PREDICTORS OF STUDENT ACHIEVEMENT IN GRADE 7: THE CORRELATIONS BETWEEN THE STANFORD ACHIEVEMENT TEST, OTIS-LENNON SCHOOL ABILITY TEST, AND PERFORMANCE ON THE TEXAS ASSESSMENT OF KNOWLEDGE AND SKILLS (TAKS) MATH AND READING TESTS

A Dissertation

Presented to

The Faculty of the School of Education

Liberty University

In Partial Fulfillment

Of the Requirements for the Degree

Doctor in Education

By

Kristen D. Karrh

August 2009
Predictors of Student Achievement in Grade 7: The Correlations between the Stanford Achievement Test, Otis-Lennon School Ability Test, and Performance on the Texas Assessment of Knowledge and Skills (TAKS) Math and Reading Tests

by Kristen DeAnne Karrh

APPROVED:

COMMITTEE CHAIR Michelle B. Goodwin, Ed.D.

COMMITTEE MEMBERS Frank Alfaro, Ed.D.
Mark Angle, Ed.D.

ASSISTANT DEAN FOR ADVANCED PROGRAMS Scott B. Watson, Ph.D.
ABSTRACT

Kristen Karrh. PREDICTORS OF STUDENT ACHIEVEMENT IN GRADE 7: THE CORRELATIONS BETWEEN THE STANFORD ACHIEVEMENT TEST, OTIS-LENNON SCHOOL ABILITY TEST, AND PERFORMANCE ON THE TEXAS ASSESSMENT OF KNOWLEDGE AND SKILLS (TAKS) MATH AND READING TESTS. (Under the direction of Dr. Michelle Goodwin) School of Education, August, 2009. State-mandated testing programs are more prevalent and have greater implications for public school systems since the passing of the No Child Left Behind (NCLB) Act of 2001 and the implementation of standards-based reform practices. The 331 grade 7 students involved in this correlation research endeavor all attended a public middle school in Texas. One evaluated and analyzed fall 2007 Stanford 10 Achievement Test and Otis-Lennon School Ability Test (8th edition) results, and spring 2008 Math and Reading Texas Assessment of Knowledge and Skills (TAKS) scale scores to determine the statistical significance and strength of the relationship and whether or not a predictive-based relationship existed. The four different research questions were examined and all revealed a positive and significant correlational and predictive relationship between TAKS Reading and Math and both the Stanford 10 Achievement Test and the Otis-Lennon School Ability Test. The resulting predictive relationship between these assessments provides the research study school district and other school systems with an additional diagnostic tool to direct classroom instructional practices and offer areas of strengths and weaknesses as students prepare for state criterion-referenced tests.
# TABLE OF CONTENTS

Signature page...........................................................................................................ii

Abstract......................................................................................................................iii

List of Tables.............................................................................................................vii

List of Figures............................................................................................................viii

Acknowledgements...................................................................................................ix

Dedication..................................................................................................................x

**CHAPTER I: INTRODUCTION**...............................................................................1

  Statement of the Problem.......................................................................................2

  Purpose Statement.................................................................................................3

  Research Questions...............................................................................................4

  Hypotheses.............................................................................................................4

  Significance of the Study......................................................................................5

  Assumptions.........................................................................................................7

  Operational Definitions.......................................................................................7

  Summary...............................................................................................................10

**CHAPTER II: LITERATURE REVIEW**.................................................................11

  Introduction.........................................................................................................11

  Testing and Accountability: A Theoretical Perspective......................................12

  Education and Testing: A Historical Perspective...........................................16

  The Standards Movement: High-Stakes Testing & Accountability................22

  The Role of Norm-Referenced and Criterion-Referenced Tests....................28
Stanford Achievement Test, Tenth Edition……………………………………30
Otis-Lennon School Ability Test (OLSAT), Eighth Edition………………33
Texas Assessment of Knowledge and Skills (TAKS)…………………..35
Summary……………………………………………………………………37
CHAPTER III: METHODOLOGY……………………………………………………39
Research Questions……………………………………………………………39
Hypotheses……………………………………………………………………40
Research Design…………………………………………………………….41
Subjects………………………………………………………………………42
Procedures……………………………………………………………………43
Instrumentation………………………………………………………………44
Data Analysis………………………………………………………………48
Summary………………………………………………………………………49
CHAPTER IV: FINDINGS…………………………………………………………51
CHAPTER V: SUMMARY AND DISCUSSION……………………………...69
Purpose of Study……………………………………………………………..69
Restatement of the Problem……………………………………………….72
Review of Methodology…………………………………………………….73
Summary of the Research Results…………………………………………74
Discussion of the Results……………………………………………………76
Implications………………………………………………………………….82
Limitations……………………………………………………………………83
## List of Tables

Table 1: TAKS Scale Score Ranges

Table 2: Description of Seventh Grade Assessment Sample

Table 3: Correlations between Stanford 10 Reading and TAKS Reading

Table 4: TAKS Reading regressed on Stanford Reading

Table 5: ANOVA: TAKS Reading and Stanford Reading

Table 6: Coefficients from TAKS Reading regressed on Stanford Reading

Table 7: Correlations between Stanford 10 Math and TAKS Math

Table 8: TAKS Math regressed on Stanford Math

Table 9: ANOVA: TAKS Math and Stanford Math

Table 10: Coefficients from TAKS Math regressed on Stanford Math

Table 11: Correlations between OLSAT Verbal and TAKS Reading

Table 12: TAKS Reading regressed on OLSAT Verbal

Table 13: ANOVA: TAKS Reading and OLSAT Verbal

Table 14: Coefficients from TAKS Reading regressed on OLSAT Verbal

Table 15: Correlations between OLSAT Quantitative and TAKS Math

Table 16: TAKS Math regressed on OLSAT Quantitative

Table 17: ANOVA: TAKS Math and OLSAT Quantitative

Table 18: Coefficients from TAKS Math regressed on OLSAT Quantitative
List of Figures

Figure 1: Scatterplot 1, The Relationship between TAKS Reading and Stanford Reading ................................................................. 55

Figure 2: Scatterplot 2, The Relationship between TAKS Math and Stanford Math ................................................................. 59

Figure 3: Scatterplot 3, The Relationship between TAKS Reading and OLSAT Verbal ................................................................. 63

Figure 4: Scatterplot 4, The Relationship between TAKS Math and OLSAT Quantitative ............................................................. 67
ACKNOWLEDGEMENTS

I give thanks to the Lord for walking beside me throughout this entire journey. He stood by me through each mountain and valley along the road and brought me safely to the end of this expedition. Even when I felt the weakest after my mom’s death, He embraced me and brought me through the storm. This research endeavor is my tribute to Him, for His unconditional grace and Eternal love.

I would never have achieved this accomplishment without the support of my family and friends. I thank my dad, Larry, who inspired me to achieve my goals and encouraged to never give up…thank you! To my sister and brother-in-law, Kara and Marc, and my brother and sister-in-law, Kevin and Jill, and their families…thanks for sending prayers of encouragement my way and lending an ear when I needed one…I am so blessed! To my extended family that always encouraged me…I thank you! And finally, to Teri, for reading my papers, listening, and offering words of wisdom as only a good friend can…God gave me a blessing when He provided you as a friend!

I appreciate the professional assistance my statistician, Dr. Phil Linerode, provided…I would never have made it through Chapters III and IV without your guidance. Finally, thank you to my committee members who encouraged me throughout the dissertation process…Dr. Michelle Goodwin, my dissertation chair, Dr. Frank Alfaro and Dr. Mark Angle…your insights, expertise, and leadership was a God-send!

Thank you all!
DEDICATION

To my mom, who literally started this journey with me by driving from Texas to Virginia for my first intensive in the summer of 2006. While the trip turned out to be our first (and only) mother-daughter trip, it was filled with special memories I will always cherish.

Though God took her home earlier than I would have liked, I am thankful for the unconditional love and endless support she gave me during her short time on earth.

In loving memory of

Patricia Louise Karrh

April 1, 1948 – March 3, 2007
CHAPTER I: INTRODUCTION

Today, American public school systems are being held accountable for student achievement and academic growth based mainly on results obtained from norm-referenced or criterion-referenced exams. Norm-referenced tests (Young and Zucker, 2004) are designed to measure individual student achievement against other students that have been subject to the same test and testing standards. While norm-referenced tests compare each student’s performance to an established norm-group, criterion-referenced exams assess the actual academic and skill level of the student. Thus, criterion-referenced tests (Young and Zucker, 2004) assess what students are expected to know based on set subject area standards, as compared to norm-referenced tests which judge a student’s performance to a set standard.

Districts are often handcuffed by state assessment results and depend on disaggregation tools and research procedures to compile comprehensive statistics to inform and drive educational and curricular decisions. With these statistical and research tools, school systems have the ability to obtain a more detailed spectrum of each student’s educational abilities and future academic promise and abilities (Young and Zucker, 2004). The norm-referenced results of Stanford 10 or OLSAT 8, coupled with performance on criterion-referenced state standardized tests, such as the Texas Assessment of Knowledge and Skills (TAKS), could generate considerable educational and instructional implications if predictive relationships were statistically determined.
Statement of the Problem

The past 30 years has brought significant changes to testing in public schools. Peter Behuniak (2004) believes two specific trends have emerged during this time period, including the expansion of achievement testing and the increased emphasis and development of educational accountability systems (p. 335). Achievement testing and accountability are pieces of the standards-based school reform movement that has gained momentum and publicity in America, especially since the passing of the 2001 No Child Left Behind legislation. Schools are giving standardized assessments to meet state and federal testing and accountability requirements and may or may not be taking the opportunity to disaggregate the results. If a norm-referenced assessment could be linked to a state criterion-referenced test, the results could impact instructional decisions, potentially predicting future academic performance on state criterion-referenced exams.

The Stanford 10 Achievement Test and the Otis-Lennon School Ability Test are two prominent norm-referenced, standardized assessment tools used to determine academic proficiency, cognitive aptitude, and general school ability level. The research study school district currently uses these two tests, in addition to other local criteria measures, to determine assignment in classes deemed as “upper placement” courses. Upper placement or upper level classes are taught at standards above general state expectations and are designed for committed, highly motivated, independent, and academically mature students who are capable of working above grade level.

In a declining economy, school systems are looking for more ways to stretch the dollar and using existing data to evaluate student progress. The research problem
was to determine whether a correlational relationship existed between Stanford 10 Math and Reading or OLSAT 8 Quantitative and Verbal norm-referenced assessments and grade 7 TAKS Reading and Math scale scores. Also, the problem was designed to investigate whether the results from the fall administration of Stanford and OLSAT could be applied to a mathematical formula capable of predicting spring performance on the grade 7 state-mandated TAKS Reading and Math exams. The null hypotheses assumed no predictive relationships between scores of the Stanford 10 Achievement Tests and/or Otis-Lennon School Ability Test (OLSAT) when compared with scale scores from the TAKS Math and Reading tests. These research questions were postulated to determine whether one or more norm-referenced tests were capable of predicting a student’s anticipated performance level on a standardized, criterion-referenced test, such as the Texas Assessment of Knowledge and Skills (TAKS).

**Purpose Statement**

The purpose of this quantitative correlational research study was to determine the relationship between the Stanford 10, OLSAT 8, and TAKS Reading and Math and to examine whether future student performance on TAKS Reading and Math could be determined when fall 2007 Stanford and OLSAT results and spring 2008 reading and math TAKS scale scores were applied to a mathematical formula. The presence of a positive relationship would impact instructional planning and have a direct correlation on future assessment procedures. Teachers and administrators would be able to confirm existing information, while examining future predictive scores in relation to a student’s current strengths and weaknesses in order to plan instruction on a more individual basis.
Research Questions

The following questions guided this study:

1. Does student performance on the norm-referenced Stanford Achievement Test Reading portion predict performance on the grade 7 Texas Assessment of Knowledge and Skills Reading test?
2. Does student performance on the norm-referenced Stanford Achievement Test Math portion predict performance on the grade 7 Texas Assessment of Knowledge and Skills Math test?
3. Does student performance on the norm-referenced Otis-Lennon School Ability Test Verbal portion predict performance on the grade 7 Texas Assessment of Knowledge and Skills Reading test?
4. Does student performance on the norm-referenced Otis-Lennon School Ability Test Quantitative portion predict performance on the grade 7 Texas Assessment of Knowledge and Skills Math test?

Hypotheses

1. There is no significant correlation between the Reading Stanford Achievement Test and student performance on the grade 7 Reading TAKS tests.
2. There is no significant correlation between the Math Stanford Achievement Test and student performance on the grade 7 Math TAKS tests.
3. There is no significant correlation between the Verbal Otis-Lennon School Ability Test and student performance on the grade 7 Reading TAKS tests.
4. There is no significant correlation between the Quantitative Otis-Lennon School Ability Test and student performance on the grade 7 Math TAKS tests.

Significance of the Study

Tests have been a standard part of classroom instruction for many years. These assessment tasks have been largely designed, according to Zucker (2003), to “measure what students know and can do, improving instruction, and helping students achieve higher standards” (p. 2). While testing has been a part of instructional practice, mass standardized testing did not emerge until federal agencies and lawmakers decided it was time to hold schools accountable for each student’s academic growth and performance.

The term accountability often referred to individuals and their responsibilities, but now this term closely links education to the expectations the nation has for American public schools (Behuniak, 2004). Educational accountability was ushered to the forefront most recently by the passing of the No Child Left Behind Act of 2001 and its demands for high levels of student achievement based on specific performance standards. NCLB expects states to ensure that school systems are accountable for student academic growth regardless of economic, gender, or ethnic status. Measured largely in part by standardized tests, the accountability or standards movement is deeply entrenched in providing equitable educational access to all students and is imbedded in the educational framework of local and state school systems. According to Zucker (2003), the mandates attached to NCLB pose a substantial challenge to
states by requiring that “all students have the same chance to be successful at showing what they know and can do it in periodic, high-stakes assessments” (p.3).

Schools which accept federal funding have been mandated to “develop challenging academic standards, test students annually using assessments aligned with those standards, and measure whether schools, districts, and states are progressing towards those high standards” (Case, 2005, p. 3). Some programs funded by the federal education agency are held to even higher stakes, as indicated by Behuniak (2004), including “allocation of financial and other resources, the security of teachers’ jobs, and even the continued existence of specific schools rested on the results of standardized tests” (p. 337). Failure to meet the mandates established by the US Department of Education (USDE) could result in significant consequences and sanctions. Therefore, school systems must use all available tools to make appropriate data driven decisions.

