THE MYTH OF BENCHMARK TESTING: ISOMORPHIC PRACTICES IN TEXAS PUBLIC SCHOOL DISTRICTS' USE OF BENCHMARK TESTING

by

Karen D. Jones, B.S., M.Ed.

A dissertation submitted to the Graduate Council of
Texas State University in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
with a Major in School Improvement
December 2013

Committee Members:

Sarah W. Nelson

Michael P. O'Malley

Larry R. Price

Patrice Holden Werner

FAIR USE AND AUTHOR'S PERMISSION STATEMENT

Fair Use

Duplication Permission

ACKNOWLEDGEMENTS

I would like to acknowledge and appreciate all who have assisted me on this journey. Thank you Dr. Sarah Nelson, my advisor and mentor. Dr. Nelson provided encouragement throughout this process. Her guidance through challenges pushed me to develop new understandings. Her counsel extended beyond academics. Dr. Nelson helped me navigate the work/life/school balance so I could preserve my sanity, my career and my personal relationships.

Thank you to my committee members, Dr. O'Malley, Dr. Price, and Dr. Werner for their valuable instruction, advice, and suggestions. Thank you, Dr. Price, for your generous assistance in working through the complex models needed for this research.

Thank you to my classmates and colleagues who listened to me, brainstormed with me, and spent hours in coffee shops thinking out loud with me.

I am grateful for my family and friends who sustained me on this journey. Without your steadfast belief in my abilities, I would not have reached this goal. You cheered me and propelled me, particularly when I was overwhelmed and under-motivated.

TABLE OF CONTENTS

Page

LIST OF TABLES

Table **Page**

LIST OF FIGURES

Figure **Page**

ABSTRACT

This dissertation examines the use of benchmark tests in Texas public schools through a quantitative study of 100 school districts using path analysis. The study examines the relationship between a district's descriptive characteristics and the number of benchmark tests they require. Districts' descriptive characteristics include district type, accountability status, percent of Limited English Proficient (LEP) students, percent of economically disadvantaged students, and percent students of color. The number of benchmark tests a district required was also compared to the percent of students passing the state required 8th grade math and reading assessments. This study found districts with certain characteristics and student populations were more likely to use benchmark tests. This study also found a small, insignificant, and negative relationship between the number of benchmark tests a district required and the percent of students passing the state tests. This suggests the greater the number of benchmark tests required by a district, the lower the percentage of students passing the state test. The results of the study are examined through the lens of isomorphism and rational myths in public education are addressed.

CHAPTER 1

INTRODUCTION

Background to the Study

American public schools administer more than 100 million standardized tests each year (Taubman, 2009). The No Child Left Behind Act of 2002 (NCLB) holds states accountable for educating all students to a high standard. Under NCLB, each state is responsible for developing an accountability system that ensures all public elementary and secondary schools make adequate yearly progress toward reaching the academic standards set forth by NCLB. Further, NCLB (2002) mandates that state accountability plans "include sanctions and rewards, such as bonuses and recognition" (NCLB, 2002: State Plans, 20 U.S.C. § 6311**)** as part of ensuring schools and districts make adequate yearly progress. NCLB requires that states define adequate yearly progress in a way that holds all students to the same high standards and measures progress primarily through academic assessments (NCLB, 2002). These assessments must be valid and reliable, consistent with "widely accepted professional testing standards (e.g., Standards for Educational and Psychological Testing, AERA, APA, NCME, 1999), and objectively measure academic achievement" (NCLB, 2002, 20 U.S.C. § 6311**).** Schools that do not make adequate yearly progress are subject to sanctions. These sanctions may include: replacing the school staff who are relevant to the failure to make adequate yearly progress, significantly decreasing management authority at the school level, and/or restructuring the internal organizational structure of the school (NCLB, 2002).

The requirements of NCLB create sanctions that are high-stakes: students may be retained in grade level, educators may lose their jobs, and schools may close based on test scores (Harrison-Jones, 2007). Pressure to succeed and have students perform well on standardized tests leads many schools and districts to require the use of benchmark, or practice tests (Valli, Croninger, Chambliss, Graeber, & Buese, 2008). These benchmark tests are viewed as a way to prepare students for the high-stakes, standardized tests required by NCLB and to predict student performance on these accountability tests.

Statement of the Problem

Throughout the twentieth and now the twenty-first centuries, assessments have been used to measure what a student knows and to help educators make instructional decisions, such as whether a student should be placed in gifted or other special classes (Delandshere, 2001; Schwandt, 2005). According to the Joint Committee on Standards for Educational and Psychological Testing of the AERA, APA, and NCME (1999), assessment is "any systematic method of obtaining information from tests and other sources, used to draw inferences about characteristics of people, objects, and programs" (p. 5). Assessment purposes have changed over the years from a focus on the individual student to a focus on groups of students, schools, and districts. The mandates and sanctions embedded in the policy NCLB (2002) have further shifted the focus of assessments from measuring learning to determining the quality of schools and the readiness of students to be promoted to the next grade (Gunzenhauser, 2003; Heilig & Darling-Hammond, 2008).

There are two main types of assessments used in schools. One type is formative assessment and the other is summative assessment. Formative assessment is used to

gather evidence of learning in an on-going progression and then adjust instruction to meet students' needs (Popham, 2006; Scriven, 1967). The second type of assessment is summative assessment, which generally occurs at the end of learning (Stiggins, 2002; Taras, 2005). Formative and summative assessments are complimentary and their combined use helps to create a more comprehensive measure of learning.

However, the high-stakes embedded in NCLB (2002) have shifted the focus of educational assessment heavily toward summative assessments. This trend is seen in the changing use of benchmark tests. When benchmark tests originated, they were intended to be formative assessments to help teachers drive instruction by showing what a student knew at a particular point in time (Black & Wiliam, 1998). Benchmark assessments were distinct from practice tests, which were used to help students become familiar with the format of the standardized tests. Since the introduction of high-stakes accountability, the nature and purpose of benchmark tests has shifted such that benchmark tests are now viewed primarily as a means to prepare students for high-stakes standardized tests. They are also used to identify students in need of additional test preparation, such as tutoring. In fact, since the inception of NCLB, the terms "practice test" and "benchmark test" have been used synonymously in the literature (Haertel, 1999; Linn, 2000; Hamilton, 2003; Trimble, Gay & Matthews, 2005).

According to Popham (2001), benchmark and other forms of practice testing disrupts instruction and impacts instruction much more than originally intended. The pressures of high-stakes testing causes teachers and schools to spend valuable time preparing students to take tests, dissuading best practices of teaching and learning (Au, 2007; Haladyna, & Allison, 1998). One study (Hoffman, Assaf et al., 2001) found that

teachers in Texas spent an average of eight to ten hours each week coaching students for the state test. Other studies have shown that students labeled as low-socio-economic status (SES) spend more time preparing for state tests than their more affluent peers (Causey-Bush, 2005; McNeil, 2000; Sheppard, 2002). Darling-Hammond (2011) found teachers working with students of color spent more time preparing their classes for the state assessment than teachers working with white students. These findings suggest that benchmark testing, which takes time away from high-quality teaching and learning, does not affect all students equitably.

Perhaps more troubling given the amount of time spent on them, the use of benchmark tests may not even help to improve student performance on standardized tests. Using a random sample of 41 public school districts in Texas, Nelson et al (2007) found that approximately 63% of districts participated in benchmark testing. The number of benchmark tests varied widely in that sample. Some districts required no benchmark tests, while other districts required students take more than 35 benchmark tests. To explore whether the practice of benchmarking effects accountability test scores, the study employed a multi-variate linear regression model. That study found that benchmark testing had little influence on standardized test scores, regardless of the number of benchmarks given. Specifically, in math they found no benefit to benchmarking and in reading the benefit was small (6%) (Nelson et al., 2007).

Purpose and Significance of the Study

In spite of scant evidence of the effectiveness of benchmark testing, many districts in Texas require some form of benchmark testing. The purpose of this exploratory study is to investigate school district characteristics and the use of benchmark

testing to determine whether there is a predictive relationship between district descriptive factors and benchmark testing and if there is a predictive relationship between the number of benchmark tests a district requires and the percent of students passing the Math and Reading TAKS tests. Specifically, this study will examine whether there is a predictive relationship between district characteristics and benchmark testing requirements. Examining the relationship between the number of benchmark tests in a district and the percent of students passing TAKS will help explain if the practice of benchmark testing is successful in helping students pass the state standardized tests. Exploring this relationship may reveal whether the use of benchmark tests is more prevalent in some kinds of districts than others and whether the use of benchmark tests is related to the organizational theory of isomorphism.

Theoretical Framework

Organizational theory will be used as the frame for the study. The contingency model of organizational theory states that there is no one best way to manage; the best way to manage is dependent on the environment (Martins, 2005; Scott & Mitchell, 1976). One type of contingency model is institutional theory. Institutional theory suggests that by examining organizations at the macro level, one can observe the *institutional rules* that are taken for granted by members of the organization and are not necessarily based on actual evidence (Lammers & Barbour, 2006; Tsouskas & Kunden, 2003). When the *institutional rules* become habitual actions and members perceive them to have value, they become *rational myths* (Burch, 2007; Meyer & Rowan, 1977). There may be no indication these systems improve performance. Often in the place of increased uncertainty, organizations will rely on rational myths for decision making. Dimaggio and

Powell (1983) expanded the ideas of institutional rules and rational myths to create the theory of isomorphism. Organizations watch other organizations in their field responding to the environment and change to adopt their practices (Burch, 2007; DiMaggio & Powell, 1983; Schelling, 1978).

Strategic isomorphism is the specific organizational theory that will be the lens for this research. Strategic isomorphism is the resemblance of an organization's policies to the policies of other organizations in its industry (Heugens & Lander, 2007; Meyer & Rowan, 1977; Abrahamson & Hegeman, 1994; Deephouse, 1996). Organizations tend to mimic other organizations in their discipline that are perceived as successful (Haberberg, 2005; DiMaggio & Powell, 1983). These perceived reasons for success may be based on institutional rules and rational myths that are assumed and not necessarily based on concrete evidence (Lammers & Barbour, 2006; Tsouskas & Snudsen, 2003; Meyer & Rowan, 1977). Strategic isomorphism is more prevalent in times of high stress, such as when organizations are competing for organizational legitimacy (Deephouse & Suchman, 2008; Aldrich & Fiol, 1994; Jepperson, 1991; Deephouse, 1996). Organizational legitimacy can come from government regulators or public opinion (Scott, 2007; Meyer & Scott, 1983). In the age of high-stakes accountability, Texas public school districts are seeking organizational legitimacy from both the government and public opinion. Government agencies bestow school ratings based on high-stakes test scores and have the power to close schools or cut funding. Public opinion and support for schools is heavily influenced by test scores and ratings that are publicized on television news, websites, banners hanging on the schools, and more (Booher-Jennings, 2005). As a result, school districts are increasingly looking to one another for strategies that will lead to success

within the accountability system. Benchmarking is one such strategy that has been adopted by many school districts in recent years. This study will employ the theory of isomorphism in examining the use of benchmark tests in public school districts in Texas.

Research Questions

One assumption guiding this study is that there is a connection between district characteristics found on the Texas Education Agency (TEA) Academic Excellence Indicator System (AEIS) report and the number of benchmark tests they require of their students. A second assumption guiding this study is that there is a relationship between the number of benchmark tests a district requires and the percent of students passing TAKS, the state standardized test. In exploring the practice of benchmark testing in Texas public schools, the following research questions will guide this study:

Primary research questions:

1. Do significant relationships (i.e regression weights) exist between a district's descriptive factors and benchmark testing practices?
2. Does a significant relationship (i.e regression weight) exist between the number of benchmark tests a district requires and the percentage of students passing the TAKS test?

Supporting questions:

1. Is the effect of AYP Status on the number of benchmark tests mediated by the percentage of economically disadvantaged students in the district?
2. Is the effect of AYP Status on the number of benchmark tests mediated by the percentage of students of color in the district?

3. Is the effect of AYP Status on the number of benchmark tests mediated by the percentage of Limited English Proficient (LEP) students in the district?
4. Is the effect of a district's TEA state accountability rating on the number of benchmark tests given mediated by the percentage of economically disadvantaged students in the district?
5. Is the effect of a district's TEA state accountability rating on the number of benchmark tests mediated by the percentage of students of color in the district?
6. Is the effect of a district's TEA state accountability rating on the number of benchmark tests mediated by the percentage of Limited English Proficient (LEP) students in the district?
7. Is the effect of TEA district type on the number of benchmark tests mediated by the percentage of economically disadvantaged students in the district?
8. Is the effect of TEA district type on the number of benchmark tests mediated by the percentage of students of color in the district?
9. Is the effect of TEA district type on the number of benchmark tests mediated by the percentage of Limited English Proficient (LEP) students in the district?
10. Is the effect of TEA district type on the percent of students passing the Math TAKS test mediated by the percentage of economically disadvantaged students, percentage of LEP students, percentage of students of color, and number of benchmarks?
11. Is the effect of TEA district type on the percent of students passing the Reading TAKS test mediated by the percentage of economically disadvantaged students,

percentage of LEP students, percentage of students of color, and number of benchmarks?

12. Is the effect of AYP Status on the percent of students passing the Math TAKS test mediated by the percentage of economically disadvantaged students, percentage of LEP students, percentage of students of color, and number of benchmarks?
13. Is the effect of AYP Status on the percent of students passing the Reading TAKS test mediated by the percentage of economically disadvantaged students, percentage of LEP students, percentage of students of color, and number of benchmarks?
14. Is the effect of a district's TEA state accountability rating on the percent of students passing the Math TAKS test mediated by the percentage of economically disadvantaged students, percentage of LEP students, percentage of students of color, and number of benchmarks?
15. Is the effect of a district's TEA state accountability rating on the percent of students passing the Reading TAKS test mediated by the percentage of economically disadvantaged students, percentage of LEP students, percentage of students of color, and number of benchmarks?
16. What is the relationship between number of benchmarks a district requires and the percent of students passing Math TAKS?
17. What is the relationship between the number of benchmarks a district requires and the percent of students passing Reading TAKS?

18. How does the organizational theory of isomorphism help to explain the use of benchmark testing in school districts and is organizational theory congruent with the path model results observed in this study?

Brief Summary of the Methods

This study will use path analysis, a multivariate regression technique, within a structural equation modeling (SEM) framework to answer the research questions. Path analysis is used to examine linear, causal relationships between observed variables (Kline, 2005; Randolph & Myers, 2013). Path analysis is the appropriate SEM method for this study because relationships between only observed variables are examined. Path analysis allows the researcher to examine indirect, mediated, effects between variables. According to Bohrnsted and Knoke (1994), path analysis is "a statistical method for analyzing quantitative data that yields empirical estimates of the effects of variables in a *hypothesized causal* system" (p. 414). Correlation between a set of independent variables and a dependent variable is a major factor in path analysis (Randolph & Myers, 2013). A path diagram is a pictorial representation of the relationships between variables (Randolph & Myers, 2013). Rectangles represent observed variables, and a circle with an arrow pointing to a dependent variable is the error term (Keith, 2006). A single-headed arrow indicates direction of the relationship, with the variable where the arrow originates is the independent variable and where the arrow terminates is the dependent variable. A mediated, or indirect, effect is facilitated by at least one intervening variable. Double-headed arrows in the model show correlation (Streiner, 2005). McNeil (2000) suggests that student demographics and district size affect the number of benchmark tests required

by the district. Using path analysis, a diagram can be constructed to show those interactions and then test the assumption. Figures 1 and 2 show the path diagrams for the Math TAKS and Reading TAKS models.

