ACADEMIC PROGRESS SCORES TO PREDICT PERFORMANCE
ON A STATE ASSESSMENT

Mitchell Curry, B.M., M.E.

Dissertation Prepared for the Degree of
DOCTOR OF EDUCATION

UNIVERSITY OF NORTH TEXAS
December 2016

APPROVED:
Frances van Tassell, Major Professor
Jane B. Huffman, Committee Member
Kelli Mallory, Committee Member
David Brackett, Committee Member
Bertina Combes, Dean of the College of Education
Victor Prybutok, Dean of the Toulouse Graduate School
This quantitative study examined seventh-grade reading scores to determine the extent to which certain demographic variables (race/ethnicity, gender, socioeconomic status) explain and MAP reading scores predict reading scores on the State of Texas Assessment of Academic Readiness (STAAR) in a selected northeast Texas public school. Standardized assessments only compare the relative performance of an individual student to other groups of students using scaled scores, which can vary from year to year and from state to state. With the advent of computer adaptive testing, this study provides information on the predictive validity of benchmark assessments. Specifically, this study looked for predictive evidence that indicates how accurately test data can predict criterion scores. Findings revealed, through a multiple regression analysis, that the fall MAP Rasch Unit (RIT) scores predicted the STAAR scale scores. Using SPSS version 22, the data were entered and analyzed in a multiple regression model to determine the presence of a statistical trend or lack thereof. Demographic data and MAP scores were entered into the regression model to examine the predictive validity of the MAP assessment in determining student performance on the STAAR seventh-grade state-mandated reading assessment. The statistical analysis revealed that MAP RIT scores explain a significant variance related to seventh-grade STAAR reading scale scores. There is a vital need for tools that improve a student’s academic development and MAP assessments have been found to predict performance on state-mandated assessments.
Copyright 2016
by
Mitchell Curry
I would like to take this opportunity to thank the entire dissertation committee for their time, effort, and willingness to serve and mentor me through this dissertation journey. Dr. van Tassell, I would like to thank you for your extra dedication in seeing me through the whole process. My admiration for you is immense and I would not be where I am if it were not for you. Dr. Huffman, Dr. Brackett, and Dr. Mallory, I have enjoyed the journey and I appreciate your knowledge and scholarship. Dr. Brackett, your expertise in quantitative data made my study much stronger and I appreciate your efforts and coaching. The different levels of defense have afforded me the opportunity to learn from all of you and I know that the final product is of high quality and scholarship because of your contributions.

Next, I would like to thank my family for the sacrifices that they have made. Valerie, Cole, and Cooper Curry, thank you for all of your love, understanding, and patience. You allowed me the resources through the gifts of time and monetary sacrifices. I would not have been able to achieve this milestone in my academic life without you. I will always cherish the support that has helped me see this journey through to the end.

Lastly, thank you to the educational colleagues that I work with for your help, guidance, support and mentorship through this study. Especially, Dr. Nathan Templeton, there have been many times throughout this dissertation journey that I needed someone to lean on and you were always there. Your help, friendship, and support have been paramount in my completion of this study.
# TABLE OF CONTENTS

## LIST OF TABLES

<table>
<thead>
<tr>
<th>Table of Contents</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>vi</td>
<td></td>
</tr>
</tbody>
</table>

## LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure of Contents</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>vi</td>
<td></td>
</tr>
</tbody>
</table>

## CHAPTERS

1. **INTRODUCTION** ..............................................................1

   - Theoretical Background ..............................................3
   - Statement of the Problem ...........................................8
   - Purpose of the Study .................................................9
   - Research Question ...................................................10
   - Significance of the Study .........................................10
   - Definition of Terms ................................................11
   - Delimitations .......................................................14
   - Organization of the Study .......................................14

2. **REVIEW OF THE LITERATURE** .......................................16

   - Literacy Achievement ..............................................18
     - Literacy Achievement in the Middle Grades ...............20
     - Literacy Achievement and Diverse Learners ..............22
     - Literacy Achievements and Universal Screening ........23
   - Computer Adaptive Testing ......................................28
     - Item Response Theory and Classic Test Theory ........30
     - Validity of Computer Adaptive Testing ....................32
     - Measures of Academic Progress ..............................33
     - Summary ..........................................................36

3. **METHOD** ..................................................................38

   - Research Questions ................................................38
   - Overview of the Study .............................................38
   - Participants .........................................................40
   - Procedure for Data Analysis ....................................41
   - Means of Conducting the Regression .........................45
   - Summary ............................................................46

4. **RESULTS** .................................................................47

   - Descriptive Statistics .............................................47
   - Results for the 2012-2013 Cohort ............................48
   - Results for the 2013-2014 Cohort ............................56
5. SUMMARY, CONCLUSIONS, AND RECOMMENDATIONS ............................65

Discussion ..................................................................................................66
Implications .................................................................................................67
Limitations ..................................................................................................68
Recommendations For Future Research ....................................................69
Summary ....................................................................................................70

REFERENCES ..............................................................................................72
<table>
<thead>
<tr>
<th></th>
<th>TABLE Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2014 Standardized and UnStandardized Coefficients</td>
<td>44</td>
</tr>
<tr>
<td>2</td>
<td>2012-2013 Demographic Percentages</td>
<td>49</td>
</tr>
<tr>
<td>3</td>
<td>2013-2014 Demographic Percentages</td>
<td>49</td>
</tr>
<tr>
<td>4</td>
<td>Significant Coefficients for 2012-2013 Seventh Grade Analysis</td>
<td>55</td>
</tr>
<tr>
<td>5</td>
<td>Research Question Coefficients for 2013-2014 Seventh Grade Students’ Fall Reading MAP RIT Results</td>
<td>55</td>
</tr>
<tr>
<td>6</td>
<td>ANOVA Table for Cohort 1 (2012-2013) Seventh-Grade Analysis</td>
<td>56</td>
</tr>
<tr>
<td>7</td>
<td>Model Summary for 2012-2013 Seventh Grade Students’ Fall Reading MAP RIT Results</td>
<td>57</td>
</tr>
<tr>
<td>8</td>
<td>Coefficients for 2013-2014 Seventh-Grade Analysis</td>
<td>62</td>
</tr>
<tr>
<td>9</td>
<td>ANOVA Table for Cohort 2 (2013-2014) Seventh Grade Students’ Fall Reading MAP RIT and Economically Disadvantaged Results</td>
<td>62</td>
</tr>
<tr>
<td>10</td>
<td>Model Summary for Cohort 2 (2013-2014) Seventh Grade Students’ Fall Reading MAP RIT Results</td>
<td>63</td>
</tr>
</tbody>
</table>
LIST OF FIGURES

1. Comparison of Classical and Response Theory Measurement Hierarchy ..................................7

2. Cohort 1 (2012-2013) Scatterplot of the STAAR Scale Scores and their Standardized Residuals and Predicted Values .................................................................51

3. Cohort 1 (2012-2013) Histogram showing the distribution of Standardized Residuals .................................................................................................................................53

4. Cohort 1 (2012-2013) Normal Probability Plot showing the outliers of the Standardized Residuals ..................................................................................................................54


6. Cohort 2 (2013-2014) Histogram showing the distribution of Standardized Residuals .................................................................................................................................59

7. Cohort 2 (2013-2014) Normal Probability Plot showing the outliers of the Standardized Residuals ..................................................................................................................60
CHAPTER 1
INTRODUCTION

The stated intention of the No Child Left Behind Act (NCLB), a sweeping reauthorization of the Elementary and Secondary Education Act (ESEA), was to level the academic standards for all students and to narrow the achievement gaps between non-traditionally disadvantaged student groups and minority students groups. The bill lists as a main purpose to “close the achievement gap between high- and low-performing children, especially the achievement gaps between minority and nonminority students, and between disadvantaged children and their more advantaged peers (NCLB, 2001, pg. 1440). When implemented, the mandate of the act was to have all students reach 100% proficiency levels in reading and mathematics by 2014, as measured by performance on state tests. NCLB gave students 12 years to demonstrate 100% proficiency on state-mandated testing in reading and mathematics. Along with the end goal of proficiency, NCLB required schools to meet Adequate Yearly Progress (AYP) proficiency targets for each student group, including English-language learners, students in special education, minority students, and low socio-economic students who traditionally achieve at significantly lower academic levels. AYP is the accountability component of Title I of NCLB that mandated criteria by which states, schools, and districts were measured. The law required states to use a single accountability system for public schools to determine if all student groups were making progress toward state academic content standards. Furthermore, AYP mandated testing for students in grades three through eight and in at least one grade in high school, in order to determine progress on these standards.

In an era of high stakes accountability, researchers have been searching for a tool that can predict the learning progress of students (Darling-Hammond & Adamson, 2014; D. Fuchs &
Fuchs, 2006; Good, Aronson, & Inzlicht, 2003; Heilig, 2008; Hew & Brush, 2007). In Texas, where the present study was conducted, the statewide assessment is the State of Texas Assessment of Academic Readiness (STAAR). In 2013, the 83rd Texas legislature passed House Bill 5 (HB 5), which became a part of the Texas Education Code. HB 5 was intended to improve student outcomes and narrow achievement gaps through appropriate interventions (Texas House Bill 5, 2013). This high-stakes test, the STAAR, administered by the State of Texas, is typically given near the end of the year and retesting opportunities are given in the summer for certain grade levels of students that did not meet standard. In 2009, the 81st Session of the Texas Legislature enacted House Bill (HB) 3. HB 3 requires that end of course (EOC) exams replace STAAR testing at the high school level as the standardized measure for high school graduation. These assessments are given in mathematics, science, English, and history. The EOC assessments are directly linked to college readiness standards deemed appropriate by the Texas Legislature. Furthermore, the STAAR assessments in grades three through eight were aligned to link to the EOC assessment and to college readiness standards developed by the state.

Among the many problems associated with high-stakes tests is their inability to measure growth and their lack of instructional intervention information pertaining to a student’s learning gaps. Central to HB 5 was limiting the number of benchmarks that can be administered to students during the school year. The State of Texas limits district-required benchmark assessments to no more than two per year in preparation for a corresponding state-mandated test. These limits do not apply to college preparatory assessments such at the Preliminary Scholastic Aptitude Test (PSAT), Scholastic Aptitude Test (SAT), American College Testing Assessment (ACT), advanced placement exam (AP) or any independent assessment designed and
administered by the classroom teacher. Therefore, educators are required to search for an improved and more efficient process to measure student progress toward grade-level mastery.

One measure of interest to educational leaders is the computer adaptive test (CAT), which is thought to provide more precise measurement of a student’s achievement level (Weiss & Kingsbury, 1984). Additionally, Millsap (2000) noted that CAT assessments provide instant access to student data, random item selection, test item banks, and quick data collection and scoring procedures.

The present study examined the ability of the CAT known as the Measures of Academic Progress (MAP), developed by the Northwest Evaluation Association (Northwest Evaluation Association, 2015), to predict scores on the reading STAAR assessment. The overall purpose of the study was to examine the predictive relationship between these two measures (STAAR and MAP) for students participating in the study and also to examine relationships associated with the demographic variables of socioeconomic status, race/ethnicity, and gender. The study used existing state data from seventh graders in a middle school located in northeast Texas.

This chapter, which introduces the study, is organized in the following sections: theoretical framework, statement of the problem, research questions, purpose of the study, significance of the study, definitions of terms, delimitations, and organization of the study.

Theoretical Background

This study is grounded in psychometrics, which is the research and theory of testing. Psychometric research has a long history dating back to Sir Francis Galton who, in the late 1880’s, developed tests intended to measure a person’s various capacities. Galton, who is widely considered the father of psychometrics, used his measurements to study differences across ordinary people, using multiple senses, such as sight, color discrimination, and pitch.
discrimination. Early psychometric work in the United States focused on measurements of intelligence, and included such well-known instruments as the Stanford-Binet test (Binet & Simon, 1916). Galton (1883) concluded that these tests did not only pertain to a person’s cognitive ability, but were indeed accurate indicators of a person’s intellectual ability as well. The Stanford-Binet test can distinguish between groups of students with a variety of intellectual abilities, such as giftedness, mental handicaps, or neurological impairments. Additionally, psychometric testing is significantly useful when assessing a student’s performance over time (Laurent, Swerdlik, & Ryburn, 1992).

Since the mid-19th century, assessment through testing has become increasingly common in education. Horace Mann introduced assessment through testing in the Boston schools. According to Gallagher (2003), these tests were intended to gain “objective information about the quality of teaching and learning in urban schools, monitor the quality of instruction, and compare schools and teachers within each school” (p. 85). Additionally, these assessments enabled researchers to glean information about learning gaps and prior knowledge to predict what information students needed for future learning. The model of standardized testing that educators are most familiar with became the standard in the 1920’s. Gallagher stated: “By 1929, more than five million tests were administered annually, and results were used to segregate those who learned from those who had not” (p. 88).