Central to meeting this balanced and equitable system of education, is a school district’s ability to create, implement, and utilize aligned instructional practices. Case (2005) states that local education agencies are expected to ensure that “curricula and instruction are aligned with content standards, performance standards, and assessment” (p. 4). In order to prepare students for these yearly, high-stakes assessments, school systems must implement an aligned curriculum that utilizes recurrent benchmarking measures to monitor, assess, and improve instructional measures for students. Schools have autonomy when it comes to determining what additional resources or tools are implemented to drive and improve instruction. Therefore, it is critical that school systems use the available results from all
assessment measures, both informal and formal, to convey meaningful information to
the teachers and instructional specialists in the classroom.

According to Whittle (2006), the NCLB act instituted some of the strictest
accountability measures in American history; therefore, applications of this study have
meaningful implications for longitudinal planning within educational school system.
Given the continued pressure facing educational institutions, educators, and students,
the outcome of this study could assist educators as they strive to bridge the academic
gap, while creating an equitable, progressive, and academically challenging
educational setting for each and every student.

Assumptions

1. Student performance data was representative of all students in the school
system.

2. Students performed with similar effort on all assessments and in the
classroom.

3. All students who took the TAKS tests had the same equal opportunities to
pass the assessment in any given school year.

4. The sample was taken from students who took the Stanford Achievement
Test and the Otis-Lennon School Ability Tests; as well as, the reading and
math TAKS. In addition, all subjects used in the sample were enrolled in
school system for the entire 2007-08 school year.

Operational Definitions

The following operational definitions are used throughout this study:
*Ability Test* is an assessment designed to investigate one’s logical reasoning and thinking performance capability.

*Accountability* (in education) is the idea that schools are responsible for student academic growth.

*Achievement Tests* are assessments that measure what a student has learned by establishing a numerical academic skill level, score, or percentile.

*Adequate Yearly Progress (AYP)* is the standard established by the federal government to monitor student progress and the educational growth of school systems.

*Assessments* are practices utilized by schools to measure specific objectives.

*Confidential Student Report* is an individual score report detailing a student’s performance on the Texas Assessment of Knowledge and Skills (TAKS) tests.

*Correlational Research* is a type of research study in which the purpose is to measure relationships between two or more variables.

*Criterion* is a reference point by which student performance is assessed.

*Criterion-Referenced Tests (CRT)* are tests that measure specific skills directly related set criteria or standards and are not designed to measure how a student compare to others.

*ESEA* refers to the Elementary and Secondary Education Act of 1965.

*High-stakes Assessments* are tests with results that have significant impact on the test-taker, teacher, and educational system.

*NCLB* is an act initiated by President Bush in 2001 and signed into law on January 8, 2002. Known as *No Child Left Behind*, NCLB became the revised version of the Elementary and Secondary Education Act of 1965 and 1994.
Norm-Referenced Tests (NRT) are tests where scores are compared based on a norm and defined group performance. Generally, results from norm-referenced test are expressed in some type of group percentile form.

Otis-Lennon School Ability Test (OLSAT) is a norm-referenced assessment designed to assess the abstract thinking and reasoning skills of children and measure their cognitive ability to learn in the school setting.

Predictive relationship is a correlational relationship capable of forecasting future performance.

Quantitative Research is a research design which emphasizes numerical and measurement outcomes.

Reliability refers to the accuracy and precision of the test scores.

School Ability Index (SAI) is a normalized standard score with a mean of 100 and a standard deviation of 16 obtained in the OLSAT 8.

Stanford Achievement Test (Stanford), Tenth Edition is a norm-referenced assessment designed to assess student academic abilities.

Student Success Initiative (SSI) is accountability legislation enacted by the Texas Legislature in 1999.

Texas Assessment of Knowledge and Skills (TAKS) is a criterion-referenced state high-stakes assessment.

Texas Essential Knowledge and Skills (TEKS) is the term used for the Texas state academic curriculum standards and objectives.

Upper Placement or Upper Level are local classes, within the research study school system, taught at standards above general state expectations and designed for
highly motivated, independent, and academically mature students who are capable of working above grade level.

USDE refers to the United States Department of Education.

Validity refers to measuring what is intended to be assessed based on research results and their interpretation.

Summary

Standardized testing is one diagnostic tool that enables educators to assess academic skills based on set standards, whether norm-referenced or criterion-referenced. School systems strive to align curriculum material to cover information assessed on these tests to ensure that students and parents are aware of expectations prior to state testing (Behuniak, 2004). Given the present state of the American education system, high-stakes, standards-based testing appears to be a present and future cornerstone that will continue to be used to determine school accountability.

The intent of this quantitative research study was to determine if Stanford 10 and OLSAT 8 results could be another tool, coupled with the TAKS Reading and Math scale scores that could help school systems improve instruction and potentially gauge future state testing performance. The following chapter will examine literature surrounding standards-based testing and accountability. It will specifically address the following topics: 1) the theoretical perspective, 2) the historical perspective of testing, 3) the standards-based reform theory and movement, and 4) the role of norm-referenced and criterion-referenced tests.
CHAPTER II: LITERATURE REVIEW

This comprehensive literature review was conducted using online documents and printed journals and books spanning numerous years. While the literature review is not exhaustive, it is representative of the research related to testing and accountability most relevant to the research study.

Introduction

Standardized testing has played a crucial role in American school systems and the educational process for many years. After World War I standardized testing became the premiere tool used to assess intellectual strengths and weaknesses. According to Jost (2001), the use of aggregated student and instructional data during the 1960s also spurred the standardized testing movement. In an education society driven by accountability, districts are striving to develop academic procedures that accurately depict student skills and performance.

State governments and local education systems utilize these standardized test results to validate student achievement and the teaching abilities of their faculty (Cearfoss, 2007). Federal and state initiatives have also dictated that districts be accountable for student performance, creating what is known as high-stakes testing. The No Child Left Behind (NCLB) Act of 2001 is the current policy driving standards-based educational reform and the accountability movement. Though high-stakes testing has been gaining momentum for many decades, it was thrust to the forefront following the passage of the NCLB legislation.
State-mandated testing programs are more prevalent and have greater implications since the passing of NCLB, a revision of the 1965 *Elementary and Secondary Education Act*, and the subsequent implementation of the *Adequate Yearly Progress* (AYP) indicator by the *U.S. Department of Education*. Passage of NCLB has increased local and state educational agencies focus on state-mandated tests and has placed a larger emphasis on the federal government’s role in public education. Bowman (2004) suggests that high-stakes, standards-based assessments give “parents information about the quality of their children’s schools, the qualification of their teachers, and their children’s progress in key subjects” (p. 14). Regardless of how the high stakes tests are used, compliance with NCLB is mandatory. Thus, *Education Week* reported in 2002 that every U.S. state education agency had some form of standards-based testing program designed to measure math and reading achievement, as well as other subject areas.

*Testing and Accountability: A Theoretical Perspective*

Testing, according to Fremer and Wall (2004), refers to “a set of questions that has been compiled to measure a specific concept such as achievement or aptitude” (p. 3). Tests are viewed as one of the top contributions generated from the field of psychology and are used widely in education. Fremer and Wall (2004) go on to say that the intent of testing is to enable others to make decisions regarding “individuals, programs, or institutions” (p.4). Theoretically, all standardized tests, whether norm-referenced or criterion-referenced, should be designed to assess the intended curriculum or performance standards.
This study uses an adaption of Vernon’s (1961) Hierarchical Structure of Human Abilities framework for standardized testing as one of the theoretical premises for assessment in schools. Harcourt, in their 2003 OLSAT Technical Manual, give a diagram illustrating the Hierarchical Structure of Human Abilities (p.6). This figure depicts a typical “hierarchical” or pyramid visual. According to Harcourt (2003), the top of the hierarchy represents Charles Spearman’s “general factor,” divided into two major ability groups: verbal-educational (v:ed) and practical-mechanical (k:m).

Each of these ability groupings of verbal-educational and practical-mechanical can be broken down further into smaller subsets (Harcourt, 2003). The verbal-educational strand of the framework is supported by verbal and numerical subtests. These subtests evaluate the heart of what individuals might perceive as the core school curriculum of reading and math. While standardized tests seem very comprehensive, norm-referenced and criterion-referenced tests are not intended to assess all of the abilities considered “verbal-educational” (Harcourt, 2003, p. 6). Therefore, educators and parents should not depend on standardized tests to solely measure all academic and ability levels (Harcourt, 2003).

Norm-referenced and criterion-referenced tests are part of an assessment structure that has contributed to the development of the standards-referenced interpretive framework (Young and Zucker, 2004). Norm-referenced framework is created when test developers design assessments to measure student achievement in relation to peers who took the same assessment instrument. According to Young and Zucker (2004), the test publisher “administers the assessment under standardized conditions to a nationally representative sample of students and analyzes the results
using rigorous statistical methods” (p. 1). Norm-referenced test results are intended to enable educators and parents the opportunity to track student achievement and pinpoint academic strengths and weaknesses. Stanford 10 and OLSAT are examples of norm-referenced tests designed around national curriculum expectations and are two of the assessments driving this research study.

Unlike the theoretical norm-referenced assessment framework, criterion-referenced tests are developed based on specific academic criteria and results generally are generally noted by scale scores or percentages. According to Young and Zucker (2004), “publishers of criterion-referenced assessments set performance levels using rigorous, widely accepted methods, such as the modified Angoff procedure, to arrive at one or more thresholds or cut scores” (p. 2). By employing multiple set scores, assessment professionals can generate a more advanced evaluation (Young and Zucker, 2004). Texas uses three cut scores for the TAKS tests, including: did not meet standard, met standard and commended performance.

The standards-referenced interpretive framework is a modern twist on the assessment process and includes elements of standards-based education reform and norm- and criterion-referenced theoretical premises (Young and Zucker, 2004). The standards-based school reform theory combines both process reform and content reform, which contribute to the standards-referenced interpretive framework. Process reform refers to the method in which educators teach and students learn, while content reform addresses what educators actually teach (Kean, 2004). The combination of these two reform theories has contributed to the current standards-based school reform framework. The general idea behind standard-based testing is that previously
published content standards, agreed upon by state education agencies, are combined with established performance expectations to ensure that students and educators are held accountable for previously agreed upon curricular standards (Koretz, 2008).

Theoretically, standards-based school reform, as indicated by Kean (2004), focuses on “improving our schools, increasing student achievement, and building accountability for results through a system with three primary components” (p. 326). The components Kean (2004) addressed include new curriculum standards, assessments modeled after new curriculum, and consequences for failure to meet the standards. Accountability, as part of standards-based school reform, is currently driven by the 2001 No Child Left Behind legislation that holds firmly to the theoretical framework that holding schools accountable for standardized testing results ultimately produces equitable educational practices for all students regardless of ethnicity, gender, or socio-economic status.

Some might consider the assessment loop theory an extension or arm of the standards-referenced interpretive framework. Assessment looping blends program evaluations and test results into a cyclic theoretical model. Erford and Moore-Thomas (2004) feel “undoubtedly” that the “assessment process is systematic, ongoing, and cyclical” in nature (p. 518). Program evaluations and the outcomes of assessment practices work hand-in-hand to refine the comprehensive educational program (Erford and Moore-Thomas, 2004). Standardized tests, whether norm-referenced or criterion-referenced, are part of this assessment loop.

The assessment loop theory is a systematic way to conceptualize program evaluation and outcome results in order to improve the overall educational program
within a school system or campus. Erford and Moor-Thomas (2004) cite five parts of the evaluation cycle:

- Defining the campus mission in the context of the program.
- Developing educational questions about program efficacy.
- Using these results for program improvement.
- Gathering evidence to answer the educational questions.
- Interpreting the evidence to determine the value and worth of educational interventions. (p. 519)

Norm-referenced and criterion-referenced test results fit into the “gathering evidence” stage of the theoretical cycle. Within this step, standards-based research data is interpreted, conclusions are drawn regarding strengths and weaknesses, and decisions are made to improve one or more sections of district or campus educational programs (Erford and Moore-Thomas, 2004). According to Young and Zucker (2004), “a standards-referenced assessment can potentially provide a more complete, efficient understanding of a student’s educational achievement and progress” for educators when assessing individual student needs (p. 4). Since standards-referenced assessments can be used for multiple purposes, they are considered more versatile than assessment counterparts designed from only one traditional framework (Young and Zucker, 2004).

*Education and Testing: A Historical Perspective*

Mehrens and Cizek (2001) described one of the earliest-accounted testing situations from the Bible:
“Are you a member of the tribe of Ephraim?” they asked. If the man replied that he was not, then they demanded, “Say Shibboleth.” But if he couldn’t pronounce the H and said Sibboleth instead of Shibboleth he was dragged away and killed. So forty-two thousand people of Ephraim died there. (Judges 12:5-6, The Living Bible)

Mehrens and Cizek’s example is an early reference of what 21st century educators’ would call high-stakes testing and demonstrating that testing has a long history, evident throughout all historical time periods. Mental and academic measurements have been part of history for many centuries, as indicated by Lee (1985). Early Chinese assessments used Confucian philosophical practices to assess the moral capacity of Chinese men. These civil exams began in 2357 BC and continued until the 1900s according to H.D. Hoover in a university presentation (Lewis, 2006). Initially the Chinese system assessed six areas: music, horsemanship, writing, archery, arithmetic, and ceremonies and art, eventually adding civil law, agriculture, military affairs, geography, and revenue (Lewis, 2006).

According to Haladyna, Haas, and Allison (1998), two primary goals have driven the American educational system, including “providing equal access to public education and efficient delivery” (p. 263). Horace Mann proposed in 1845 that students in Boston Public Schools begin taking written exams in lieu of traditional oral exams (Gallagher, 2003). According to Gallagher (2003), Mann hoped to provide objective information about the quality of teaching and learning occurring in urban schools. It was Horace Mann’s desire that schools become places for kids to learn skills that helped them advance in society (Gallagher, 2003).
Fisher’s Scale Books for handwriting, spelling, math, scripture knowledge began being published in 1864, and the standardized testing initiative was born (Lewis, 2006). Not long after these books were published, one of the first high school admissions tests was created and utilized. This test, called the New York Regents’ Exam, officially was put into practice in 1865 and led to what is now known as college admission tests (Lewis, 2006).