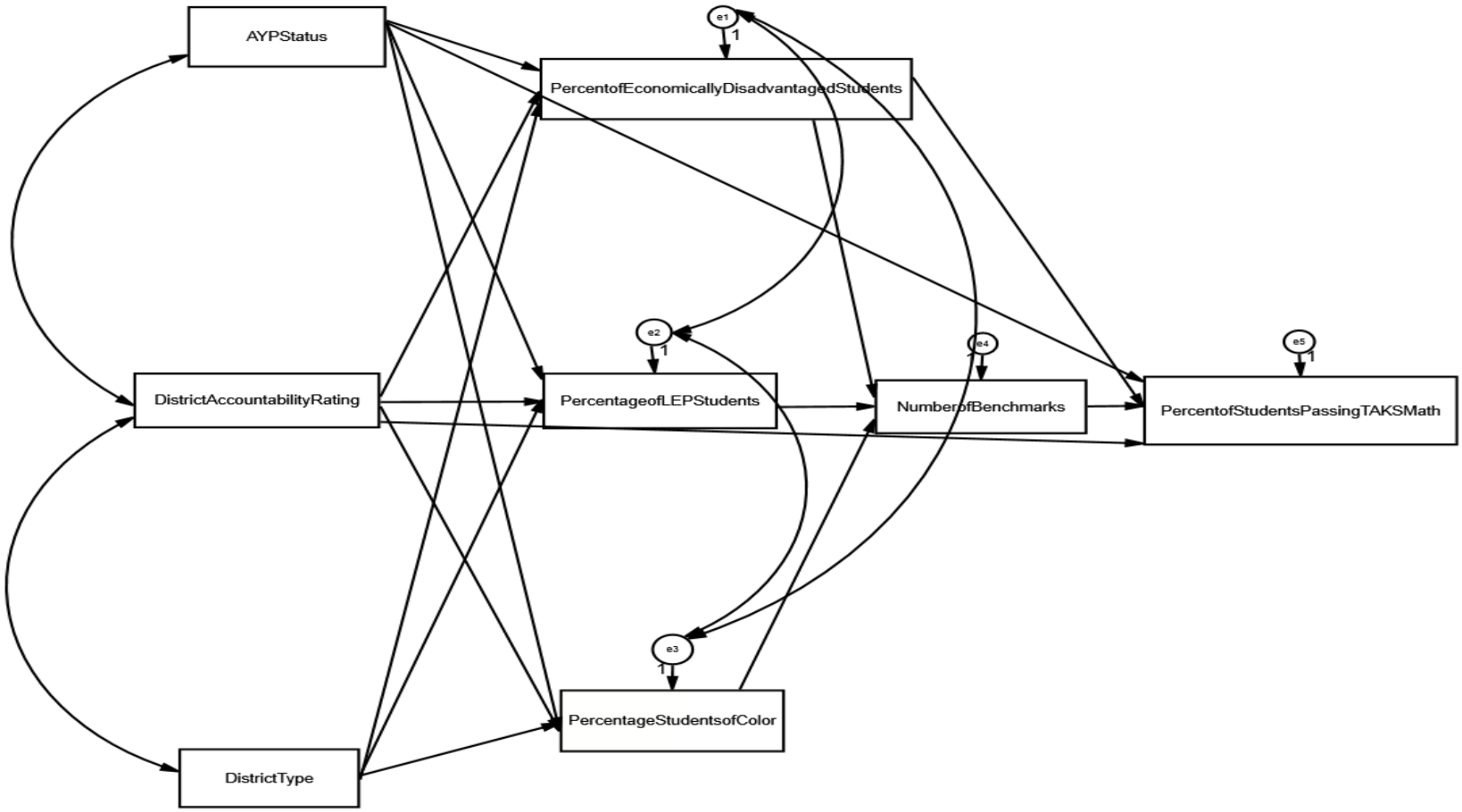

Figure 1. Conceptual Path Model Math. This conceptual path model depicts the structural equation model developed for the study for Math TAKS scores. Arrows drawn between variables specify mediating effects. The rectangles represent observed variables. Arrows point to the dependent variable on the right, the percent of students passing Math TAKS.

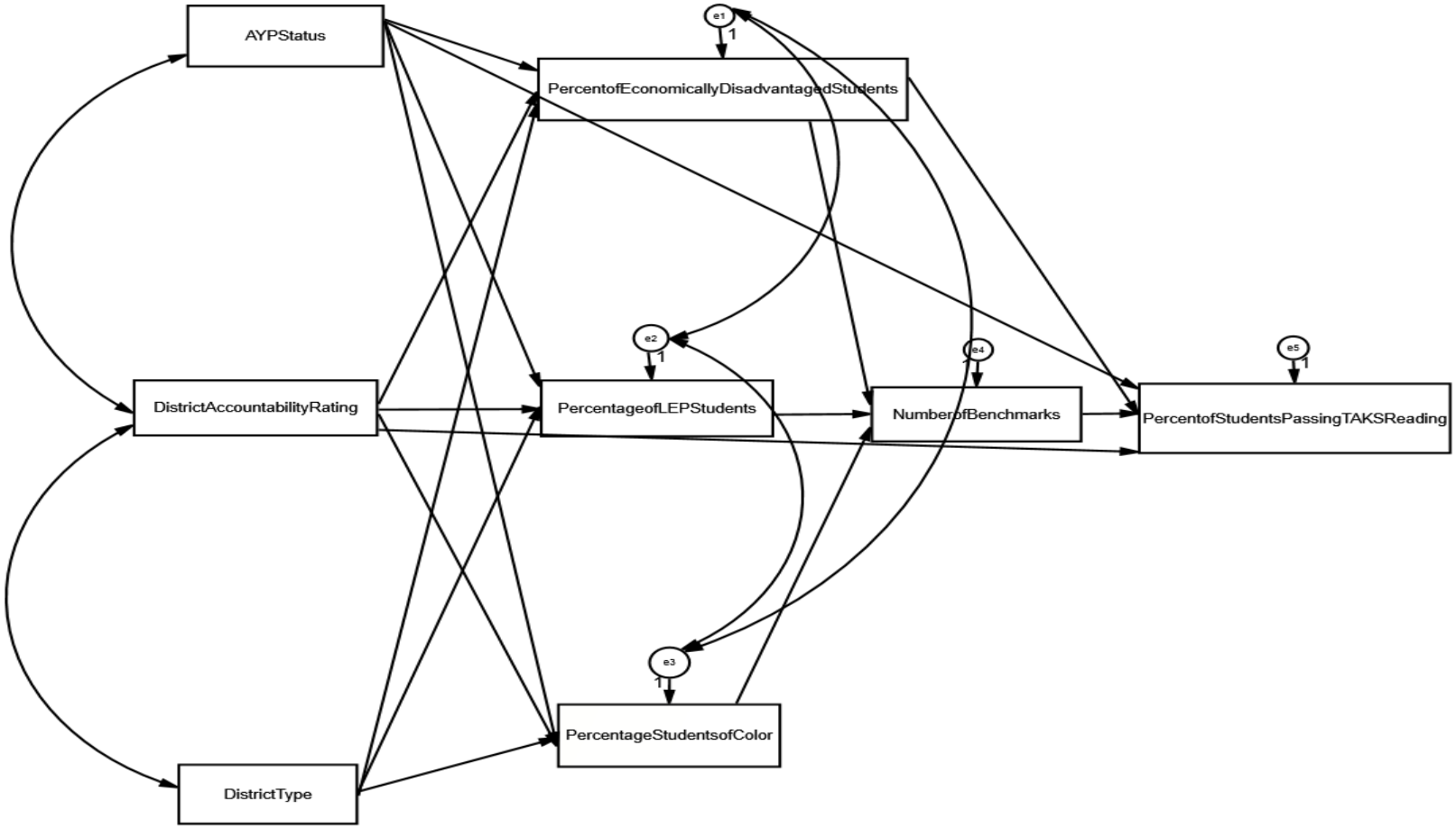

Figure 2. Conceptual Path Model Reading. This conceptual path model depicts the structural equation model developed for the study for Reading TAKS scores. Arrows drawn between variables specify mediating effects. The rectangles represent observed variables. Arrows point to the dependent variable on the right, the percent of students passing Reading TAKS.

Conclusion

Hours and days of instruction are lost each year to benchmark testing. It is important to determine if the benchmark testing makes a significant difference in students' scores on the high-stakes standardized tests. The results of this research can influence district curriculum leaders who determine the number of benchmark tests students take each year. The results of this study will help school district officials reflect on their testing practices. They can consider if their benchmark practices are based on the specific needs of their student population, or if their benchmark testing practices are based on district characteristics they share with other school districts. The scholarly literature illustrates that benchmark testing has seemingly little influence on test scores (Hoffman, et al., 2001; Schmidt, 2009) and that students of color and low SES students spend more time taking benchmark tests to prepare for the state assessment than their white, higher SES peers (Heiling & Darling-Hammond, 2008; McNeil, 2000; Sheppard, 2002). Texas Education Agency (TEA) describes school districts based on their size, location, wealth, and percentage of students identified by different racial/ethnic groups. These district descriptive factors are reported in the Texas Education Agency's annual Academic Excellence Indicator System (AEIS) Reports.

CHAPTER 2

REVIEW OF THE LITERATURE

This chapter explores the history of assessment and curriculum in the United States. The literature review begins with a look the historical development of curriculum in the United States. Curriculum history is relevant to this study as the drive to standardize curriculum lead to tests to determine the effectiveness of that curriculum. Then this review examines the role of curriculum leaders who often make decisions in districts about the number and types of benchmark tests that are required. This literature review describes different types of educational assessment including standardized tests. The use of standardized tests as high-stakes accountability is reviewed. Finally, benchmark tests, as a way to prepare for the high-stakes standardized tests, their history, benefits and costs are examined.

Curriculum

Definition

Curriculum is the "intentional experiences within the school planned for students" (Gress & Purpel, 1988, p. 495). Curriculum takes content from external standards and molds it into a plan for teaching and learning (Wiggins & McTighe, 2005). Schwab (1969) describes curriculum as the ideas of what should be taught, the order in which children can learn those ideas developmentally, and the experiences used for students to associate, organize, and apply those ideas. The history of curriculum in the United States shows how the attempts to develop a standardized curriculum lead to the development of a standardized testing system to measure the effectiveness of curriculum delivery. "As

one begins to see the extent of testing today, what becomes clear is the enormous influence tests have not only in determining the future of students, teachers, administrators, and schools, but also in shaping curriculum and classroom practice" (p. 52, Taubman, 2009).

History

The history of curriculum in the United States can be viewed as four common eras (Janesick, 2003; Kliebard, 1995). The Progressive Era began in the late 1800s and lasted until about 1930. During this time, there was a large growth in the number of students served by the public education system. Beginning around 1920 the Social Efficiency Era began. This era was marked by the use of IQ tests and training students for the workforce. The Technical Scientific Era in education was a way of planning education for the success. The Reconceptualist Era began around 1960 in the United Stated and focused on popularizing educational content for all students. These Eras are not discrete movements with definitive beginning and ending times, but thoughts and ideas that shifted the movement of education reform in the United States.

Progressive era. The Progressive Era of curriculum development in the United States began in the mid-1800s when William Harvey Wells divided students in Chicago public schools into grades and then created a unique course of study for each subject at each grade level (Lieberman, 2005; Tyack, 1974). The 1870s marked the beginning of state compulsory attendance laws and a standardized curriculum was a way of managing the large number of students now in schools (Gress & Purpel, 1988; Williamson, 2008). As a result of Wells' work and varied competing academic philosophies in practice

across the United States, the National Education Association appointed a committee of ten educators, known as the Committee of Ten to establish a standard curriculum.

Convening in 1892, the Committee of Ten built on Wells' work and was charged with developing a plan to standardize high school curriculum (Kelting-Gibson, 2005). This standardized high school curriculum was broken into distinct subjects that intended to prepare high-school students for college (Noddings, 2005; Kliebard, 1995). Some of the committee's suggestions such as 12 years of schooling, 8 elementary and 4 high school, were implemented by school districts across the U.S. (National Education Association, 2009). The Committee of Ten recommended that all students be taught a curriculum that focused on grammar and arithmetic regardless of their future plans (Janesick, 2003).

The Progressive Era of education expanded the ideas of the Report of the Committee of Ten. Progressivism lasted from about 1890-1920 (Gress & Purpel 1988; Janesick, 2003; Gill & Schlossman, 2004). During this era, John Dewey encouraged problem solving in addition to rote lessons in areas of grammar and arithmetic (Gress & Purpel, 1988; Labaree, 2005). Other key ideas of Progressivism were the search for understanding how people know and learn and how to introduce new curriculum ideas to effect educational practice (Bellack, 1972; Labaree, 2005). At the end of this era, Franklin Bobbitt from the University of Chicago published the first general book on curriculum in 1918. His book *The Curriculum*, laid out procedures for curriculum planning consisting of identifying learning objectives and creating experiences to enable students to learn these objectives (Gress and Purpel 1988; Hlebowitsh, 2005). Bobbitt's ideas lead to a shift in thinking about curriculum and education.

Social efficiency era. John Franklin Bobbitt, an American educator who taught in the Philippines, enhanced the standardized curriculum by his book publications: *The elimination of waste in education* (1912); and *How to make a curriculum* (1924), to name a few. Bobbitt believed that education should include classical subjects and those topics that will prepare students for their roles in industrial society. Through the early to mid 1900's, curriculum continued to be shaped by principle educational influences such as John Dewey, Ralph Tyler and Benjamin Bloom, but Fenwick W. English (mid 1970's) was the first educator to introduce the concept of curriculum mapping. Mapping, the idea that college preparatory curriculum should be discipline-oriented and that curriculum planning consists of subject naming, specifying content, and ordering a course of action for instruction survives to this day (McKenney, Nieveen, & van den Akker, 2006; Walker & Soltis, 1986). This pragmatic approach to curriculum development has continued to be validated throughout the years by philosophical educational theorists such as, Madeline Hunter (1970's-1990's), Heidi Jacobs (1997), Wiggins and McTighe (1998) and H. Lynne Erickson (2002).

In 1922 Denver public schools implemented a system-wide program of curriculum revision (Broom, 2011; Caswell, 1988). Also in the early 1920s Winnetka, Illinios shifted their curriculum to focusing on individual student progress (Broom, 2011; Caswell, 1966; Spring, 2008). St. Louis also began change in the 1920s. In 1925 St. Louis began a 2-year effort to create new courses of study for all elementary and secondary school subjects (Broom, 2011; Caswell, 1988). These changes in the Progressive Era moved the definition of curriculum in U.S. public schools. According to Caswell (1988), "the traditional concept of the curriculum as consisting of a group of

courses of study, leaders of state programs came to view the curriculum operationally, considering it to be composed of the experiences pupils actually had under the guidance of the school" (p. 33). This moved curriculum from not only *what* is taught, but *how* it is taught. Another shift in U.S. curriculum is attributed to the Soviet Union.

