyzing the correlation between the Math I benchmarks and the North Carolina Math I EOC assessment for predictive validity is evident.

This study supports the determinations from previous research that benchmark assessments can be used to accurately predict student performance on summative state assessments (Brown & Coughlin, 2007). In addition, this study has revealed that the economical East County Math benchmark assessments can produce accurate predictions of student performance. This fact is particularly important to economically challenged districts like East County.

Limitations of the Study

Several limitations, defined as influences the researcher could not control, were indicated. The researcher is a math teacher in the district where this study took place. This limitation was controlled by requesting that the data received from the Executive Director of Instructional Services be de-identified. Further, the study is limited to a single district with locally developed benchmark assessments and may not be easily generalized to other settings. In addition, the initial data size in this study was 618 North Carolina Math I EOC assessment scores; but after data screening, this number was reduced to 524 assessment scores. Consequently, the data size reduction made the study less representative of the actual student population. Furthermore, the type of questions on the benchmarks and state assessment in North Carolina Math I are objective, reducing the impact of subjective scoring inconsistencies. Assessments containing subjectively scored responses, such as English II, may not be a viable fit for the methodology utilized in this study.

Delimitations of the Study

Delimitations, defined as boundaries set by the researcher, also existed in this study. North Carolina Math I was selected as the primary focus, rather than including all three EOC tested areas (Math I, Biology, and English II). Since this study sought to investigate the link between East County Math I benchmark performance and North

Carolina Math I EOC assessment performance, grade level was not considered. In addition, demographic data and teacher data were not considered as factors in this study because the study focused on determining the predictability of the North Carolina Math I benchmark regardless of student race or ethnicity or teacher-student assignment. In addition, the study focused on proficiency, not growth.

Scholarly Significance

The significance of this research study rests solely upon the application of the findings. Black and Wiliam (1998a) stated,

> So changing teachers' practice cannot begin with an extensive program of training for all; that could be justified only if it could be claimed that we have enough "trainers" who know what to do, which is certainly not the case. The essential first step is to set up a small number of local groups of schools–some primary, some secondary, some inner-city, some from outer suburbs, some rural–with each school committed both to a school-based development of formative assessment and to collaboration with other schools in its local group. (p. 10)

Math I is the foundation for the subsequent high school math courses required by the state of North Carolina. The perpetual improvement of student learning through improved instruction and planning is directly driven by implementing research-based strategies. Both the teacher and the student will benefit from receiving knowledge about content mastery in Math I. As stated by Earl (2012),

> There is no doubt that teachers will have to work within systemic guidelines, but when the day-to-day work in classrooms is about learning, teachers go deeper and think about assessment and how they can intentionally move assessment so that it provides them and their students, with insights about what is being learned and

> how to move the learning forward better and faster. (p. 121)

Consequently, assessment is driven by the expectations for learning and reveals student understanding of the content.

Recommendations for Future Research

The results of this study yield multiple future research needs regarding the use of benchmark assessments, particularly in high school mathematics. More research pertaining to high school benchmarks in all content areas is necessary. A large portion of existing benchmark research is focused on the elementary school level, and it gradually decreases through the higher grade levels (Thompson, 2016). The push towards national-based assessments like the SAT, ACT, or WorkKeys for college admission or the job force requires research on how benchmarks can be utilized to predict and prepare students for their future endeavors.

This research study did not consider diversity; however, additional benchmark research that is representative of student diversity, specifically at the high school level, is warranted. According to Thompson (2016),

> Data derived from prediction models may also be used to identify whether certain subgroups of students are more likely to perform poorly on a Common Core based assessment, thus enabling teachers and schools to target resources towards lower performing subgroups in order to address learning deficiencies. Given the importance of closing achievement gaps among subgroups, the use of a method which predicts future performance similar to the method used in this study may provide justification for school improvement teams to develop strategies which target specific weaknesses within subgroups. Thus, the value of prediction models is not limited to anticipating how students will perform months down the

> road, but also for identifying students who need to receive extra help before they fall further behind. (p. 91)

Several studies have focused on small, low socioeconomic school districts that implement low cost, locally developed benchmarks. A larger magnitude study of multiple districts that use locally developed benchmarks may add justification to the utilization of these low cost alternative benchmark assessments.

The North Carolina EVAAS houses an enormous amount of valuable student and teacher educational data. While this data is not available for immediate instructional interventions as the benchmark data used in this study, research into the practical applications of the currently housed data, or common formative assessments, as a predictive element is warranted.