By the end of the nineteenth century many scientists, psychologists, and professors were experimenting with the use of objective tests, scientific scales, and surveys including Charles Darwin and E.L. Thorndike (Gallagher, 2003). Thorndike is often referred to as the father of “test development” as several of his students went on to create and generate prominent educational assessments (Lewis, 2006). Thorndike’s research on assessments and the establishment of standard scales has led to what many refer to as standards-based, criterion-referenced tests (Janesick, 2001).

In addition to objective tests, schools began utilizing intelligence tests in the early part of the twentieth century thanks to the research efforts of Frenchman Alfred Binet and Theodore Simon. Their research teamwork in 1905 led to the development of the first intelligence test designed to measure whether or not a student was “mentally defective” (Janesick, 2001; Jost, 2001). The contributions of Binet and Simon, to the field of intelligence testing, provided schools with another source of information when engaging in the decision-making process and introduced the process of assessing the mind’s ability (Gallagher, 2003). A little over a decade later, Stanford University psychologist Lewis Terman adjusted and revised the work of Binet and Simon to create an IQ assessment for the US Army (Jost, 2001). Terman believed that
the assessment tool was beneficial for determining educational placement and tracking students for career purposes, in addition to evaluating students for mental retardation (Terman, 1916).

During World War I the U.S. Army administered tests to narrow their pool of recruits and created what is commonly referred to as the Army Alpha and Beta tests (Gallagher, 2003; Kennedy, 2003). These large-scale standardized tests began surfacing when excessive numbers of people volunteered (Kennedy, 2003). To minimize these large numbers, the military began investigating scientific methods to maximize efficiency and assign recruits to specific military positions (Janesick, 2001). The outcomes of their investigation resulted in the development of the Army Alpha and Army Beta tests. The Army Alpha assessment was written for those citizens who could read and the Army Beta was for illiterate individuals (Kennedy, 2003). Feedback obtained from the Army Alpha and Beta tests helped define the age of assessment.

In 1926 the Scholastic Aptitude Test (SAT) was developed and implemented to assist with the college admissions process (Gallagher, 2003; Jost, 2001). This assessment is recognized as one of the first commercially published tests according to Haladyna, Haas, and Allison (1998). Numerous achievement tests can trace their origin back to this time period, including the California Achievement Test and in 1933, the Iowa Test of Basic Skills (Haladyna et al., 1998; Jost, 2001). The development of achievement and college admission assessments helped usher in the standards-based educational testing movement.
Haladyna, Haas, and Allison (1998) cite specific causes impacting the progression of assessments, including “changing demographics caused by immigration, technological challenges introduced during the Cold War and exacerbated by the computer age and racial inequality” (p. 264). Kennedy (2003) noted, “Driven by legislation which mandated schooling of immigrants, standardized tests played an increasing role in the educational process” (p. 3). Additionally, the 1950’s were fueled by the Cold War and civil rights issues laying the framework for improving and restructuring the American public education school system (Jost, 2001).

In 1965 the United States Congress passed the *Elementary and Secondary Education Act* (ESEA) to help increase the Title I funds available to schools with disadvantaged students (Jost, 2001). In addition, the National Assessment of Educational Progress (NAEP) was formed to monitor and measure the national progress of American students (Jost, 2001). By the mid-to-late 1970’s, minimum competency tests began emerging to establish basic curriculum standards for high school students (Jost, 2001).

During the 1970’s, Howard Gardner started his brain research and the theory of multiple intelligences was born (Janesick, 2001). These investigations led to his theory on multiple intelligences and revolutionized the way schools looked at educating students. Through his research, educational institutions and state education agencies began to see the need for educating the whole child. According to Janesick (2001), Gardner’s research almost “single-handedly started the assessment movement” (p. 92).
During this same period, the National Institute of Education formally became the U.S. Department of Education (Janesick, 2001). Additionally, schools saw the rate of standardized tests grow substantially with the invention of personal computers for educational use (Kennedy, 2003). Cizek and Burg (2006) state that in the “not-too-distant past, testing in schools was largely classroom-focused, comparatively informal, and had only mild consequences associated” (p. 2). Standardized testing was initially intended to give local school systems the ability to gather district-wide educational information on students. States were not originally concerned with how individual districts were performing, rather that they were meeting the educational needs of the community.

The Reagan administration, in the 1980’s, published a report titled “A Nation at Risk” and publicly criticized schools for their lagging scores and demanded educational reform (Janesick, 2001; Jost, 2001). This report called for more student assessments but provided limited funds. During this time period, standardized test makers heard Washington’s cry for more tests resulting in a deluge of new achievement and ability assessments (Janesick, 2001). By 1989, President George Bush convened the first academic summit, involving state governors to examine the educational crisis. During this summit, six national goals for education were defined (Jost, 2001).

The 1990’s ushered in a full standards-based educational movement with a goal to improve academic curriculum and performance. The assessment movement was geared towards state-based testing and was given a financial boost with the reauthorization of the Elementary and Secondary Education Act in 1994 (Jost, 2001).
Within the next decade, *No Child Left Behind Act* of 2001 was signed into law requiring the assessment of students in grades 3-8 in the subject areas of math and reading. While the federal government required the specific testing, states were still given the autonomy to develop assessment instruments.

Haladyna, Haas, and Allison (1998) cite two conclusions that can be drawn from the past century:

- Testing has been and will always remain a basis for knowing about how schools affect students, despite the potential for misinterpretation and misuse of test scores;
- Two governing principles continue to influence school testing: 1) all students must be given equal opportunities regarding their education and 2) schooling must be offered in an efficient manner. (p. 264)

American education is rooted in objective and subjective standards-based testing and school systems must find a way to use common testing practices to improve student learning.

*The Standards-Based Movement: High-Stakes Testing and Accountability*

Wiggins (1991) believes that the term standards can be interpreted as a “passion for excellence and habitual attention to quality. A school has standards when it has high and consistent expectations of all learners in all courses” (p. 18). These standards establish a goal for all students, though there are times the goal may seem out of reach for some (Wiggins, 1991). Therefore, these curriculum and performance standards are specific measures that hold campuses, school systems, and state education agencies accountable. According to Janesick (2001), high-stakes testing is
the term “used to label tests for which the consequences of a student’s score are extremely serious” (p. 112). High-stakes testing and the practice of setting curriculum standards may have a new educational term, but in retrospect, standards and tests have dictated educational outcomes for almost 50 years.

In 1959, author S. Sarason wrote that “we live in a test-conscious, test-giving culture in which the lives of people are in part determined by their test performance” (p.26). Many individuals question the necessity of exposing students to such high-stakes environments and wonder whether the test results obtained outweigh the pressures induced by the situation as a whole. As educational standards become more rigorous and as high stakes testing situations become more frequent, researchers must address the potential implications for students in order to provide the most appropriate educational curriculum.

Since the launching of Sputnik, Americans have sensed an urgency to catch up with other leading scientific and academia nations. During the 1960’s the public began to demand a higher quality education for their children (Kennedy, 2003). The Elementary and Secondary Education Act of 1965 was passed and designed to address the discrepancies in the American education system (Kennedy, 2003). This act provided designated funds for schools serving low-income students and campuses that were struggling academically.

Few Americans would argue that educational reform was necessary, but shared differing opinions on how to improve public education. During the 1960’s significant data emerged showing “test score differences between different racial and ethnic groups” (Airasian, 1988, p. 301). These achievement gaps gave way to the need for
compensatory academic and educational initiatives. Similar gaps also surfaced in the 1970’s with special need students, launching the request for new special education reforms (Airasian, 1988). By the 1980’s it was apparent that schools were lacking in quality and students were falling behind their international peers. The National Commission on Educational Excellence published a report in 1983 that documented these specific educational gaps (Airasian, 1988). The highlights of this federal report indicated students were steadily declining in the classroom, as well as on college admission exams like the SAT.

School reforms generally rise and fall like that of a tide depending on educational need and political influence. Janesick (2001) states that there are eight characteristics generally associated with school reform, specifically:

1. The reforms are politically inspired and coerced by state governments.
2. The stress on higher student achievement is based on standards-based reports that were prepared by professional associations, not by local school boards.
3. Content standards tend to be collections of outcomes or student behaviors, assembled in a nonsystematic manner and without content hierarchies clearly shown.
4. Cost benefit analyses are lacking from the reports on state school reforms.
5. Control on education has shifted to the national and state levels and away from localities.
6. The reform agendas, though fragmentary, are broad in scale and encompass most of the fifty states.
7. Politically inspired as the education reform movement has been, it must still be classified as being theoretical, that is, its basic premises are grounded not in empirically sound studies but rather in political enthusiasms and intuitions.

8. Implied within these reforms is the conclusion that, as a consequence of standards and high-stakes state testing and assessment programs, there should be dramatic increase in student achievement. (p. 96-97)

Airasian (1988) believes the current standards-based, state-mandated tests are very different than previous assessments for three specific reasons:

• The new state tests are mandated for all schools and virtually all pupils within a state.

• The mandate eliminates most of the local district discretion in test, selection, administration, content coverage, scoring, and interpretation. A single, state-approved tests that is administered, scored, and interpreted according to state guidelines is used across school districts.

• The tests have built-in sanctions or a so-called high stake associated with specific levels of performance. (p. 305)

With in-state test standardization, school systems are now able to compare apples to apples and oranges to oranges when it comes to evaluating each student’s academic growth and progress within the state in which they reside. The purpose behind standards-based, high stake assessments, according to Airasian (1988), is to “force an
instructional response to the test so that the test content will, in current parlance, drive instruction” (p. 305).

Accountability as a whole often carries a negative connotation, though holding someone accountable for their actions is generally considered important for one’s character. Educational accountability, according to Friedman (2001), generally holds to punishing school systems who fail, instead of rewarding institutions that succeed. According to Herman (2007), “ pressured by fear of sanctions-and less often by rewards-teachers and students are motivated to teach/learn the expected standards and to use the information from the assessment to improve their efforts” (p. 3). These high-stakes tests and standards-based performance measures often cause significant stress on students and teachers, directly impacting the educational environment.

The icing on the accountability and standards-based reform movement came with the passing of No Child Left Behind Act in 2001 ushering in the 21st century educational accountability era. The No Child Left Behind Act was a revision of the Elementary and Secondary Education Act, focusing largely on the equitable access for all children to a quality education (Mohr, 2006). The NCLB legislation focused on four basic principles, according to Hamilton, Stecher, and Klein (2002), including: “stronger accountability for results, increased flexibility and local control, expanded options for parents, and an emphasis on teaching methods that have been proven to work” (p.6).

Spurred by the NCLB legislation, states have designed and implemented standards-based assessments that correspond to performance standards by which each grade level and curriculum area are measured. The federal NCLB legislation is a
structured accountability system that includes consequences for those school systems or campuses that fail to meet minimum performance standards on required state assessments. NCLB expects school systems to provide quality instruction for all students regardless of their background and educational obstacles. The federal government demands that all students be able to read by grade 3 and perform at or above grade level each year on state standard-based assessments. In addition to the high academic standards, NCLB insists that yearly assessments align to the approved curriculum standards. By demanding high academic standards, the federal government hopes achievement gaps will decrease in at-risk and minority subgroups.

To adhere to federal requirements, all states must implement an approved assessment program that measures adequate yearly progress or AYP. It is crucial that the testing system enacted by the different states evaluate the progress of every student regardless of ethnicity, program, SES, or gender. Specifically, the job of the USDE is to ensure that disadvantaged students in each state are granted an equal education as obtained by their peers. AYP is designed to hold schools accountable for successes or failures and make certain restructuring occurs when low-performing campuses are identified. By closing the achievement gap, the federal government hopes more minorities and impoverished students will attend institutions of higher learning, strengthening the nation as a whole.

President George W. Bush left the frameworks of his educational legislation firmly intact when he transitioned from his role as Texas Governor to president of the United States of America. Prior to leaving for the White House, President Bush proposed an educational reform bill known as the Student Success Initiative. Passed by
the Texas Legislature in 1999, the initiative required additional promotional requirements for students in grades 3, 5, and 8 (TEA, 2004). According to the legislation, students can only be promoted provided they meet the minimum proficiency standards on specific state assessments as outlined in the educational accountability law. The goal of SSI is to ensure that all students receive the instructional assistance necessary to perform on grade level, specifically in the subject areas of reading and math.

While curriculum and performance standards are at the forefront of educational politics, it is important to note that there is a place for standards-based assessment if it meets the needs of the students and is presented in an unbiased manner. Despite the fact that state assessment and accountability systems differ slightly from state to state, research does indicate that standards-based performance testing does increase educator and student motivation, positively impacting curriculum and instructional practices (Herman, 2007). High-stakes testing, performance-based accountability and school reform is a fixture in the American education system. Communities, schools, and lawmakers must advocate for testing practices designed to assess the curricular content without creating unnecessary stress on students and educators.

The Role of Norm-Referenced and Criterion-Referenced Tests

Norm-referenced tests have long been used for comparing one student’s score with other individuals who took the same assessment known as a “norm group” (Taylor and Walton, 2001). These norm groups are created when publishing companies develop the assessment and administer the test to large samples of students (Taylor and Walton, 2001; Young and Zucker, 2004). Janesick (2001) states that
norm-referenced tests are “designed to rank order the test takers to accommodate a bell curve” where a few will score high, some will score low, and the remaining scores will be concentrated in the middle (p. 102). These particular norm-referenced assessments are intentionally designed to ensure that half of the testing population fall in the mid-range or below (Janesick, 2001).

The use of a set score distribution is often difficult for individuals to comprehend. The basic premise is similar to a normal bell curve where it is impossible for all students to score above average. The score distribution will always have scores ranging from low to high, which can sometimes be alarming to parents when they are reviewing their child’s test results. According to Taylor and Walton (2001), “test questions on a norm-referenced test are chosen for the express purpose of comparing children, not to determine how much of the school curriculum children have mastered” (p. 18).