Technical scientific era. Progressivism gave way to the Technical Scientific Era in education from about 1920-1950; a key reason for this shift was World Wars I and II (Caswell, 1966; Gress & Purpel, 1988; Janesick, 2003). This was a more basic, technocratic period than the Progressive Era, with curriculum based on grammar and arithmetic (Goodson, 2005; Janesick, 2003). This was a more traditional view that worked under the assumption that "all children can learn all data at the same time" (Janesick, 2003, p. 95). These ideas led to an organized curriculum movement and "made the curriculum a field of established professional performance" (Bellack, 1972, p. 253). Caswell (1966) found three central concerns of curriculum specialists that began in the Technical Scientific Era and continue, 1) assure continuity in the curriculum, 2) establish relationships between general goals of education and specific objectives to guide teaching, and 3) design curricula that balance emphasis in various areas of study. Kliebard (1968) also views this era as the beginning of the professional curriculum movement when efficiency was most important. Curriculum specialists at this time were trying to be efficient in educating large numbers of students while wars were disrupting the country. Two dichotomies emerged at this time: 1) the school subject dichotomy – academic vs. practical, and 2) the school population dichotomy – college preparatory vs. non-college preparatory (Goodson, 2005; Kliebard, 1968). Testing, as discussed above was one method for navigating these dichotomies.

Some educational theorists believe the task of the school was to convey a prearranged body of subject matter rooted in idealist and/or realist views of knowledge (Dittmar 1993; Popa, 2009). "Adding to the philosophical justification for curriculum according to individual subject areas were those educational philosophers who favored technical approaches to curriculum development," (Kelting-Gibson, 2005, p. 27). The technical-scientific approach is a way of thinking and planning curricula in subject areas to optimize student learning. Since the 1920s, curriculum has been developed using the technical-scientific approach (Kelting-Gibson, 2005). Models of technical-scientific curriculum planning developed in the 1950s and 1960s have directed curriculum developers and teacher in planning curricula for years and continue to present day (Tyler, 1950; Taba, 1962). These models follow four similar steps: 1) define goals or objectives, 2) define experiences related to the goals, 3) organize the experiences, and 4) evaluate the goals (Kelting-Gibson, 2005).

The 1950s in U.S. education were marked by the launch of Sputnik. Sputnik I, the first Earth-orbiting satellite, was launched October 4, 1957 by the Soviet Union. "During the post-Sputnik fifties, energies were mobilized to redress presumed failures of public schooling in the areas of science and technology" (Alexander & Pallas, 1984, p. 391). Because of this, the National Defense Education Act (NDEA) was passed in 1958 by the United States Congress. There were two main parts to the act. One was to provide financial assistance for students attending college. The second main section provided funding to all levels of U.S. education for increases in foreign language, math, and science curriculum (Davies, 2007; Gress & Purpel, 1988). Curriculum in U.S. schools changed to focus on foreign language, math, and science due to NDEA. Another change

in public school curriculum due to Sputnik was the increase in homework for students (Gill & Schlossman, 2004; Urban 2010). Both parents and educators agreed that more homework would better prepare students academically. During the late 1950s and 1960s fear of U.S. schools' inferiority brought on by Sputnik influenced curriculum.

Reconceptualist era. The 1960s-1970s in U.S. education was marked by change. Changes in the civil rights movement, feminist movement, and gay rights movement altered the way people thought about the world and lead to the Reconceptualist Era in education (Janesick, 2003; Spring, 2008). According to Alexander and Pallas (1984), "throughout the sixties and much of the seventies, equity issues were dominant" (p. 391). The main concern was to guarantee equal educational opportunities to racial and ethnic minorities. There is a "continuity between the curriculum movement of the 1960s and its counterpart during the Progressive Era, in that both aimed ultimately at humanizing knowledge so that it could be popularized" (Cremin, 1975, p. 26). Popularizing knowledge meant it was for all students, not a select group. This movement continued to the 1980s when A Nation at Risk was published.

In 1983 the National Commission for Excellence in Education published A Nation at Risk. This report touted the failings of U.S. schools (1983). These failings included declining test scores, proliferation of remedial mathematics courses at the postsecondary level, and high levels of functional and scientific illiteracy (Alexander & Pallas, 1984; Davies, 2007). A few responses to A Nation at Risk include: the Pennsylvania State Board of Education tripled the amount of science and mathematics required for graduation, the Ypsilanit, Michigan school board lengthened the school day for elementary students and increased high school graduation requirements, the

superintendent in Tulsa, Oklahoma outlined the standing of schools in that district compared to the recommendations in A Nation at Risk (Goldberg & Harvey, 1983; Rothstein, 2008). These and other changes came about quickly due to the severe nature of the report. The report stated that the poor quality of education in the United States would jeopardize America's position in the international economic order (Alexander and Pallas 1984; Rothstein, 2008). "Not since the heady days following the launching of Sputnik I has U.S. education been accorded so much attention" (Goldberg & Harvey, 1983, p. 14). According to Gill and Schlossman (2004), curriculum and education changes in the 1980s were motivated by fear of economic competition from around the world in the same way changes in the 1950s were spurred by the fear of Soviet dominance with Sputnik. These fears lead to the desire to closely measure and monitor student success and failures using standardized tests. This focus on curriculum tied to measurable tests continues today.

School District Curriculum Leaders

The focus on curriculum in U.S. school districts has led to the creation of curriculum directors, administrators responsible for overseeing curriculum. A curriculum director is a district level administrator responsible for curriculum and instruction (Hamm, 1993; Honig & Coburn, 2008). There are differing views of a curriculum director's role in the literature. Some see curriculum leaders as middle management (Garman, 2006). They must navigate between superintendents, principals, and teachers, similar to mid-level management in business (Wraga, 2006). Often, curriculum leaders must communicate directives to principals and teachers. They are often in the middle of district mandates, principals, teachers, and the community (Garman, 2006).

Another view of a curriculum director's role is that of a scholarly leader. "Curriculum leaders who are public intellectuals must possess the capacity to identify, analyze, and resolve public curriculum problems" (Wraga, 2006, p. 83). These individuals "seek to create conditions that will improve learning" (Gress & Purpel, 1988, p. 30). Curriculum directors must be able to communicate with a variety of groups, such as the superintendent, school board, principals, teachers, and parents on issues of curriculum and instruction (Hamm, 1993; Honig & Coburn, 2008). Curriculum practices must be enacted to suit local circumstances, such as state curriculum standards. Leaders must understand concepts such as the official curriculum, written curriculum, taught curriculum, and hidden curriculum (Glatthorn, Boschee, & Whitehead, 2005; Wraga, 2006). The official and written curriculums of schools are explicit, mandated curriculum (Au, 2010; Bloom, 1971; Giroux, 1978). The taught curriculum is what is actually taught directly to students (Wiggins & McTighe, 2005). The taught curriculum may or may not match the official and written curriculums. The hidden curriculum refers to unstated norms, values, and beliefs transmitted to students (Freire, 1973; Glatthorn, Boschee, & Whitehead, 2005). A curriculum leader must understand these various facets to curriculum to ensure what is being taught meets state and federal curriculum guidelines and the hidden curriculum does not interfere with teaching and learning. The responsibilities of curriculum leaders may vary in different school districts in order to meet the variety of needs in different areas.

Roles and Responsibilities

It has been difficult to define a clear list of curriculum directors' duties and responsibilities with any specificity (Babcock, 1965; Eye, Netzer et al., 1971; Honig,

2006). States and school districts view the position differently. According to Garman (2006), pre-NCLB curriculum leaders concentrated on aligning goals, standards, and evaluation. Archbald and Porter (1990) state, "As state authorities in the late 70's and early 80's became more involved in school reform and as higher test scores became increasingly the avowed goal of reform, attention to curriculum alignment grew (p.26)." As high-stakes testing became the focus with NCLB, "alignment takes on a new meaning – aligning high stakes test scores to teacher performance evaluation as well as school performance assessments" (Archbald & Porter, 1990, p. 74). Rogers (1999) asserts that though curriculum decisions are made using a variety of sources, a primary source is standardized tests mandated under NCLB. The high stakes tests and students' scores on the tests are key elements curriculum directors use for curriculum alignment under NCLB (Honig, 2006). "Using such tests to make choices about curriculum often means 'aligning' classroom curriculum with the skills and content to be tested" (Rogers, 1999). Simply stated in the words of Archbald and Porter (1990), "If the goal is to improve test scores, then instruction should focus on what is tested" (p. 26). As the instructional focus shifted with NCLB to improving test scores, so did the role of the curriculum leader.

The role of the curriculum leader it to build curriculum for the district based on the needs of the teachers and students. The development of curriculum usually involves preparing, organizing, applying, and assessing curriculum (Wraga, 2006). Curriculum leaders must navigate a landscape of state and district policy while concentrating on best educational practices in order to direct curriculum focus in their districts. The curriculum director coordinates all phases of curriculum development including selecting materials appropriate for students' use and evaluating the effectiveness of instructional programs

(Babcock, 1965; Honig & Coburn, 2008). Some of these instructional programs are commercially available curriculums are marketed to help districts align learning, but are not oriented toward any particular district or state objectives (Archbald, 1990; Honig, 2006). Curriculum leaders often grasp best practices of teaching and learning, but are pulled in two directions when considering the many decisions they are making and the impact of high stakes testing on what is taught.

Directors must reconcile the high stakes testing curriculum as well as state and national curriculum standards. Curriculum directors navigate federal and state policies that require evidence, data, and research used for decision making be documented (Honig & Coburn, 2008). NCLB (2001) requires that all programs funded with this initiative are "scientifically based" and "data-driven." Districts gather program evaluation data, school and student performance data, and school improvement plans to meet policy and funding requirements (Honig & Coburn, 2008). The work of curriculum directors includes teaching and learning along with navigating standardized testing and policy issues for the school district.

Educational Assessment

Purpose

Assessment is the systematic method of gathering information from tests and other sources to draw inferences about people, objects and programs and using the information to aid educational decision making to determine if students learn what is expected (Joint Committee on Standards for Educational and Psychological Testing of the AERA, APA, and NCME, 1999).

The term assessment is also used to describe a conclusion that can be validated according to specific objectives (Scriven, 1967; Taras, 2005). And while there is agreement among educators about the concept of assessment, exact definitions of terms related to assessment vary.

At recent meetings of the American Educational Research Association assessment researchers were brought together to find common definitions of the term assessment. Their task was to find common definitions that could be used throughout the United States to facilitate discussions among educators. They were unable to reach agreement on meanings of central assessment terms such as performance assessment and authentic assessment (Frey & Schmitt, 2007). As a result of this lack of agreement, scholars provide their own definitions of key assessment vocabulary. According to Newton (2007), the purpose of assessment can be interpreted in different ways – at the judgment, decision, and impact levels. Newton's purposes for assessment are used in this study because they describe the work of curriculum directors who make decisions that impact numerous students and teachers. The judgment level is concerned with the (Delandshere, 2001; Long, Wood, Littleton, Passenger, & Sheehy, 2010) technical aim of the assessment, such as to find a standards-referenced judgment (Newton, 2007). The second level is the decision level when the assessment's purpose is to support a decision or action, such as the entry to a college or university (Newton, 2007). The third level is the impact level which concerns the impacts of running an assessment system (Newton, 2007). Examples of this third level are students remain motivated, or students learn a common curriculum for each subject. If the different meanings of assessment purpose are not distinguished clearly, any debate will be unfocused and ineffective (Newton, 2007).

Though definitions of assessment terms vary, assessment purposes have not significantly changed for the better part of the twentieth century. "Assessment is mainly used for placement, selection, and certification decisions, based on measures of what individuals know" (Delandshere, 2001, p. 114). For example, teachers use assessment to determine students' reading levels. Students are selected and placed in groups for reading instruction based on assessment results. Students are placed in groups for remedial work, such as in after school tutorials, based on their assessment scores. In the last 50 years policy makers who are dissatisfied with education have increasingly demanded educational testing for placement, selection, and certification of students and programs (Delandshere, 2001; Long, Wood, et al, 2010). Policy makers seem to view testing as the one way to make decisions about student placement. Delandshere (2001) suggests assessment is viewed as a technology developed by experts that others can use to make specific decisions about students and programs.

The decisions based on assessment scores are often shared by the state to the public regarding fair educational opportunities for all students (Au, 2007; Delandshere, 2001). Test results are made public to the schools, communities, and families as a way of sharing information about student, teacher, school, and district achievement. These results may be used for instructional and political decisions, such as funding or sanctions. This is one way political decisions drive state assessment systems. Many state assessment practices are based on the assumption that a summary measure of students' achievement is a significant indicator for the educational possibilities in a school (Delandshere, 2001; Watson & Robbins, 2008). There are different types of student assessments that fall under the purpose of measuring information about schools to share with the public.

Types of Assessments

There are two main types of assessments used in classrooms and schools. One type is formative assessment, "a systematic process to continuously gather evidence about learning" (Heritage, 2007, p. 141). Michael Scriven (1967) and Benjamin Bloom (1969) were early proponents of formative assessment as a method of on-going improvement in teaching and learning (Wiliam, 2006). A crucial feature of formative assessment is that the information is used to make changes with the intent of benefiting the student (Scriven, 1967; Popham, 2006). A teacher identifies a student's needs, provides feedback, and works with a student to create future goals with formative assessment (Heritage, 2007). Formative assessment is sometimes known as assessment *for* learning because the goal of the assessment is to guide student learning (Stiggins, 2002; Taras, 2008). According to Sadler (1989), formative assessment is used to shape and improve students' competence. Often formative assessment occurs in the daily processes and flow of the classroom. It does not need to be formal or done with paper and pencil; formative assessment includes teacher observation, oral response, and student performance (Stiggins, 2002). Formative assessment provides the teacher with information that allows her to adapt teaching to meet students' needs (Newton, 2007). After formative assessment and learning has taken place, summative assessment occurs.

The second primary type of assessment is summative assessment. According to Taras (2005), summative assessment is "a judgment which encapsulates all the evidence up to a given point. This point is seen as finality at the point of the judgment" (Taras, 2005, p. 468). Summative assessment is also known as assessment *of* learning because it occurs after learning (Stiggins, 2002). This type of assessment is often formal, paper and

pencil assessment. Most standardized tests and unit tests fall into this category because they assess learning after the learning occurs in a formal assessment format. Both formative and summative assessments are necessary in the progression of teaching and learning as they collect information at different points in the learning process.

According to Stiggins (2002; Stiggins & DuFour, 2009) both formative and summative assessments are vital in education because they serve unique purposes. Scriven (1967) sees formative and summative assessments as having different roles, but believes one directly leads to the other. Most often, formative assessment precedes summative assessment. In contrast, Sadler (1989) presents formative and summative assessments as having separate practices and values that are not always connected. Whether the practices are connected or not, many scholars believe a balance between formative and summative assessments would meet the needs of students and create the greatest opportunities for learning (Sadler, 1989; Stiggins, 2002; Taras, 2008). Despite scholars' beliefs, however, Stiggins and Chappuis (2005) explain that interest in summative assessment has far outweighed formative assessment, as these tests are being used for classroom grading as well as state testing for accountability. Stiggins and Chappuis state, "The demands of No Child Left Behind have intensified the use and attention given to summative assessment because states are required to articulate their achievement standards and report annual evidence of the proportion of students meeting those standards." (2005). This type of summative assessment, standardized testing, is at the forefront of the education debate in the United States due to government policies that have mandated their use.

Standardized Testing

Defining Standardized Testing

There are a variety of definitions regarding standardized tests in the literature. According to Popham (1999, p. 8-9), "a standardized test is any examination that's administered and scored in a predetermined, standardized manner." Another definition of a standardized test comes from Haladyna, et al. (1998) that states it is designed to provide norm-referenced analysis of student achievement in particular content areas. Standardized tests compare a student's score to a predetermined, normative score (Duckworth, Quinn & Tsukayama, 2012). There are two major types of standardized tests: aptitude tests and achievement tests. The SAT and ACT are examples of standardized aptitude tests that attempt to predict how students will achieve in some subsequent educational venue (Popham, 1999). Standardized achievement tests attempt to make a statement about an individual student's knowledge or skills in a particular content area (Duckworth, et al. 2012; Popham, 1999). Since the test can measure an individual student's achievement in a certain area, the definition of standardized tests can also include which groups of learners are required to take the tests.