Some states use annually-administered criterion-referenced tests to assess student mastery of learning objectives. Criterion-reference testing measures a student’s ability based on a fixed set of criteria or fixed learning standards. Furthermore, criterion-reference testing assumes students have a common knowledge base or reference that can be used for purposes of comparison (Calhoun, 2004). STAAR is a criterion-referenced test that measures mastery at the
higher cognitive levels of Bloom’s taxonomy. However, STAAR testing is not the first version of standardized testing to be administered by the State of Texas. Texas Assessment of Basic Skills (TABS), Texas Educational Assessment of Minimum Skills (TEAMS), Texas Assessment of Knowledge and Skills (TAKS) and Texas Assessment of Academic Skills (TAAS) were previous criterion-referenced standardized tests that required students to demonstrate mastery, albeit at lower cognitive levels of the tested skills and knowledge (Texas Education Agency, 2014b).

Classical psychometric theory is the objective measurement of skills and knowledge, abilities, attitudes, personality traits, and educational achievement. There are two distinct branches of psychometric theory, the Item Response Theory (IRT) and the Classic Test Theory (CTT) (Markus & Borsboom, 2013). Most tests are based on the CTT. Standardized tests tend to be based on a sophisticated and modernized version of the CTT, and interpretations of scores provided in manuals and interpretation rubrics, as well as by practicing psychologists, are used to interpret the results of standardized assessments (Irby, Brown, Lara-Alecio, & Jackson, 2013). CTT is designed to account for errors in the data. Errors are found, indexed, and analyzed, based on their randomness, and are correlated to other data points being studied.

In 1904, Charles Spearman developed a method to account for measurement error and to calculate the error into the data as a correction, therefore leveling the data by removing random outliers. Although it is assumed that each of the characteristics of the participants and the test is statistically equal and has the same standard error of measurement, one of the shortcomings of these assessments is their characteristic dependence on one another. Additionally, when checking for reliability, it is feasible to compare the assessment to a parallel form of itself. It should be noted, this is only to check for reliability. This has made comparing students, school
districts, and states very difficult, as most states give a different variation of assessments each year.

The STAAR test uses Item Response Theory (IRT) and is an advanced application of the classical test theory. IRT is a modern form of the CTT and assumes that all items are identical and considered parallel. IRT is different from CTT by taking into account the test items to be incorporated in scaling items. The items have an item characteristic curve (ICC), which is designed to give the questions a set of qualities. These qualities can be rated as difficult or simple. IRT focuses on the theory of the questions as opposed to focusing on the final measurement or outcome of the testing. Additionally, the STAAR criteria are aligned with the Texas grade level standards known as the Texas Essential Knowledge and Skills (TEKS) (Texas Education Agency Student Assessment Division, 2013).

IRT is an advanced form of the CTT that has become the standard for the development of tests that measure the ability levels of the examinee by adapting the level of the test to mirror the proficiency level of the student. Specifically, IRT affords the means to measure the ability level of the test taker and the difficulty level of the test on the same scale. This outcome gives the examiner the ability to predict the probable proficiency of an examinee of known ability on items of known difficulty (Hsu, Wang, & Chen, 2013). More importantly, IRT allows test administrators to determine the likelihood of student proficiency on grade-level or age-appropriate tasks. Additionally, the newer developed tests still provide traditional data, such as percentile ranks, standard scores, and age/grade equivalent response outcomes.

Figure 1 shows the difference in the way that Classic Test Theory (CTT) and Item Response Theory (IRT) differ. CTT assumes that true scores and error scores are uncorrelated. Additionally, CTT assessments begin with the assertion that all testing participants begin with an
equal level of knowledge and ability. IRT assumes that the test is unidimensional and local independence is present. Furthermore, IRT focuses more on the item and the probability factor associated with the item.

*Figure 1.* Comparison of Classical and Item Response Theory Measurement Hierarchy shows the difference in the way that Classic Test Theory (CTT) and Item Response Theory (IRT) differ. CTT assumes that true scores and error scores are uncorrelated. Additionally, CTT assessments begin with the assertion that all testing participants begin with an equal level of knowledge and ability. IRT assumes that the test is unidimensional and local independence is present.

IRT has increased benefits for educators through Computer Adaptive Testing (CAT). As shown in Figure 1 above, each time a question is answered by the examinee, the IRT allows the adaptive testing model to level the questions to the ability of the examinee. Therefore, the test
needs fewer items to get a reliable measurement of the student’s age/grade level (Dahl & Woodcock, 1971). This allows the test to hover around the ability of the examinee by eliminating the items that are too easy or too hard. This design uses an assessment that is shorter and more efficient (Wainer, 1990). Additionally, IRT allows for the calculation of separate reliability and standard errors of measurement for different question banks (Wright & Stone, 1979). For example, if a test question is poorly represented or inappropriate for the examinee, the testing application allows for a recalculation of a normative score and the removal of the item.

The Measures of Academic Progress assessment (MAP) is an example of a Computer Adaptive Test (CAT) that is designed to be an objective tool that provides more precise measurement for all students. CAT assessments, specifically the MAP assessment, provide achievement and growth data that can be used immediately by the teacher and student to identify learning gaps and mastery of knowledge levels. Brown and Coughlin (2007) noted several studies in which MAP assessments validated strong interrelationships among test items and the state assessments to which they were designed to correlate. The MAP administration manual for teachers (Northwest Evaluation Association, 2006) and the MAP technical manual (Northwest Evaluation Association, 2003) provide a clear understanding of these correlations and help the teacher in assessing the student’s strengths and weaknesses.

Statement of the Problem

Since 2000, CAT has taken precedence in the measurement of student progress and learning, as well as becoming the standard as a benchmarking device in student accountability (Huang, Lin, & Cheng, 2009). By means of CAT data, researchers, educators, and
administrators gain information on what rate students are growing in their learning over a period of years.

The majority of data about effective instructional practices and student performance is based on fixed-form standardized tests. The fixed-form tests give researchers a plethora of data in regards to both effective instructional strategies and students’ performance, based on passing standardized tests (McCall, Hauser, Cronin, Kingsbury, & Houser, 2006). These tests have a predetermined passing rate that gives the researcher an artificial scale score and is unreliable when measuring a student’s academic growth from year to year. The primary data that are used by practitioners to measure student academic growth consist of standardized testing data. Adaptive tests measure the student’s academic growth and give researchers more precise data in each content area over both long and short periods of time. This study investigated the predictive relationship between adaptive testing in the form of students’ reading scores on the MAP and their performance on a more traditional test, the seventh-grade STAAR reading assessment.

Purpose of the Study

At the present time, there is limited empirical evidence that demonstrates the value of CAT data in measuring the success of student scores related to instructional strategies. The purpose of this predictive study was to examine student data over a two-year period to determine if adaptive testing is a statistically valid predictor of student performance on the reading STAAR test. The adaptive testing model provides a multi-year profile for children and their expected growth index. Adaptive tests have the capacity to predict future learning growth as well as to reveal when students do not achieve at the rate expected, as compared to standardized testing that reports only pass or fail. There seems to be little research on the predictability of CAT for future
testing achievements on standardized testing assessments (Wainer, Dorans, Flaugher, Green, & Mislevy, 2000). Therefore, this study was designed to add to the research literature related to the value of adaptive benchmark testing.

Research Question

The following question guided this study:

To what extent do certain demographic variables (race/ethnicity, gender, socioeconomic status) explain, and MAP reading scores predict, reading scores on the State of Texas Assessment of Academic Readiness (STAAR) for seventh-grade students in a selected northeast Texas public school?

As explained in more detail in Chapter III, the study employed a multiple regression procedure to determine how well seventh-grade students’ MAP reading scores predict whether they would meet standard or not meet standard on the reading portion of the STAAR test. In that analysis, various demographic variables, including socioeconomic status, race/ethnicity, and gender, were used as possible predictors.

Significance of the Study

As the nation struggled to define proficiency standards in reading over the last decade, the common measurement tool was the state-created standardized fixed-form test. These standardized assessments only compare the relative performance of an individual student to other groups of students using scaled scores for their specific state assessment, which can vary from year to year and from state to state (Cronin, Dahlin, Adkins, & Kingsbury, 2007). Due to the decades of readily available standardized test data, the overwhelming majority of research done on teaching methods and treatments has used this bank of millions of tests as its data reference. With the advent of computer adaptive testing, which is potentially more informative, this study
will provide additional information on the predictive validity of benchmark assessments. Additionally, this study looked for predictive evidence that indicates how accurately test data can predict criterion scores, or scores on other tests.

The findings could significantly impact how district personnel make judgments about student performance and how accurately test data can predict criterion scores on state-mandated standardized testing. According to Ujifsa (2015), the United States spent $1.7 billion on standardized testing in 2012. Further research and additional information on the predictive validity of benchmark assessments, as well as computer adaptive testing, are needed in making sound decisions for states to have a reasonable chance of success. This study should contribute to a growing body of research, such as that by McCall et. al (2006), who concluded that MAP data provided useful growth model estimates in a particular content area over a period as short as four weeks.

Definition of Terms

The following terms are defined as they are used in this study and for consistency and clarity throughout the study.

*Academic Excellence Indicator System (AEIS)* consists of a wide range of information on the performance of students in each school and district in Texas that is available each year, reflecting the previous year’s performance indicators. The indicators include State of Texas Assessment of Academic Readiness (STAAR) met standard data, attendance information, dropout rates, and percent of students in advanced courses. Performance on each of those indicators is shown disaggregated by ethnicity, special education, and income status. AEIS reports also provide extensive information on school district staff, finances, programs, and demographics.
Achievement Level Test (ALT) is a paper-pencil assessment that measures a student’s general knowledge in reading, language usage, mathematics and science.

Adaptive testing is testing that adjusts to the test-taker’s ability. It provides questions that are tailored in difficulty to the test-taker’s performance on previous questions. It is often administered on a computer, and, in that case, is known as computer-adaptive testing (CAT) (Wainer, 2000).

DesCartes is a learning continuum resource aligned to state standards. It is designed to help teachers translate the raw data from students' assessments into actionable plans for instruction, grouping and more.

Gender is a reference to physical attributes used to assign sex; that is, to identify an individual as a dichotomous variable male or female. In this study, male and female participants were a dichotomous variable. Additionally, the researcher disaggregated data based on gender.

Item Response Theory (IRT) is a modern form of Classic Test Theory (CTT) and uses a method to account for measurement error and to calculate the error into the data as a correction, therefore leveling the data by removing random outliers in the data.

Learning continuum uses Rasch Unit (RIT) levels on the MAP assessment to show teachers what skills and concepts students are ready to learn next. The RIT is a fixed equal-interval scale used to indicate student achievement and growth. Additionally, the learning continuum recommends student-specific information that teachers can use to drive appropriate instructional decisions.

Measures of Academic Progress (MAP) is a norm-referenced measure of student growth over time. MAP assessments, joined with other data points, provide detailed, actionable data about where each child is academically in comparison to other students nationally. For this
study, the specific MAP assessment used was seventh-grade reading, administered in the Fall, Winter, and Spring of the 2011-12 and 2012-2013 school years.

*Race* refers to a group of people distinguished by genetic traits that are common with other members of that group. In this study, the United States Department of Education Race Reporting Standards found within the Texas Education Agency guidelines (Texas Education Agency, 2014a) were used to classify the following five racial categories:

- Hispanic/Latino
- Asian
- Black or African American
- White
- Two or More Races

*Rasch Unit (RIT)* is a fixed equal-interval scale used to indicate student achievement and growth. Developed to simplify how scores are interpreted, RIT scores on the MAP range from 140 to 300. Students typically start at the 140 to 190 levels in the third grade and progress to the 240 to 300 levels by high school. RIT scores are placed along a linear growth model, which was previously called DesCartes, but is now called the *learning continuum*, both of which are defined in this section (Yugi, 2000).

*Socioeconomic status* (SES) references the economic status of the student’s family and generally is used to reference students of concern in a low SES. In this study, SES is a dichotomous variable, high and low, with a student’s qualification for free or reduced lunch noted as an indicator of low.

*State of Texas Assessment of Academic Readiness* (STAAR) is a series of state-mandated standardized assessments used in Texas public primary and secondary schools to assess a student's achievement and knowledge learned in the grade level.
*STAAR met standard* is the STAAR assessment performance standard as it relates to the expectations defined in the state-mandated curriculum standards known as the Texas Essential Knowledge and Skills (TEKS). Cut scores established by the state agency distinguish between performance levels, or categories. The process of establishing cut scores that define performance levels for an assessment is standard setting. Standard setting is also used to classify students into an appropriate performance category. In this study, the Level II standard for satisfactory academic progress was used.

*Texas Education Agency (TEA)* is the administrative unit in Texas for primary and secondary education. The agency is comprised of the Commissioner of Education and agency staff, with duties that include guiding and monitoring activities and programs related to public education. The stated mission of the agency includes the goal of providing all students with a quality education.

**Delimitations**

For the study, the dependent variable was a dichotomous categorical variable, met standard/did not meet standard on the STAAR test. Even though standard scores might be more precise, the met standard/did not meet standard criterion is of major importance to educational administrators. The standard score that determines that criterion is not consistent from year to year across grade levels.