Conclusion

Chapter 5 provided a summary of the findings of this study and discussed connections to existing research on benchmark assessments. The findings from this quantitative research study support the following conclusion for the first research question: East County Math I benchmark assessments are accurate predictors of student performance on the North Carolina Math I EOC assessment. The findings from this quantitative research study support the following conclusion for the second research question: The sensitivity and specificity of the East County Math I benchmark assessments can be determined with greater than 70% accuracy for each benchmark based on the identified cut-off scores.

The connection to previous research on benchmark assessments was described. The results associated with Research Question 1 parallel the findings of previous studies analyzing the relationship between benchmark assessments and EOC assessments. The results showed that a moderate to strong correlation exists between East County Math I benchmarks and the North Carolina Math I EOC assessment, supporting similar results from Brown and Coughlin (2007). The results for Research Question 2 paralleled the findings from the ROC analysis conducted by Hintz and Silberglitt (2005) and Thompson (2016).

The implications of this study for student assessment were presented. This study revealed that the economical East County Math benchmark assessments can produce accurate predictions of student performance. This fact is particularly important to economically challenged districts like East County.

The limitations and delimitations of the methodology utilized to complete the study were described and recommendations for future research were posed. Further, the scholarly significance of the results of this study were discussed and indicated the research-based benefit of utilizing the information yielded by this study.

Finally, the recommendations were presented. More research pertaining to high school benchmarks in all content areas is necessary. Also, benchmark research that is representative of student diversity, specifically at the high school level, is warranted. Last, research focusing on the predictive value of North Carolina’s EVAAS data for instructional improvements is recommended.

References

Abdi, H. (2007). Signal detection theory (SDT). In *Encyclopedia of Measurement and Statistics*. Retrieved September 4, 2017, from https://pdfs.semanticscholar.org/dfd3/2eb45dae167e9cfef2649e9de294ad0ab753.pdf

About the Standards (2017). *Common core state standards initiative*. Retrieved May 31, 2017, from http://www.corestandards.org/about-the-standards/

Abrams, L. M., & Madaus, G. F. (2003). The lessons of high stakes testing. *Educational Leadership, 61*(3), 31-35.

Ainsworth, L. (2010). *Rigorous curriculum design: How to create curricular units of study that align standards, instruction and assessment*. Lanham, MD: Advanced Learning Press.

Ainsworth, J. M. (2016). *The predictive validity of coordinate algebra common district assessments on high-stakes coordinate algebra end of course assessment.* (Doctoral dissertation). Retrieved from ProQuest (UMI 10109224).

Alber, R. (2014). Why formative assessments matter. *Edutopia*. Retrieved May 5, 2017, from http://www.edutopia.org

Anderson, D., Alonzo, J., & Tindal, G. (2010). A cross-validation of easy CBM mathematics cut scores in Washington State: 2009-2010 test (Technical Report No. 1105). Retrieved June 15, 2017, from http://files.eric.ed.gov/fulltext/ED531930.pdf

Ansley, T. (2000). The role of standardized tests in grades k-12. In A.D. Trice (Ed.) *Handbook of classroom assessment* (pp. 265-285). New York, NY: Longman.

American Educational Research Association, American Psychological Association, & National Council of Measurement in Education (2003). *Standards for educational and psychological testing*. Washington, DC: AERA

American Educational Research Association, American Psychological Association, & National Council of Measurement in Education (2014). *Standards for educational and psychological testing*. Washington, DC: AERA

Bailey, K., & Jakicic, C. (2012). *Common formative assessment: A toolkit for professional learning communities at work*. Bloomington, IN: Solution Tree.

Bennett, R. E. (2011). Formative assessment: A critical review. *Assessment in Education: Principles, Policy & Practice, 18*(1), 5-25. Retrieved from http://dx.doi.org/10.1080/0969594X.2010.513678

Bidwell, A. (2014). The history of common core state standards. *U.S. news and world report.* Retrieved May 30, 2017, from https://www.usnews.com/news/special-reports/articles/2014/02/27/the-history-of-common-core-state-standards

Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2003). *Assessment for learning: Putting it into practice*. Berkshire, UK: Open University Press.

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

Black, P., & Wiliam, D. (1998b). *Inside the black box: Raising standards through classroom assessment*. London: School of Education King's College.