Norm-referenced tests have given education a means to assess standards-based achievement and ability levels, but should not be the sole method for determining student potential. Since the test items are generally formatted as multiple choice questions, the results only reflect a snapshot of information and should only be viewed as a general estimate of a child’s promise. Janesick (2001) argues that one of the “major criticisms of standardized testing are that it is biased” (p. 103). These biases may produce significant discrepancies in scores depending on one’s gender, ethnicity, and SES status.

Unlike norm-referenced tests, criterion-referenced assessments are intended to measure whether a child meets a particular performance standard and is not designed
to be compared to his or her peers (Taylor and Walton, 2001). The majority of state 

exams, especially those for the purpose of state and federal accountability, are 
criterion-referenced tests intended to test curriculum and performance standards 
established by the state education agency. These standards-based, criterion-referenced 
tests are important for determining whether or not campuses and school systems are 
successfully teaching fundamental state curriculum expectations. It is also worth 
noting that students may perform differently on criterion-referenced tests as compared 
to norm-referenced exams due to varying test structures, expected results, and level of 
difficulty (Taylor and Walton, 2001).

*Stanford Achievement Test, Tenth Edition.* Since the first printing of the Stanford Achievement Test in 1923, nine additional updates have been published in the years 1929, 1940, 1953, 1964, 1973, 1982, 1989, 1996, with the most recent edition in 2003 (Harcourt, 2004). Each subsequent revision of the Stanford Achievement Test series was written and developed to:

- Update content to align the test with current educational and curriculum trends,
- update the normative information to make score interpretation more valid,
- increase and improve the kinds of information available for testing, and revise the look of the test to make it more relevant to students (Harcourt, 2004, p. 5).

The Stanford 10 is designed to estimate academic school achievement in reading, math, language arts, science, and social studies (Harcourt, 2004). The Stanford 10 is a multiple-choice examination developed around national and state curriculum standards as well as those trends promoted by national professional educational groups (Harcourt, 2004). Unlike other types of exams arranged in an easy
to hard format, the Stanford 10 varies the levels of difficulty to diminish student frustration during the assessment. The designers of the Stanford 10 test believe this random arrangement of difficulty enables students to focus longer and stay more motivated during the examination (Harcourt, 2004).

Stanford 10 was formatted and designed after comprehensive reviews by professionals, examination of current textbook series, analyses of state and district curriculum and performance standards, and national professional and educational initiatives (Harcourt, 2004). According to Harcourt (2004), after completion of the comprehensive review process, Stanford 10 test blueprints were developed that contained the following critical components:

- Number of test levels necessary for complete coverage across the elementary and secondary grades
- Test content areas
- Instructional standards coverage in each content area
- Articulation of the content and instructional standards across the grade levels
- Approximate number of items necessary for breadth of content and reliability of assessment
- Proportions of test items devoted to each of the two cognitive levels: Basic Understanding and Thinking Skills (p. 8)

Each test item was written to meet the achievement parameters specified within the content clusters, process clusters, cognitive levels, and instructional standards (Harcourt, 2004). Harcourt (2004) obtained content validity by asking content area
specialists, test construction professionals, and Harcourt editors nationwide to review and evaluate the test questions.

In a diverse society, Harcourt gave specific attention to procedures designed to eliminate bias and it is the belief of Harcourt (2004) that test content “should neither be offensive to members of a particular group nor unfairly disadvantage the performance of a particular group because of extraneous factors irrelevant to the constructs the test intends to measure” (p. 16). Throughout the test development process, measures were implemented to reduce the use of test items which “inadvertently reflected ethnic, gender disability socioeconomic status, cultural, or regional bias or stereotyping, or that included content that disadvantaged a group because of differences in culture and familiarity” (p. 16).

The Stanford 10 is available in the traditional format and an abbreviated version. The abbreviated test has roughly 20 to 30 test items, compared to the full version that has upwards of 50 or more items per subject test (Harcourt, 2004). The Stanford 10 reports an array of scores, including raw scores, scale scores, percentile ranks, and grade equivalents, just to name a few. The concepts assessed throughout the exam represent the general curriculum taught at schools across the nation.

Stanford 10 uses a vertical scaling system, as compared to an equal-interval scale, across all grade levels (Jorgensen, 2004). This developmental scale is helpful to school systems as they monitor student progress. While the Stanford 10 does not replace the required standards-based, criterion-referenced assessment in Texas required for state and federal accountability, it can be used to compare local students
to a national population of peers. The statistical information obtained may also serve as valuable feedback when evaluating general academic achievement abilities.

*Otis-Lennon School Ability Test (OLSAT), Eighth Edition.* The OLSAT 8 is the newest version of the Otis-Lennon ability test. OLSAT 8 is designed to evaluate verbal, quantitative, and figural reasoning skills. The OLSAT series originated from a simple intelligence scale, known as the *Otis Group Intelligence Scale*, to a series of mental ability assessments. Based on the 1918 graduate work of Dr. Arthur Otis at Stanford University, the test series was originally designed to examine students’ abilities to cope with academic learning tasks (Harcourt, 2003). Through the assessment of these learning tasks, educators are able to develop curriculum and instructional practices based on student needs and abilities.

The OLSAT series is “based on the idea that to learn new things, students must be able to perceive accurately, and to recognize and recall what has been perceived” (Harcourt, 2003, p.5). Additionally students must be able to “think logically, to understand relationships, to abstract from a set of particulars, and to apply generalizations to new and different contexts” (Harcourt, 2003, p.5). Thus, the OLSAT was written to evaluate specific nonverbal and verbal abilities that are linked to academic achievement and/or the ability to learn. The OLSAT 8 contains multiple-choice questions, such as analogies, sequencing, classifying, and recognizing differences.

Each portion of the test was written to encompass a wide range of item difficulty in order to effectively examine a student’s spectrum of abilities. Comprised of seven levels, OLSAT 8 is capable of assessing students from kindergarten through
grade 12. According to Harcourt (2003), each level of OLSAT 8 has been constructed to “be an accurate measurement instrument for most students in the grade(s) for which it is recommended” (p. 6). OLSAT results in two subset scores, Verbal and Nonverbal. These subset scores combine to form a Total score, often reflective of a student’s intellectual ability or general IQ score. The Total score is considered to be the best “overall indicator of school learning ability” (Harcourt, 2003, p. 6).

The test items and questions reflect Arthur Otis’s original intent that the most effective manner to evaluate cognitive ability is through a varied group of reasoning tasks (Harcourt, 2003). Harcourt (2003) publishing states that “data amassed over the last 70 years attest to the general usefulness of these reasoning tasks, as do the various analyses and classifications of the items that have occurred over the years” (p. 8). Items selected for OLSAT 8 were reviewed by staff members, measurement specialists, and psychologists to establish content validity and overall general quality.

Like the Stanford 10, also published by Harcourt, extensive measures were taken to eliminate bias during test development. The Harcourt (2003) Advisory Panel believe that test content “should neither be offensive to members of a particular group nor unfairly disadvantage the performance of a particular group because of extraneous factors irrelevant to the constructs the test intends to measure” (p. 13). Throughout the entire test development process, procedures were used to minimize the use of items “inadvertently reflected ethnic, gender disability socioeconomic status, cultural, or regional bias or stereotyping, or that included content that disadvantaged a group because of differences in culture and familiarity” (p. 13). Each level of the OLSAT 8
test was designed to distribute item difficulty across the test in order to reduce the frustration level of students.

Standardization is a critical part of the testing process. By testing thousands of students across the country, Harcourt (and other publishers) are able to establish scores that reflect all students regardless of demographic group or socioeconomic status. At Level F, the focus of this study, all of the items are arranged in a “spiral omnibus format” according to Harcourt (2003, p. 16). Again, this format arranges questions according to type and difficulty on a rotation basis with the intent of minimizing student frustration during the assessment (Harcourt, 2003). Regardless of test level, the word levels are written below recommended reading levels to ensure that the questions truly examine a child’s reasoning skills instead of their reading ability (Harcourt, 2003).

*The Texas Assessment of Knowledge and Skills (TAKS).* In 1999, the 76th Session of the Texas Legislature passed *Senate Bill 103* which required the implementation of a new comprehensive and statewide standards-based assessment program. The *Texas Assessment of Knowledge and Skills* (TEA, 2007) or TAKS was designed and implemented in the spring of 2003 in accordance with state law and legislation. According to the *Texas Education Agency Technical Digest* (2004), the TAKS test is designed to “measure the extent to which a student has learned and is able to apply the defined knowledge and skills at each tested grade level. In addition, every TAKS test is directly aligned to the respective *Texas Essential Knowledge and Skills* or TEKS content area” tested (p. 11). A standards-based, criterion-referenced
test, the TAKS test is intended to provide students, parents, and educators with direct information about what the student has learned over the course of a school year.

Over the years, Texas has developed and administered numerous standards-based assessments, from Texas Assessment of Basic Skills in the early 1980’s, to Texas Educational Assessment of Minimal Skills in the late 1980’s, to Texas Assessment of Academic Skills in the 1990’s, with the Texas Assessment of Knowledge and Skills being the most recent criterion-referenced test (TEA, 2002). The creation and design stage for the TAKS test was a three-year process dating back to 1999. In September 2001, the State Board of Education adopted the new rules for the TAKS assessment program (TEA, 2002). The state testing program requires assessment of students in grade 3 through grade 11 (Exit Level) in reading and math each year with additional tests in writing, science, and social studies designated in specific grade levels. These adopted assessment rules were signed into law and became effective on November 15, 2001.

Under the direction of the Texas Education Agency, thousands of highly qualified educators and TEA staff members collaborated in numerous committee sessions to draft objectives and standards to be tested on the new TAKS assessment (Parks, 2007). The committees were charged with creating an assessment that tested the state Texas Essential Knowledge and Skills (TEKS) in an unbiased fashion, while holding school systems accountable for yearly academic performance and growth of students. The draft version of the TAKS test was reviewed by numerous sub-committees and tested with students throughout the state in field-testing environments. Data was analyzed by the TEA Research Division and by Pearson Measurement
consultants. In addition, university professors reviewed test questions to establish subject-specific, content validity (Parks, 2007).

Summary

When it comes to education, academic testing is as important as any type of health-related test your physician might order. These assessments ensure that students are academically healthy and capable of actively contributing to society (Gandal and McGiffert, 2003). While the use of high-stakes testing is a valuable performance accountability tool, it is important to remember that multiple sets of data, from a variety of sources should be utilized to drive curriculum decisions. This may include everyday assignments, activities, benchmarking, and the use of other norm-referenced assessments, such as Stanford and OLSAT. Gandal and McGiffert (2003) do believe educators are “delinquent if we pass students through the grades and award them diplomas even if they are unprepared for the opportunities and challenges that await them” (p. 41).

According to Airasian (1988), standards-based tests “symbolize order and control, a focus on important educational outcomes, and basic, traditional moral values” (p. 311). State accountability measures are directly tied to performance-based test results; thus, high-stakes testing outcomes have significant implications for educators. Local, state, and federal accountability measures are not likely to disappear, so it is critical that school systems use all available assessment, benchmarking, and instructional measures to make curriculum decisions. These standards-based school reform measures are representative of public expectations and motivate school
systems to analyze educational practices to meet the changing needs of the school, students, and community. Thus, Airasian (1988) summarizes:

Regardless of the actual impact that state-mandated high-stakes tests have on pupils, teachers, and the curriculum, they are likely to have an important perceptual impact on the public at large. This perceptual impact underlies the social consensus or social validation that provides the legitimization of testing as a workable and desired reform strategy (p. 311-312).

Our students deserve to have the greatest instructional opportunities available to them and analyzing classroom practices and testing data are the first steps to ensuring students have access to the most comprehensive academic and assistive programs. Utilizing norm-referenced tests, such as Stanford and OLSAT, are just two of the available standardized assessment tools capable of potentially gauging future performance and should be considered when gathering data and information for curricular decisions.
CHAPTER III: METHODS

Given the economic crisis and continued emphasis on standards-based school reform and high-stake accountability measures, schools must continue to find creative ways to analyze academic student progress with assessment looping to avoid placing additional financial burdens on the school system and community. This study may uncover new correlations that can be used for instructional purposes by analyzing the potential predictive relationship between Stanford, OLSAT, and the TAKS reading and math assessments.

Approval for this research endeavor was received from the Liberty University Internal Review Board (IRB) prior to the collection of data as reflected in Appendix B. Chapter III will outline the procedures for examining the prospective correlational and predictive relationship between these norm- and criterion-referenced tests and state assessments comparable to the Texas Assessment of Knowledge and Skills in the subject areas of math and reading. This chapter highlights the setting, methodology, measurement instruments, and data analysis processes that were employed by this investigation.

Research Questions

As stated in Chapter 1, there were four research questions which drove this research endeavor.

1. Does student performance on the norm-referenced Stanford Achievement Test Reading portion predict performance on the grade 7 Texas Assessment of Knowledge and Skills Reading test?
2. Does student performance on the norm-referenced Stanford Achievement Test Math portion predict performance on the grade 7 Texas Assessment of Knowledge and Skills Math test?

3. Does student performance on the norm-referenced Otis-Lennon School Ability Test Verbal portion predict performance on the grade 7 Texas Assessment of Knowledge and Skills Reading test?

4. Does student performance on the norm-referenced Otis-Lennon School Ability Test Quantitative portion predict performance on the grade 7 Texas Assessment of Knowledge and Skills Math test?

Hypotheses

1. There is no significant correlation between the Reading Stanford Achievement Test and student performance on the grade 7 Reading TAKS tests.

2. There is no significant correlation between the Math Stanford Achievement Test and student performance on the grade 7 Math TAKS tests.

3. There is no significant correlation between the Verbal portion of the Otis-Lennon School Ability Test and student performance on the grade 7 Reading TAKS tests.

4. There is no significant correlation between the Quantitative portion of the Otis-Lennon School Ability Test and student performance on the grade 7 Math TAKS tests.
Research Design

A correlational research design and multiple regression formula were used to examine any potential correlational relationships between seventh grade Texas Assessment of Knowledge and Skills Reading and Math scale scores, Reading and Math Stanford Achievement Test percentiles, and Verbal and Quantitative Otis-Lennon School Ability Test SAI scores. The dependent variables were the Reading and Math TAKS scores, while the independent variables were the Stanford Reading and Math percentiles and the Verbal and Quantitative OLSAT SAI scores.