**Organization of the Study**

This study is organized into five chapters. This chapter presents an introduction to the study. Chapter II reviews related literature focusing on high stakes testing and its importance in the current political climate, the development of tests that are better measures of growth, and the new emphasis on adaptive testing and growth models. Chapter III presents the design of the
study, including the research questions, a description of data collected (the independent and dependent variables), and the procedure for data analysis technique. Chapter IV is a report of the findings in terms of general patterns and contribution of various demographic variables, and Chapter V discusses those findings and explains the potential significance of the study.
CHAPTER 2
REVIEW OF THE LITERATURE

The effectiveness of public education continues to be a significant concern in the United States. As a result, the ability to measure school effectiveness has become the pivotal issue in education policy and legislation (Fry, 2008; Kena et al., 2015). The No Child Left Behind Act (NCLB, 2001) included provisions for accountability by mandating states to develop assessments of basic skills for all students. NCLB aimed to “close the achievement gap between high and low performing children, especially the achievement gaps between minority and nonminority students, and between disadvantaged children and their more advantaged peers” (2001, p. 16). Under NCLB, states are required to report achievement gaps between demographic subgroups, including those delineating race, to help schools, districts, and states decrease achievement gaps over time (Berends, Lucas, & Penaloza, 2008).

Paradoxically, the accountability and reporting mandates of NCLB created, and even increased, the racialized achievement gaps in many school districts (Magnuson & Waldfogel, 2008). Specifically, standardization, accountability, and transparency have done little to alleviate the racial achievement gap between White students and Black and Latino students (Hursh, 2007), thus continuing and expanding racialized and class gaps in schools (Lipman, 2011).

By 2013, 34 states had been granted waivers from NCLB regulations. These waivers were granted because current laws were not focused on fostering necessary reading skills, thus constraining state and district efforts to reform public schools through innovation (Department of Education, 2013). By late 2015, Congress had moved beyond NCLB toward comprehensive legislation that would give states more latitude to determine how test results will be used for
accountability purposes. The new framework is called the Every Student Succeeds Act (ESSA) and is the reauthorization of the Elementary and Secondary Education Act (ESEA) of 1965. The ESSA legislation gives districts and their respective states more autonomy in regards to accountability requirements and school improvement. Moreover, the new framework will allow states and local educational leaders to use block grants to allocate funding, based on the needs of their schools and communities (ASCD, 2015). While the framework will continue to maintain annual testing, decision-making authority on school improvement and achievement gap closures will be relegated to the states.

In 2009, educational leaders from 48 states launched the Common Core State Standards (CCSS). Although NCLB focused educational efforts on all students acquiring basic literacy skills, the Common Core State Standards were not intended to supplant the intent of the legislation (Hiebert & Pearson, 2013). Building on the foundational skills of decoding and letter recognition inherent in NCLB, the CCSS were intentionally designed to prepare students to become prolific readers, writers, and thinkers through a more prominent focus on deeper learning and more advanced literacy. According to Camera (2015), as of September 2015, many states had discontinued using the CCSS that began in 2009. These states fear low scores as a result of administering a more difficult assessment without time for teachers and students to adjust to these new standards. Many of these same states have teacher evaluation systems that tie teacher pay to test scores and fear a drop in scores will result in punitive repercussions. Additionally, the U.S. House and Senate are moving toward proposals that will give states more localized autonomy of their own testing system.

Furthermore, as states continue to search for measures to address academic disparity of all students through early identification and intervention, demands have increased for progress-
monitoring strategies that reliably predict outcomes on statewide assessments (Weiss, 2011). According to Thompson and Weiss (2011), Computer Adaptive Testing (CAT) may be a viable option. Therefore, the purpose of this study was to investigate the relationship between the predictive correlation of MAP and student performance on seventh-grade reading scores on the State of Texas Assessment of Academic Readiness (STAAR). The review of literature presented in this chapter addresses concepts in reading; discusses the characteristics of STAAR as a formative assessment tool; and examines the effectiveness of a specific Computer Adaptive Test, the Measures of Academic Progress (MAP), in predicting student performance on high-stakes testing.

Literacy Achievement

Enactment of the No Child Left Behind Act (NCLB, 2001), along with the reauthorization of the Individuals with Disabilities Education Act (IDEA, 2004), significantly changed the assessment landscape for all students in public schools and, in particular, those at risk for unsatisfactory reading outcomes. One of the more significant provisions of NCLB was the requirement that states adopt and conduct annual assessments to gauge school districts’ progress in improving students’ academic achievement. Additionally, states are responsible for holding schools accountable for the performance of all students in all groups. As a result, states need to know how to identify students who struggle with reading, using fair and valid assessments (NCLB, 2001).

As Reynolds, Creemers, Stringfield, Teddie, and Schaffer (2002) discussed, students who have deficiencies in reading face potentially lifelong challenges. “Students reading below grade level cannot understand the text at a literal level, make connections between the text and their own experiences, and cannot make inferences from the text” (p. 277). Thus, children who do
not read well are at higher risks of functioning below grade level and exhibiting behaviors that do not lead to future academic success. Thus, identifying and remediating reading difficulties has potential long-term advantages to the individual and to the greater society.

In related research, Berman and Biancarosa (2005) suggested that while most adolescents can read simple texts, many “frequently cannot understand specialized or more advanced texts” and “are unprepared to meet the higher literacy demands of today’s colleges and workplaces” (p. 6). Conversely, according to the National Reading Panel (2000), literacy instruction has progressed sufficiently to ensure that 95% of the children in the United States can be taught to read at a level of proficiency to engage in independent, age-appropriate reading activities. In fact, researchers (Reynolds et al., 2002) suggest that targeted instruction in reading strategies (predicting, questioning, thinking aloud, note taking, and recognizing text structure) can improve students’ ability to comprehend texts in content areas. Moreover, schools should consider building students’ subject area knowledge by providing the background knowledge they need in order to comprehend more difficult content in later grades (Alvermann & Moore, 2002; Beck, McKeown, & Kucan, 2002; Pressley, 2004; Stockard, 2010; U.S. Department of Health and Human Services, 2000). Pressley et al. (2003) suggested that effective teaching practices include those that use instructional materials based on the abilities of the students, which are determined by formative assessments and the learning objectives set for each student. Despite the inclusion of evidence-based recommendations to improve reading instruction, there is little evidence of improved literacy outcomes in America during the last 30 years (Kamil et al., 2008; Perie, Moran, & Lutkus, 2005). The National Association of Educational Progress (NAEP) report (2015) confirmed that reading scores showed improvement from 1971-2012, but have improved very little since 2012.
Literacy Achievement in the Middle Grades

Kamil (2003) reported that the crucial litmus test for evaluating improved literacy outcomes is whether or not children are reading at grade level at the beginning of the fourth grade. It is important to note that while overall NAEP results for reading in 2002 were not encouraging, the grade four average score in 2002 was higher than in 1994, 1998, and 2000, but was not found to be significantly different from 1992. Likewise, Kena et al. (2015) noted that while statistically significant gains have not been made in this age group over the past three decades, the report by the NAEP (2015) notes achievement gains in 13-year-olds in reading performance from 2008 to 2012.

Most fourth graders have mastered skills to analyze and manipulate simple texts, such as fictional narratives and uncomplicated informational books. Furthermore, most can navigate through simple procedures and processes, such as following directions for constructing models and working through simple word problems. These skills afford students to have confidence as independent readers and consumers and producers of written materials (Robinson, McKenna, & Conradi, 2011). However, Perie, Moran, and Lutkus (2005) intimated that the gap between students from higher- and lower-income families is growing wider, with 17% improvement seen among the former group compared to a marginal 6% improvement among their lower-income peers. Additionally, significant improvement has not been sustainable beyond grade four and into the middle grades, as evidenced by the recent National Assessment of Educational Progress (NAEP) (Kena et al., 2015).

The middle grades are a critical transition period for developing literacy skills, as students shift from learning to read to reading to learn (Goodwin, 2011; Kamil, 2003). In addition to an introduction to high-stakes tests that hold students and teachers accountable for
learning, during the middle years, students are required to manipulate more abstract, complex and unfamiliar texts than in the earlier grades (Alvermann, 2002). Moreover, since 2001, research has identified inconsistencies between instructional practices and literacy research. Marzano (2001) noted that instruction failed to account for developmental and cognitive differences in students. Students were expected to encounter complex thinking without being taught how to strategically encounter the material. Instruction also was not aligned to content objectives, as teachers continued to expose students to knowledge-based competencies – teaching content versus teaching content literacy. Finally, Marzano found that instruction did not build on the knowledge and dispositions that students had already acquired; but, as literacy instruction became more experiential, aligned and student-centered, so did the progress toward narrowing gaps in achievement.

In contrast, the National Center for Education Statistics (2012) reported that long-term trend assessments show some progress toward narrowing achievement gaps, with average reading scores for students in the middle grades trending 8% to 10% higher in 2012, after sluggish gains from the period of 1973-1986. During the 2012 reporting period, student performance in grades four and eight in the state of Texas largely mirrored that of the national scene; however, certain populations in grade seven had significantly higher numbers of students reading below grade level. Yet, the average 2015 NAEP student showed little or no gain since 2012. According to the Texas Education Agency’s 2013 Annual Summary Report for seventh graders (Texas Education Agency, 2014b),

- 34% of African American, 32% of Latino, and 25% of Native American students read below grade level compared to only 13% for white students and 10% for Asian students.
- 35% of seventh graders eligible for free and reduced lunch read below grade level.
• 10% of African American, 11% of Latino, and 19% of Native American students are reading above proficiency levels compared to 31% for white and 45% for Asian students.
• 47% of students with special needs read below grade level.
• 57% of English language learners read below grade level.

Literacy Achievements and Diverse Learners

Despite efforts by educators and policymakers during the last several decades, achievement gaps between certain groups of students stubbornly persist (Kena, Musu-Gillette, Robinson, Wang, Rathbun, Sidney, Wilkinson-Flicker, Barmer, & Velez, 2015). This issue is further exacerbated by the population shift in Texas. In 2008, more than 800,000 students (nearly 15% of K–12 students in Texas’ public schools) were English Language Learners (Texas Education Agency Student Assessment Division, 2009).

In 2014, Texas school enrollment exceeded five million, and Hispanics surpassed Whites as the largest ethnic group in Texas, comprising 51.3% of public school enrollment (Texas Education Agency, 2014b). As recorded by the Texas’ Public Education Information Management System (PEIMS), a vast majority of these students (over 900,000) were identified as English Language Learners (ELLs), or have Limited English Proficiency (LEP). For clarity, in the State of Texas, the terms are used interchangeably.

The United States Department of Education (USDOE) and the Texas Education Agency (TEA) worked to improve instructional practices for ELLs, including requiring Texas teachers who instruct ELLs to implement both the state’s English Language Proficiency Standards (ELPS) and its academic standards, the Texas Essential Knowledge and Skills (TEKS). Texas also aligned its English language and academic standards with the Texas English Language Proficiency Assessment System, used to assess a student’s English language proficiency level and progress in the domains of speaking, listening, reading, and writing (USDOE, 2010). The
results determine the linguistic accommodations students received during instruction and on assessments (Wackwitz & Burniske, 2009).

Additionally, Wackwitz and Burniske (2009) discussed the manner in which Language Proficiency Assessment Committees help local districts and campuses determine the needs of ELLs, select instructional interventions, monitor student progress, make assessment decisions, and maintain required documentation. Likewise, the use of reading comprehension measures, such as word reading and oral language demands, were replaced with more central comprehension processes, such as the Spanish Woodcock Tests of Cognitive Abilities, so that students’ comprehension skills were assessed more accurately. Even with this guidance, however, Texas educators face a number of challenges in meeting the needs of ELLs. Many content-area teachers are unfamiliar with the ELPS and have only a superficial knowledge of the types of linguistic accommodations needed to meet them. Furthermore, while the linguistic accommodations provided on state assessments are fairly consistent among teachers, those provided during instruction vary widely (Wackwitz & Burniske, 2009).

Literacy Achievement and Universal Screening

Wide ranges of assessments have been developed for use with diverse learners. Universal screening is a principal means of identifying students, regardless of ethnicity or linguistic deficiency, who require early reading intervention (Glover & Albers, 2007). Response to Intervention (RTI) is based on students who do not respond to instruction, based on data from assessment results. RTI interventions move from general (Tier 1) to moderate (Tier 2) to intensive (Tier 3). For clarity, Tier 2 intervention is delivered in small groups while Tier 3 is individualized 1:1 intervention. The purpose of these intervention strategies is to trigger intervention protocols prior to the onset of significant problems (Compton et al., 2010).
Moreover, universal screening involves measures of early literacy and foundational reading skills, including phonemic awareness, letter-naming fluency, concepts about print, word reading, and oral language ability, including vocabulary. As a matter of best practice, a score point is established where students who scored below the cut-point are considered at-risk for reading difficulties and in need of additional intervention (Jenkins, Hudson, & Johnson, 2007).

While most researchers argue for administering screening protocols in the early grades, predicting which preliterate students are at-risk for developing reading disabilities proves problematic. Torgesen (2002) discussed the disparity in results, arguing that preschool and kindergarten screening tools have identified too many students as in need of intervention who did not actually develop reading problems later in life. These cases are called false-negatives, with percentages ranging from 20% to 60%.