Another description of standardized tests comes from Wang, Beckett, and Brown (2006), who maintain that a standardized test is one that "a) is externally imposed by the state government; b) assesses state-prescribed content standards; c) follows a uniform procedure in administering, scoring, and interpreting the test; and d) the results are often used to determine rewards and sanctions for students, teachers, schools, or districts" (p. 307). This definition explicitly incorporates government standards as part of the criteria for a standardized test. Standards-based assessment, a type of standardized test, relates a

student's achievement to a prescribed set of content standards and not to a norm group of peer students (Wang, et al. 2006). This is in contrast to Haladyna and Haas (1998) who state that a standardized achievement test is designed to present norm-referenced interpretations of how a student achieves compared to other students in the nation. For this study, a combination of the above definitions will be used. A standardized test is a) an externally imposed test, b) uniformly administered, scored, and interpreted, c) used to relate a student's knowledge against state content standards, and d) the results are used to reward or sanction students, teachers, schools, or districts. The standardized test used in this study is the Texas Assessment of Knowledge and Skills (TAKS) test. TAKS meets the criteria set out in the definition above: a) TAKS is imposed by the State of Texas' Education Agency, b) it is uniformly administered and scored, c) relates a student's knowledge against the Texas Essential Knowledge and Skills (TEKS) state curriculum standards, and d) the results of TAKS are used to reward or sanction students, teachers, schools, and/or districts.

When standardized tests compare an individual student to a larger group and are used to determine the quality of schools, promotion to the next grade, or the governance of a school, they are known as high-stakes tests (Gunzenhauser, 2003; Madaus, Russell & Higgins, 2009). While standardized tests have been a part of public education in the United States for more than a century, they have not always been of the high-stakes nature as they are now.

History of the Use of Standardized Testing

For more than 150 years, attempts have been made to regulate the way students are educated and measured. In the 1840s, Horace Mann changed the way students were

assessed in the United States with standardized tests. In contrast to the oral exams widely used at the time, Mann led reform with the creation of the Boston Survey, a 1-hour written exam given simultaneously to 7,000 students, based on information in student textbooks (Crocker, 2003; Gallagher, 2003). The written exam stressed application of facts, not simple recall. The results of the Boston Survey were presented to the Boston School Board as an instrument to improve educational quality. The Boston School Board rank ordered schools based on the results. The Smith School in Boston, serving primarily children of freed slaves, was at the bottom of the list and the school board believed the students' failure the fault of the teacher (Crocker, 2003; Ladson-Billings, 2006). These standardized tests and school rankings came at a time in American history when the focus was shifting from educating the elite to educating the masses (Haladyna, Haas et al., 1998; Spring, 2008). There were now many more students to be educated and managed and standardized tests played an important role in this shift.

Between 1890 and 1918 the high school student population grew at a rate more than 10 times the growth of the population of the United States (Linn, 2001; Spring, 2008). Testing was a way to manage the large influx of students into the education system. Columbia University professor E. L. Thorndike experimented with quantifiable scales, objective tests, and efficient surveys in the 1890s to measure a student's ability. The attention shifted from testing a student's knowledge to testing a student's ability based on Thorndike's work. This information could be used to segregate students according to their potential for academic success (Haladyna, Haas et al., 1998; Sears, 2007). Schools in Pennsylvania, New Jersey, New York, Massachusetts, Michigan,

Kansas, and California began using Dr. Thorndike's measurement tools in their public schools from 1900 to 1910 (Gallagher, 2003; Sears, 2007).

The intelligence test is one specific type of ability test developed by French researcher Alfred Binet, the Binet Scale, in the early 1900s was used to individually test children to determine "slow children who would not profit significantly from schooling" (Walsh & Betz, 1995). H. H. Goddard brought Binet's model from France to the United States in 1911 and worked to convince public school officials to use intelligence testing when making decisions regarding individual students (Gallagher, 2003; Kamphaus, Winsor, Rowe & Kim, 2005). In 1912 William Stern created the modern IQ formula based on Binet's model, making it easier to compare students' intelligence test results. These positivistic ideas of measuring mental capacity dominated thought in U.S. public schools. Positivism is the belief that there is a real world with verifiable patterns that can be observed and measured (Kamphaus, , et al, 2005; Patton, 2002). Lewis Terman, a positivist, of Stanford University revised the IQ test and renamed it the Standford-Binet Test of Intelligence in 1916. Its use was expanded from simply identifying "feebleminded" students and was used for educational placement and career tracking in U.S. schools (Kamphaus, , et al, 2005; Terman, 1919). Educational tracking was one way to manage the large number of students being required to attend school.

In the 1920s, school compulsory attendance laws were implemented, forcing a great number of students into public schools, and the United States' education system looked for an efficient way to educate the large and diverse student population (Solley, 2007; Stiggins, 1991). Arthur Otis and Robert Yerkes developed the Army Alpha Test, an efficient paper-and-pencil, multiple-choice test to measure soldiers' mental abilities

during World War I (Gallagher, 2003; Merenda, 2005). This design was viewed as the most effective way to test large groups of people, and has been used in almost all later standardized tests (Rothman, 1995; Solley, 2007). Terman transformed the standardized Army Alpha into the National Intelligence Tests for schoolchildren in 1919 and more than 400,000 copies were sold the in the following year (Merenda, 2005; Terman, 1919). The drive for greater efficiency in assessing large numbers of students pulled schools away from essay tests to more cost-effective multiple-choice tests that could offer information about a large number of students for a small charge (Haladyna, Haas et al., 1998). Standardized test scores were used to track students of perceived differing capacities into different academic tracks, and thereby restrict their social and curricular choices (Solley, 2007; Zanderland, 1998). "As well-intentioned as some motivations for the IQ and other tests may have been, they were not actually measures of innate ability, and their use sometimes caused harm" (Heubert & Hauser, 1999, p. 32). When these tests are used to limit a student's access to academic choices or curriculm, they can be detrimental. Often this damage to students was in practices that sorted and labeled them on the basis of one test measure. The prevalence of standardized scores were understandable to researchers, but teachers and parents needed help understanding the growing number of standardized assessments.

Monroe, DeVoss, and Kelly wrote *Educational Tests and Measurements* in 1917 to help parents and teachers understand standardized tests. Their purpose was to educate teachers about different types of assessments. They discuss the limitations of written essay tests due to the subjective nature of scoring. In contrast, the authors believe standardized tests "are ways of securing more accurate measures of the achievements of

school children" (Monroe, DeVoss, & Kelly, 1917, p. 11). Standardized tests are more accurate, according to Monroe, DeVoss, and Kelly (1917), because they are constructed to eliminate or reduce the bias of teacher-developed tests. The tests are more objective because "there is uniformity in the administration of the test and also in the basis of interpreting the measure which it yields" (Monroe, et al. 1917, p. 12). The authors do not believe standardized tests to be perfect, but more reliable than almost all teacher-developed written tests. It was believed that a collection of standardized assessments provide a more complete view of a student as opposed to one test measure.

One such assessment is the Stanford Achievement Tests, a collection of tests in different content areas for elementary students, that were first published in 1923 (Domino & Domino, 2006; Gallagher, 2003). The results identified students who had learned content from those who had not (Thorndike & Bergman, 1934). The scores were also used to sort schools based on instructional effectiveness. The University of Iowa created the Iowa Test of Basic Skills (ITBS) and the Iowa Test of Educational Development (ITED), standardized achievement tests. The ITBS and ITED were developed in response to the popularity of a statewide scholastic competition sponsored by the State University of Iowa (Domino & Domino, 2006; Peterson, 1983). They were the first set of tests administered statewide on a voluntary basis to public school students in grades 1-8; for more than 50 years they were the most often used commercially available achievement tests in the United States (Jeynes, 2007; Peterson, 1983). The use of standardized tests in public schools to measure a student's ability and what she had learned led educators to question whether standardized tests could also be used to predict how well a student

would function in college. Groups of educators were concerned that educational opportunities were available to students who were likely to be successful in college.

College admissions groups formed the College Entrance Examination Board (CEEB) in 1923 to create an examination based on common admission standards to address concerns in the admission process. The CEEB was to address the concern that equal education opportunity was available to all students, and the right opportunities were available to the appropriate individuals (Gallagher, 2003; Jeynes, 2007). The test that was created was an intelligence test, designed to ensure students of high intelligence had opportunities to attend college. The CEEB test was refined in 1925 and renamed the Scholastic Aptitude Test (SAT). In 1990 the SAT name was changed to Scholastic Assessment Test due to uncertainty that the test functioned as an intelligence test. Finally, in 1993, the name was changed to SAT Reasoning Test, with the letters SAT not standing for anything. This test then began to define the content of college preparatory instruction because it was used by almost all colleges and universities in the United States when considering which students would be admitted to college (Jeynes, 2007; Walsh & Betz, 1995). Standardized tests were being used widely in K-12 schools and for college entrance in the twentieth century; if they were biased or inequitable, many students would unfairly be denied educational opportunities.

Equity and Assessment

The Civil Rights movement during the mid-1960s heightened awareness of testing inequities, denied educational opportunities, and unfair testing practices along lines of social class and cultural background (Hall, 2005; Sacks, 1999). Testing inequities could be seen with IQ and other tests that were used by Southern schools to segregate African

Americans and other groups into lower educational tracks (Heubert & Hauser, 1999; Magnuson & Valdfogel, 2008). Use of these tests yielded unfair treatment of students of certain social-economic backgrounds. For example, often the results of these standardized tests were used to justify segregation. The Elementary and Secondary Education Act (Title 1) passed in 1965 mandated schools to administer standardized tests and present results to qualify for federal funds in subsequent years (Chapman, 1988; Magnuson & Valdfogel, 2008). Thus, standardized testing was widespread in the 1960s. The intent of the law was to ensure equity by having all students take the same assessment. In 1962 three-fourths of all high school students in the United States took standardized tests (Hawes, 1964; Spring, 2008). One hundred forty-three million standardized test booklets and answer documents were sold in 1962 to public school systems; this total is "several times the total number of persons tested, because people are most often given a combination of several tests at one time" (Hawes, 1964, p. 3). In 1964 Texas, California, New York, Iowa, Florida, Minnesota, and Virginia tested all students in certain grades annually with standardized tests (Hawes, 1964; Spring, 2008). In 1969 the United States government expanded the National Assessment of Educational Progress (NAEP), a national standardized test, to test samples of students in all subject areas from various states. This test was to gauge national achievement of students and schools and was nicknamed the "Nation's Report Card" because it compared state and district performances in almost every state and established a national score used for international ratings (Gallagher, 2003; Yeager, 2007). By having students in every state take the same standardized assessment, the government was trying to preserve educational opportunities equal because these scores allowed comparisons across state assessments. Standardized

assessments were used throughout the U.S. If the assessments were biased, then how accurate were they to measure student success, and who is responsible for student achievement?

Accountability as Student Achievement

According to Stiggins (1991), in the late 1960s the focus began to shift from the idea that schools were accountable for only presenting educational opportunities to the notion that schools were accountable for a student's attainment of educational results. Equity was not simply providing opportunities for education, but helping students attain educational goals. It was not enough that teachers taught the curriculum – students had to learn. The National Center of Education Statistics commissioned the Coleman Report in 1966 to study equity issues in standardized testing. One discovery of this report said a student's home background and neighborhood factors were the most important predictor of school achievement (Berliner & Biddle, 1995; Ladson-Billings, 2006). These claims were later found invalid due to data analysis flaws, but for many years groups used these findings to their advantage. Some groups believed this proved intelligence was fixed and schools have little impact on academic achievement. Other groups claimed the Coleman Report proved there were no design flaws in standardized tests because home environments were responsible for the lower test scores of some student groups (Gallagher, 2003; Ladson-Billings, 2006). As a result of the Coleman Report, schools began to take responsibility for students' academic achievement by setting standards for what is taught.

During the 1970s and 1980s educational accountability grew and the index for measuring schools' success in teaching students was standardized test scores (Stiggins,

1991). If standardized test scores were the measure of schools' accountability, standards of what should be taught and learned were necessary. A movement towards defining uniform, high standards began with mathematics teachers (Borman, 2005; Jennings, 1998). In the late 1980s the National Council of Teachers of Mathematics (NCTM) began to develop standards for mathematics education (Zemelman, Daniels, & Hyde, 2005). Also in the early 1980s, the state of California developed curricular frames for basic subjects (Borman, 2005; Jennings, 1998). These frames were standards of what would be taught throughout California in basic subject areas. This major shift to creating uniformity in what is taught "is generally labeled standards-based reform" (Jennings, 1998, p. 6). The standards-based reform movement focuses on the idea that "setting clear, high standards for what children are supposed to learn and then holding students – and often educators and schools – to those standards" (Heubert & Hauser, 1999, p. 13). In 1994 Congress passed a law to begin the process of setting national content standards (Borman, 2005; Ravitch, 1996). Proponents of national content standards believe if educators do not identify what students should learn, the decision will be "left to the marketplace – textbook publishers, testmakers, and interest groups" (Ravitch, 1996, p. 134). The reform focused on outcomes-based education to ensure equity in learning for all students. Setting high standards would ensure that the outcomes of all students' learning should be the same. Teachers will know the standards they are to teach from kindergarten through high school graduation (Edwards, 2006; Jennings, 1998). The standards movement had to be measured if leaders were to know if it was successful.

Test scores became a school's critical weapon in protecting against loss of students, programs, and funding (Sacks, 1999). Federal policy mandated that schools use

standardized test scores in connection with receiving federal funds. Schools and districts use test scores to maintain programs and funding. Congress changed the design of Title 1 in 1974 so progress goals were measured and schools funded using standardized test scores (Gallagher, 2003). Standardized tests went hand-in-hand in measuring the effectiveness of teaching and learning in an atmosphere of standards-based curriculum. When the test scores did not show successful student learning, the National Commission on Excellence in Education commissioned a study. This group was created by the U.S. Secretary of Education and they created the report, A Nation At Risk, on the quality of education in the U.S. The group was made up of university faculty, district superintendents, principals, and others in education.

In 1983 A Nation at Risk was released by the National Commission on Excellence in Education and warned "the educational foundations of our society are presently being eroded by a rising tide of mediocrity that threatens our very future as a Nation and a people" (1983, p. 5). Based on this report, federal and state policymakers sought ways to increase student achievement and reform and fragmented educational system (Croninger, et al., 2003). The debate that followed A Nation at Risk shifted the focus of federal and state education policies from looking at inputs, such as providing schools with more resources to address inequalities, to looking at outputs and performance-based accountability (Goetz, 2001). The report recommended increasing statewide standardized testing programs to monitor student outputs. By 1989, 47 states had implemented policies that would expand testing from their previous levels in the early 1980s. The U.S. government seemed to believe standardized tests were the way to measure a school's success and states complied.

In 1989, President George H. W. Bush convened U.S. governors to discuss education (Jennings, 1998; Naftali, 2007). President George H. W. Bush and the governors adopted six national education goals in 1990. The goals encouraged the expansion of more sophisticated standardized tests to accurately describe student achievement amidst concerns of educational quality. Standardized tests were also a prominent part of President Bill Clinton's 1994 Goals 2000: Educate America Act. Test-based graduation requirements were adopted by more than 35 states (Gallagher, 2003). These national education talks, led by U.S. presidents, moved testing from the state to the national level. The tide began to shift from states creating their own testing policies, to the U.S. government setting laws to regulate standardized tests.