Blankstein, A., Houston, P., & Cole, R. (2010). *Data enhanced leadership*. Thousand Oaks CA: Corwin Press.

Bloom, B. S. (1977). Favorable learning conditions for all. *Teacher, 95*(3), 22-28.

Bloom, B. S., Hastings, J. T., & Madaus, G. F. (1971). *Handbook on formative and summative evaluation of pupil learning.* New York: McGraw-Hill.

Brown, R. S., & Coughlin, E. (2007). The predictive validity of selected benchmark assessments used in the Mid-Atlantic Region. *Issues & Answers Report*, *REL 2007(17)*. Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory MidAtlantic. Retrieved June 5, 2017, from https://ies.ed.gov/ncee/edlabs/regions/midatlantic/pdf/REL_2007017.pdf

Creswell, J. (2013). *Research design qualitative, quantitative, and mixed methods approaches* (4th ed.). Thousand Oaks, CA: Sage.

Development Process. (2017). *Common core state standards initiative*. Retrieved June 25, 2017, from http://www.corestandards.org/about-the-standards/development-process/

Drago-Severson, E. (2009). *Leading adult learning: Supporting adult development in our schools.* Thousand Oaks, CA: Corwin Press.

Dunn, K. E., & Mulvenon, S. W. (2009). A critical review of research on formative assessment: The limited scientific evidence of the impact of formative assessment in education. *Practical Assessment, Research & Evaluation, 14*(7). Retrieved July 1, 2017, from https://www.researchgate.net/profile/Karee_Dunn/publication/237409416_A_Critical_Review_of_Research_on_Formative_Assessment_The_Limited_Scientific_Evidence_of_the_Impact_of_Formative_Assessment_in_Education/links/54723a070cf2d67fc035c4f3.pdf

Earl, L. (2012). *Assessment as learning: Using classroom assessment to maximize student learning* (2nd ed.). Thousand Oaks, CA: Corwin.

Elementary and Secondary Education Act. (1965). Retrieved August 26, 2017, from http://education.laws.com/elementary-and-secondary-education-act

Every Student Succeeds Act. (2015). Retrieved August 25, 2017, from https://ed.gov/policy/elsec/leg/essa/essafactsheet170103.pdf

Fuch, L. S., & Fuch, D. (1986). Effects of systematic formative evaluation: A meta-analysis. *Exceptional Children, 53*(3), 199-208. Retrieved from https://doi.org/10.1177/001440298605300301

Goertz, M., Nabors Oláh, L., & Riggan, M. (2009). *Can interim assessments be used for instructional change?* (Policy Brief RF-51). Philadelphia, PA: Consortium for Policy Research in Education. Retrieved May 31, 2017, from http:// www.cpre.org

Graham, P., & Ferriter, W. (2010). *Building a professional learning community at work*. Bloomington, IN: Solution Tree Press.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.

Greenstein, L. (2010). *What teachers really need to know about formative assessment*. Alexandria, VA: ACSD.

Gruber, S. S. (2006, February). *High stakes proficiency testing: Is it good for education?* Paper presented at the 9th Annual Meeting of the American Association of Behavioral and Social Sciences, Las Vegas, NV.

Hintze, J., & Silberglitt, B. (2005). A longitudinal examination of the diagnostic accuracy and predictive validity of r-cbm and high-stakes testing. *School Psychology Review, 34*(3), 372-386.

Improving America’s Schools Act. (1994). Retrieved June 5, 2017, from https://www2.ed.gov/offices/OESE/archives/legislation/ESEA/brochure/iasa-bro.html

Jorgensen, M. A., & Hoffmann, J. (2003). *History of the No Child Left Behind Act of 2001 (NCLB).* San Antonio, TX: Pearson Inc.

Leandro v. State of North Carolina, 488 S.E.2d 249 (N.C. 1997). Retrieved August 24, 2017, from https://www.courtlistener.com/opinion/1210770/leandro-v-state/

Lewis, L. M. (2010). *Teachers' perceptions of the effectiveness of benchmark assessment data to predict student math grades.* (Doctoral dissertation). Retrieved from ProQuest (UMI 3436603)

Madaus, G. F. (1988). The distortion of teaching and testing: High stakes testing and instruction. *Peabody Journal of Education, 65*(3), 29-46.