Conversely, Compton et al. (2010) argued that reading assessment improves as students encounter more reading instruction. For this reason, these authors advocated screening students at the beginning of first grade rather than in preschool or kindergarten. However, accuracy did not improve statistically, as false negatives remained low but well outside the acceptable range. A false negative would be recorded if a student was scored in the average range, but indeed had a reading problem. Even so, screening all students at the beginning of the school year still remained a valid method to identify students at-risk for poor reading outcomes, as false negative rates may have actually reflected children’s response to the intervention received.

In an effort to improve screening accuracy of preliterate children, Wilson and Lonigan (2009) administered the Get Ready to Read! and the Individual Growth and Development Indicators (IGDI) screening measures. The former measures phonological awareness while IGDI focuses more on reading. Both systems showed promise as early intervention tools, but still
failed to reach the level of accuracy recommended by Jenkins (2007). Consequently, Compton et al. (2006) explored ways to improve the efficiency of universal screening procedures, increase overall classification rates while eliminating false positives, and accelerate the movement of the most at-risk readers to more intensive levels of services. The authors reported a multivariate screening battery that contained measures of phonological awareness, rapid naming, oral vocabulary, and word identification fluency skills. The results produced classification accuracies (sensitivity of .90 and specificity of .83) consistent with the recommendations of Jenkins (2007). Compton et al. (2010) replicated the 2006 study, but added progress monitoring (direct assessment), phonemic decoding, and forecasting (dynamic assessment). The authors reported significant increases in classification accuracy (sensitivity of .90 and specificity of .91) and the elimination of false positives by 40%.

Furthermore, with regard to progress monitoring, Compton et al. (2012) found that when working with students unresponsive to instruction in a small group setting (RTI Tier 2), response data were not necessary to accurately predict a group of students for whom Tier 2 interventions were likely to be unsuccessful. Rather, by using norms on first-grade word identification fluency growth and linking those with distal outcomes of reading disability at the beginning of 3rd grade, the team was able to accurately predict students who would not respond to intervention. This suggested that students could be accurately identified for Tier 3 (1:1) intervention without participating in and failing to benefit from Tier 2 (small group) interventions.

Regarding intervention on the national level, the cornerstone of alternative approaches to learning disabilities identification, as outlined in the reauthorization of IDEA (2004), was the provision requiring the measurement of a student’s academic outcomes, according to research and scientifically based strategies. In Response to Intervention, progress monitoring is used to
move students between levels of intervention. As previously discussed, the fluidity between levels can be somewhat ambiguous. Therefore, Fuchs, Fuchs, and Compton (2004) examined two reading intervention protocols, Word Identification Fluency and Nonsense Word Fluency, and their impact on assessing early reading development skills. With Word Identification Fluency, students have one minute to read isolated high-frequency words presented in a list containing 50 words. By contrast, Nonsense Word Fluency presents students with a single page of 50 consonant-vowel-consonant or vowel-consonant-pseudo-words. Fuchs et al. (2004) monitored students for 20 weeks and found Word Identification Fluency a better intervention tool, based on the ability to predict end-of-year reading skill.

In a follow-up study, Zumeta, Compton, and Fuchs (2012) examined whether sampling procedures for developing Word Identification Fluency had an effect on growth parameter estimates and the correlation between student outcome growth estimates and subsequent reading skills. Three samples of students were taken from the overall pool, a representative sample which reflected the distribution of readers in the study, and included low, average, and high achieving students; a second sample that included all students with low reading achievement; and a third sample with high/average achievement. Word Identification Fluency data were collected weekly for 15 weeks using two different lists – broad lists and narrow lists. Sampling from 500 high-frequency words revealed broad lists, whereas sampling the 133 words from the Dolch preprimer, primer, and first-grade word lists created narrow lists. Overall, narrow sampling was found to be better for screening the representative group and the high/average subgroup. Broad sampling was superior for screening the low-achieving subgroup and for progress monitoring student groups, making Word Identification Fluency well suited as a
screening and progress-monitoring tool for making accurate decisions regarding movement within a tiered RTI model.

Similarly, Universal Screening is highlighted by dynamic assessment, defined as supporting learner development through understanding learner deficiencies. Dynamic assessment was traditionally used as a short but focused session of instruction intended to gauge whether a particular intervention proved useful to the student, often yielding varied results (Elleman, Compton, Fuchs, Fuchs, & Bouton, 2011).

In related research, Bridges and Catts (2011) developed and examined the predictive validity of a dynamic assessment screening of phonological awareness in two samples of students who were administered the dynamic assessment at the beginning of the kindergarten year and standardized measures of reading achievement at the conclusion of the year. In the first sample, the predictive utility of dynamic assessment was compared to a static version of the same screening assessment where no feedback or support was provided. Positive results were evident as dynamic assessment accurately forecasted the reading ability of developing readers. By comparison, in the second study, the predictive utility of dynamic assessment was compared to Dynamic Indicators of Basic Early Literacy Skills (DIBELS) Initial Sound Fluency, a commonly used screening measure for beginning readers. Findings here also revealed that dynamic assessment predicted kindergartners’ end-of-year reading skills over and above what was measured by Initial Sound Fluency alone. Using duplicated methodology, similar results were reported with second-grade and fourth-grade students. Likewise, revisiting the use of the Individual Growth and Development Indicators (IGDI) progress monitoring system, modified with a web-based decision making tool for practitioners, may improve the ability to implement effective early literacy intervention with at-risk students.
Computer Adaptive Testing

Iterative processes continue to be part of the instructional cycle during direct instruction and guided practice. The exceptional educator seeks to make connections with what students already know, and understanding the case for digital natives and their craving for lightning swift technology, educators turned to technology integration to enhance and remediate learning (Friedman, 2005). However, a continued national focus on accountability forced instructional leaders to continuously investigate how technology could improve teaching and learning. Within the last five years, numerous technology platforms have focused on using technology to predict student results on standardized measures. As such, the use of computer-assisted programs gained popularity. One such method, Computer Adaptive Testing (CAT), is a form of computer-based assessment that adapts to the ability level of the examinee.

In the state of Texas, educators are admonished to teach higher order cognitive skills rather than teaching students how to take tests or teaching to a test. As a result, in 2015 the 83rd regular session of the Texas Legislature enacted House Bill 5 (Texas House Bill 5, 2013) to overhaul the state’s curriculum and graduation requirements. Additionally, the bill substantially changed the accountability landscape in Texas by reducing the number of end-of-course examinations and limiting the number of campus-wide benchmark assessments administered to students. Since the stated purpose of benchmarks is to inform instruction, instructional leaders debated how to better identify students in need of remediation and if such remediation could accurately predict subsequent STAAR scores.

Weiss and Kingsbury (1984) provided the framework for the application of CAT to educational problems. Affirmed by the work of Thissen and Mislevy (2000), CAT maximizes assessment precision by using prior test performance to tailor subsequent test constructs. For
clarity, first-testing administrations set baseline data by affording the facilitator the right to select ranges of difficulty (medium or medium-easy). From the perspective of the examinee, CAT tailors progression through the test, based on the test-taker’s ability level. For example, grade-level performance on a subset of questions would bring about a more difficult set. Likewise, poor performance would bring questions of equal or lesser difficulty for the purpose of remediation. In contrast to traditional multiple-choice formats, CAT does not use a fixed set of questions that are administered to all test takers and requires fewer questions to provide accurate feedback. As Green (2000) noted, due to adaptive administration of CAT, there is a probability that each examinee receives different test versions with each assessment.

Piton-Gonçalves and Aluisio (2012) discussed the basic CAT methodology, using an iterative algorithm, as in the following steps:

1. The pool of available items is searched for the optimal item, based on the current estimate of the examinee’s ability.
2. The chosen item is presented to the examinee, who then answers it correctly or incorrectly.
3. The ability estimate is updated, based upon all prior answers.
4. Steps 1–3 are repeated until a termination criterion is met.

CAT is advantageous for intervention because scores are immediately available after testing. Additionally, since each test taker theoretically receives a different test, cheating and fraud become less of a consideration. Moreover, when used as an intervention tool, CAT builds teacher capacity for differentiated instruction, a cornerstone of reform efforts. Likewise, an adaptive test can be shortened by as much as 50% and still maintain a higher level of accuracy. That is, a person’s ability can be measured with relatively few items (Halama, 2005; Weiss, 2004). Moreover, CATs are excellent tools for formative assessments that guide subsequent instruction, as they utilize Item Response Theory to model the relationships between ability (cognitive attributes) and item responses.
While the benefits of CAT are numerous, research supports some disadvantages worth noting, including calibrating the item pool, referencing the starting point, and setting termination criteria. Haley et al. (2011) elaborated on these challenges. Regarding the item pool and in order to categorize the characteristics of the items, all items must be piloted by being administered to a large sample pool and then analyzed, including new items with operational items. Secondly, items in CAT are selected based on the examinee's prior performance. However, CAT cannot make a grounded estimate of examinee ability when no items have been administered. Therefore, some other initial estimate of examinee ability is necessary, such as classroom formative assessments or a pre-test. Finally, termination criteria must be set. CATs are designed to repeatedly administer items and update user ability. This method will continue until the pool is exhausted; therefore, criteria must be set to terminate testing when the user either reaches a preset reliability level or the standard error of measure falls below a given value. Best practice dictates that the latter value should closely correlate with scores deemed as failing on corresponding state assessments.

Item Response Theory and Classical Test Theory

As discussed in the first chapter, Item Response Theory (IRT) is the psychometric technology that allows equitable scores to be computed across different sets of items (F. Baker, 2001; Rudner, 2001). More specifically, Rudner (2001) discussed Item Response Theory “as the study of test and item scores based on assumptions concerning the mathematical relationship between abilities (or other hypothesized traits) and item responses” (n. p.). Other names and subsets include Item Characteristic Curve Theory, Latent Trait Theory, Rasch Model, 2PL Model, 3PL model and the Birnbaum model. Modeling the relationships between ability and a set of items provides the basis for numerous practical applications, most of which have
advantages over their classical measurement theory counterpart (Li & Sireci, 2005; Rudner, 2005).

By contrast, the Classic Test Theory (CTT) refers to older models of measurement and achievement. Gregory (2011) explained how the CTT was born only after the following three achievements or ideas were conceptualized: a recognition of the presence of errors in measurements, a conception of that error as a random variable, and a conception of correlation and how to index it. Charles Spearman (1904) postulated a formula to correct for measurement error due to issues with the correlation coefficient and this was thought to be the beginning of CTT. More importantly, Spearman’s work accounted for obtaining the index of reliability needed in making the correction. Essentially, unlike IRT, CTT assumes that each person has a true score that would be obtained if there were no errors in measurement. A person's true score is defined as the expected number-correct score over an infinite number of independent administrations of the test. In educational assessment discussions, a true score is not a consideration (most standardized measures instead use field questions for construct reliability), as observed score is what is reported. Therefore, CTT has reliability issues for achievement reporting, since a true score cannot be obtained.

Although there are many negative aspects to standardized testing measures, such as the narrowing of the curriculum, exclusion of non-tested subject areas, excessive test preparation time, adapting teaching styles to testing formats, disproportionate impact on disadvantaged students, and misleading testing results, there are positive outcomes that are being recognized by many practitioners (Baker & Johnston, 2010; Blazer, 2011). Baker and Johnson (2010) further explained that positive effects on education encompass better teaching strategies, continued study of high reliability remediation of low-socioeconomic students, better teacher professional
development opportunities, and a more data driven culture. Though these improvements have
given educators a more specific and focused goal, there have been no gains in closing
achievement gaps and only marginal improvements in testing scores overall, as confirmed by the
2015 NAEP report (Kena et al., 2015).

Validity of Computer Adaptive Testing

Technology has played an integral role in school achievement tests since the mid-20th
century. Many achievement and aptitude tests have computerized versions, but current
educational policy reforms place emphasis on methodological improvements to increase
effectiveness. To achieve precision and efficiency in assessments, CAT was reviewed due to its
ability to precisely estimate respondents’ level of function. It is essential to mention that CAT
does not measure construct, so it is not intended to yield generalizable results between computer-
based (CB), computerized adaptive (CA), or paper-pencil (PP) formats. However, results on
CATs are psychometrically comparable to other modes (Butcher, Perry, & Hahn, 2004).

Zitny (2012) analyzed 15 studies of ability testing in clinical psychology in order to
explore the reliability, utility (comparability), and validity of CAT. Summarized findings
indicated CAT provided an effective means of gaining an optimal amount of information needed
to answer the assessment question, while keeping time needed to gather that information to a
minimum. CAT scores correlated significantly with the full item bank (r = 0.83 – 0.99) and
moderately with the established measures (range r = 0.58 – 0.83), providing the evidence for
reliability, validity, and comparability of adaptive tools.

In a related study, Zitny, Halama, and Jelinek (2012) examined correlations between
school achievement and standardized test scores on the Tests of Intellectual Potential (TIP) and
the Vienna Matrices Test (VMT). The authors administered paper and pencil and computerized
adaptive testing with 567 participants. The results demonstrated a significant difference only between the correlations for students’ school achievement and mathematics (paper and pencil $r = -0.33$) and computerized adaptive ($r = -0.15; x=2.08, p=0.038$). No other significant differences between the correlations across modes for student achievement were found.