Thus, the correlational research \( (r) \) design tested the relationship between the measured variables. Using inferential statistics, this design provided judgment about a larger population and a means to infer correlational relationships. When assessing independent and dependent variables, it is common to plot the values in a dimensional manner through the use of a scatterplot. While determining whether or not a correlation is present does not imply a specific causal-type relationship, it does suggest there is some type of significant correlation between the variables. The scatterplot graphics provide a visual representation of the data and gave this researcher an opportunity to determine whether or not a relationship truly exists between the variables.

If there appeared to be some type of relationship or association between the variables, this researcher then applied a linear regression model to the data to determine predictive ability. Correlation and regression analysis are often used together since correlation analysis has the ability to measure the strength of the dependent and independent variables \((X \text{ and } Y)\) being assessed in the linear
regression. The initial step in the multiple linear regression analysis was to calculate the correlation coefficient (R) for each of the four research questions. A perfectly positive relationship between two variables would yield a 1.00, at the same time; a perfectly negative relationship would be represented by a coefficient of 0.00. If a value was determined that fell between 0.00 and 1.00, one could deem that the variables were positively related and a value between 0.00 and -1.00 would be indicative of a negative relationship.

After the correlation coefficient was established, R\(^2\) was calculated for each research question. The multiple correlation coefficient (R\(^2\)) was utilized to describe the percentage of TAKS Math or Reading variance attributed to an increase in a student’s Stanford Math, Stanford Reading, OLSAT Verbal, or OLSAT Quantitative percentage/score. The final step of the linear regression analysis was to establish a mathematical equation that could be used to predict the values of a specific dependent variable (TAKS Reading or Math score) when based on the values of a given independent variable (Stanford Reading percentile, Stanford Math percentile, OLSAT Verbal score, or OLSAT Quantitative score). The linear regression analysis highlighted the straight line that visually fit the observed data.

**Subjects**

The research was correlational because there was no manipulation or control of variables. The students involved in this correlational research endeavor all attended a Texas public middle school. Grade 7 students who took both assessments (OLSAT and Stanford) and were enrolled at the entire 2007-2008 school year were the subjects for this study (N = 331). Two-thirds of the subjects were Caucasian, with the
remaining third primarily of Hispanic decent. Of the 331 identified participants, almost 20% participate in the federal Free and Reduced Lunch Program.

Procedures

Permission to Collect Data. Once the research proposal was approved for study by the IRB (Appendix B), the following information was disseminated to central office and school administrators: a) the approved research proposal, b) copy of IRB approval letter, and c) letter of request (Appendix A) to Assistant Superintendent for Curriculum and Instruction for a blind-copy of 7th grade students scores (2007-2008 school year) on the following assessments: Stanford, OLSAT, and TAKS Math and Reading. The research study school district granted their approval (Appendix C) for blind data collection, using local identification numbers only to match data.

TAKS Administration. TAKS administrations and procedures were established by the Texas Education Agency and could not be altered or manipulated. This research study utilized the math and reading TAKS results obtained during the 2007-2008 standards-based, state-mandated testing season to determine potential correlations with Stanford 10 and OLSAT 8. Student information was matched utilizing local student identification numbers to protect the confidentiality of the students. The researcher did not physically handle any of these assessments and only used the data collected during the testing session for predictive comparison.

Stanford Achievement Test and the Otis-Lennon School Ability Test. The Stanford 10 and OLSAT 8 were given district-wide to grade 7 students in the fall of the 2007-2008 school year. The scores were used for upper placement decision making and impacted the particular level of class a student could choose. Taken in
classroom groups of about 25, the tests were administered over multiple days in November of 2007. Results were obtained in late December and utilized to determine academic placements for the 2008-2009 school year. The researcher gathered the blind/masked results and data obtained during the fall 2007 administration for each of these norm-referenced standardized tests as part of the research endeavor.

Instrumentation

*Texas Assessment of Knowledge and Skills.* In 1999, the 76th Session of the Texas Legislature passed Senate Bill 103 which required the implementation of a new comprehensive and statewide assessment program. Based on legal requirements and governmental expectations, the *Texas Assessment of Knowledge and Skills* or TAKS was designed and implemented in the spring of 2003. According to the *Texas Education Agency Technical Digest* (2004), the TAKS test is intended to “measure the extent to which a student has learned and is able to apply the defined knowledge and skills at each tested grade level. In addition, every TAKS test is directly aligned to the respective *Texas Essential Knowledge and Skills* or TEKS content area” (p. 11).

Under the direction and approval of the Texas Education Agency, Pearson Educational Measurement used “various statistical analyses, including classical measurement theory and item response theory” to confirm the reliability and validity of the TAKS assessment (TEA, 2004). Test items included in the TAKS assessment were calibrated and figured using the *Rasch Partial-Credit Model*, also known as RPCM. According to the *TEA Technical Digest* (2007), the RPCM was chosen because “of its flexibility in accommodating multiple-choice data as well as multiple-response category data, and for its ability to maintain a one-to-one relationship
between derived scores and the raw scores” (p. 138). According to the most recent online *Texas Education Agency TAKS Technical Digest* (2007), the TAKS reading assessment has a reliability coefficient of .90, while the TAKS math test has a reliability coefficient of .91. These statistical reliability results were determined based on the application of the Kuder Richardson Formula 20, or KR20, to the complete subtest for each subject area.

TAKS scores for the 2007-2008 school year were reported as scale scores and represent specific performance levels as shown on Table 1. These uniform scale scores were standardized across all subject areas in grades 3 through 11 who were administered the yearly spring TAKS tests. A guide was sent home with the results to assist in the understanding of the *Confidential Student Report*. Scores below 2100 “Do Not Meet Standard,” scores between 2100 and 2399 were considered to have “Met Standard,” and scores 2400 or above were noted as “Commended Performance.”

<table>
<thead>
<tr>
<th>TAKS Scale Score</th>
<th>Does Not Meet Standard</th>
<th>Met Standard</th>
<th>Commended Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Below 2100</td>
<td>2100-2399</td>
<td>2400 or above</td>
</tr>
</tbody>
</table>

*Stanford Achievement Test.* The Stanford 10, a multiple-choice test, was developed around the national and state instructional standards as well as those trends outlined by national professional educational groups (Harcourt, 2004). The test items used in the final forms were balanced based on basic understanding and thinking cognitive skills (Harcourt, 2004). Unlike other types of assessments arranged in an easy to hard format, the Stanford 10 mixed the levels of difficulty to avoid student
frustration during the test (Harcourt, 2004). For this particular research study, results from the abbreviated form of the Stanford 10 were utilized. The Form A abbreviated version consisted of questions taken from the full battery assessment (Harcourt, 2004).

The Stanford scale developed by the Pearson assessment division meets “rigorous industry standards, including suggestions in *Standards of Educational and Psychological Testing* that test publishers conduct periodic checks of the stability of the scale on which scores are reported” (Jorgensen, 2004, p. 4). According to Jorgensen (2004), the Stanford scale was designed based on “a common person equating design in which a large number of on-level and off-level test items are administered to several hundred thousand students during periodic national research programs” (p. 4). The results of this comprehensive testing process enabled Pearson to establish vertical linking of the test levels creating a common measured scale.

Although the abbreviated version of Stanford 10 was a shorter test, it still measured the core concepts and subjects found within the full-length version (Harcourt, 2004). Harcourt (2004) indicates that the KR20 was used to “overcome the possibility of non-equivalent portions” (p. 49). The KR20 generated high internal consistency scores with Stanford 10 reading showing a .92 reliability coefficient and Stanford 10 math indicating a .91 reliability coefficient (Harcourt, 2004). Additionally, Stanford 10 Abbreviated (Harcourt, 2004) used the same item response theory (Rasch model) used by the full-length version to establish and create the specific scales and norm data. Therefore, both the Stanford 10 full-length and the Stanford 10 Abbreviate reflect identical ability scales (Harcourt, 2004). According to Harcourt (2004), “while the relationship of the raw score to ability may differ from
one test to another, the relationship of ability (scaled score) to percentile rank is the
same” (p. 46). Therefore, information applying to the full-length version of the
Stanford 10 was also applicable to the Stanford 10 Abbreviated battery.

*Otis-Lennon School Ability Test.* This study focused on the grade 7 OLSAT,
Level F, assessment. This test used a “spiraling item structure design” to keep students
from becoming “discouraged by finding increasingly difficult items” (Harcourt, 2003,
p. 16). Additionally, the test word levels were written below the student’s current
grade level in order to measure the reasoning ability, rather than the student’s specific
reading ability (Harcourt, 2003).

Harcourt (2003) publishing stated that “data amassed over the last 70 years
attest to the general usefulness of these reasoning tasks, as do the various analyses and
classifications of the items that have occurred over the years” (p. 8). Items selected for
the OLSAT 8 were reviewed by staff members, measurement specialists, and
psychologists to establish content validity and test quality. Test reliability was
calculated using the Kuder-Richardson formulas to determine internal consistency.
OLSAT Verbal scores had a .86 reliability coefficient, while Nonverbal scores showed
a .88 reliability measure (Harcourt, 2003). Multiple measures were taken during the
test development process to reduce the use of items that “inadvertently reflected
ethnic, gender disability socioeconomic status, cultural, or regional bias or
stereotyping, or that included content that disadvantaged a group because of
differences in culture and familiarity” (p. 13). This study used OLSAT 8 (Harcourt,
2003) School Ability Index (SAI) scale scores for purposes of analysis, comparison,
and predictability which contained a mean of 100 and a standard deviation of 16.
**Data Analysis**

This research study produced four sets of data: two sets of math data including scores from the TAKS Math scale score and the Stanford Achievement Test or the OLSAT quantitative portion. Additionally, two sets of reading data were obtained consisting of the reading scale scores from TAKS Reading and from the Stanford Achievement Test or the OLSAT verbal portion. The Pearson product moment correlation coefficient, also commonly referred to as the Pearson $r$, provided a statistical means to determine the strength of the relationship between the Stanford, OLSAT, and performance on the math and reading TAKS assessments.

Results obtained, using the correlation research statistic, were analyzed using the *Statistical Package for the Social Sciences (SPSS)* to determine potential relationships. This researcher compared the level of significance generated by the inferential procedure against the critical level of significance (in this case 0.05); then based on the comparison, made a decision in regards to whether or not one failed to reject the null hypothesis or rejected the null hypothesis and what the decision inferred about the overall relationship.

The data was subsequently organized using a graphical procedure called a scatterplot. Ary et al. (2006) stated that in the research situation the scores on the horizontal axis will be those of the independent variable, with the lowest score on the left and the highest score on the right. The scores on the vertical axis will be those of the dependent variable ($y$) with the lowest score at the bottom and the highest score at the top. (p. 151)
The type of relationship was then determined based on the distribution of the data points on the scatterplot.

Once the data had been plotted, this researcher was able to examine the association between the variables by drawing a “best-fit” straight line through the series of data points. Mathematically, the linear regression line had an equation of \( Y = a + bX \), where \( X \) was considered the independent variable, \( Y \) was the dependent variable, the intercept was \( a \) and the slope of the line was \( b \). If it appeared that a linear relationship existed, the linear regression model would be applied to determine the relationship. Through the analysis of the linear regression results, important statistics were obtained known as correlation coefficients. Additionally, the predictive strength was examined. Valued between -1 and 1, each correlation coefficient was representative of the strength of the association between the variables. The intent of the linear regression analysis was to generate a mathematical equation that could be used to predict the values of a precise dependent variable when based on the values of a set independent variable.

**Summary**

Due to the accountability measures enacted with the passage of NCLB, districts are working diligently to maximize student performance and academic growth. Assessment loop theory stresses the importance of examining all available types of academic, ability, and achievement data to inform decisions. Standards-based, norm-referenced tests, like the Stanford 10 or OLSAT 8, could provide school systems with additional academic indicators of future student performance. The outcomes of this study could also provide school systems with an additional
benchmarking tool to determine future student progress on state criterion-referenced exams.
CHAPTER IV: FINDINGS

As stated in Chapter I, the study discussed in this chapter evaluated whether the norm-referenced Stanford 10 Achievement Test and/or Otis-Lennon School Ability Test were correlated and capable of predicting performance on the Texas Assessment of Knowledge and Skills (TAKS) Math and/or Reading criterion-referenced tests. The results of the 2007-2008 Stanford Math and Reading, OLSAT Verbal and Quantitative, and TAKS Reading and Math scores are reported. Organized in three sections, this chapter summarizes the research sample, descriptive findings, and statistical results obtained through research process.

Research Sample

The students involved in this research endeavor all attended a public Texas middle school. Grade 7 students who took both assessments (OLSAT and Stanford) and were enrolled for the entire year were the subjects for this study (N = 331). The sample was comprised of grade 7 students all housed on the same middle school (grades 6 through 8) campus. Two-thirds of the subjects were Caucasian, with the remaining third of primarily Hispanic decent. Of the 331 identified participants, nearly 20 percent participated in the federal Free and Reduced Lunch Program. Those few students with special needs were eliminated from the study because they took alternate approved state assessments with different tested objectives.

Descriptive Findings

The standards and performance-based TAKS Reading and Math assessments were administered to grades 3 through 11 as required by state law. The “Passing” scale
score (SS) for these assessments, regardless of grade or subject area, was 2100. Students were deemed to be at the “Commended” level when a scale score of 2400 or higher was obtained. The rounded mean scale score (SS) for the research school district’s seventh graders on the 2008 TAKS Reading state assessment was 2381, well above the state passing standard of 2100. For TAKS Math, the seventh grade students had a rounded mean scale score of 2329 which again surpassed the state passing level of 2100. The assessment statistics presented in Table 2 were based on the blind data sets obtained from the research study school system for grade 7 students during the 2007-2008 school year.