National government regulation of testing in U.S. schools marked a change in thinking from the 1800s. At the beginning of the 20th century the goal of testing was to differentiate the large masses from small group of elite students; as the 21st century began the shift was to creating high educational standards for all students (Linn, 2001). National legislation was a sign of the shift of thinking about testing. In 2001 President George W. Bush's Elementary and Secondary Education Act, or No Child Left Behind (NCLB, 2001) passed into law, mandating annual testing of students in grades 3 through 10. According to this law, states must put into action an accountability system to make certain all districts demonstrate adequate yearly progress in achievement (2001).

Before NCLB (2001), all most Americans expected of its schools were compulsory attendance, access to high-quality programs for all students, and high academic achievement for all students. (Valli, et al. 2008). NCLB added accountability to the list of expectations (Valli, et al., 2008). Federal legislation now held "schools, local

education agencies, and States accountable for improving the academic achievement of all students, and identifying and turning around low-performing schools" (NCLB, 2001). Low-performing schools that do not show progress may face actions such as replacement of school staff or school restructuring. Most states use standardized tests to measure a school's performance. NCLB does not require standardized testing; it calls for academic assessment that is valid and reliable (2001). The standardized tests used by most states for complying with NCLB are referred to as high-stakes tests because of the consequences attached to them.

High-Stakes Testing

High-stakes tests are tests used to make high-stakes decisions that have important consequences for individual students (Heubert & Hauser, 1999; Nichols, 2007). These decisions can include tracking students to particular programs of study, whether a student will be promoted to the next grade, or if a student will receive a high school diploma. The Committee on Appropriate Test Use, part of the National Research Council in 1999, recognized there are not only consequences for individual students, but the high-stakes also relate to accountability for educators and school systems (Heubert & Hauser, 1999; Hursh, 2005). High-stakes tests are a form of accountability to ensure schools are meeting goals of student achievement (Hamilton, et al. 2002; Nichols, 2007). NCLB (2001) has increased the high-stakes nature of standardized testing in the United States and these high-stakes can have negative repercussions on students and schools.

According to Gallagher, "the high stakes nature of such tests produces anxiety in students, parents, and educators; a test score remains a valued piece of information to be considered during decision-making processes" (2003, p. 95). Test scores are used as

unbiased evidence of student achievement and school officials regularly use test scores to determine retention, promotion, and graduation (Heubert & Hauser, 1999; Wilson, 2007). Haladyna & Haas (1998) state, "Standardized testing is entrenched in American education. The public continues to support testing because it perceives that test scores are valid indicators of children's learning" (p. 264).

Supporters of high-stakes testing believe the publication of test scores, particularly by subgroups, pressure schools to help disadvantaged students (Harris & Herrington, 2006). Carnoy and Loeb (2003) found that strict systems of government-based accountability have a positive and statistically significant effect on achievement for all students, but larger effects for minorities compared with whites. Promotion and graduation requirements tied to standardized test scores have a positive impact on minority students more than their white counterparts (Harris & Herrington, 2004). A study of Chicago Public Schools found that math and reading scores on the high-stakes test increased significantly following the introduction of the accountability policy, with low-achieving schools showing substantially larger gains than other schools (Jacob, 2005). Others believe high-stakes tests have negative consequences for schools and students.

Opponents of high-stakes testing "argue that such policies, and indeed the entire standards movement, are based on faulty assumptions about human motivation" (Natriello & Pallas, 2001, p. 22). Ravitch (2010) states that standardized tests are imprecise and the level of education many students receive is extremely low when teachers teach to the test. Many teachers feel pressured to rush through teaching content that will appear on the test superficially, without teaching deeply (Darling-Hammond,

2011). Sheldon and Biddle (1998) believe the standards movement's narrow accountability, rigid standards, and sanctions may reduce the motivation of both teachers and students. According to Heubert and Hauser (1999) not enough information has been collected to argue for or against high-stakes testing. However, what is clear from the research is that the consequences of high-stakes tests do not affect all students equally.

Examining testing in Texas, New York, and Minnesota, it was found that consequences of high-stakes testing were not uniform across racial, ethnic, and social class lines (Natriello & Pallas, 2001). According to Natriello and Pallas (2001):

> If defenders of the current arrangements for schooling rely on the results of high-stakes tests to define any and all patterns of educational deficits as originating and residing in the backgrounds and individual capacities of students alone, then we should be concerned that these tests will be used to justify the maintenance of an educational system that only appears to provide fundamental educational right, while denying those rights in defiance of state and federal constitutional provisions. (p. 38)

According to Darling-Hammond (2011), NCLB's complex system for a school to show Adequate Yearly Progress (AYP) allows for the chances that a school will be designated as failing increase in proportion to the number of demographic groups served by the school" (p. 41). Two separate studies found that schools teaching limited-English proficiency, poor, and minority students are disproportionately likely to be identified as "needing improvement" under NCLB (Novak & Fuller, 2003; Sunderman & Kim, 2004). There are differing opinions between education leaders, policy makers, and the public as to how standardized tests scores should be used with students and schools.

The tension between policymakers and education experts is central to two dilemmas posed by standardized tests when they are used as policy strategy (Heubert & Hauser, 1999; Valenzuela, 2005; Wilson, 2007). First, public and policy expectations of testing usually surpass the technical capacity of the tests themselves (Heubert & Hauser, 1999; Wilson, 2007). This often occurs because policymakers use existing tests for purposes for which they were not designed or validated. This may happen when a test created to produce valid measures of performance at the school or classroom level is used to report on individual students. The second dilemma comes from tensions between the motives of fairness and the impulse to sort students (Heubert & Hauser, 1999; Valenzuela, 2005). Relying on standardized tests is seen as fair because the tests are uniform and all students are tested over the same material. Often, though, these same tests are used to sort students in ways that are xenophobic or racist (Heubert & Hauser, 1999; Valenzuela, 2005). Educational experts and policymakers have different ideas of how standardized tests could and should be used in U.S. education.

Since World War II there have been a number of waves of reform involving test use (Linn, 2001). The Elementary and Secondary Education Act (ESEA) of 1965 was created to even the disparities in educational opportunities throughout the United States. Congress demanded accountability for the Title I ESEA funds to be distributed (Linn, 2001). This demand for accountability led to an expansion in the use of norm-referenced tests. Minimum-competency testing using norm-referenced tests, as a prerequisite for high school graduation was established by states during the 1970s and 1980s (Linn, 2001). The high-stakes nature of standardized tests and the potential consequences for

schools has led to a shift in instructional practices. Practicing for high-stakes, standardized tests is a part of U.S. public education.

Benchmark Tests

Definition

State standardized tests are powerful motivators when the scores are made public in local newspapers and are linked to funding. Administrators and teachers want data on students who are not making progress in order to make adjustments in teaching and learning before they take the standardized test to influence the final scores. Benchmark assessments measure students' progress throughout the year to provide educators with information on how to adjust instruction for the high-stakes state assessments (Olson, 2005). According to Bancroft (2010), regularly scheduled benchmark tests throughout the school year are "utilized as a means to have greater surveillance of teaching and learning, with the ultimate goal of closing achievement gaps" (p. 59). Screenings on benchmark tests help educators identify students who may need additional or differentiated instruction (Nese, Park, Alonzo, & Tindal, 2011). "Vendors and service providers have jumped in to fill this gap with a variety of products and services known by such names as *benchmark tests, progress monitoring systems,* and *formative assessment"* (Herman & Baker, 2005, p. 48). These locally developed and vendor-developed testing systems coordinate with state standards and are administered multiple times throughout the year (Herman & Baker, 2005; Olson, 2005; Bulkley, Nabors Olah & Blanc, 2010). Discussions of *interim assessments* are also found in the literature. Interim assessments are defined as assessments that "(1) evaluate students' knowledge and skills relative to a specific set of academic goals, typically within a limited time frame, and (2) are designed

to inform decisions at both the classroom and beyond the classroom level" (Perie Marion, Gong, & Wurtzel, 2007, p. 4). Benchmarking "refers to the practice of giving students periodic assessments that simulate state accountability tests for the purpose of predicting how students will perform on those tests" (Nelson, et al., 2007, p. 3). In this study, interim assessments are viewed as a type of benchmark assessment. Benchmarking is a relatively new practice that came after the mandates of high-stakes standardized tests.

History of Benchmark Tests in Schools

Black and Wiliam's (1998) review of empirical studies revealed well-conceived classroom assessments benefited student learning. They concluded that when teachers and schools adjust ongoing classroom instruction based on the results of formative classroom assessments, students mastered content and their performance on external achievement tests improved (Black & Wiliam, 1998). In order for formative assessments to be successful teachers must interpret student results and modify instruction (Heritage, et al, 2009; Perie, Marion & Gong, 2009). Improved student learning and performance is likely to occur when teachers provide feedback with clear guidance for improvement (Nichols, Myers & Burling, 2009; Shepard, 2009). Students who develop reflective thinking about their own strategies are prone to improve their performance (Bangert-Downs, Kulik, Kulik & Morgan, 1991). The pressure for students to perform well on standardized tests is great and the idea that classroom assessment could lead to higher test scores enticed many school officials. "Test publishers began to re-label many of their tests as 'formative.' This name-switching sales ploy was spurred on by the growing perception among educators that formative assessments could improve their students' test scores and help their schools dodge the many accountability bullets being aimed their

way" (Popham, 2006, p. 86). There is debate as to the effectiveness of these practice tests.

Benefits and Costs

According to Herman and Baker (2005), well-designed, formative, benchmark tests can contribute to student learning, but tests that are not well-designed can waste students' and teachers' valuable time and energy. Evidence shows that formative assessments that are quickly used to change instruction do benefit students; however, most benchmark test systems fail to get significant results to teachers soon enough for meaningful instructional modifications to be made (Popham, 2006). A variety of practice tests, both well-designed and no well-designed, are developed and sold to schools.

State mandated standardized tests given annually monitor progress towards state standards, but do not provide ongoing data regarding a student's progress. Vendors and school districts have stepped in to fill this gap with products with names such as benchmark tests, progress tests, progress monitoring systems, and formative assessments (Herman & Baker, 2005). These test products "are designed to coordinate with state standards and assessment and are administered regularly – often quarterly – to gauge student progress" (Herman & Baker, 2005, p. 48). One factor that can change a student's score on an assessment is known as the practice effect. "Practice effects occur simply because of the students' exposure to a test" (Kulik, Kulik, & Bangert, 1984, p. 438). Working under the assumption of the practice effect, the more times a students takes a benchmark test, the better the student's results on the state standardized assessment. Many schools and districts spend money on benchmark tests hoping to benefit from the practice effect, though they are a limited view of a student's knowledge (Bancroft, 2010).

One downfall to benchmark tests is that they only demonstrate where a student is performing at one point in time. Teachers and administrators need continuous information about the activities, curriculum, and programs that provide the maximum student development in order create enduring changes that will benefit students (Olson 2007). Benchmark tests given a few times a year do not fill the need for continuous feedback on teaching and learning. Another downfall to benchmark testing is that high-stakes testing disrupts instruction and the tests have a greater impact than originally intended (Popham 2001). State testing pressures teachers to spend valuable time preparing students to take tests, discouraging sensible teaching and learning (Haladyna, Haas et al. 1998; Wilson, 2007). In Texas, teachers spent an average of eight to ten hours each week preparing students for the state test (Heilig and Darling-Hammond, 2008; Hoffman, et al. 2001). Low-socio-economic status (SES) students spend more time preparing for state tests than students in affluent schools (McNeil, 2000; Sheppard, 2002; Valenzuela, 2005). Texas's standardized test, TAAS, supplanted a more varied curriculum, but only in low-performing schools (McNeil and Valenzuela 2001). Schools attended by low-income and nonwhite students spend more time preparing for the test. This time was spent modeling how to mark the answer document, reviewing topics that would be on the test, and taking tests from previous years as practice, or benchmark tests (Hoffman, et al., 2001). Many students spend large amounts of time practicing for high-stakes tests, though it may not affect their scores.

Nelson, et al (2007) researched the use of benchmark testing in Texas public schools and whether the use of such tests affected state standardized test scores in English language arts (ELS) and math. Using a random sample of public school districts in Texas,

Nelson et al found that approximately two-thirds (63%) of the 41 participating school districts engaged in benchmark testing. It was unclear what factors led to a district's decision to benchmark test. Further the authors found "that regardless of the extent of practice, benchmarking had little effect on accountability test scores" (Nelson, et al. 2007, p. 23). For English language arts there was a small benefit of 6%. In math there was no benefit found. This study suggests that while many Texas schools and districts are requiring the use of benchmark tests, such tests may not be effective.

Organizational Theory

Definition

There are different definitions regarding what constitutes an organization in the literature. An early definition comes from Barnard (1937), where an organization is defined as any consciously coordinated system of cooperative activities. Presthus (1958) defined an organization as a system of structural interpersonal relations where individuals are differentiated in terms of authority, status, and role with the result that personal interaction is prescribed. For the purpose of this study, an organization is defined as "a system of coordinated activities of a group of people under authority and leadership" (Scott & Mitchell, 1976, p. 27).

Texas school districts meet Scott & Mitchell's definition of an organization, based on the description of school district found in the Texas Education Code (TEC). The first part of the definition of an organization is that it is a system of coordinated activities of a group of people. Title 2, Subtitle A, Chapter 4, Section 4.001, states the goal of a school district "to ensure that all Texas children have access to a quality education that enables them to achieve their potential and fully participate now and in the future in the social,

economic, and educational opportunities of our state and nation." The definition of an organization describes it as working under authority and leadership. School districts work under authority and leadership of the Board of Trustees and Superintendent. TEC states the Board of Trustees and Superintendent work together to "provide educational leadership for the district, including leadership in developing the district vision statement and long-range educational plan; and establish district-wide policies and annual goals that are tied directly to the district's vision statement and long-range educational plan" Title 2, Subtitle C, Chapter 11, Subchapter D, Section 11.1512(b)(3-4)). According to Scott and Mitchell's definition of an organization, and the TEC definition of a school district, a school district is an organization.

History

Early organizational theory was focused around improving industrial efficiency, and the focus was on individuals within the organization (Tosi, 2009). Frederick Winslow Taylor focused attention on work design and functional management in Shop Management, a paper read before the American Society of Mechanical Engineers in New York in 1903. Taylor (1903) proposed managers with specialized jobs, and employees would have multiple supervisors for different aspects of their tasks. Kimball (1913) and Elborune (1914) wrote textbooks explaining alternate ways to organize administrative hierarchies and standardize procedures. In the 1920s and 1930s Weber and Michels began studying organizations as administrative hierarchies with well-defined tasks to perform (Touskas and Knudsen, 2003). Organizational study began to move from looking at individual organizations, to comparing effectiveness of multiple organizations. In 1931 Mooney and Reily emphasized establishing a universal set of management principles that

could be applied to all organizations. Organization theory gained scholarly legitimacy when Princeton University held the Business Concentration and Price Policy Conference in June 1952. More than 30 scholars from a dozen universities gathered to discuss morale, leadership, decision-making, and effects of organizations on their members (Daft, 2008; Touskas and Knudsen, 2003).