While research reveals disagreement over the use and legal consequences of CAT in organizations and executive assessment (Schmidt, 2012), its use in education reform has been applauded due to the ability to measure individual progress over time. This fact not only aligns with the specifications of NCLB (NCLB, 2001) but also is a provision of best practice. Continued interest abounds for educators wishing to understand whether a student’s level of achievement, understanding, or performance has changed due to instructional interventions (Wackwitz & Burniske, 2009).

Measures of Academic Progress

Educators continue to base instructional interventions on scores from benchmark measures. Benchmarks, or interim assessments, serve as progress monitoring tools, allowing educators to gauge whether such progress will allow the student to reach end-of-year goals (Torgesen & Miller, 2009). Furthermore, the use of standardized benchmark measures is used, in part, to differentiate instruction through individualized intervention. Effective differentiation involves the use of many data sets, including recognizing and teaching to student interests and understanding prior readiness levels (McTighe & Brown, 2005). While teachers have traditionally used benchmarks to monitor learning, it is challenging for them to equate performance on classroom measures as a predictor of performance on state achievement tests or nationally-normed standardized tests. Benchmark assessments reflective of such external tests are potentially more useful in helping teachers make decisions about differentiating instruction,
which, in turn, can lead to student learning gains, higher scores on state standardized tests, and improvements in school-wide achievement (Baenen, Ives, Warren, Gilewicz, & Yaman, 2006; Baker & Linn, 2003).

Ash (2008) reported that one of the most widely used benchmark assessments is the Measures of Academic Progress (MAP). The MAP program consists of Computer Adaptive Testing (CAT) administered multiple times during the year as well as training for educators on effectively using benchmark measures for student growth. While research has created dialogue regarding the effectiveness of MAP, little research has examined how MAP influences achievement outcomes (Ash, 2008; Clark, 2005; Olson, 2007).

However, empirical studies investigating the effects of benchmark assessment on student achievement have begun to emerge. Borman, Carlson, and Robinson (2010) reported the results of a multistate district-level cluster randomized trial investigating the impact of benchmark assessment on student achievement. In a sample of 509 schools across 56 districts in seven states (Alabama, Arizona, Indiana, Mississippi, Ohio, Pennsylvania, and Tennessee), results showed significant positive effects of the intervention on students’ state test scores in mathematics (d = 0.21) but not in reading (d = 0.14; p-value = .10).

Additionally, Dahlin and colleagues provide evidence that MAP assessments accurately predict scores on state standardized tests, using growth data collected from hundreds of thousands of students across the United States. This information was gleaned from The Kingsbury Center at the NWEA data repository (Dahlin, Xiang, Durant, & Cronin, 2010). Their study consisted of more than 30 schools focused on reading and language arts in grades four and five. Using a control group and a study group, half of the study population received the MAP intervention while the control group did not. The program was implemented with moderate
fidelity and subsequent program results indicated that the MAP program did not have a statistically significant impact on student achievement in grades four and five. However, MAP results did reveal positive outcomes for students identified as academically at-risk (Tomlinson & McTighe, 2006).

Furthermore, studies in Georgia and Delaware have found reliability in using CTT models of measurement and achievement (Hall-Michalcewiz, 2008; Jones, 2015). Both studies showed the ability of the Measurement of Academic Progress (MAP) assessment to diagnose and support the improvement of student achievement levels in reading and mathematics. Additionally, both studies noted the improvement in teaching strategies and professional development gleaned from skillfully using, interpreting, and analyzing the data MAP assessment provided.

The MAP assessment includes an interactive tool for teachers, called the Learning Continuum. Teachers that utilize the Learning Continuum can see where the students are academically and at what RTI level they are performing. Furthermore, the teacher can determine whether the student is ready to learn, using the learning statements within the continuum to guide individual learning and differentiated instruction for all students. Together with other assessment measures, such as formative assessments, informal observations, class work, quizzes, and standardized tests, the teacher can design instruction that will best meet the needs of each individual student (Northwest Evaluation Association, 2015).

Summary

The effectiveness of public education continues to be a critical concern in the United States. As a result, the ability to measure school effectiveness has become the pivotal issue in education policy and legislation (Fry, 2008; N. A. Thompson & Weiss, 2011). The No Child
Left Behind (NCLB, 2001) act aimed to close the achievement gap between high and low performing children, especially the achievement gaps between minority and nonminority students, and between disadvantaged children and their more advantaged peers.

This gap is exacerbated in literacy. Students from diverse populations, including English Language Learners (ELLs), continuously lag behind their peers when assessing reading on grade level. Particularly, according to the 2015 NAEP (2015) report, there have been no significant gains in literacy education over the past four decades. Students in grade four reading showed marginal gains; however, these gains did not continue through the middle grades. Even so, students from traditionally underrepresented populations continue to lag behind academically.

Consequently, as states continue to search for measures to address academic disparity of all students through early identification and intervention, demands have increased for progress-monitoring strategies that reliably predict outcomes on statewide assessments (McGlinchey & Hixson, 2004). According to Kingsbury and Hauser (2004), Computer Adaptive Testing (CAT) may be a viable option. CAT, as an intervention tool, affords the opportunity for educators to build teacher capacity for differentiated instruction, a cornerstone of reform efforts. Likewise, an adaptive test can be shortened by as much as 50% and still maintain a higher level of accuracy. That is, a person’s ability can be measured with relatively few items (Halama, 2005; Weiss, 2004). Moreover, CAT assessments are excellent tools for formative assessments that guide subsequent instruction, as they allow for immediate reporting of student achievement and the measuring of progress over time. CAT assessments adapt to and are tailored for the intervention needs of each student.

While teachers have traditionally used benchmarks to monitor learning, it is challenging to equate performance on classroom measures as a predictor of performance on state
achievement tests or nationally normed standardized tests. Benchmark assessments reflective of such external tests are potentially more useful in helping teachers make decisions about differentiating instruction, which, in turn, can lead to student learning gains, higher scores on state standardized tests, and improvements in school-wide achievement (Baenen et al., 2006; Baker & Linn, 2003). One of the most widely used computer adaptive tests, Measures of Academic Progress (MAP), shows progress as a benchmark tool, thus is the focus for this study.
CHAPTER 3

METHOD

The purpose of this predictive study was to determine if the Measures of Academic Progress (MAP) assessment is a statistically valid predictor of student performance on the State of Texas Assessment of Academic Readiness (STAAR) seventh-grade reading assessment. A predictive research design seeks to find relationships between independent and dependent variables (Brewer & Kubn, 2010).

Research Question

The following question was addressed in this study:

To what extent do certain demographic variables (race/ethnicity, gender, socioeconomic status) explain, and MAP reading scores predict, reading scores on the State of Texas Assessment of Academic Readiness (STAAR) for seventh-grade students in a selected northeast Texas public school?

Overview of the Study

This research study was an exploration of various predictive factors related to student performance on the STAAR reading test. This predictive design used a parametric analysis technique and multiple regression as the method. Predictors were scores on the MAP Reading Rasch Unit (RIT) assessment as well as three demographic factors, gender, race/ethnicity, and socioeconomic status (SES).

A multiple regression model allows an outcome by studying the relationship between a dependent variable and one or more independent variables or predictors, whose contribution can be assessed collectively and individually. Specifically, the contributions of the independent variables to predict the dependent variable is what the researcher is trying to measure. Simply
stated, this design was capable of predicting which among the independent variables are related to the dependent variable and to explore these relationships in the regression (Field, 2005). Field noted,

Correlations can be a very useful research tool but they tell us nothing about the predictive power of variables. In regression analysis, we fit a predictive model to our data and use that model to predict values of the dependent variable from one or more independent variables. (p. 144)

The dependent variable for this study was based on a scale score from one assessment to another. The multiple regression model was used to predict scores on the STAAR reading exam, based on the predictor variables and to identify what relationship, if any, exists between the independent variables and scores on the STAAR reading exam. In a multiple regression, a relationship between predictor variable and an outcome variable can be estimated (Field, 2005; Kedem & Fokianos, 2002).

Permissions to obtain data were garnered from the participating school district and the University of North Texas Institutional Review Board (Appendix). State archival test data were used in the sample when a student had a Spring 2013 MAP reading score and Spring 2013 STAAR score for cohort one. The same method was used for cohort two when the student had both a Spring 2014 MAP reading score and Spring 2014 STAAR score. Demographic information and test-score data were collected by the state from students in grade seven during the 2012-2013 and 2013-2014 school years; these archived state data were accessed for the study participants who were from a district located in a northeast Texas public school.

Two cohorts of participants were included in this study. Scores for the participants were included only if the student was present during the spring testing of the 2013 seventh-grade STAAR reading assessment for cohort one and for the 2014 seventh-grade STAAR reading assessment for cohort two. Additionally, the student’s MAP score was considered only if the
student had both fall and spring reading assessment scores for the 2012-2013 school year for cohort one or 2013-2014 school year for cohort two. Furthermore, only the standard STAAR seventh-grade reading assessment was considered in the data. Alternative tests such as STAAR alternative, STAAR A, STAAR M, and STAAR L were not used.

Participants

This was a convenience sample because of my relation with the school district and, thus, my ability to acquire this information. The participants in the study included two cohorts of students enrolled in grade seven in a selected northeast Texas school who were administered the STAAR reading assessment during the 2012-2013 and 2013-2014 academic years.

Of the 2012-2013 cohort of 240 students, 49.2% were male (n = 118), while 50.8% were female (n = 122). Further, the racial composition of the participants includes 2.8% (n = 7) Asian or Other Pacific Islander, 20.0% (n = 50) Black, 35.2% (n = 88) Hispanic, 4.0% (n = 10) Two or More Races, and 38.0% (n = 95) White. Other variable groups include the following categories: age, in which 57.6% (n = 144) were 13-year-old students, 39.6% (n = 99) were 14-year-old students, and 2.8% (n = 7) were 15-year-old students; special education, in which 6.4% (n = 16) of participants fell in this category; English language learners, in which 5.6% (n = 14) are LEP; and socio-economic status, in which 47.6% (n = 119) were classified as low socio economic.

Of the 2013-2014 cohort of 234 students, 50.4% (n = 123) were male, while 49.6% (n = 121) were female. Further, the racial composition of the participants includes 23.3% (n = 8) Asian or Other Pacific Islander, 19.7% (n = 48) Black, 34.4% (n = 84) Hispanic, 2.0% (n = 5) Two or More Races, and 40.6% (n = 99) White. Other variable groups include, Age, 58.6% (n = 143) 13-year-old students, 39.3% (n = 96) 14-year-old students, and 2.0% (n = 5) 15-year-old
students; 4.5% (n = 11) SPED; 7.0% (n = 17) LEP; and 51.2% (n = 125) Low Socio Economic Status.

Students who did not have both a Spring 2013 and Spring 2014 MAP score and a Spring 2013 and Spring 2014 STAAR reading score were excluded for the study. Additionally, students who participated in STAAR modified of STAAR Alternative in Spring 2013 or Spring 2014 were also excluded from the study. STAAR Modified and STAAR Alternative are designed for students who have significance cognitive disabilities and are receiving special education services.

Procedure for Data Analysis

This section provides a description of the variables, both dependent and independent, and the procedures employed in the regression. SPSS version 22 was used for all analyses.

Dependent Variables

The dependent variable in this study was the met/not met standard on the State of Texas Assessment of Academic Readiness (STAAR) reading test, the state’s current student achievement assessment program. The STAAR scores were treated as a continuous numerical value in the current study, as each student receives a scaled score assessing performance in reading (Texas Education Agency, 2015).

Independent Variables

Reading score. The score for reading was a dichotomous categorical determination: met/not met. This refers to either meeting or not meeting grade level achievement on the Measures of Academic Progress (MAP). This particular test adapts to meet the student’s learning level and gives the teacher a personal snapshot of the student’s growth and progress. The MAP, developed by the Northwest Evaluation Association (NWEA), is a computer adaptive
test that is designed to regulate itself to match the performance of the student after a response is
given by the student (Northwest Evaluation Association, 2015). Psychometric theory is the basis
for the MAP assessment and stems from the use of IRT as a framework in designing the
assessment. The level of the question is adjusted according to the answers. The assessment then
levels the next question, based on the student’s academic mastery. This process is repeated until
the assessment is consistently at the student’s appropriate mastery level (Northwest Evaluation

Additionally, the MAP assessment uses psychometric theory to show evidence of validity
for measuring the achievement level and academic progress of the student. The MAP
assessment continues to use psychometric testing and computer adaptive testing processes
developed by the Northwest Evaluation Association (NWEA) to refine the quality and efficiency
of the MAP assessments. Although these standards were not initially developed by NWEA,
NWEA researchers continue to use the standards established by professional organizations in the
field (American Educational Research Association, American Psychological Association, &
National Council on Measurement in Education, 1999). Furthermore, the psychometric
soundness of the assessment is its primary source of reliability and is expressed on the MAP
assessment in the same manner as a correlation coefficient. Student achievement levels of the
Spring 2013 and Spring 2014 tests were used to determine the RIT cut line scores that denote
students that are on or above the appropriate seventh-grade reading level. Using these RIT cut
scores, each participant received a code based on his or her success or failure to meet a grade-
level rating.