Martin, P. L. (2012). *A preliminary examination of the intended purpose, actual use, and perceived benefit of district-led interim assessments on student achievement in North Carolina schools.* (Doctoral dissertation). Retrieved from ProQuest (UMI 3550261).

McFall, R. M., & Treat, T. A. (1999). Quantifying the information value of clinical assessments with signal detection theory. *Annual Review of Psychology, 50*(1), 215-241.

McNeese, B. (2016). *Are the skewness and kurtosis useful statistics*. Retrieved January 18, 2018, from https://www.spcforexcel.com/knowledge/basic-statistics/are-skewness-and-kurtosis-useful-statistics

Nese, J. F. T., Park, B. J., Alonzo, J., & Tindal, G. (2011). Applied curriculum based measurement as a predictor of high-stakes assessment. *The Elementary School Journal, 111*(4), 608-624.

NIST/SEMATECH (2013). *e-Handbook of statistical methods*. Retrieved August 1, 2017, from http://www.itl.nist.gov/div898/handbook/eda/section3/eda35b.htm

No Child Left Behind Act of 2001, Pub. L. No. 107-110, §3, 115 Stat. 1425 (2002).

North Carolina Department of Public Instruction. (n.d.a). *About home base.* Retrieved May 19, 2017, from http://www.ncpublicschools.org/homebase/about/

North Carolina Department of Public Instruction. (n.d.b). *Home base overall.* Retrieved May 19, 2017, from http://www.dpi.state.nc.us/homebase/faq/overall/

North Carolina Department of Public Instruction. (2013a). *North Carolina home base guidelines for building benchmark assessments.* Retrieved May 30, 2017, from https://1pdf.net/north-carolina-home-base-guidelines-for-building-benchmark-_585f94b2e12e898676474bd3

North Carolina Department of Public Instruction. (2013b). *North Carolina testing program: Technical information 2013-14 and beyond.* Retrieved July 21, 2017, from http://www.ncpublicschools.org/accountability/testing/technicalnotes

North Carolina Department of Public Instruction. (2016a). *NC standard course of study (NCSCS) for mathematics 2016–17 end-of-course assessment of NC Math 1 North Carolina assessment specifications.* Retrieved July 20, 2017, from http://www.ncpublicschools.org/docs/accountability/testing/technotes/math1specs16.pdf

North Carolina Department of Public Instruction. (2016b). *North Carolina end-of-course assessment of NC math 1*. Retrieved February 4, 2018, from http://www.ncpublicschools.org/docs/accountability/testing/achievelevels/math1achvlvl16.pdf

North Carolina Department of Public Instruction. (2017). *Consolidated state plan*. Retrieved February 4, 2018, from http://www.ncpublicschools.org/docs/succeeds/essa-state-plan.pdf

North Carolina Public Schools. (2011). *Released tests frequently asked questions.* Retrieved May 31, 2017, from http://www.ncpublicschools.org/docs/accountability/testing/releasedforms/faqs92110.pdf

North Carolina School Report Cards. (2016). Retrieved May 20, 2017, from https://ncreportcards.ondemand.sas.com/snapshots/070342_2016_9-12-School.pdf

North Carolina READY Accountability Background Brief. (2016). Retrieved August 24, 2017, from http://www.ncpublicschools.org/docs/accountability/reporting/16bckgrndbrf.pdf

Olson, L. (2005). Benchmark assessments offer regular checkups on student achievement. *Education Week, 25*(13), 13-14. Retrieved July 20, 2017, from http://www.edweek.org/ew/articles/2005/11/30/13benchmark.h25.html?print=1

Oswalt, S. G. (2013). *Identifying formative assessment in classroom instruction: Creating an instrument to observe use of formative assessment in practice* (Doctoral dissertation). Retrieved June 1, 2017, from http://scholarworks.boisestate.edu/td/753/

Payne, B. (2013). *The nature and predictive validity of a benchmark assessment program in an American Indian school district*. (Doctoral dissertation). Retrieved from ProQuest (UMI 3567959).