In the 1950s organization theory shifted from organizations as settings where individuals make decisions to organizations as decision-making systems. March and Simon (1958) described organizations themselves as information processors. Organizations were viewed as systems that learn and have their own decision-making processes (Cyert and March, 1963; Daft, 2008). Ideas Cyret and March (1963) presented in their work, A Behavioral Theory of the Firm, directly influence modern organizational learning theory (Argote, 1999; Huber 1991; Levitt & March, 1988). Learning at the level of the overall organization is the focus of organizational learning theory. Research is this area looks at how organizations learn from experience (Agote and Greve, 2007). Organizational learning theory is a basic assumption of modern organization theories (Agote & Greve, 2007; Tsouskas & Knudsen, 2003).

Modern organization theory can be broken into a systems model and a contingency model. The systems model began in the 1960s and the focus is on the interdependence of parts within the organization (Tsouskas and Knudsen, 2003). Scott and Mitchell (1976) described organizations as social systems where the various elements do not act as isolated segments with separate functions. The systems model sees the organization as a whole, dynamic process and does not focus on the parts (Daft, 2008;

Senge, 1990). Appropriate actions can be taken when one appreciates the organization's systems. In contrast to the systems model is the contingency model.

The contingency model is another type of modern organization theory. "Contingency approach is an approach where the behavior of one subunit is dependent on its environmental relationship to other units or subunits that have some control over the consequences" (Tosi and Hamner, 1974, 1). There is no one best way to manage or organize; what managers do is dependent on circumstances and the environment (Miner, 2005; Scott and Mitchell, 1976). Fiedler's (1967) contingency model states that a leader's effectiveness is based on leadership style and situational favorableness; there is no ideal leader. Under a contingency model, any leader can be effective if his/her leadership style meets the situation. A type of contingency model is institutional theory.

Institutional theory examines populations of organizations at the macro level. This theory states that organizations function under *institutional rule*s that are taken for granted and not necessarily based on concrete evidence. (Tsouskas and Knudsen, 2003). *Institutional rules* usually form through habitual actions that members of an organization perceive to have value; at some point these actions assume an objective quality (Tsouskas and Knudsen, 2003; Meyer and Rowan, 1977). This is when the rules become *rational myths*, and become an external constraint on the behavior of individuals. There may not be evidence that prescribed practices increase performance. Organizations that conform to these institutional rules are rewarded with increased legitimacy and reduced uncertainty (Aldrich and Fiol, 1994; Meyer and Rowan, 1977; Miner, 2005). Institutional isomorphism is built on the idea of institutional rules and rational myths.

Institutional Isomorphism

Dimaggio and Powell (1983) built on the ideas of institutional rules and rational myths to examine how organizations are the same. *Isomorphism* is the resemblance of an organization to other organizations in its environment or field (DiMaggio & Powell, 1983, Mizruchi & Fein, 1999; Deephouse, 1996). An organization's field is made up of the key suppliers, consumers, regulatory agencies, and other organizations that produce similar services or products (Dimaggio and Powell, 1983; Jones, 2009). Organizations may adopt practices of other organizations in the field. These new practices can be given value beyond the requirements of the task at hand (Jones, 2009; Selznick, 1957). According to Schelling (1978), organizations watch other organizations in their field responding to challenges in their environment, which consists of organizations responding to an environment of organizations' responses. In this loop, institutional rules and rational myths become driving forces of individuals' behaviors.

One type of isomorphism is *strategic isomorphism*, the similarity of an organization's strategy to the strategies of other organizations in its industry (Meyer and Rowan, 1977; Abrahamson and Hegeman, 1994; Deephouse, 1996). According to DiMaggio and Powell (1983), "organizations tend to model themselves after similar organizations in their field that they perceive to be more legitimate or successful" (p. 152). Organizations attempt to find organizational legitimacy through isomorphism (Deephouse, 1996; Jones, 2009). This legitimacy is demonstrated from an evaluative perspective, signifying the organization's desirability and normatively compared with other organizations in its field (Aldrich & Fiol, 1994; Jepperson, 1991; Deephouse, 1996). Organizational legitimacy can be defined as a status conferred on an organization

by social actors (Ashforth & Gibbs, 1990; Pfeffer & Salancik, 1978). Social actors are groups or individuals relevant to the organization's field, with the standing to confer legitimacy. One group of social actors includes government regulators that govern an organization and its field (Baum & Oliver, 1991, Meyer & Scott, 1983). Another social actor that influences an organization's legitimacy is public opinion, "which has the important role of setting and maintaining standards of acceptability" (Deephouse, 1996, p. 1025). To be considered a legitimate organization, it must have values and actions congruent with a social actor's expectations (Galaskiewicz, 1985; Deephouse, 1996). Similarity among organizations in the same field does not develop due to competition or objective analysis, but instead on an organization's quest to attain legitimacy (Meyer & Rowan, 1977; DiMaggio & Powell, 1983, Mizruchi & Fein, 1999).

As Texas school districts have been defined previously in this paper as an organization, the theory of strategic isomorphism will be examined in relation to school districts. This study will explore predictive relationships between district factors and their benchmarking practices.

CHAPTER 3

RESEARCH METHODOLOGY

This chapter describes the design of the present research project, including an introduction to path analysis, the associated key terms and an introduction to mediation analysis. Path analysis is an analytic technique useful for identifying relationships in multivariate data structures (e.g., data structures with more than two variables).

The design of the present study is non-experimental or observational because it uses existing data to answer research questions. Furthermore, this study is exploratory in that the primary goal of the analysis is to uncover patterns among variables in the data (Tukey, 1977; Behrens, 1997). For example, one aim of exploratory techniques is to observe relationships among different factors and outcomes as they exist in the *real world* (Boslaugh, 2012). The phrase *real world* does not reflect an entirely objective reality; rather the outcomes (i.e. variables) examined in this study are also generated at least in part as social constructions (Gergen & Gergen, 2003) in our education system. In this sense, the present study operates within the epistemological framework of critical realism (Manicas, 2006). In this epistemology, although measured variables are generated within social systems (i.e. are constructions within systems), the real world *talks back* or provides feedback these constructions. In exploratory research that examines relationships among variables, correlations are examined without necessarily determining causal relationships – as would be the case in experimental research. Causal relationships exist if (a) the cause preceded the effect and there is no plausible alternative explanation for the effect other than the cause (Shadish, Cook & Campbell, 2002) or (b) random

assignment to a study condition excludes rival explanations or hypotheses about a relationship(s). This study is correlational, not causal (i.e. experimental), because it is not possible to determine which variables came first or isolate the effects due to individual variables (Boslaugh, 2012; Shadish, Cook & Campbell, 2002).

Path Analysis

Structural Equation Modeling (SEM) consists of a set of linear equations that tests two or more relationships at the same time (Shook, Ketchen, Hult, & Kacmar, 2004). SEM identifies the relationships between observed and latent variables (Bauer, 2003; Byrne, 2001; Kline, 2005) and at its core is the measurement model. For example, a measurement model is comprised of a single latent variable (a.k.a. a construct) and a collection of observed variables, for example items on an instrument or test, that are indicators of the construct of interest. Path analysis is a form of SEM (and an extension of multiple regression) that focuses on the relationships between observed variables (Randolph & Myers, 2013) rather than latent and observed variables together such as there is no measurement model involved. Wright (1921) pioneered the use of path analysis as a method to measure the influence of independent variables on a dependent variable in a more flexible manner than was available with multiple linear regression. The present study uses only observed variables, which is why path analysis was the analytic technique of choice.

In path analysis, the path diagram is a graphical model that represents the relationships between observed variables (Kline, 2005; Shadish, Cook & Campbell, 2002). Observed variables are represented with rectangles (Kline, 2005). Each line with a single arrowhead, known as a path, shows a hypothesized direct effect one variable on

another (Kline, 2005). Independent variables are assumed to influence other variables (Boslaugh, 2012). Dependent variables are assumed to be influenced by other, independent variables in the study (Boslaugh, 2012); however dependent variables are also able to influence other dependent variables. Indirect effects occur due to the mediating influence of a single variable existing (or positioned) between two other variables and pass the effect of one to the other (Kline, 2005; Jose, 2013). A variable mediates a relationship when the basic relationship between variables is reduced when the mediating variable is present (Baron & Kenny, 1986; Jose, 2013). Figure 3 shows an example of mediation.

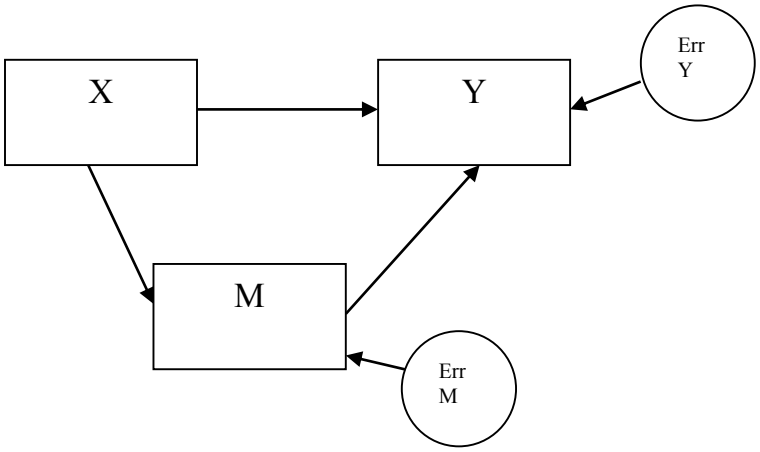

Figure 3. Sample Mediation Model. Sample statistical mediation. ErrY and ErrM = error terms.

Because the unit of analysis in this study is the school district, data were collected at the district-level, rather than the individual student level.

A final reason for selecting path analysis is that it is a statistical tool that provides a formal mechanism for both confirmation and exploration (Kline, 2005). As a confirmatory tool, the researcher creates a model at the beginning (*a priori*) based on a comprehensive literature review then fits the model to actual data and evaluates how

closely the hypothesized model fits the actual data. Path analysis works as an exploratory tool when revised models are used with the same data. The focus of exploratory statistical techniques is to find patterns in the data (Tukey, 1977; Behrens, 1997). To this end, the present study is driven by the literature review but is also exploratory because the mediating relationships between independent and dependent variables have not been previously investigated using path analysis.

Key Terms

While much of the terminology used in path analysis is common to other statistical methods, some terms and concepts are unique. Both common and unique terms are defined below. The following list of terms and definitions are provided to enhance understanding of the analytic technique used in this study.

1. Alpha level (α level). This is sometimes referred to as *p* level. A value of typically either .01 or .05 that indicates the probability of misinterpreting analytical results to indicate a genuine effect on the population when in fact there is no effect. This is referred to as a Type I error (i.e. a false positive or falsely rejecting the null hypothesis when in fact no difference or effect exists). At $p \leq .05$ there is a 5% or less chance of making a Type I error. This means there is a 95% or greater chance of the results explaining the noted effect (Cohen and Cohen, 1984). In path analysis, hypothesis tests specific to path coefficients test the hypothesis that the regression weight, path coefficient, is significantly different from zero.
2. Causal arrow. A straight, single-headed arrow in a structural equation diagram, indicating a hypothesized direct effect on one variable from another. The arrow

points to the variable that is believed to be affected, and originates from the variable presumed to be the cause (Loehlin, 2004; Kline, 2005).

3. Chi-Square test of Independence. When data are non-parametric and not normally distributed, it is a statistical test to determine goodness-of-fit for a hypothetical model in relation to a set of observed data (Pett, 1997; Schumaker and Lomax, 2010). A statistically significant Chi-Square statistic signifies that the observed are *not independent* of the expected data (according to probability theory). Interpreted practically, a non-significant Chi-Square test means that the observed data fit the hypothesized data as *expected by probability theory*.
4. Correlation. Pearson's *r* is a generally accepted measure of correlation, demonstrating the strength of the relationship between the variables. The strength of correlation over covariance is that Pearson's *r* is based on standard deviation units, allowing an estimation of effect size (Field, 2005). For example, Pearson correlation coefficients are in of themselves effect size coefficients ranging from small (0.0 - .29), medium (.30 - .59) and large (.60 – 1.0).
5. Degrees of freedom. The number of variables that are free to take on any value in an equation (Field, 2005).
6. Dependent variable. The variable whose outcome is influenced by other variables in the model (Schumaker & Lomax, 2010).
7. Effect size. Magnitude of an independent variable's effect, usually expressed as a proportion of explained variance in the dependent variables (Cohen, 1990; Weinfurt, 2000). Pearson's *r* is the most common method of interpreting effect

size. An *r* value of around 0.2 shows a small effect, a value of 0.5 is a medium effect, and 0.8 is a large effect size (Norman & Streiner, 2003).

8. Endogenous variable. Dependent variable in a structural equation model, having straight, single-arrows pointing to them, signifying they are influenced by independent variables (Byrne, 2001).
9. Exogenous variable. Independent variable in a structural equation model, it has straight arrows leading away from it, demonstrating its influence on other variables. They never have arrows leading to them (Loehlin, 2004).
10. Independent variable. A variable under influence of the researcher and is affected directly or indirectly to forecast the change in dependent variables (Norman & Streiner, 2003).
11. Latent variable. A latent variable is not directly observed or measured. It is estimated by the researcher in terms of components assumed to construct the variable's meaning (Schumacker & Lomax, 2004).
12. Non-parametric tests. These tests are used to analyze ordinal and nominal data, and are based on frequencies and rank orders. They can be used with small sample sizes because they make few assumptions about the population's distribution (Pett, 1997).

Introduction to Mediation Analysis

Mediation is a distinctive type of path model that emphasizes the variables between predictor variables and dependent variables (Jose, 2013). A mediator variable stands between two other variables and passes the effect of the one on to the other (Jose, 2013; Kline, 2005). According to Baron and Kenny (1986), a mediator variable "accounts

for the relation between the predictor and the criterion" (p. 1176). Mediators help explain how or why effects occur between variables (Baron & Kenny, 1986; Jose, 2013). Mediation is said to occur when a causal effect of some independent variable, X, on a dependent variable, Y, can be explained by an intervening variable, M (Shrout & Bolger, 2002; see Figure 4). Mediating variables describe the relationship of two other variables. Figure 4 shows the direct effect of independent variable X on dependent variable Y, path *c*, as well as the direct effect of X on mediator variable M, path *a*, and the direct effect of M on Y, path *b*.

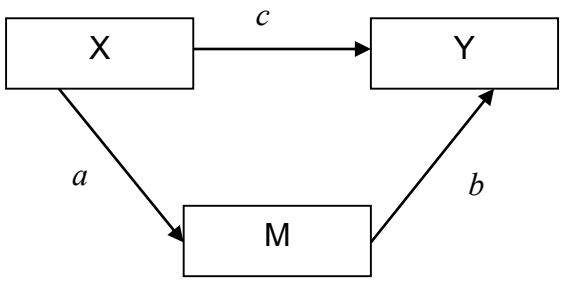

Figure 4.Mediation. Sample statistical mediation

Mediation analysis is useful in observational studies, such as in the present research, to clarify the relationship between variables. Using a deductive approach, the researcher can construct a hypothetical model of the mediating effects and then test the model with the existing data set (Jose, 2013). Kenny (2008) describes this as, "a mediational analysis provides the researcher with a story about a sequence of effects that leads to something" (p. 2). The mediation path analysis research design *cannot explain causality*; instead it explains to what extent a mediation pattern affects the dependent variable (Jose, 2013; Shrout & Bolger, 2002). Mediation analysis may suggest a causal

relationship among variables, but it is not able to offer definitive evidence regarding causality. Path analysis allows the researcher to look at the effects of multiple mediators on a dependent variable. Multiple mediators can be tested simultaneously to learn if the mediation is independent of the other mediators (Kline, 2005). In the present study mediated path analysis allows for in-depth exploration of a complex model.