Gender. For gender, the categorical data drawn from the state assessment data were
dichotomous in nature. Data were coded as a dummy variable with 1 for males and 0 for
females.

Race/Ethnicity. For race/ethnicity, data came from state assessment data. Categories included white, Hispanic, African American, and Asian. Each category received its own column and the following scheme was employed for entering data as dummy variables in the regression. A person who originates from the Far East, Southeast Asia, the Indian subcontinent, or the Pacific Islands was considered Asian or Other Pacific islander. These participants were coded 1 and all others 0. A person who originates from peoples of the black racial groups of Africa was considered Black and were coded 1 and all others 0. A participant of Mexican, Puerto Rican, Cuban, Central or South American, or other Spanish culture or origin was considered Hispanic and was coded 1 and all others 0. A participant having origins that fit into more than one racial group was considered two or more and was coded 1 and all others 0. A participant having origins in any of the original peoples of Europe, the Middle East, or North Africa (National Center for Education Statistics, NCES, 2002) was considered White and was coded 1 and all others 0.

Socioeconomic status (SES) is represented in this study by those students who qualified for the National School Lunch Program, either through free or free-and-reduced lunch prices. The National School Lunch Program is a federally-assisted meal program that provides low-cost or free lunches to eligible students (United States Department of Agriculture, 2015). The program is referred to the as the free/reduced-price lunch program and is offered to those students whose family incomes are at or below 130% of the poverty level; also, reduced-price lunches are offered to those students whose family incomes are between 130% and 185% of the poverty level (United States Department of Agriculture, 2008). In the target school, the information is used as a valid indicator of SES because each student that meets the criteria fills
out a free/reduced lunch application. In the present study, SES was measured as a categorical variable with students identified as low SES coded 1, while students not identified as low SES were coded 0.

*Independence of these variables.* These independent measures were checked for multicollinearity using the procedure for cleaning and screening data set forth by Mertler and Vannatta (2010), which included determination of tolerance and VIF statistics. Results of this check are proved in Table 1 below.

Table 1

**Research Question Coefficients for 2013-2014 Seventh-Grade Students’ Fall Reading MAP RIT Results**

<table>
<thead>
<tr>
<th>Model</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
<th>Correlations</th>
<th>Collinearity Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
<td>Std. Error</td>
<td>Beta</td>
<td>t</td>
</tr>
<tr>
<td>1</td>
<td>(Constant)</td>
<td>260.490</td>
<td>147.044</td>
<td>1.772</td>
</tr>
<tr>
<td></td>
<td>Age</td>
<td>.671</td>
<td>8.201</td>
<td>.003</td>
</tr>
<tr>
<td></td>
<td>Sped</td>
<td>12.443</td>
<td>23.257</td>
<td>.023</td>
</tr>
<tr>
<td></td>
<td>LEP</td>
<td>19.194</td>
<td>18.754</td>
<td>.047</td>
</tr>
<tr>
<td></td>
<td>EconDis</td>
<td>-27.006</td>
<td>10.537</td>
<td>-.127</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>4.680</td>
<td>8.878</td>
<td>.022</td>
</tr>
<tr>
<td></td>
<td>AfAm</td>
<td>-.7287</td>
<td>13.492</td>
<td>-.028</td>
</tr>
<tr>
<td></td>
<td>Hisp</td>
<td>-1.451</td>
<td>11.462</td>
<td>-.007</td>
</tr>
<tr>
<td></td>
<td>MIXRACE</td>
<td>-33.926</td>
<td>20.281</td>
<td>-.066</td>
</tr>
<tr>
<td></td>
<td>RitFall</td>
<td>6.381</td>
<td>.398</td>
<td>.748</td>
</tr>
</tbody>
</table>

Dependent Variable: Scale Score

* Significant at the p > .05 level
** Significant at the p > .001 level

In this standard multiple regression, all independent variables were entered into the analysis simultaneously. The effect of each independent variable on the dependent variable is
assessed as if it had been entered in the equation after all other independent variables had been entered. Each independent variable is then evaluated in terms of what it adds to the prediction of the dependent variable, as specified by the regression equation.

Multicollinearity happens when moderate to high intercorrelations occur among independent variables used in a regression analysis. Multicollinearity is not a problem as the tolerance for Age, Sped, LEP, EconDis, Female, Asian, AfAm, Hisp, and RitFall do not exceed 1.00 and are not less than .1; additionally, the VIF does not exceed 10 for any of the variables as well (Mertler & Vannatta, 2010). To assess for multicollinearity, the tolerance statistics and the variance inflation factor were reviewed. Tolerance values less than 0.1 and variance inflation factors greater than 10 indicate issues with multicollinearity. Tolerance values for each independent variable are all greater than 0.1; variance inflation factors for each independent variable are all less than 10; therefore, multicollinearity is not a concern (see Table 1). All assumptions were met; therefore, a multiple regression technique is an appropriate analysis.

Means of Conducting the Regression

The data were entered into SPSS using a regression model to determine the presence of a statistical trend or lack thereof. Demographic data were entered first in the regression model and MAP scores were entered second to examine the predictive validity of the MAP assessment in determining student performance on the seventh-grade state-mandated reading assessment, the STAAR. During this examination of data, patterns and trends were observed. Any pattern and trend in the quantitative data sets that could be organized systemically to identify and characterize the trends that are statistically significant were analyzed. Additionally, a student’s successful performance on the STAAR seventh-grade reading assessment was noted (Irby et al., 2013).
Once a multiple regression model was identified as a good fit for the study, the model was conducted on both cohorts of data using SPSS version 22 to run the analysis. The model employs the use of an external replicability analysis. This allows the researcher to derive a more accurate analysis from these two demographically similar participant cohorts (B. Thompson, 2006). The purpose of replicating the model on the second set of data was to ensure the model holds and was indeed reliable in predicting student success on such high-stakes exams. Thompson (1994) recommended replicating models involved in high-stakes decisions to ensure that decisions are not based on erroneous models, which can have detrimental effects on students and their future. Using multiple regression allowed prediction of the outcome that most often occurs when nothing is known other than the values of the outcomes of 1 (the outcome did occur) or 0 (the outcome did not occur). Additionally, it was possible to observe and compare the predicted values with the outcomes, using the Pearson correlation coefficient. Using the Pearson correlation statistic for comparing the outcome allowed analysis of independent variables and their statistical significance using the regression model Beta Weights.

Summary

The purpose of this chapter was to report an overview of how the study was conducted. In this chapter, the research questions and an overview of the study, participants, and procedures for data analysis were described. Multiple regression analysis was well suited for a study of this nature to determine the effectiveness of the MAP assessment, in concert with various demographic factors, in predicting STAAR assessment passing rates on seventh-grade reading scores. The study was intended to examine relationships among variables. Chapter 4 provides the results of the data analysis.
CHAPTER 4

RESULTS

The purpose of this study was to examine student data over a two-year period to determine if a computer adaptive test such as the MAP assessment is a statistically valid predictor of student performance on the state-required reading STAAR test. For this study, as part of a multiple regression analysis technique, the Pearson $r$ correlation coefficient was used. The Pearson $r$ (reported as adjusted $R^2$) is considered to be the most stable measure of correlations and is the most accepted measure in correlating variable data that are either interval or ratio (Field, 2005). Statistical analyses were conducted using SPSS, Version 22. Chapter IV includes descriptive statistics, the results of the multiple regression analysis, and the answer to the research question in this study. The chapter concludes with a summary of the results.

Descriptive Statistics

Data used in this study include seventh-grade students’ reading STAAR scale scores, reading MAP RIT scores, and characteristics of race, gender, and economically disadvantaged status. Information from the 2012-2013 cohort of 250 students and the 2013-2014 cohort of 244 students from a selected northeast Texas school were included in the analyses. Table 2 represents the percentage of students included in the analyses by race, gender, and economically disadvantaged status for the 2012-2013 cohort and Table 3 represents the same for the 2013-2014 cohort. Both tables have adjusted sample sizes after the elimination of univariate and multivariate outliers, as recommended by Mertler and Vannatta (2010), which is outlined in greater detail below. Descriptive statistics were used to quantitatively characterize the features of the sample data.
Table 2

2012-2013 Demographic Percentages, N = 240

<table>
<thead>
<tr>
<th>Cohort</th>
<th>Asian</th>
<th>Black</th>
<th>Hispanic</th>
<th>Two or More Races</th>
<th>White</th>
<th>Male</th>
<th>Female</th>
<th>Economically Disadvantaged</th>
</tr>
</thead>
<tbody>
<tr>
<td>12-13</td>
<td>0.0*</td>
<td>20.8</td>
<td>35.8</td>
<td>4.2</td>
<td>39.2</td>
<td>49.2</td>
<td>50.8</td>
<td>46.7</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SpEd</th>
<th>LEP</th>
<th>Age</th>
</tr>
</thead>
<tbody>
<tr>
<td>12-13</td>
<td>5.8</td>
<td>5.0</td>
</tr>
</tbody>
</table>

*Due to the small number of Asian students and the variability associated with this classification in relation to the other independent and dependent variables, this demographic category was eliminated from the analysis as it appeared in the analysis as an outlier.

Table 3

2013-2014 Demographic Percentages, N = 234

<table>
<thead>
<tr>
<th>Cohort</th>
<th>Asian</th>
<th>Black</th>
<th>Hispanic</th>
<th>Two or More Races</th>
<th>White</th>
<th>Male</th>
<th>Female</th>
<th>Economically Disadvantaged</th>
</tr>
</thead>
<tbody>
<tr>
<td>13-14</td>
<td>1.7</td>
<td>20.5</td>
<td>35.5</td>
<td>0.0*</td>
<td>40.6</td>
<td>50.9</td>
<td>49.1</td>
<td>50.9</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SpEd</th>
<th>LEP</th>
<th>Age</th>
</tr>
</thead>
<tbody>
<tr>
<td>13-14</td>
<td>3.8</td>
<td>7.3</td>
</tr>
</tbody>
</table>

*Due to the small number of students who reported being Two or more Races and the variability associated with this classification in relation to the other independent and dependent variables, this demographic category was eliminated from the analysis as it appeared in the analysis as an outlier.

Results for the 2012-2013 Cohort

To answer the research question (Can reading scores on the Measures of Academic Progress (MAP) and demographic characteristics significantly predict reading scores on the State of Texas Assessment of Academic Readiness (STAAR) for seventh-grade students in a selected northeast Texas public school?), a multiple regression analysis was performed using students’
STAAR scale scores as the dependent variable and students’ fall reading MAP RIT scores, as well as race, gender, special education status, language status, and economically disadvantaged status as the independent variables.

When using multiple regression analysis, each categorical variable is limited to having no more than two categories. There were five categories (Asian, African American, Hispanic, Two or More Races, and White) in the independent variable of ethnicity; therefore, the variable was dummy-coded. When including a categorical variable with more than two categories in a multiple regression analysis, the researcher must recode the categorical variable into separate dichotomous values. The ethnicity variable was dummy coded into five separate dichotomous variables: Asian, African American, Hispanic, Two or More Races, and White. The category of White was used as the comparison variable because the number of participants who were White was significantly larger than the other four recoded variables.

A multiple regression was performed with the 2012-2013 student cohort, which was comprised of 240 seventh-grade students after eliminating both univariate and multivariate outliers, as recommended by Mertler and Vannatta (2010). To determine the presence of a linear relationship among the variables included in the model, the Standardized Residuals were plotted against the Unstandardized Predicted Values. A scatterplot, histogram, and normal probability plot were used to confirm the assumption of no multicollinearity. Figure 2 shows that there is no collinear relationship between any of the variables. The mass forms a circle denoting the relationship between the scale score and the other variables for the 2012-2013 cohort, which confirms the assumption of no multicollinearity (Mertler & Vannatta, 2010). Visual inspection of the scatterplot in Figure 2 reveals that the dependent variable and independent variables are linearly related.
Figure 2. The 2012-2013 Cohort Scatterplot of the STAAR Scale Scores and their Standardized Residuals and Predicted Values reveals the lack of multicollinearity. The mass forms a circle denoting the relationship between the scale score and the other variables for the 2013-2014 cohort verifying the assumption of no multicollinearity (Mertler & Vannatta, 2010).

Further examination of the scatterplot reveals the lack of a multicollinear relationship between variables, that is, the variables are not alike enough that combining variables or recoding is necessary (Mertler & Vannatta, 2010). There was independence of residuals, as reported by the Durbin-Watson statistic of 2.053. The Durbin-Watson was used to determine the correlation of the variables. Field (2005) noted that a Durbin-Watson statistic between one and
three indicates a positive correlation of the variables being studied. Additionally, homoscedasticity was used to verify that the predictor variables had similar variances in order to compare them. To confirm homoscedasticity, the Regression Standardized residual was plotted against the Regression Standardized Predictive value. Visual inspection of the scatterplot reveals that the spread of residuals is constant and there is homoscedasticity.

Two methods were used to determine the residuals for the normal distribution: histogram and normal probability plot. Histograms and normal probability plots indicate if there is a normal distribution and a univariate normality among variables. Multivariate normality and homoscedasticity were examined through the generation of a residual plot with another preliminary regression. The residual plot was normal and contained very few outliers. Thus, multivariate normality and homoscedasticity were assumed.