Table 2. Description of the Grade 7 Assessment Sample

<table>
<thead>
<tr>
<th>Assessment</th>
<th>Mean</th>
<th>Standard Deviation</th>
<th>Median</th>
<th>Mode</th>
</tr>
</thead>
<tbody>
<tr>
<td>OLSAT Verbal</td>
<td>100.03</td>
<td>14.29</td>
<td>102</td>
<td>107</td>
</tr>
<tr>
<td>OLSAT Quantitative</td>
<td>105.14</td>
<td>14.9</td>
<td>108</td>
<td>111</td>
</tr>
<tr>
<td>Stanford Reading %ile</td>
<td>74.30</td>
<td>23.70</td>
<td>82</td>
<td>87</td>
</tr>
<tr>
<td>Stanford Math %ile</td>
<td>78.59</td>
<td>20.45</td>
<td>84</td>
<td>99</td>
</tr>
<tr>
<td>TAKS Reading SS</td>
<td>2381</td>
<td>178.74</td>
<td>2431</td>
<td>2532</td>
</tr>
<tr>
<td>TAKS Math SS</td>
<td>2329</td>
<td>192.72</td>
<td>2437</td>
<td>2593</td>
</tr>
</tbody>
</table>

While the grade 7 students generally performed well above the passing standard for both subject areas, the standard deviation for Math and Reading, when applied to the assessment mean scores, indicated that some students were just meeting the recommended passing standard for Reading and Math.
Research Question #1

Does student performance on the norm-referenced Stanford Achievement Test Reading portion predict performance on the grade 7 Texas Assessment of Knowledge and Skills Reading test?

To frame the first question, the Pearson’s correlation was applied to determine if a relationship existed between grade 7 norm-referenced Stanford 10 Reading test ($M = 74.30$, $SD = 23.70$) and the scale score on the grade 7 criterion-referenced Texas Assessment of Knowledge and Skills Reading test ($M = 2380.69$, $SD = 178.74$). A significant positive correlation was revealed, $r = .664$, $p = .000$ (Table 3).

Table 3. Correlations between Stanford 10 Reading and TAKS Reading

<table>
<thead>
<tr>
<th></th>
<th>TAKS Reading Scale Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stanford Reading %ile</td>
<td></td>
</tr>
<tr>
<td>Pearson Correlation</td>
<td>.664**</td>
</tr>
<tr>
<td>Sig. (2-tailed)</td>
<td>.000</td>
</tr>
<tr>
<td>N</td>
<td>331</td>
</tr>
</tbody>
</table>

**. Correlation is significant at the 0.01 level (2-tailed).

Since the effect was significant (less than the .05 significance level), the null hypothesis was rejected. It is concluded that the correlation between the grade 7 Stanford 10 Reading test and the grade 7 TAKS Reading test was greater than zero.

Given the significant positive correlation, the statistical calculations were taken one step further with the application of the linear regression model. The linear regression model was used to determine the predictive value of the Stanford 10 Reading test for future TAKS Reading performance. The strength of the correlation was assessed on a scale of -1 to 1. Given the correlation of .664 (R) between Stanford
Reading and TAKS Reading it was concluded that there was a positive, strong correlation as depicted in the following Model Summary (Table 4).

Table 4. *TAKS Reading regressed on Stanford Reading*

<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>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.664*</td>
<td>.441</td>
<td>.439</td>
<td>133.915</td>
</tr>
</tbody>
</table>

*a. Predictors: (Constant), Stanford Reading %ile*

Additionally, the coefficient of determination (or R Square) value of .441 signified that 19.4% of the variability or change in TAKS Reading scale scores could be directly attributed from the changes in a student’s Stanford Reading percentiles.

Further testing with ANOVA assessed the linear relationship between the Stanford Reading and TAKS Reading assessments. Since the significance factor was 0.000 (Table 5) and fell between .00 and .05 it was concluded that a linear relationship did in fact exist between the two variables because the best-fitting line had a straight appearance, as indicated in Figure 1.

Table 5. *ANOVA: TAKS Reading and Stanford Reading*

**ANOVA°**

<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></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>4645322.668</td>
<td>1</td>
<td>4645322.668</td>
<td>259.033</td>
<td>0.000*</td>
</tr>
<tr>
<td></td>
<td>Residual</td>
<td>5900054.770</td>
<td>329</td>
<td>17933.297</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>1.055E7</td>
<td>330</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*a. Predictors: (Constant), Stanford Reading %ile*

*b. Dependent Variable: TAKS Reading SS*
Figure 1. Scatterplot 1: The Relationship Between TAKS Reading and Stanford Reading

Since a linear relationship was present between the two variables, a linear equation could be obtained using coefficient data from Table 6. Thus, the equation became $y = ax + b$ where $y$ was equal to future TAKS Reading scale score, $a$ was the Stanford Reading percentile coefficient, $x$ was the most recent Stanford Reading percentile, and $b$ was the TAKS Reading defined constant. Therefore, the prediction equation can be written:

$\text{Predicted TAKS Reading Scale Score} = 5.097(\text{Stanford Reading %ile}) + 1999.408$
Table 6. Coefficients from TAKS Reading regressed on Stanford Reading

<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>1999.408</td>
<td>24.838</td>
<td>80.499</td>
</tr>
<tr>
<td></td>
<td>Stanford Reading %ile</td>
<td>5.097</td>
<td>.317</td>
<td>.664</td>
</tr>
</tbody>
</table>

a. Dependent Variable: TAKS Reading SS

Based on the data it was concluded that as a student’s Stanford Reading Percentile increases, so does his or her TAKS Reading Scale Score. More precisely, for every one point difference in Stanford Reading Percentile, one can predict a 5.097 point difference in TAKS Reading Scale Score. This prediction creates the rate at which predicted TAKS Reading Scale Scores change in relation to Stanford Reading Percentiles.

Research Question #2

Does student performance on the norm-referenced Stanford Achievement Test Math portion predict performance on the grade 7 Texas Assessment of Knowledge and Skills Math test?

Pearson’s correlation was used to test the second research question to determine if a relationship existed between grade 7 norm-referenced Stanford 10 Math test ($M = 78.59$, $SD = 20.45$) and the scale score on the grade 7 criterion-referenced Texas Assessment of Knowledge and Skills Math test ($M = 2328.74$, $SD = 192.72$). A significant positive correlation was revealed, $r = .709$, $p = .000$ (Table 7).
Table 7. *Correlations between Stanford 10 Math and TAKS Math*

<table>
<thead>
<tr>
<th>Correlations</th>
<th>TAKS Math Scale Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stanford Math %ile</td>
<td>Pearson Correlation</td>
</tr>
<tr>
<td></td>
<td>Sig. (2-tailed)</td>
</tr>
<tr>
<td></td>
<td>N</td>
</tr>
<tr>
<td></td>
<td>.709**</td>
</tr>
<tr>
<td></td>
<td>.000</td>
</tr>
<tr>
<td></td>
<td>332</td>
</tr>
</tbody>
</table>

**. Correlation is significant at the 0.01 level (2-tailed).

Since the effect was significant (less than the .05 significance level), the null hypothesis was rejected. It was concluded that the correlation between the grade 7 Stanford 10 Math test and the grade 7 TAKS Math test was greater than zero.

Given the significant positive correlation, the statistical calculations were further analyzed with the application of the linear regression model. The linear regression model was used to determine the predictive value of the Stanford 10 Math test for future TAKS Math performance. The strength of the correlation was assessed on a scale of -1 to 1. Given the correlation of .709 (R) between Stanford Math and TAKS Math it was concluded that there was a positive, strong correlation as depicted in the following Model Summary (Table 8).

Table 8. *TAKS Math regressed on Stanford Math*

<table>
<thead>
<tr>
<th>Model Summary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
</tr>
<tr>
<td>-------</td>
</tr>
<tr>
<td>1</td>
</tr>
</tbody>
</table>

a. Predictors: (Constant), Stanford Math %ile
Additionally, the coefficient of determination (or R Square) value of .502 signified that 25.2% of the variability or change in TAKS Math scale scores was directly attributed to changes in a student’s Stanford Reading percentiles.

Further testing with ANOVA assessed the linear relationship between the Stanford Math and TAKS Math assessments. Since the significance factor was 0.000 (Table 9) and fell between .00 and .05 it was concluded that a linear relationship does in fact exist between the two variables since the best-fitting line had a straight appearance, as indicated in Figure 2.

Table 9. ANOVA: TAKS Math and Stanford Math

<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>Regression</td>
<td>6176914.064</td>
<td>1</td>
<td>6176914.064</td>
<td>333.224</td>
<td>0.000</td>
</tr>
<tr>
<td>Residual</td>
<td>6117155.174</td>
<td>330</td>
<td>18536.834</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>1.229E7</td>
<td>331</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

a. Predictors: (Constant), Stanford Math %ile

b. Dependent Variable: TAKS Math SS
Since a linear relationship was present between the two variables, a linear equation could be obtained using coefficient data from Table 10. Thus, the equation was expressed \( y = ax + b \) where \( y \) was equal to future TAKS Math scale score, \( a \) is the Stanford Math percentile coefficient, \( x \) was the most recent Stanford Math percentile, and \( b \) was the TAKS Math defined constant. Therefore, the prediction equation reads:

\[
\text{Predicted TAKS Math Scale Score} = 7.001(\text{Stanford Math \%ile}) + 1773.127
\]
Table 10. *Coefficients from TAKS Math regressed on Stanford Math*

<table>
<thead>
<tr>
<th>Model</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
<td>Std. Error</td>
<td>Beta</td>
<td>t</td>
</tr>
<tr>
<td>(Constant)</td>
<td>1773.127</td>
<td>31.341</td>
<td>56.575</td>
<td>.000</td>
</tr>
<tr>
<td>Stanford Math %ile</td>
<td>7.001</td>
<td>.383</td>
<td>.709</td>
<td>18.254</td>
</tr>
</tbody>
</table>

*a. Dependent Variable: TAKS Math SS*

Based on the data it was concluded that as a student’s Stanford Math Percentile increases, so does his or her TAKS Math Scale Score. More precisely, for every one point difference in Stanford Math Percentile, one can predict a 7.001 point difference in TAKS Math Scale Score. This prediction creates the rate at which predicted TAKS Math Scale Scores change in relation to Stanford Math Percentiles.

*Research Question #3*

Does student performance on the norm-referenced Otis-Lennon School Ability Test Verbal portion predict performance on the grade 7 Texas Assessment of Knowledge and Skills Reading test?

To frame the third question, a Pearson’s correlation was applied to determine if a relationship existed between grade 7 norm-referenced Otis-Lennon School Ability Test Verbal ($M = 100.03, SD = 14.29$) and the scale score on the grade 7 criterion-referenced Texas Assessment of Knowledge and Skills Reading test ($M = 2380.69, SD = 178.74$). A significant positive correlation was revealed, $r = .455, p = .000$. 
Table 11. *Correlations between OLSAT Verbal and TAKS Reading*

<table>
<thead>
<tr>
<th>Correlations</th>
<th>TAKS Reading Scale Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>OLSAT Verbal</td>
<td>Pearson Correlation .455**</td>
</tr>
<tr>
<td></td>
<td>Sig. (2-tailed) .000</td>
</tr>
<tr>
<td></td>
<td>N 331</td>
</tr>
</tbody>
</table>

**. Correlation is significant at the 0.01 level (2-tailed).

Since the effect was significant (less than the .05 significance level), the null hypothesis was rejected. It was concluded that the correlation between the grade 7 Otis-Lennon School Ability Test Verbal and the grade 7 TAKS Reading test was greater than zero.

Given the significant positive correlation, the statistical calculations were taken one step further with the application of the linear regression model. The linear regression model was used to determine the predictive value of the OLSAT Verbal test for future TAKS Reading performance. The strength of the correlation was assessed on a scale of -1 to 1. Given the correlation of .455 (R) between OLSAT Verbal and TAKS Reading it was concluded that there was a positive, medium correlation as depicted in the following Model Summary (Table 12).

Table 12. *TAKS Reading regressed on OLSAT Verbal*

<p>| Model Summary |
|---------------|----------------|---------------|----------------|</p>
<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>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.455**</td>
<td>.207</td>
<td>.205</td>
<td>159.405</td>
</tr>
</tbody>
</table>

a. Predictors: (Constant), OLSAT Verbal
Additionally, the coefficient of determination (or R Square) value of .207 signified that 4.2% of the variability or change in TAKS Reading scale scores was directly attributed to changes in a student’s OLSAT Verbal score.

Further testing with ANOVA assessed the linear relationship between the OLSAT Verbal and TAKS Reading assessments. Since the significance factor was 0.000 (Table 13) and fell between .00 and .05 it was concluded that a linear relationship does in fact exist between the two variables since the best-fitting line had a straight appearance, as indicated in Figure 3.

Table 13. *ANOVA: TAKS Reading and OLSAT Verbal*

<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>2189016.281</td>
<td>1</td>
<td>2189016.281</td>
<td>86.148</td>
</tr>
<tr>
<td></td>
<td>Residual</td>
<td>8385242.764</td>
<td>330</td>
<td>25409.827</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>1.057E7</td>
<td>331</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

a. Predictors: (Constant), OLSAT Verbal

b. Dependent Variable: TAKS Reading SS
Since a linear relationship was present between the two variables, a linear equation could be obtained using coefficient data from Table 14. Thus, the equation became $y = ax + b$ where $y$ was equal to a future TAKS Reading scale score, $a$ was the OLSAT Verbal coefficient, $x$ was the most recent OLSAT Verbal score, and $b$ was the TAKS Reading defined constant. Therefore, the prediction equation reads:

$$\text{Predicted TAKS Reading Scale Score} = 5.767(\text{OLSAT Verbal}) + 1801.116$$
Table 14. *Coefficients from TAKS Reading regressed on OLSAT Verbal*

<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>1801.116</td>
<td>63.053</td>
<td>28.565</td>
</tr>
<tr>
<td></td>
<td>OLSAT Verbal</td>
<td>5.767</td>
<td>.621</td>
<td>.455</td>
</tr>
</tbody>
</table>

*a. Dependent Variable: TAKS Reading SS*

Based on the data it was concluded that as a student’s OLSAT Verbal score increases, so does his or her TAKS Reading Scale Score. More precisely, for every one point difference in OLSAT Verbal score, one can predict a 5.767 point difference in TAKS Reading Scale Score. This prediction creates the rate at which predicted TAKS Reading Scale Scores change in relation to OLSAT Verbal Scores.

*Research Question #4*

Does student performance on the norm-referenced Otis-Lennon School Ability Test Quantitative portion predict performance on the grade 7 Texas Assessment of Knowledge and Skills Math test?