The amount of mediation in a model is known as the indirect effect. The indirect effect is measured with the regression coefficients of the paths in the diagram. The effect size of the indirect effect is the product of the direct effects. In Figure 4, the indirect effect is the product of *a* times *b* (Jose 2013; Kline, 2005). The strength of the mediation effect can be examined by looking at the regression coefficient of the indirect effect. The direct effect provides the size of the indirect effect of X on Y through M (Jose, 2013). The indirect effect not only explains the size of the effect of one variable on another through a mediator, it can explain the type of effect. If the indirect effect regression weight is positive, an increase in X will lead to an increase in Y (Keith, 2006; Randolph & Myers, 2013). If the indirect effect size is negative, an increase in X will lead to a decrease in Y (Keith, 2006; Randolph & Myers 2013). The *p*-value of the regression weight tells whether the indirect effect size is significant or not. According to Fisher (1971), a *p*-value of less than 0.01 is highly significant, *p*-values of 0.05 to 0.01 are marginally significant, and *p*-values of greater than 0.05 are not statistically significant.

Mediated path analysis allows for thorough understanding of complex models. It is appropriate for the present study because a variety of variables exist in the data, working together simultaneously. Instead of trying to find which variables cause one

another, the researcher can focus on how and to what extent the mediator variables affect the interaction between the independent and dependent variables.

Population and Sample

The target population of interest in this study consisted of public school districts in Texas. During the 2010-2011 school year there were 1,029 public school districts in Texas. The analytic sample consisted of a stratified random sample of Texas school districts from which benchmark testing calendars could be obtained.

To obtain a stratified random sample, the 1,029 districts in the population were separated into the nine District Types for Texas public schools, a classification used by the Texas Education Agency (TEA). District Type takes into account a district's size, location, and student enrollment characteristics (http://www.tea.state.tx.us/acctres/analyze /1011/gloss1011.html). Each District Type group was randomized and emails sent to school districts in order that the districts appear on the randomized list. Emails were sent to districts in a District Type group until a 10% rate of return was reached from each group.

A school district may give multiple benchmark tests to different students throughout the year. Which test a student is required to take is often based on grade level and status in particular groups, such as if the student is receiving special education services or is an English Language Learner (ELL). In order to provide a common point of reference within the data that was returned, this study will focus on the number of benchmark tests an eighth grade general education student, not receiving special education or ELL services, is required to take by the school district. Only tests required by the district, but not mandated by the State, will be counted for this study.

Data Collection

Using the randomized list of districts by District Type, emails were sent to request district's benchmark calendar. The request was emailed to someone in a position to have the district benchmark calendar. Due to differences in districts, employee titles, and district sizes the request may not always be sent to individuals with similar titles. Email requests were sent to curriculum directors, testing coordinators, superintendents, and principals, among others. The researcher collected email addresses for district personnel on each school district's website. The email stated the researcher is completing a study on benchmark testing and ask that a copy of the district's testing calendar to be sent electronically to the researcher.

To test the viability of this data collection method, a pilot test was conducted. Districts in the pilot test sample self-reported and replied to the email in various ways. More than half of the districts contacted did not reply. The testing calendars were submitted to the researcher in a variety of formats. Some were attachments of calendars with testing dates marked on them. Others were attachments of spreadsheets and lists of tests that are given throughout the year. Still other school district personnel responded simply with a number in the text of the reply email (i.e., "we give 4 benchmarks a year"). For the purposes of this study, the reported data will be assumed to accurately reflect the number of benchmark tests given by a district. In other words, the researcher will make no attempt to verify the accuracy of the data reported by each district.

Variables Used in the Study

The analytic models developed in this study consisted of two separate path diagrams, one for Math TAKS and one for Reading TAKS. The models include one

endogenous variable, three exogenous variables, and four mediator variables (see Figure 5). The endogenous (dependent or outcome) variable, "Percent of Students Passing TAKS" indicates the percent of students passing the Reading or Math TAKS test in spring 2011 as reported on TEA's Academic Excellence Indicator System (AEIS). The three exogenous variables "School District's State Accountability Rating,", "Annual Yearly Progress (AYP) Status", and "Texas Education Agency (TEA) District Type" are also found on the 2011 AEIS reports. Three of the mediator variables, "Percent of Limited English Proficient (LEP) Students," "Percent of Students of Color," and "Percent of Economically Disadvantaged Students," are gathered from the AEIS reports. The mediator variable, "Number of Benchmarks" was gathered through email requests sent to districts as described earlier. Each of these variables and their component indicators are summarized in Tables 1 – 3.

Table 1

Construction of the "School District's State Accountability Rating" Variable

Accountability Rating	Response Category
Exemplary	4
Recognized	3
Academically Acceptable	2
Academically Unacceptable	1

Table 2

Construction of the "Annual Yearly Progress (AYP) Status" Variable

Accountability Rating	Response Category
Met AYP	4
Missed AYP – Stage 1	3
Missed AYP – Stage 2	2
Missed AYP – Stage 3	1

Table 3

Description of Variables

Variable	Description
Percent of Limited English Proficient (LEP) Students	Students identified as limited English proficient by the Language Proficiency Committee (LPAC) according to criteria established in the Texas Administrative Code.
Percent of Economically Disadvantaged Students	Students eligible for free or reduced-price lunch or eligible for other public assistance.
Percent of Students of Color	Students reported as Hispanic and/or African American. Student ethnicity is reported by the student's parent/guardian upon registration.

The variable "TEA District Type" was combined to create composites. TEA distinguishes nine separate district types – Major Urban, Major Suburban, Other Central City, Other Central City Suburban, Independent Town, Non-Metropolitan Fast Growing, Non-Metropolitan Stable, Rural, and Charter School Districts. This study did not examine Charter School Districts because they are not regulated in the same way as other public

school districts in Texas. For the purpose of this study the following TEA District Types are combined: Major Urban and Major Suburban, Other Central City and Other Central City Suburban, Independent Town and Non-Metropolitan Fast Growing; Non-Metropolitan Stable; and Rural. Composite variables are used here as a means of data reduction and also to maximize the information in the available data. Since composite variable formulation involves taking a large number of variables and combining them into a single "composite", the process reduces the sample size used in an analysis. To this end, composite variables influence sample size and statistical power (Rowe, 2006). Sample size and power calculations were conducted during the planning and analysis phases of this study to ensure that sample size was adequate given the hypothesized path models. The results of the power analysis revealed an adequate sample size for the models used in this study so that the sample size was large enough to ensure that parameter estimates and hypothesis test were not biased. Testing calendars were collected from 10% of each TEA District Type before the District Types are combined for analysis. Table 4 demonstrates the composite TEA district type variables for this study.

Table 4

Construction of the "Texas Education Agency District Type" Variable

Variable Name	Variable Description	Variable Label
Major Sub/Urban	District is located in a county with a population of at least 775,000 and has an enrollment of at least 75 percent of the largest district in the county, and at least 35 percent of the enrolled students are economically disadvantaged and/or its enrollment	5

Table 4 (continued)

Variable Name	Variable Description	Variable Label
	Is at least 15 percent that of the nearest major urban district	
Other Central City/ Other Central City Suburban	District located in a county with a population between 100,000-749,999 and its enrollment at least 75 percent the largest district enrollment in the county, and/or it is located in a county with a population of between 100,000-774,999 and its enrollment is at least 15 percent of the largest district enrollment in the county	4
Independent Town/Non-Metropolitan Fast Growing	District located in a county with a population of between 25,000-99,999 and its enrollment is at least 75 percent the largest district enrollment in the county; or it has an enrollment of at least 300 students and its enrollment has increased by at least 20 percent over the last five years	3
Non-Metropolitan Stable	District's enrollment exceeds the median district enrollment for the state	2

Table 4 (continued)

Variable Name	Variable Description	Variable Label
Rural	District has an enrollment of between 300 and the median district enrollment for the state and enrollment growth rate of less than 20 percent during last 5 years, or an enrollment of less than 300 students	1

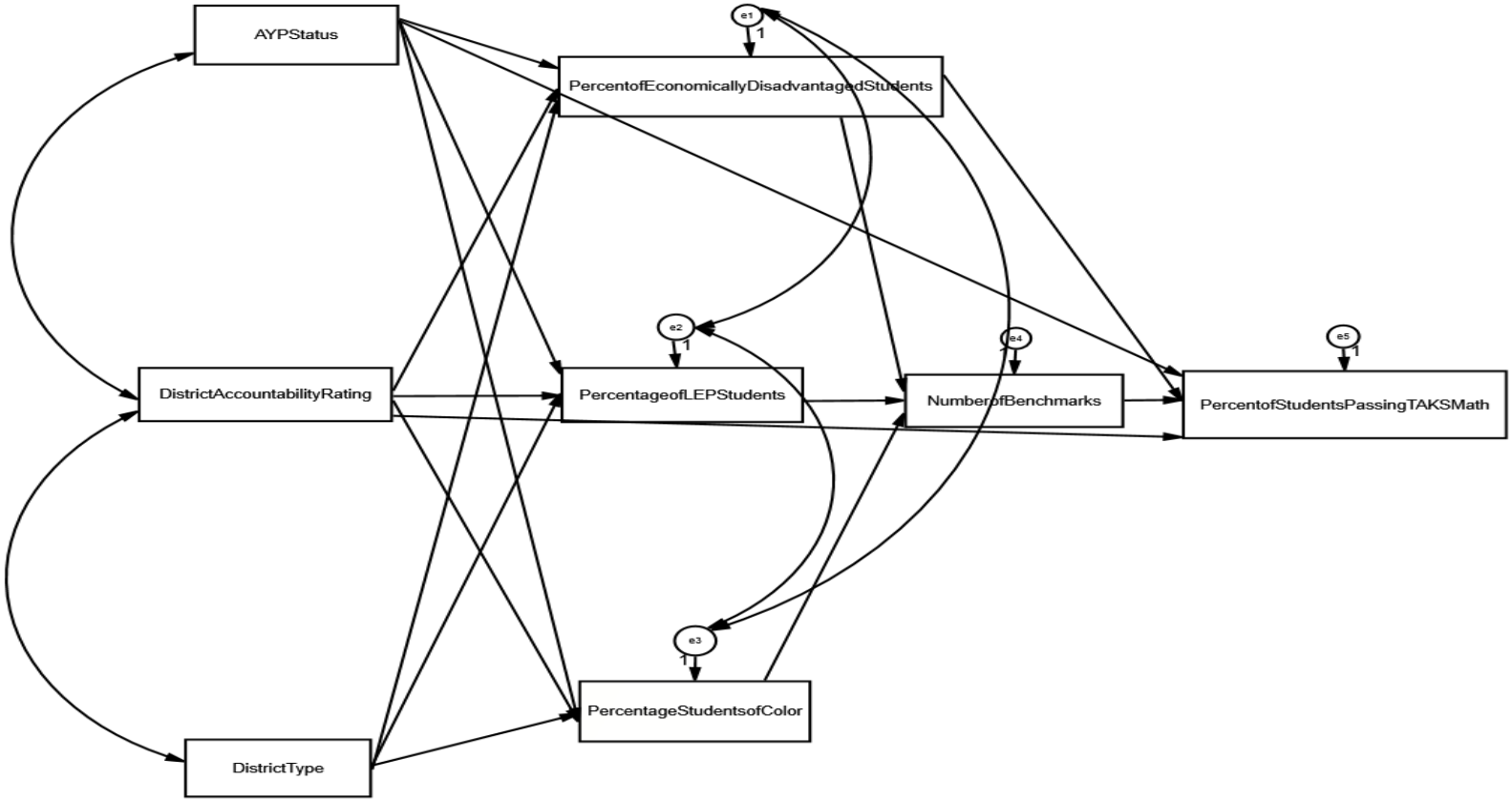

Figure 5. Proposed Analytical Model Math. This conceptual path model depicts the structural equation model developed for the study for Math TAKS scores. Arrows drawn between variables specify mediating effects. The rectangles represent observed variables. Arrows point to the dependent variable on the right, the percent of students passing TAKS.

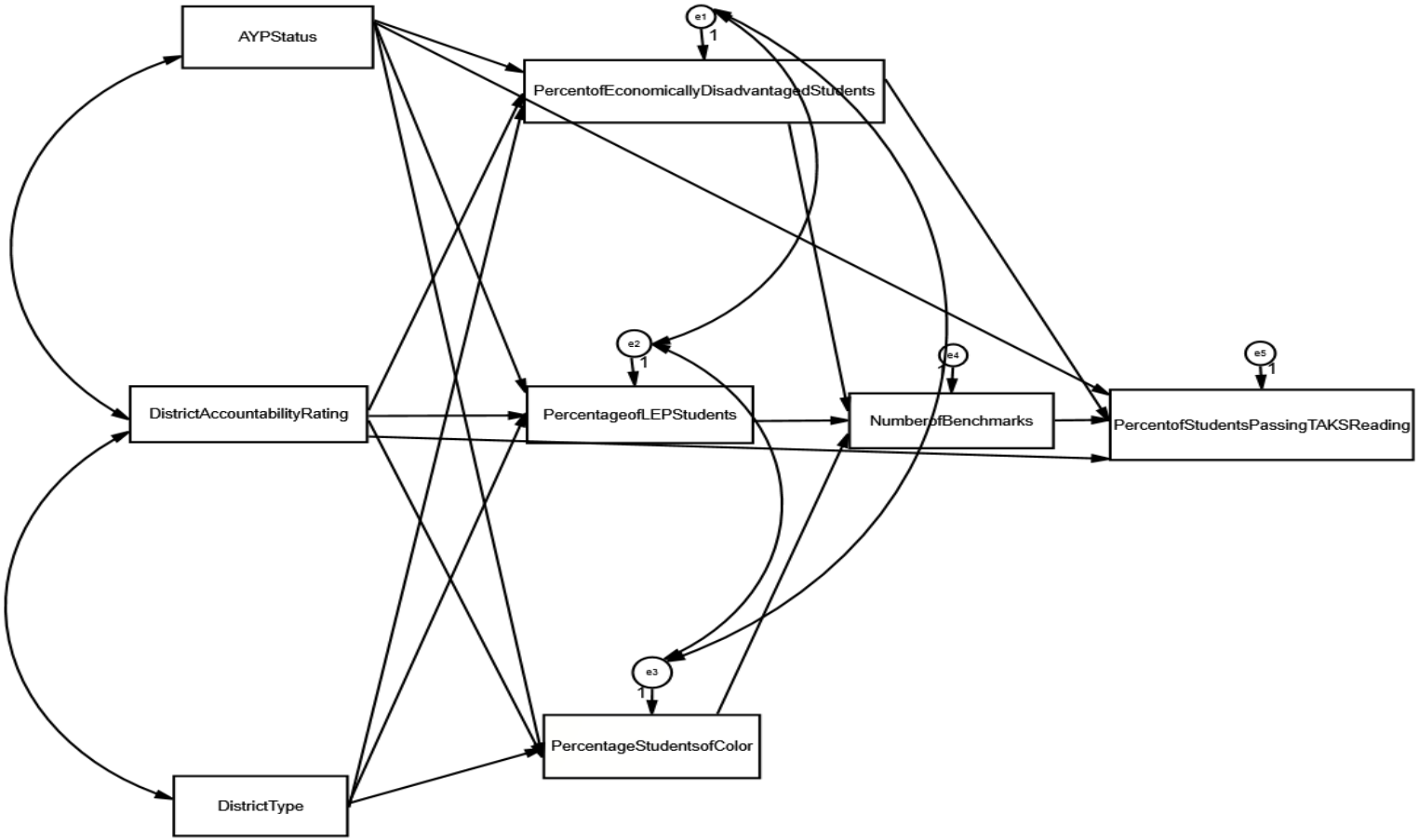

Figure 6. Proposed Analytic Model Reading. This conceptual path model depicts the structural equation model developed for the study for Reading TAKS scores. Arrows drawn between variables specify mediating effects. The rectangles represent observed variables. Arrows point to the dependent variable on the right, the percent of students passing TAKS.