The data were cleaned and screened, including the elimination of univariate and multivariate outliers, according to the procedures recommended prior to analysis by Mertler and Vannatta (2010). Data were first screened for missing data and outliers and then examined to test assumptions. Outliers were identified by calculating Mahalanobis distance in a preliminary regression procedure, as outlined by Mertler and Vannatta (2010). Running the Explore function in SPSS 22 for quantitative variables on the Mahalanobis variable (MAH_1) determined which cases exceeded the chi-square ($X^2$) criteria at the $p < .001$. Elimination of cases that met this very strict criteria effectively eliminated any univariate and multivariate outliers.

Visual inspection of the histogram (Figure 3) reveals that the standardized residuals appear to be normally distributed and the normal probability plot (Figure 4) confirms that the residuals are normally distributed. A normal probability plot shows deviations from normality, if present. A straight line demonstrates a normal distribution and the points represent the observed
residuals and confirms that the data is ready to be run in a statistical tool such as a multiple regression.

Figure 3. This histogram expresses the normality of the residuals for the 2012-2013 cohort. The normal distribution (bell-shaped curve) lies between ±3 standard deviations.

The residuals were analyzed to determine if SPSS identified any outliers. SPSS identified one outlier with standardized results greater than ±3 standard deviations. A visual inspection of both the histogram and normal probability plot testing confirms there are no significant outliers. According to Thompson (2006), larger sample sizes will likely have a few outliers. However, these anomalies are not bad and should be included in the analysis. Analyses were conducted with and without the outliers and no appreciable differences were found in the
results. All assumptions were met for a multiple regression analysis with the data in its natural form, using the aforementioned coding scheme.

The normal probability plot in Figure 4 reveals a normal distribution of the 2012-2013 data set. The linearity of the data is demonstrated by the observed residuals and their proximity to the line throughout the plot and the absence of any severe curvature.

Table 4 shows the MAP reading RIT and met standard coefficients for the 2012-2013 cohort. These variables are a significant predictor of the STAAR scale score ($t(229) = 12.063,$...
\[ p < .001 \text{ and } t(229) = 8.653, p < .001 \]. In Table 5, the other variables (Age, Sped, LEP, EconDis, Female, AfAm, Hisp, and MIXRACE) were not significant contributors to the model.

Table 4

**Significant Coefficients for 2012-2013 Seventh-Grade Analysis**

<table>
<thead>
<tr>
<th>Model</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
<th>t</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
<td>Std. Error</td>
<td>Beta</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>(Constant)</td>
<td>636.255</td>
<td>127.195</td>
<td>5.002</td>
</tr>
<tr>
<td></td>
<td>MetStandard</td>
<td>110.047</td>
<td>12.718</td>
<td>.392</td>
</tr>
<tr>
<td></td>
<td>RITFALL</td>
<td>4.385</td>
<td>.363</td>
<td>.556</td>
</tr>
</tbody>
</table>

Dependent Variable: SCALESCORE

Table 5

**Research Question Coefficients for 2012-2013 Seventh-Grade Students' Fall Reading MAP RIT Results**

<table>
<thead>
<tr>
<th>Model</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
<th>Collinearity Correlations</th>
<th>Collinearity Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
<td>Std. Error</td>
<td>Beta</td>
<td>t</td>
</tr>
<tr>
<td>1</td>
<td>(Constant)</td>
<td>636.255</td>
<td>127.195</td>
<td>5.002</td>
</tr>
<tr>
<td></td>
<td>Age</td>
<td>-2.577</td>
<td>7.491</td>
<td>-.014</td>
</tr>
<tr>
<td></td>
<td>Sped</td>
<td>-5.481</td>
<td>18.687</td>
<td>-.013</td>
</tr>
<tr>
<td></td>
<td>LEP</td>
<td>28.349</td>
<td>20.355</td>
<td>.060</td>
</tr>
<tr>
<td></td>
<td>EconDis</td>
<td>-2.553</td>
<td>9.360</td>
<td>-.012</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>1.659</td>
<td>8.102</td>
<td>.008</td>
</tr>
<tr>
<td></td>
<td>AfAm</td>
<td>-7.561</td>
<td>11.840</td>
<td>-.030</td>
</tr>
<tr>
<td></td>
<td>Hisp</td>
<td>-15.961</td>
<td>10.007</td>
<td>-.075</td>
</tr>
<tr>
<td></td>
<td>MIXRACE</td>
<td>-33.926</td>
<td>20.281</td>
<td>-.066</td>
</tr>
<tr>
<td></td>
<td>Met</td>
<td>110.047</td>
<td>12.718</td>
<td>.392</td>
</tr>
<tr>
<td></td>
<td>RitFall</td>
<td>4.385</td>
<td>.363</td>
<td>.556</td>
</tr>
</tbody>
</table>

Dependent Variable: Scale Score
The result of the multiple regression produced a significant model. This model revealed that there was a statically significant relationship between seventh-grade students’ 2012-2013 fall reading MAP RIT scores, whether these scores were at the met standard level, and their STARR reading scale scores \((F(10, 239) = 28.747, p < .001)\). Fall reading MAP RIT scores and whether those score were at the met standard level accounted for 65.5% of the variance in the STAAR reading scale scores. Table 6 details the seventh-grade 2012-2013 student cohort results of the linear regression analysis for the research question. The regression equations follow:

The following gives the standardized regression equation for the model:

\[
Z_{\text{ScaleScore}} = 0.556Z_{\text{RitFall}} + 0.392Z_{\text{MetStat}}
\]

Unstandardized regression equation:

\[
\text{ScaleScore} = 4.385\text{RitFall} + 110.047\text{MetStat} + 636.255
\]

Table 6

<table>
<thead>
<tr>
<th>Model</th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Regression</td>
<td>10</td>
<td>146664.255</td>
<td>28.747</td>
<td>&lt;.001</td>
</tr>
<tr>
<td></td>
<td>Residual</td>
<td>239</td>
<td>5415.465</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>2686010.176</td>
<td>249</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

a. Dependent Variable: SCALESCORE
b. Predictors: (Constant), RITFALL, ASIAN, AGE, MIXRACE, AFRIAMER, FEMALE, SPED, LEP, ECONDIS, HISP

The multiple regression on fall reading MAP RIT scores and student characteristics was significantly related to STAAR reading scale scores, \((F(10, 239) = 28.747, p = <.001)\). The multiple correlation coefficient was .651, indicating that approximately 65.1% of the variance of STAAR reading scores can be accounted for by the linear combination of students’ seventh-grade fall reading MAP RIT scores and their met standard status. Table 7 details the Cohort 1
(2012-2013) seventh-grade students’ fall reading MAP RIT results of the analysis. A summary model shows the correlation coefficient squared (known as the coefficient of determination, $R^2$). The adjusted $R^2$ in Table 7 shows the amount of variance the predictors have on the STAAR scale scores. Of the variance, 65.1% of the 2012-2013 STAAR reading scores can be explained by the fall MAP RIT and met standard status.

Table 7

*Model Summary for 2012-2013 Seventh-Grade Students’ Fall Reading MAP RIT Results*

<table>
<thead>
<tr>
<th>Model</th>
<th>R</th>
<th>R Square</th>
<th>Adjusted R Square</th>
<th>Std. Error of the Estimate</th>
<th>Durbin-Watson</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.816</td>
<td>.665</td>
<td>.651</td>
<td>60.713</td>
<td>2.053</td>
</tr>
</tbody>
</table>

a. Predictors: (Constant), RITFALL, ASIAN, TWO OR MORE RACES, AFRIAMER, FEMALE, ECONDIS, HISP  
b. Dependent Variable: SCALESCOR

In Table 7, the Durbin-Watson statistic 2.053 demonstrates an assumption that errors in regression are independent. The Durbin-Watson statistic is used to detect an autocorrelation in the residual from a statistical regression analysis (Mertler & Vannatta, 2010). Additionally, the Durbin-Watson is a simple numerical method for checking serial dependence between residuals. Scores close to two (between one and three) demonstrate that the assumption is likely to be met.

Results from the 2013-2014 Cohort

As with the 2012-2013 cohort, a multiple regression model was performed with the 2013-2014 cohort. This cohort was comprised of 244 seventh-grade students. After the elimination of univariate and multivariate outliers, according to the procedure set forth by Mertler and Vannatta (2010), the sample consisted of 234 students. Again, to check for linearity, a Standardized Residuals was plotted against the Unstandardized Predicted values.
Visual inspection of the scatterplot in Figure 5 reveals that the dependent variable and independent variables are linearly related. Like in Figure 2, the scatterplot in Figure 5 reveals the lack of multicollinearity; that is, the variables are not so alike that they require combination or recoding (Mertler & Vannatta, 2010), which is a vital assumption in a multiple regression as multicollinearity can artificially affect the slope of the regression line. There was independence of residuals, as assessed by the Durbin-Watson statistic of 2.021.

Figure 5. The 2013-2014 Cohort Scatterplot of the STAAR Scale Scores and their Standardized Residuals and Predicted Values shows that there is a linear relationship between the variables. The mass forms a circle denoting the relationship between the scale score and the other variables for the 2013-2014 cohort verifying the assumption of no multicollinearity (Mertler & Vannatta, 2010).
As in the previous scatterplot, a Durbin-Watson statistic between one and three indicates a positive correlation of the variables being studied. Additionally, homoscedasticity was used to check that the predictor variables had similar variances such that their relationships could be analyzed. To confirm homoscedasticity, the Regression Standardized residual was plotted against the Regression Standardized Predictive value. Figure 6 supports the required assumption of normality and reveals a normal distribution of residuals.

**Figure 6.** This histogram shows the distribution of standardized residuals expressed in the normality of the residuals for the 2013-2014 cohort. There is a normal distribution (bell-shaped curve) between ±3 standard deviations.
After analyzing Figures 5, the scatterplot supported the assumption of homoscedasticity. Together the histogram (Figure 6) and normal probability plot (Figure 7) were the methods used to determine whether the residuals show a normal distribution. Additionally, a visual inspection of the histogram reveals that the standardized residuals appear to be normally distributed and the normal probability plot confirms that the residuals are normally distributed. The histogram and normal probability plot both confirm there are no significant outliers. Analyses were conducted with and without the outliers and no appreciable differences were found in the results. All assumptions have been met; therefore, again multiple regression is an appropriate analysis technique.

Figure 7. The 2013-2014 Cohort normal probability plot shows the outliers of the standardized residuals. The line and the points represent the observed residuals in this normal probability plot. In a normally distributed data set; all points are on or near the line throughout and no severe curvilinear pattern is observable.
As before, the normal probability plot in Figure 7 shows a lack of deviation of the residuals from normality. The straight line represents a normal distribution, and the points represent the observed residual data from the 2013-2014 cohort.

The multiple regression procedure produced a significant model that revealed a statically significant relationship between seventh-grade students’ 2013-2014 fall reading MAP RIT scores, their SES, and their STAAR reading scale scores \((F(9, 224) = 41.902, p < .001)\). Fall reading MAP RIT scores and student SES status accounted for 62.7% of the variance in the STAAR reading scale scores. The model produced the following regression equations:

The standardized regression equation

\[ Z_{ScaleScore} = 0.748Z_{RitFall} - 0.127Z_{EconDis}. \]

The unstandardized regression equation

\[ ScaleScore = 6.381RitFall - 27.006EconDis + 260.490. \]

Tables 8 and 9 detail the seventh-grade 2013-2014 student cohort results of the multiple regression analysis.

As shown in Table 8, the significant variables \(RitFall\) and \(EconDis\) created the overall significant model (Table 7) to predict \(ScaleScore\) \((RitFall \ t(224) = 16.027, p < .001 & EconDis \ t(224) = -2.563, p = 0.011)\) and there is no notable collinearity issues as the tolerance and VIF statistics are within range (discussed above).

Furthermore, multicollinearity happens when moderate to high intercorrelations occur among independent variables used in a regression analysis. To assess for multicollinearity, the tolerance statistics and the variance inflation factor were reviewed. Tolerance values less than 0.1 and variance inflation factors greater than 10 indicate issues with multicollinearity. Tolerance values for each independent variable are all greater than 0.1; variance inflation factors
for each independent variable are all less than 10; therefore, multicollinearity is not a concern (see Table 5). All assumptions were met; therefore, a multiple regression technique was the appropriate analysis.