Pearson’s correlation was used to test the fourth research question to determine if a relationship existed between grade 7 norm-referenced Otis-Lennon School Ability Test Quantitative ($M = 105.14$, $SD = 14.90$) and the scale score on the grade 7 criterion-referenced Texas Assessment of Knowledge and Skills Math test ($M = 2328.74$, $SD = 192.72$). A significant positive correlation was revealed, $r = .541$, $p = .000$. 
Table 15. *Correlations between OLSAT Quantitative and TAKS Math*

<table>
<thead>
<tr>
<th></th>
<th>TAKS Math Scale Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>OLSAT Quantitative</td>
<td><strong>.541</strong></td>
</tr>
<tr>
<td>Pearson Correlation</td>
<td>.000</td>
</tr>
<tr>
<td>Sig. (2-tailed)</td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>332</td>
</tr>
</tbody>
</table>

**. Correlation is significant at the 0.01 level (2-tailed).

Since the effect was significant (less than the .05 significance level), the null hypothesis was rejected. It was concluded that the correlation between the grade 7 Otis-Lennon School Ability Test Quantitative and the grade 7 TAKS Math test was greater than zero.

Given the significant positive correlation, the statistical calculations were taken one step further with the application of the linear regression model. The linear regression model was used to determine the predictive value of the OLSAT Quantitative test for future TAKS Math performance. The strength of the correlation was assessed on a scale of -1 to 1. Given the correlation of .541 (R) between OLSAT Quantitative and TAKS Math it was concluded that there was a positive, strong correlation as depicted in the following Model Summary (Table 16).

Table 16. *TAKS Math regressed on OLSAT Quantitative*

<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>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.541**</td>
<td>.293</td>
<td>.291</td>
<td>162.269</td>
</tr>
</tbody>
</table>

a. Predictors: (Constant), OLSAT Quantitative
Additionally, the coefficient of determination (or R Square) value of .293 signified that 8.5% of the variability or change in TAKS Math scale scores was directly attributed to changes in a student’s OLSAT Quantitative score.

Further testing with ANOVA assesses the linear relationship between the OLSAT Quantitative and TAKS Math assessments. Since the significance factor was 0.000 (Table 17) and fell between .00 and .05 it could be concluded that a linear relationship did in fact exist between the two variables because the best-fitting line had a straight appearance, as indicated in Figure 4.

Table 17. ANOVA: TAKS Math and OLSAT Quantitative

<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>Regression</td>
<td>3604725.546</td>
<td>1</td>
<td>3604725.546</td>
<td>136.899</td>
<td>0.000*</td>
</tr>
<tr>
<td>Residual</td>
<td>8689343.692</td>
<td>330</td>
<td>26331.345</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>1.229E7</td>
<td>331</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

a. Predictors: (Constant), OLSAT Quantitative

b. Dependent Variable: TAKS Math SS
Since a linear relationship was present between the two variables, a linear equation could be obtained using coefficient data from Table 18. Thus, the equation became $y = ax + b$ where $y$ was equal to future TAKS Math scale score, $a$ was the OLSAT Quantitative coefficient, $x$ was the most recent OLSAT Quantitative score, and $b$ was the TAKS Math defined constant. Therefore, the prediction equation reads:

$$\text{Predicted TAKS Math Scale Score} = 7.264(\text{OLSAT Quantitative score}) + 1560.635$$
Table 18. Coefficients from TAKS Math regressed on OLSAT Quantitative

<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>1560.635</td>
<td>66.250</td>
<td>23.557</td>
</tr>
<tr>
<td></td>
<td>OLSAT Quantitative</td>
<td>7.264</td>
<td>.621</td>
<td>.541</td>
</tr>
</tbody>
</table>

Based on the data it was concluded that as a student’s OLSAT Quantitative score increases, so does his or her TAKS Math Scale Score. More precisely, for every one point difference in OLSAT Quantitative score, one can predict a 7.264 point difference in TAKS Math Scale Score. This prediction creates the rate at which predicted TAKS Math Scale Scores change in relation to OLSAT Quantitative scores.

Results Summary

The 2007-2008 seventh grade test results presented in this chapter demonstrate significant correlations and predictive strength, regardless of which norm-referenced assessment is compared to the TAKS Reading or Math test. With a statistically significant (two-tailed) correlation of .000 for each research question, it was clear that a predictive relationship did in fact exist for the questions investigated. More detailed descriptions of the findings are found in chapter five.
CHAPTER V: DISCUSSION

The purpose of this final chapter is to summarize the research study presented in the previous four chapters and discuss the results of the study. The chapter is divided into the following sections: (a) purpose of the study; (b) problem statement; (c) methodology review; (d) research study results summary; (e) discussion of results; (f) implications; (g) limitations; and, (h) suggested recommendations for further research.

Purpose of the Study

Testing has experienced significant changes in the public school sector during the past 30 years. Peter Behuniak (2004) cited two specific changes that altered the face of testing, including the expansion of achievement testing and the increase emphasis and development of educational accountability systems (p. 335). Achievement testing and accountability are pieces of a theoretical-based standards movement that has continued to gain political and public momentum since the passing of No Child Left Behind in 2001. While many educators, parents, and communities are still unsure as to whether or not NCLB is the answer to improving public education, most agree some type of standards-based, performance measure is necessary to further the educational gains of American children.

According to Zucker (2003), standards-based assessments have largely been used to “measure what students know and can do, improving instruction, and helping students achieve higher standards” (p.2). Educators rely on carefully developed tests to provide valuable information for school personnel as they plan and guide student
instruction. Fremer and Wall (2004) state that testing design is “a special way of collecting information used to help make decisions about individuals, programs, or institutions” (p. 4). Additionally, Fremer and Wall (2004) state that “NCLB legislation requires regular accountability testing as a prerequisite for receiving federal funds. The legislation puts pressure on states to identify poorly performing schools and seek remedies to the situation” (p. 8). The pressure that has been placed upon school systems has led to a practice by state education agencies to design and implement standards and performance-based, “high-stake” examinations.

Schools are giving these standard-based assessments to adhere to state and federal testing and accountability measures. Countless school systems are redesigning their testing practices to ensure curricular alignment and gathering comprehensive data to inform decisions. Given that there is no set national curriculum that is taught in every state, states have the flexibility of designing their own “high-stake” assessments to comply with federal testing guidelines.

High-stakes testing is often viewed as a negative part of current standards-based school reform theories; however, performance-based testing actually carries both positive and negative effects for students, teachers, and administrators. Though the tendency is to voice the negatives of high-stakes testing, Stecher (2002) indicated that students were granted “better information about their own knowledge and skills” and obtain “clearer signals as to what to study” (p. 86). He went on to mention that teachers were able to “support better diagnosis of individual student needs…and the strength and weakness” of the classroom curriculum, causing school administrators to “examine school policies related to curriculum and instruction” (p. 86). The positive
effects of standards or performance-based testing, as previously mentioned, enables educators to better assess the personal academic needs of students to ensure a challenging and comprehensive education for all.

While accountability often drives a school’s use of testing results, standards-based reform theory, assessment looping, and Vernon’s (1961) Hierarchical Structure of Human Abilities theoretical framework are capable of signaling the effectiveness of educational programs when skillfully interpreted and applied to curricular decisions. Hence, school systems should investigate correlations between norm-referenced assessment results and state criterion-referenced standardized testing outcomes to further evaluate student strengths and weaknesses and to gauge future performance.

To assist in meeting the NCLB regulations and to familiarize students with standardized testing structure and format, many campuses frequently use research-based, norm-referenced ability and/or achievement tests to benchmark student progress and growth prior to “high-stakes testing” situations. The Stanford Achievement Test and the Otis-Lennon School Ability Test are two prominent norm-referenced, standardized assessment tools frequently administered to determine student academic skill and ability level. Despite the fact that these exams may not be completely aligned to each state’s standards and objectives, they closely resemble general curriculum expectations and offer valid and reliable feedback to educators and parents. Additionally, these exams are a powerful tool for school systems that use data to help facilitate curriculum decision making.

The purpose of this quantitative correlational research study was to determine if the Stanford 10 and OLSAT could predict an anticipated student performance level
on the Texas state criterion-referenced reading and math assessments known as the TAKS test. Results of the Stanford, OLSAT, and the TAKS reading and math assessments from the 2007-2008 school year were used to determine if a significant relationship exists. The presence of a significant correlation would be indicative of the potential predictive value of the Stanford and OLSAT when projecting future performance on TAKS assessments. This positive relationship would have the capability to impact instructional planning and have a direct correlation on future assessment procedures by enabling schools to systematically gather yearly and historical data results for disaggregation and interpretation. In addition, merging data results would create comprehensive tools for tracking, appraising and monitoring student progress.

Problem Statement

School systems are looking for more ways to stretch the dollar and using existing data to evaluate student progress, especially given recent nation-wide financial woes. The research problem for this study was to determine whether a correlational relationship existed between the Stanford 10 Achievement Test or the OLSAT 8 and grade 7 TAKS Reading and/or Math scale scores. In addition, the problem was also designed to investigate whether the results from the Stanford and OLSAT administrations could be applied to a mathematical formula capable of predicting anticipated performance on the grade 7 state-mandated reading and math exams known as the Texas Assessment of Knowledge and Skills (TAKS). The null hypotheses assumed no significant correlational relationships between scores of the Stanford Achievement Tests and/or Otis-Lennon School Ability Test (OLSAT) when
coupled with TAKS math and reading results. These research questions were postulated to conclude whether one or more norm-referenced tests were capable of predicting a student’s probable performance level on a criterion-referenced test, such as the Texas Assessment of Knowledge and Skills (TAKS).

Methodology Review

This correlational research endeavor involved grade 7 students from a Texas public middle school. The research sample was comprised of 331 grade 7 students who were enrolled for the entire 2007-2008 school year and took the OLSAT and Stanford, as well as the spring administration of TAKS Reading and Math. Archived test results from the 2007-2008 school year were obtained from the research study school system for the indicated grade 7 students in the areas of TAKS Reading and TAKS Math, as well as from the Stanford Reading, Stanford Math, OLSAT Verbal, and OLSAT Quantitative subtests.

A correlational research design and multiple regression formula were applied to study any potential correlational relationships between the Texas Assessment of Knowledge and Skills Reading and Math scale scores, Reading and Math Stanford Achievement Test percentiles, and Verbal and Quantitative Otis-Lennon School Ability Test SAI scores. The dependent variables were the Reading and Math TAKS scores, while the independent variables were the Stanford Reading and Math percentiles and the Verbal and Quantitative OLSAT SAI scores.

Using scores from each of the previously mentioned assessments or subtests, the correlation research design (r) was applied to test the relationship between the measured variables. In addition, regression analysis was also used to determine the
ability to measure the strength of the dependent and independent variables (X and Y) being assessed in the linear regression. The goal of the linear regression analysis was to establish a mathematical equation that could be used to predict the values of a specific dependent variable when based on the values of a given independent variable. In its simplest form, linear regression analysis highlighted the straight line that best fit the data evaluated. All statistical analyses were performed using the Statistical Package for Social Sciences software and addressed Research Questions 1, 2, 3, and 4.

Research Study Results Summary

Research Question #1. The first research question addressed whether a relationship existed between the Stanford Achievement Reading Test and the TAKS Reading Test. Additionally, the question investigated whether Stanford Reading was capable of predicting future TAKS Reading performance. For Research Question #1, the Null Hypotheses 1-\(H_0\) was rejected when a significant positive correlation was revealed. Finding were: \(r = .664\) and \(p = .000\). It was concluded that the correlation between the grade 7 Stanford 10 Reading test and the grade 7 TAKS Reading test was greater than zero.

Given the significant positive correlation relationship, the linear regression model was applied to determine the predictive value of the Stanford 10 Reading test on future TAKS Reading scores. The linear regression model yielded a \(R^2\) value of \(.441\) signifying that 19.4% of the variability or change in TAKS Reading scale scores could be directly attributed to the changes in a student’s Stanford Reading percentiles.

Research Question #2. The second research question postulated whether a relationship existed between the Stanford Achievement Math Test and the TAKS...
Math Test. Additionally, the question evaluated whether Stanford Math was capable of predicting future TAKS Math performance. For Research Question #2, the Null Hypotheses $H_{01}$ was rejected when a significant positive correlation was revealed. Findings were: $r = .709$ and $p = .000$. It was found that the correlation between the grade 7 Stanford 10 Math test and the grade 7 TAKS Math test was greater than zero.

With the presence of a positive correlational relationship, the linear regression model was applied to investigate the predictive value of the Stanford 10 Math test on future TAKS Math scores. The linear regression model yielded a $R^2$ value of .502 signifying that 25.2% of the variability or change in TAKS Math scale scores could be directly attributed to the changes in a student’s Stanford Math percentiles.

*Research Question #3.* The third research question investigated whether a relationship existed between the OLSAT Verbal and the TAKS Reading Test. Additionally, the question analyzed whether OLSAT Verbal was capable of gauging future TAKS Reading performance. For Research Question #3, the Null Hypotheses $H_{01}$ was rejected when a significant positive correlation was revealed. Findings were: $r = .455$ and $p = .000$. It was determined that the correlation between the grade 7 OLSAT Verbal test and the grade 7 TAKS Reading test was greater than zero.

The linear regression model was applied to determine the predictive value of the OLSAT Verbal test on future TAKS Reading scores since a significant positive correlational relationship was found. The linear regression model yielded a $R^2$ value of .207 signifying that 4.2% of the variability or change in TAKS Reading scale scores could be directly attributed to the changes in a student’s OLSAT Verbal score.
Research Question #4. The final research question addressed whether a relationship existed between OLSAT Quantitative and the TAKS Math Test. In addition, the question investigated whether OLSAT Quantitative was capable of predicting future TAKS Math performance. For Research Question #4, the Null Hypotheses 1-\(H_0\) was rejected when a significant positive correlation was revealed. Findings were: \(r = .541\) and \(p = .000\). It was statistically concluded that the correlation between the grade 7 OLSAT Quantitative test and the grade 7 TAKS Math test was greater than zero.

The presence of a significant positive correlation relationship enabled this researcher to apply the linear regression model in order to determine the predictive value of the OLSAT Quantitative test on future TAKS Math scores. The linear regression model yielded a \(R^2\) value of .293 signifying that 8.5% of the variability or change in TAKS Math scale scores could be directly attributed to the changes in a student’s OLSAT Quantitative score.