Pearson. (2017). *Schoolnet*. Retrieved August 25, 2017, from http://www.pearsonassessments.com/largescaleassessment/products-services/schoolnet/schoolnet-assessment.html

Pearson product-moment correlation. (2013). Retrieved January 19, 2018, from https://statistics.laerd.com/statistical-guides/pearson-correlation-coefficient-statistical-guide.php

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

Public Schools/Testing Schedule, North Carolina H687 § 1 (2015). Retrieved February 15, 2016 from http://www.ncleg.net/Sessions/2015/Bills/House/PDF/H687v2.pdf

Race to the Top. (2017). Truth in American education. Retrieved June 2, 2017, from https://truthinamericaneducation.com/race-to-the-top/

SAS. (2016). *White paper SAS EVAAS for K-12 statistical models*. Retrieved August 21, 2017, from https://www.sas.com/content/dam/SAS/en_us/doc/whitepaper1/sas-evaas-k12-statistical-models-107411.pdf

Scriven, M. (1967). *The methodology of evaluation*. Washington, DC: American Educational Research Association.

Statistics How To. (2018). *What is a "high" range?* Retrieved January 31, 2018, from http://www.statisticshowto.com/sensitivity-vs-specificity-statistics/

Stockman, M. (2016). *Predictive validity of local algebra benchmark assessments for Maryland algebra high school assessment*. (Doctoral dissertation). Retrieved from ProQuest (UMI 10130169).

The Bill. (2002). *Frontline: The new rules*. Retrieved May 31, 2017, from PBS Online http://www.pbs.org/wgbh/pages/frontline/shows/schools/nochild/nclb.html

Thompson, S. A. (2016). *Benchmark assessment as predictors of success on end-of-course standardized tests in algebra 1*. (Doctoral dissertation). Retrieved from ProQuest (UMI 10109217)

Tuzlukov, V. P. (2001). *Signal detection theory*. NY: Springer Science + Business Media.

Urdan, T. (2010). *Statistics in plain English* (3rd ed.). New York: Routledge

U.S. Department of Education. (1983). *A nation at risk.* Retrieved July 2, 2017, from https://www2.ed.gov/pubs/NatAtRisk/risk.html

U.S. Department of Education. (1994). *The improving America's schools act of 1994: Reauthorization of the elementary and secondary education act.* Retrieved July 3, 2017, from https://www2.ed.gov/offices/OESE/archives/legislation/ESEA/brochure/iasa-bro.html

Appendix

Math I Raw Scores to Scale Scores

Raw to Scale Score Conversion Table
End-of-Course Assessment of Math I Form A (paper format)

Form	Raw Score	Scale Score	Standard Deviation
A	0	227	5
A	1	228	6
A	2	228	6
A	3	229	6
A	4	230	6
A	5	231	6
A	6	232	6
A	7	233	6
A	8	234	5
A	9	236	5
A	10	237	5
A	11	238	5
A	12	239	5
A	13	241	5
A	14	242	4
A	15	243	4
A	16	244	4
A	17	246	4
A	18	247	3
A	19	248	3
A	20	249	3
A	21	250	3

A	22	251	3
A	23	251	3
A	24	252	3
A	25	253	3
A	26	254	3
A	27	255	3
A	28	256	3
A	29	256	3
A	30	257	3
A	31	258	3
A	32	259	3
A	33	260	3
A	34	260	3
A	35	261	3
A	36	262	3
A	37	263	3
A	38	264	3
A	39	265	3
A	40	266	3
A	41	267	3
A	42	268	3
A	43	270	3
A	44	271	3
A	45	272	4
A	46	274	4

A	47	276	4
A	48	278	4
A	49	281	5

Raw to Scale Score Conversion Table
End-of-Course Assessment of Math I Form B (paper format)

Form	Raw Score	Scale Score	Standard Deviation
B	0	229	5
B	1	229	5
B	2	230	6
B	3	230	6
B	4	231	6
B	5	232	6
B	6	233	6
B	7	234	6
B	8	235	6
B	9	236	5
B	10	237	5
B	11	238	5
B	12	240	5
B	13	241	5
B	14	242	4
B	15	243	4
B	16	245	4
B	17	246	4
B	18	247	3
B	19	248	3
B	20	249	3
B	21	250	3

B	22	251	3
B	23	251	3
B	24	252	3
B	25	253	3
B	26	254	3
B	27	255	3
B	28	256	3
B	29	256	3
B	30	257	3
B	31	258	3
B	32	259	3
B	33	260	3
B	34	260	3
B	35	261	3
B	36	262	3
B	37	263	3
B	38	264	3
B	39	265	3
B	40	266	3
B	41	267	3
B	42	268	3
B	43	270	3
B	44	271	3
B	45	273	4
B	46	275	4

B	47	277	4
B	48	279	5
B	49	281	5