Table 8

*Coefficients for 2013-2014 Seventh-Grade Analysis*

<table>
<thead>
<tr>
<th>Model</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
<th>Correlations</th>
<th>Collinearity Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
<td>Std. Error</td>
<td>Beta</td>
<td>t</td>
</tr>
<tr>
<td>1 (Constant)</td>
<td>260.490</td>
<td>147.044</td>
<td></td>
<td>1.772</td>
</tr>
<tr>
<td>EconDis</td>
<td>-.27.006</td>
<td>10.537</td>
<td>-.127</td>
<td>-2.563</td>
</tr>
<tr>
<td>RitFall</td>
<td>6.381</td>
<td>.398</td>
<td>.748</td>
<td>16.027</td>
</tr>
</tbody>
</table>

Dependent Variable: Scale Score

Table 9

*ANOVA Table for Cohort 2 (2013-2014) Seventh-Grade Students’ Fall Reading MAP RIT and Economically Disadvantaged Results*

<table>
<thead>
<tr>
<th>Model</th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Regression</td>
<td>1655867.336</td>
<td>9</td>
<td>183985.260</td>
<td>41.902</td>
<td>&lt;.000&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td>Residual</td>
<td>983557.318</td>
<td>224</td>
<td>4390.881</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>2639424.654</td>
<td>233</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

a. Dependent Variable: ScaleScore
b. Predictors: (Constant), RitFall, Asian, Age, AfAm, Female, Sped, LEP, EconDis, Hisp

A multiple regression also was conducted to evaluate how well Cohort 2 (2013-2014) seventh-grade students’ fall reading MAP RIT scores, as well as student characteristics of gender, ethnicity, and economically disadvantaged status, predicted reading achievement as shown by the STAAR reading scale scores. The variables RitFall and EconDis create the model to predict ScaleScore (F(9,224) = 41.902, p <.001). A review of the beta weights specifies that
only these two variables, $RitFall (t(224) = 16.027, p < .001)$ and $EconDis (t(224) = -2.563, p = 0.011)$ significantly contributed to the model. The variable that is the best predictor of ScaleScore is $RitFall$ because it has the highest Beta weight and the smallest $p$-value (significance).

The multiple correlation coefficient was .612, indicating that approximately 61.2% of the variance of STAAR reading scores can be accounted for by the students’ seventh-grade fall reading MAP RIT scores and economically disadvantaged status. The adjusted $R^2$ in Table 10 shows the amount of variance the predictors have on the STAAR scale scores ($R^2 = .627, R^2_{adj} = .612, F(9, 224) = 41.902, p < .001$).

Table 10

*Model Summary for Cohort 2 (2013-2014) Seventh-Grade Students’ Fall Reading MAP RIT Results*

<table>
<thead>
<tr>
<th>Model</th>
<th>R</th>
<th>R Square</th>
<th>Adjusted R Square</th>
<th>Std. Error of the Estimate</th>
<th>Durbin-Watson</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.792&lt;sup&gt;a&lt;/sup&gt;</td>
<td>.627</td>
<td>.612</td>
<td>66.264</td>
<td>1.824</td>
</tr>
</tbody>
</table>

<sup>a</sup> Predictors: (Constant), RITFALL, ASIAN, AGE, MIXRACE, AFRIAMER, FEMALE, SPED, LEP, ECONDIS, HISP

b. Dependent Variable: SCALESCOR

In Table 10, the Durbin-Watson statistic 1.824 demonstrates an assumption that errors in regression are independent. The Durbin-Watson statistic is used to detect an autocorrelation in the residual from a statistical regression analysis (Mertler, C. A., & Vannatta, R. A., 2010). Additionally, the Durbin-Watson is a simple numerical method for checking serial dependence between residuals. Scores close to two (between one and three) demonstrate that the assumption is likely to be met.
Summary

Chapter IV describes the result of the multiple regression analysis used to answer the research question that guided this study. For the 2012-1013 seventh-grade cohort, fall MAP scores and student met standard status accounted for 65.1% of the variance in STAAR reading scores. In the 2013-2014 seventh-grade cohort, fall MAP scores and students’ SES status accounted for 61.2% of the variance in STAAR reading scores. This means that using the multiple regression model and resulting regression equations for each cohort, and possible future cohorts with similar demographic compositions, can significantly predict, with more than 60% accuracy, students’ STAAR reading scale scores. The implications of these findings are discussed in Chapter V.
CHAPTER 5

SUMMARY, CONCLUSIONS, AND RECOMMENDATIONS

The purpose of this predictive study was to examine a two-year period of student data to determine if computer adaptive testing, specifically the MAP reading assessment, is a statistically valid predictor of student performance on the STAAR seventh-grade reading assessment. Moreover, the purpose was to determine whether student characteristics of gender, economically disadvantaged status, and ethnicity have any significant statistical impact on STAAR seventh-grade reading performance. Data from 474 students enrolled in grade seven in a selected northeast Texas school were included in the analysis. Data used in the study included the MAP reading RIT scores, and the STAAR reading scale scores from two seventh-grade student cohort groups. The MAP RIT scores and the STAAR scale scores used were from the 2012-2013 and the 2013-2014 academic years. A multiple regression research design was used to answer the question below.

To what extent do certain demographic variables (race/ethnicity, gender, socioeconomic status) explain, and MAP reading scores predict, reading scores on the State of Texas Assessment of Academic Readiness (STAAR) for seventh-grade students in a selected northeast Texas public school?

This chapter provides the discussion, implications, limitations, and recommendations for future research.

Discussion

Chapter five presents quantitative evidence that answers the research question that guided this study. Evidence reveals that there is a strong relationship between MAP seventh-grade reading scores and the seventh-grade STAAR reading scores for both the 2012-2013 and the
2013-2014 cohorts. Additionally, there was evidence that there were no statistically significant two-way interactions among any of the predictors. Although the conclusions do not show that one assessment can predict another, a statically significant percentage of the variability of the STAAR seventh-grade reading scale scores can be explained by the seventh-grade MAP RIT scores. Furthermore, the use of the MAP reading assessment, as well as other student variables, may provide prescriptive and diagnostic information to teachers, principals, parents, school testing administrators, and other educators who may use the information to prepare students for successful achievement on state-mandated standardized assessments.

NCLB established mandates with the objective of ensuring that all students receive a high-quality education, as measured by standardized test scores (U.S. Department of Education, 2002a). Smith (2007) noted that NCLB required schools to identify and provide remediation to students that are at-risk of academic failure and further suggested that states, districts, schools and educators have made continuous efforts to ensure students are achieving on-grade-level performance on state-mandated assessments. Their efforts have included prescriptive and diagnostic assessments to guide instructional practices. It is imperative that educators have tools that enable them to understand and to identify the students who are at risk of failing state-mandated assessments.

The results of the current study indicate that MAP assessments can predict student performance at a statistically significant level on STAAR testing; therefore, there are positive implications for using MAP assessment data to provide early intervention screening and intervention strategies to students who are at risk of not obtaining a favorable score on the seventh-grade reading STAAR assessment. Additionally, MAP assessments can be given in the fall, winter and spring to guide instruction, to make continuous decisions about student
performance, and to identify learning gaps. Moreover, the STAAR assessments are only available after the school year is concluded, making the actionable student data inaccessible for use during the school year when the teacher is making important instructional decisions.

Implications

The results of the statistical analysis indicate that the fall MAP RIT scores predicted the STAAR scale scores. Specifically, the fall MAP RIT scores explained 50% of the variability in the STAAR reading scale scores for the 2012-2013 cohort and 59% could be explained in the 2013-2014 cohort. In support of these findings, Jones (2015) found statistically significant correlations between MAP reading scores and the Georgia Criterion Referenced Competency Tests (CRCT) for seventh- and eighth-grade students. Shields (2008) found that the MAP RIT scores were valid predictors for reading achievement, as measured by the Missouri Assessment Program of students in grades six, seven, and eight. Additionally, Andren (2010) and Michalcewiz-Hall (2008) found that MAP reading RIT scores are a valid predictor of the state-mandated New England Common Assessments Program (NECAP) and the Delaware State Testing Program (DSTP). Andren (2010) and Michalcewiz-Hall (2008) further noted that MAP assessments provide teachers, principals, and district testing officials with instruments to help impact instruction in a way that would aide students predestined to do poorly on state-mandated assessments. Furthermore, these authors suggested an increased focus on professional development specifically intended to increase the proficiency of teachers and principals in interpreting and analyzing the data points that MAP provides. Thus, teachers and principals could utilize those data points to increase student assessment performance on standardized assessments.
Although MAP provides much information and guidance on instructional methods and deficits for specific students, it is important for districts to know if funding MAP assessments and spending valuable time out of instruction to administer them is truly worth the information gleaned from these assessments. Also, districts need to know how MAP assessments are being used as a diagnostic tool. Furthermore, national, state, and district funding continues to diminish and there is a great need for assessments that are not only relevant, but diagnostically accurate. There is a vital need for tools that improve a student’s academic development and MAP assessments have been found not only to predict performance on state-mandated assessments, but also to give diagnostic data and tools to improve student academic progress. MAP assessments have been shown to identify students at risk of failing state-mandated assessments and to provide prescriptive and diagnostic information. Additionally, MAP assessments may also be used for Response to Intervention (RTI) early identification and screening purposes. Jenkins, Hudson, and Johnson (2007) recommended using RTI and an early identification and prevention framework to identify at-risk students and provide interventions in a systematic and timely manner. Additionally, Glovers and Albers (2007) suggested that computer-adaptive assessments are a reliable and viable option for universal screening for RTI purposes. Thus, once at-risk students are identified, MAP assessment tools and prescriptive information can be used for diagnostic implementation. Moreover, these tools can be used to help guide instructional decisions and intervention practices in helping struggling students that are not performing on grade level.

Limitations

In any research design, there are threats to internal and external validity. Internal validity refers to the extent to which changes in the dependent variable are directly related to the
independent variable and not due to some other variable, while external validity refers to whether or not the results of research are generalizable to other settings and groups outside the research setting (Fraenkel et al., 2012). The examples of internal validity include: (a) subject characteristics, (b) mortality, (c) location, (d) instrumentation, (e) data collector characteristics, (f) data collector bias, (g) testing, (h) history, (i) maturation, (j) regression, and (k) implementation. Threats to external validity include the ability to make generalizations from the results of the study. The limited population of this study and sample size could be examples of a threat to the external validity. The target population included in this study is composed of seventh-grade students from a northeast Texas school district comprised of two cohorts, one from 2012-2013 and one from 2013-2014, who took the MAP and the STAAR seventh-grade reading assessment. Thus, the findings and results may not be applicable to other schools that are not identically matched demographically and who do not give the same assessments.

Secondly, the findings are not binding to any school or institution, including the one whose data were used in this study. Only the state, federal, and local governments can establish laws, rules, and regulations that bind schools and their operations. This study was intended for academic and professional interest as well as adding to the literature on this subject and was not conceived with the intent to regulate or dictate actions beyond the scope of this study.

**Recommendations for Future Research**

Further research is needed to determine what professional development teachers and administrators need to skillfully utilize the MAP reports that are provided by the Northwest Evaluation Association (NWEA). According to NWEA, the main focus of the MAP assessments is to provide tools and accurate information regarding student growth and achievement. Additionally, MAP can be used to identify skills and concepts students have learned, determine
instructional needs, monitor growth over time, make data-driven decisions regarding students, 
schools, and districts, and determine appropriate placement for students into these programs 
(Northwest Evaluation Association, 2015). It is imperative for schools and school districts to 
have extensive knowledge about how teachers use the data provided by the MAP as well as the 
specific uniformity to which teachers use these reports. This information will be valuable when 
planning professional growth opportunities for teachers and will directly impact students’ 
academic growth.

Another recommendation for further research is to examine scores for students in third, 
eighth, and tenth grades to evaluate performance on the STAAR at those grade levels. Since 
MAP is an assessment of students on a continuum basis, it may assist teachers in identifying 
students at risk of not meeting grade level achievement, at an earlier grade level. As students 
advance in grades, it appears the gaps among high and low achieving students increases (Hall-
Michalcewiz, 2008). If the MAP benchmark assessment provides a valid prediction of student 
performance, this would give teachers ongoing information through tenth grade, thus earlier 
identifying students at risk of not meeting grade level expectations and increasing their chance to 
meet graduation requirements.

Summary

The goal of this quantitative research study was to examine the predictive ability of 
computer adaptive assessments on standardized assessments and the relevance of testing in 
regards to student achievement, growth, and academic development. Additionally, the relational 
roles of the variables and their inevitable factor in student academic achievement and growth 
were analyzed.
Through the analysis of the data, the study constitutes a meaningful contribution to the body of research in regards to predicting student achievement through frequent formative analysis of student achievement using computer adaptive testing measures. Moreover, this study will give credibility to administrators, teachers, and district personnel that use these assessment measures as a resource in guiding instructional decisions for students, specifically, students that are performing below grade level. Additionally, districts that invest in professional development tools that accompany MAP assessment will be focused on mastering classroom formative assessment practices, thus improving assessment performance by shrinking the gap in knowledge and academic skills.

Through the analysis and research invested in this study, convictions about student learning and the use of MAP assessments as a vital tool for improving assessment performance are confirmed. It is anticipated that districts will use this information to continue to harness the information that MAP testing can provide. Specifically, it is hoped that districts will invest in the professional development and data analysis technique that accompany MAP assessments, consequently improving teaching and learning for all students.
REFERENCES


Dahlin, M., Xiang, Y., Durant, S., & Cronin, J. (2010). *State standards and student growth: Why state standards don't matter as much as we thought.* Portland, OR: NWEA.


Texas Education Agency Student Assessment Division. (2009).


Texas House Bill 5, 83rd Legislature, §74.11 (2013).


U.S. Department of Health and Human Services. (2000). *Teaching Children to Read: An Evidence-Based Assessment of the Scientific Research Literature on Reading and Its Implications for Reading Instruction.* (Report of the National Reading Panel No. 00-4769). Washington, DC: National Institutes of Health.