Discussion of Results

The ability to reject each of the null hypotheses is significant for educators. By establishing a mathematical equation capable of predicting future TAKS performance, school administrators and teachers can more accurately plan instructional activities. Schools are constantly looking for evaluative tools to assess student progress, creating the theoretical assessment loop. The assessment loop theory holds that the use of assessments is never-ending. Erford and Moore-Thomas (2004) believe the assessment loop cycle represents a “process in which assessment results are interpreted and fed back into the improvement process” (p. 520). Many campuses may need the
information to enhance curriculum or may desire the information to respond to
individual needs of students. Norm-referenced and criterion-referenced tests create
that quantitative data that provides reliable and valid evidence for decision making.

Interest in this research was motivated by the desire to obtain an additional use
for existing TAKS, Stanford 10, and OLSAT evidence-based data. The data obtained
in this research study was important because it helped generated a specific
mathematical formula for predicting a student’s future performance on either the
TAKS Reading or TAKS Math test and confirmed correlational relationships between
the analyzed assessments and subtests. This equation gives educators a better
predictive tool for gauging a student’s future academic performance on state
standards-based, criterion-referenced assessments versus the traditional “guessing”
method. The additional statistical evidence obtained provides another statistically
analyzed tool for educators as they examine and disaggregate existing standards and
performance-based testing data.

No prior research studies were found comparing student performance on the
Texas Assessment of Knowledge and Skills Reading and/or Math test and either the
Stanford 10 Achievement Math and Reading Test or the Otis-Lennon School Ability
Verbal and Quantitative Test. However, Bowman (2004), in a comparable study using
the previous version of the Texas state assessment, found that student’s with a low IQ
or School Ability Index on the OLSAT had the lowest TAAS (Texas Assessment of
Academic Skills) scores. The TAKS test replaced the TAAS test in 2003 and requires
higher level thinking and curricular standards.
While there is not an exact replica study, a comparable study using the Otis-Lennon Mental Ability Test, Stanford Achievement Test (SAT), and three demographic variables were investigated as potential predictors of achievement in second and fourth grades. The 1982 study, conducted by Antonak, King, and Lowy, was designed to determine that relationships existed between the previous mentioned variables and student achievement. The researchers’ found that the best predictor for math achievement occurred from grade 2 Stanford Math to grade 4 Stanford Math. Though the OLMAT, predecessor of OLSAT, did show prediction capabilities within a given grade level, the SAI ultimately became a minimal factor when forecasting future academic achievement. The outcomes of the Antonak, King, and Lowy (1982) study confirmed similar results found by Churchill and Smith in 1966.

The results reported in one’s research study reflect similar conclusions as those found by Antonak, King, and Lowy. Though designed slightly different, one’s research endeavor also concluded that Stanford Math was the best predictor of future math performance, but on the TAKS Math assessment, instead of a different grade level of the Stanford Math assessment. This researcher’s study found that Stanford Math Achievement Test was most highly correlated to TAKS Math performance (R = .709) for grade 7 students. Stanford Reading Achievement Test also showed a strong correlation to TAKS Reading performance (R = .664). Both of these studies confirm the strong relationships between achievement tests and future academic performance. With these findings, school systems might consider reevaluating the frequency of administering both the SAT 10 and OLSAT.
The OLSAT provides valuable insight into a child’s general scholastic ability, but weak correlations were concluded when compared with TAKS. The weakest correlation found existed between the OLSAT Verbal and TAKS Reading ($R = .455$), though statistically it is still considered an average positive correlation. This finding was a bit surprising as most people might assume that IQ is a better predictor for academic success over results obtained from achievement tests. This researcher believes that while IQ may initially impact ability to learn, a student’s work ethic and commitment to learning is often not reflected in an IQ score and may account for some of the differences between the Stanford and OLSAT results. While this researcher was initially surprised that the achievement test was a better correlational predictor of TAKS performance, Antonak, King, and Lowy (2006) established similar results and even recommended in their study that school systems discontinue the use of ability tests. Their study suggested that “school districts with a comprehensive achievement testing program gained nothing by adding a mental ability test to the program” (p. 372).

Regardless of the strength of correlations, once the linear regression model was applied to each research question, this researcher was able to determine the predictive value and impact on future TAKS Reading and Math scale scores. Stanford Achievement Test, whether Math or Reading, indicated significant variability or change to a student’s TAKS Math or Reading score with each Stanford percentile increase. The same held true for the OLSAT Verbal and Quantitative scores, as they increased, so did a student’s TAKS Reading or Math score. Based on study results and the strength of the correlations, SAT 10 should give schools a clearer snapshot of
current achievement levels and future TAKS performance; information critical to
driving curricular decisions.

stated that “curriculum and assessment standards have been in place in Texas long
before the passage of NCLB,” which “makes the State of Texas both an expert and a
target in the area of standardized assessments” (p. 211). Texas schools face enormous
pressure to make achievement gains each year on both the state and federal levels.
Although schools utilize Stanford 10 and OLSAT 8 for multiple different purposes,
campuses may not fully understand the potential predictive power of the norm-
referenced data when coupled with the state criterion-referenced scores and
statistically examined.

Given that the correlations exist between the two norm-referenced tests and
TAKS, Texas schools could input the necessary scores into the mathematical equation
to generate a more precise prediction of future TAKS performance and narrow the
pool of “bubble kids.” Bubble kids are generally perceived as those students who
“might” pass the state assessment or fail the state assessment…it could go either way.
Most of the time bubble students are discovered from previous testing results,
benchmarking outcomes, and classroom performance. Though not foolproof, the
mathematical equation generated from this study applies the most statistical backing
behind classifying a student as a bubble kid.

Bubble kids and students underperforming are the perfect group for response to
intervention (RTI) services. RTI developed from the Individuals with Disabilities
Education Act of 2004. IDEA, as it is commonly known, is the governing legislation
that assists children with disabilities. RTI is a tiered intervention process that systematically provides research-based programs to students with the intent of closing deficient academic, emotional, or behavioral gaps. Educators now have the option of using RTI to provide early interventions to students at risk in school (Fuchs and Fuchs, 2006). Additionally, according to Fuchs and Fuchs (2006), schools may use “as much as 15% of their special education monies to fund early intervention activities” (p. 93).

One of the first steps of the RTI process is to identify at risk students. The outcomes of this study provide educators with a means to statistically generate a pool of students who are predicted to have academic gaps based on current achievement information. Fuchs and Fuchs (2006) indicate that the “best strategy is to assess every student in the grade on a screening tool with a benchmark that demonstrates utility for predicting end-of-year performance on high-stakes tests” (p.93). Stanford 10 and OLSAT both demonstrated in this study that one or both instruments are capable of serving as a screening tool for future academic success.

Progress monitoring of students is a foundational piece of RTI. Teachers evaluate the data, according to Fuchs and Fuchs (2006), to “determine whether they (the educators) need to change their curricula, materials, or instructional procedures” (p. 94). School systems that utilize the RTI process generally establish two to four tiers of intervention, with the fourth tier being the most intensive and direct. A RTI committee is established by the campus to:

Determine the magnitude of the problem, analyze its causes, design a goal-directed intervention, conduct it as planned, monitor student progress, modify
the intervention as needed, and evaluate its effectiveness and plot future actions (Fuchs and Fuchs, 2006, p. 95).

RTI relies on data to drive campus and curriculum decisions. Stanford 10, OLSAT, and TAKS all provide valuable academic and ability information. These critical pieces of information give RTI teams the necessary data to monitor student progress and modify campus interventions to meet individual student needs. Regardless of how the data is utilized, the results of this study have several implications for educators and students.

Implications

Based on the findings in this study, several implications can be noted. First, regardless of whether Stanford or OLSAT was used, both had the ability to help predict future TAKS Reading or Math performance. The predictive equation that developed from this research study was significant because it gives schools a statistically proven method for finding borderline students. Bridging the achievement gap would be significant for those bubble students who might have been overlooked by teachers or other educators. Finding a means to bridge the gap in grade 7 is especially critical in Texas because in grade 8 students are required by law to pass math and reading state assessments prior to promotion to grade 9. These findings suggest that school personnel should be evaluating fall performance on these norm-referenced tests to better reveal a student’s potential score range on the spring TAKS Reading and Math assessments.

A second significant implication includes using the results to assist in identifying those students in need of “response to intervention” services. Standards-
based assessments, like TAKS, provide important curriculum documentation of a student’s academic strength and weaknesses. The ability to make probable predictions of future TAKS performance would give RTI committee’s two advantages. First, the committees would be able to pre-target students in need of interventions prior to obtaining actual standards-based score information, delivering support from a proactive approach. The second benefit for RTI committee’s includes the ability to track academic gains of an intervention program based on increases in a student’s anticipated TAKS Math or Reading performance level. Regardless of the specific implication, having these predictive findings as part of the assessment loop will enable school personnel the opportunity to narrow educational deficiencies in students.

**Limitations**

While the comparisons of norm-referenced and criterion-referenced tests are one tool for identifying the strengths and weaknesses of students, one must not assume that the predictive mathematical formulas are statements of cause and effect. The norm-referenced scores are said to account for a specific variability in a student’s standards-based, criterion-referenced performance (like that of the TAKS Reading and Math test) and provide school leadership with additional data to drive educational and curricular decisions.

An additional limitation of the study was that the data collected in the study reflected student performance on the Texas Assessment of Knowledge and Skills which is specific to curricular standards designed and required by the Texas Education Agency and may not reflect other state curricular standards. It should also be noted
that the results in this study are limited to information collected from one Texas public school system containing one junior high campus for the 2007-2008 school year.

Given that school populations are very transient and in constant flux, one is cautioned when generalizing these findings past the studied school year. While the research study revealed a predictive mathematical formula for the given school year, it is not known whether the predictive ability can be future-cast into multiple school years without recalculation of new OLSAT, Stanford, and TAKS scores.

**Suggested Recommendations for Further Research**

Based on this and the limited research available linking norm-referenced and state criterion-referenced tests, additional research is recommended to examine whether a more precise relationship exists when both Stanford Reading and OLSAT Verbal scores are combined to predict future TAKS Reading performance or when Stanford Math and OLSAT Quantitative scores are combined to gauge future TAKS Math performance. These larger-scale research studies could potentially further the predictive use of norm- and criterion-referenced assessments in a campus’ or school system’s assessment loop process.

An additional study is recommended to determine if a student’s predicted TAKS Reading and Math performance could be applied to future school years when more than one year of historical TAKS, Stanford, and OLSAT scores were evaluated for a given sample of students. This researcher also wonders whether the predictive value of TAKS, Stanford, or OLSAT scores would vary by ethnicity, gender, or economic status. It is also suggested that the initial research study be replicated using Stanford Science and Social Studies percentiles and TAKS Science and Social Studies
scale scores to determine their respective correlational and/or predictive relationship. Finally, a longitudinal study of three or more years investigating the predictive value of norm-referenced on state standards-based, criterion-referenced assessments is advised to establish historical trends between these examinations.

Conclusion

Standards-based education is a fixture in the current public school arena. Regardless of whether one believes this is the answer to improving American schools or not, performance-based assessment has the ability to improve academic practices and ensure accountability for those responsible for teaching our children. It is imperative that schools investigate and analyze all data sources to appropriately design and implement academic intervention and classroom instructional programs. The results of this study provide another avenue for educators to explore as they disaggregate data to balance, inform, and drive curriculum decision making processes. Central to creating this balanced and equitable system of education, is a school district’s ability to revise and implement sound curriculum, intervention, and instructional practices. Standards-based, criterion-referenced test prediction is a means for school systems to improve curriculum, align instructional practices, bridge achievement gaps, and forecast the approaching scholastic needs of their students.
REFERENCES


APPENDIX A

Permission Request to Research Study School System’s Assistant Superintendent
XXX Broadway
XXXXXXX, TX XXXXX
November 3, 2008

Dear Mrs. XXXXXX,

I hope you are doing well! As you know, I am currently a doctoral candidate in Educational Leadership at Liberty University in Lynchburg, VA. The purpose of my dissertation is to determine the correlation between the Otis-Lennon School Ability Test or the Stanford Achievement Test and the Texas Assessment of Knowledge and Skills Math and Reading Tests. It is also my intent to investigate whether or not a predictive relationship exists between these variables.

I would like to utilize the following seventh grade 2007-2008 school year test results:

- OLSAT Verbal Score
- OLSAT Quantitative Score
- Stanford Reading Percentile
- Stanford Math Percentile
- TAKS Reading Scale Score
- TAKS Math Scale Score

I respectfully request your permission to obtain the above data sets from the district data system.

I appreciate your continued support of my educational goals. At your request, I will share the results with you and any other school personnel at the conclusion of the research study.

I look forward to hearing from you and again, thank you for your consideration. If you have any questions, Please feel free to contact me at (XXX) XXX-XXXX.

Sincerely,

Kristen Karrh
APPENDIX B

Institutional Review Board (IRB) Approval
IRB Approval 685.022509: Kristen Karrh

Predictors of Student Achievement in Grade 7

Dear Kristen,

We are pleased to inform you that your above study has been approved by the Liberty IRB. This approval is extended to you for one year. If data collection proceeds past one year, or if you make changes in the methodology as it pertains to human subjects, you must resubmit the study to the IRB. See the IRB website for appropriate forms in these cases.

Thank you for your cooperation with the IRB and we wish you well with your research project.

Sincerely,

Fernando Garzon, Psy.D.
IRB Chair, Liberty University
Center for Counseling and Family Studies
Liberty University
1971 University Boulevard
Lynchburg, VA 24502-2269
(434) 592-4054
Fax: (434) 522-0477
APPENDIX C

Permission Letter from Research Study School District’s Assistant Superintendent
March 6, 2009

To Whom It May Concern:

Kristen Karrh has been granted permission to access Independent School District 2007-2008 seventh grade Stanford, Otis-Lennon School Ability Test, TAKS Reading Test, and TAKS Math Test results for the purpose of her dissertation research study. All information obtained by Ms. Karrh will be in a blind-format to protect the confidentiality of students. Please let me know if you need additional information.

Sincerely,

[Signature]

Assistant Superintendent of Curriculum and Instruction

Independent School District