Data Analysis

All data was entered into SPSS version 21.0 (IBM, 2012) and subsequently screened for assumptions of the normality, linearity, outliers and missing values. Next, Analysis of Moment Structures (AMOS) version 21.0 was used to create and analyze the path models.

Summary

The methodology provided a framework to answer the following questions:

Primary research questions:

1. Do significant relationships (i.e regression weights) exist between a district's descriptive factors and benchmark testing practices?
2. Does a significant relationship (i.e regression weight) exist between the number of benchmark tests a district requires and the percentage of students passing the TAKS test?

Supporting questions:

1. Is the effect of AYP Status on the number of benchmark tests mediated by the percentage of economically disadvantaged students in the district?
2. Is the effect of AYP Status on the number of benchmark tests mediated by the percentage of students of color in the district?
3. Is the effect of AYP Status on the number of benchmark tests mediated by the percentage of Limited English Proficient (LEP) students in the district?
4. Is the effect of a district's TEA state accountability rating on the number of benchmark tests given mediated by the percentage of economically disadvantaged students in the district?

5. Is the effect of a district's TEA state accountability rating on the number of benchmark tests mediated by the percentage of students of color in the district?
6. Is the effect of a district's TEA state accountability rating on the number of benchmark tests mediated by the percentage of Limited English Proficient (LEP) students in the district?
7. Is the effect of TEA district type on the number of benchmark tests mediated by the percentage of economically disadvantaged students in the district?
8. Is the effect of TEA district type on the number of benchmark tests mediated by the percentage of students of color in the district?
9. Is the effect of TEA district type on the number of benchmark tests mediated by the percentage of Limited English Proficient (LEP) students in the district?
10. Is the effect of TEA district type on the percent of students passing the Math TAKS test mediated by the percentage of economically disadvantaged students, percentage of LEP students, percentage of students of color, and number of benchmarks?
11. Is the effect of TEA district type on the percent of students passing the Reading TAKS test mediated by the percentage of economically disadvantaged students, percentage of LEP students, percentage of students of color, and number of benchmarks?
12. Is the effect of AYP Status on the percent of students passing the Math TAKS test mediated by the percentage of economically disadvantaged students, percentage of LEP students, percentage of students of color, and number of benchmarks?

13. Is the effect of AYP Status on the percent of students passing the Reading TAKS test mediated by the percentage of economically disadvantaged students, percentage of LEP students, percentage of students of color, and number of benchmarks?
14. Is the effect of a district's TEA state accountability rating on the percent of students passing the Math TAKS test mediated by the percentage of economically disadvantaged students, percentage of LEP students, percentage of students of color, and number of benchmarks?
15. Is the effect of a district's TEA state accountability rating on the percent of students passing the Reading TAKS test mediated by the percentage of economically disadvantaged students, percentage of LEP students, percentage of students of color, and number of benchmarks?
16. What is the relationship between number of benchmarks a district requires and the percent of students passing Math TAKS?
17. What is the relationship between the number of benchmarks a district requires and the percent of students passing Reading TAKS?
18. How does the organizational theory of isomorphism help to explain the use of benchmark testing in school districts and is organizational theory congruent with the path model results observed in this study?

Path analysis is a multivariate regression technique and is used to determine the relationships between independent, mediating, and dependent variables in this study.

This study examines the relationships of different variables on the number of benchmarks a district requires and the relationships of variables on the percent of students passing the Math TAKS and Reading TAKS tests.

CHAPTER 4

RESULTS

Chapter four provides the results of the analyses conducted in this study. Specifically, this chapter provides the results from analyses examining the relationship between a district's descriptive characteristics, its benchmark testing practices, and the impact of selected mediating variables on benchmark testing practices and TAKS Math and Reading scores. This chapter is organized by (a) descriptive characteristics of the sample, (b) model fit statistics, (c) statistical and practical evaluation of the research questions, and (d) limitations of the study.

Data Descriptive Characteristics

Sample data used are from Texas Education Agency's Academic Excellence Indicator System (AEIS) Reports from the 2010-2011 school year. The Number of Benchmarks data was collected from a stratified random sample of school districts in Texas from the 2010-2011 school year. Prior to conducting data analyses, the distributional characteristics of the data for each variable must be known in order to ensure that violations of statistical tests are not biased. For example, all statistical techniques rely on underlying assumptions regarding the shape of the distribution of the data comprising the variables under study. Standard screening techniques for data comprising variables includes evaluation of (a) the mean, (b) variance, (c) range, (d) identification of missing data points, (e) skewness, and (f) kurtosis. Knowing this information, data analyses can proceed in consideration of the distributional characteristics of the variables. The mean is the measure of central tendency or average

of the data (for data measured on an ordinal or interval level) and it is calculated by finding the sum of all values and dividing by the number of values (Jose, 2013; Welkowitz, Cohen, & Lea, 2012). Using the mean, the variance can be found. Variance examines the dispersion of that data, how far data points are from the mean. To calculate variance, first subtract the mean from each data point and then square the difference (Welkowitz, Cohen, & Lea, 2012; Hanneman, Kposowa, & Riddle, 2012; Wilcox, 2009). As the variance increases, the dispersion of the individual data values increases. The standard deviation is the square root of the variance (Jose, 2013; Welkowitz, Cohen, & Lea, 2012; Wilcox, 2009). The formula for finding the standard deviation is the square root of $(x\text{-}m)^2 / n$, taking the square root of the average of the squared differences from their average value. A small standard deviation indicates that the data points are close to the mean such as individual data point exhibit less dispersion; large standard deviation indicates the data points are dispersed over a large range. To determine if the distribution of data is skewed, it is compared to a normal bell-shaped curve (Hanneman, Kposowa, & Riddle, 2012; Wilcox, 2009). Skewness describes an asymmetrical statistical distribution, where the curve is distorted to the left or right (Hanneman, Kposowa, & Riddle, 2012; Randolph & Myers, 2013). Skewness if found with the formula $(y_1 – y)^3 / (n – 1)^3$ or the sum of the deviations from the mean, raised to the third power, divided by the number of data points minus 1, multiplied by the standard deviation raised to the third power (Randolph & Myers, 2013). Kurtosis examines if the frequencies of the sample data decline more or less rapidly than a normal curve (Hanneman, Kposowa, & Riddle, 2012). Kurtosis is how flat or peaked the frequency distribution curve is around the mode (Hanneman, Kposowa, & Riddle, 2012; Randolph & Myers, 2013). The formula for

kurtosis is the sum of the deviations from the mean raised to the fourth power, with the quantity divided by the standard deviation raised to the fourth power or ${}_{i}^{n} - 1(y_i - y)^4 / (n-1)s^4$. In a normal, bell-shaped distribution the skewness and kurtosis are zero. The further away from zero the skewness and kurtosis are, the more the data distribution deviates from normal (Randolph & Myers, 2013). The standard score, also known as the Z-score, is the number of standard deviations a data point is from the mean (Randolph & Myers, 2013; Welkowitz, Cohen, & Lea, 2012). The Z-score is calculated by subtracting the population mean from an individual raw score and then dividing the difference by the population standard deviation (Hanneman, 2013; Welkowitz, Cohen, & Lea, 2012). A positive Z-score indicates data is above the mean, and a negative Z-score represents data below the mean. Approximately 95% of cases in a normally distributed Z-score for skewness are between -2.00 and +2.00 (Hanneman, 2013; Randolph & Myers, 2013). Z-scores for kurtosis are between -6.00 and +6.00 in a normal distribution (Hanneman, 2013; Randolph & Myers, 2013). In the present data set, skewness z-scores for some variables do not fall within the normal distribution. The skewness Z-scores are not normally distributed for the variables of AYP status (-3.23), percent LEP students (6.83), number of benchmarks (7.28), and percent of students passing Reading TAKS (-3.55). The kurtosis Z-scores for all variables fall within a normal range. Table 5 provides a summary of the sample characteristics. Figures 7-10 and 12-16 show the histograms with the data distributions for each variable in the study.

Table 5

Descriptive Statistics

	Minimum	Maximum	Mean	SD	Skewness	Kurtosis	Z-score (Skewness)	Z-score (Kurtosis)
District Type	1.0	7.0	2.92	1.37	.12	-1.23	0.51	-2.57
District Accountability Rating	2.0	5.0	3.42	0.64	0.07	-0.17	0.29	-0.36
AYP Status	3.0	7.0	5.75	1.34	-0.78	-0.66	-3.23*	-1.39
Percent Economically Disadvantaged	2.3	99.8	54.11	18.65	-0.31	-0.09	-1.29	-0.20
Percent Students of Color	7.3	97.8	44.16	23.32	0.31	-0.77	1.27	-1.61
Percent LEP Students	0.2	37.8	9.36	8.45	1.65	2.28	6.83*	4.77
Number of Benchmarks	0	28.0	4.83	6.82	1.76	2.55	7.28*	5.34
Percent of Students Passing Math TAKS	71	100.0	88.53	6.96	-0.44	-0.50	-1.81	-1.05
Percent of Students Passing Reading TAKS	80	100.0	94.21	4.16	-0.86	1.13	-3.55*	2.36

Note. * indicates the Z-score is significant at $p < 0.05$.

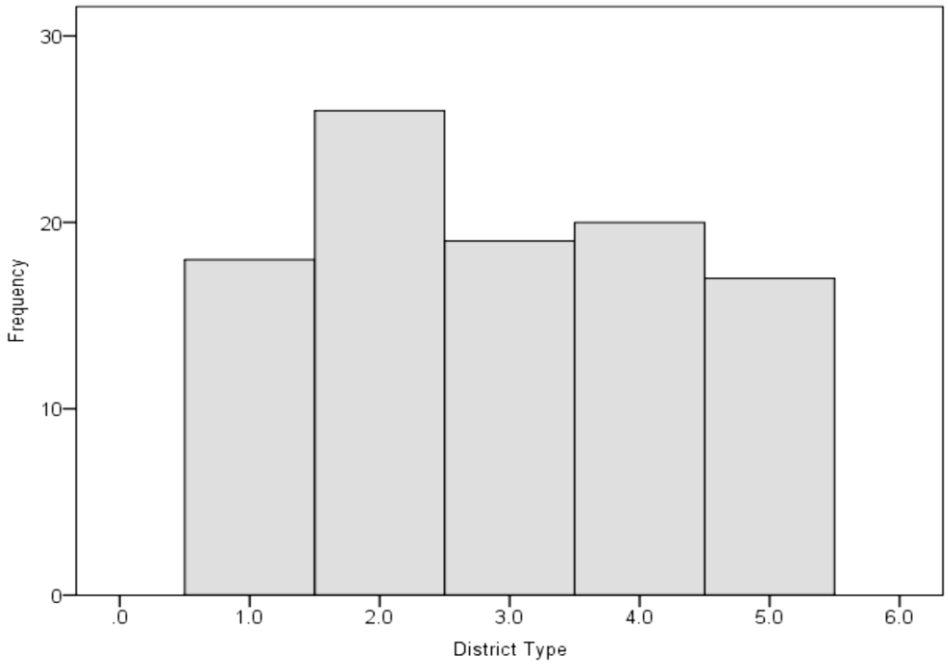

Figure 7. Types of School Districts. Histogram showing frequency distribution of districts.

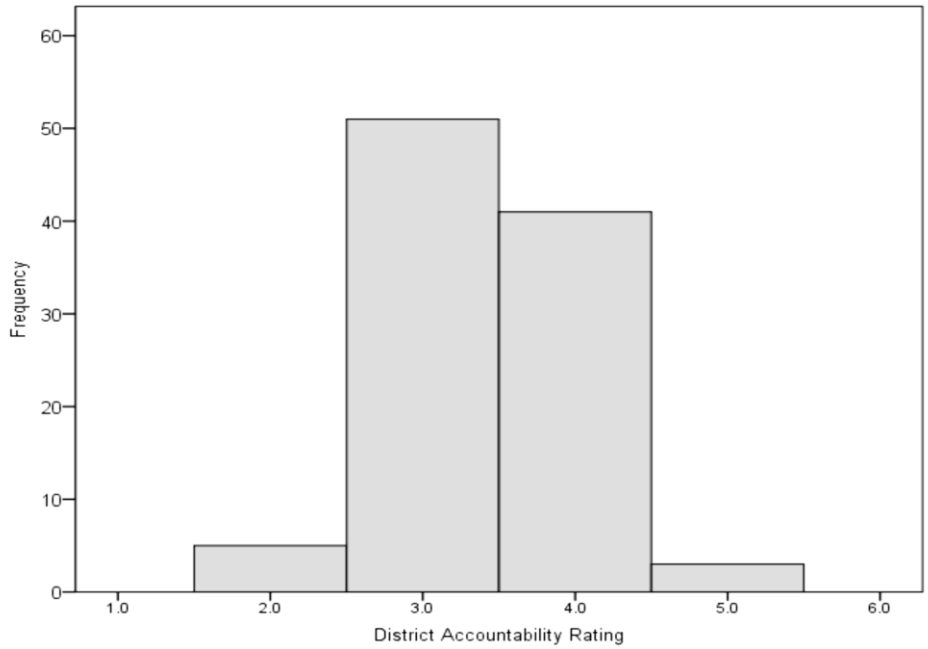

Figure 8. Accountability Ratings. Histogram showing frequency distribution of district accountability ratings.

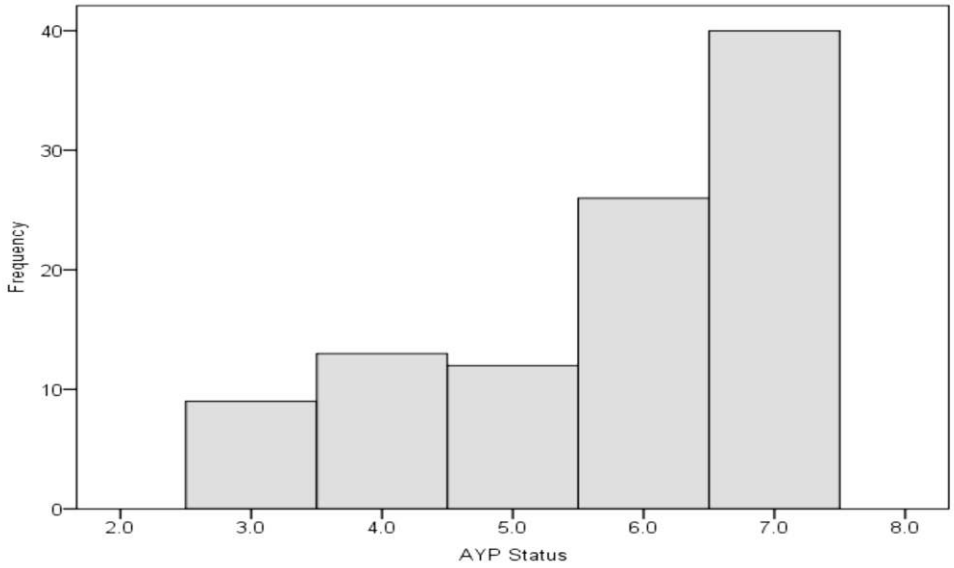

Figure 9. AYP Status. Histogram showing frequency distribution.

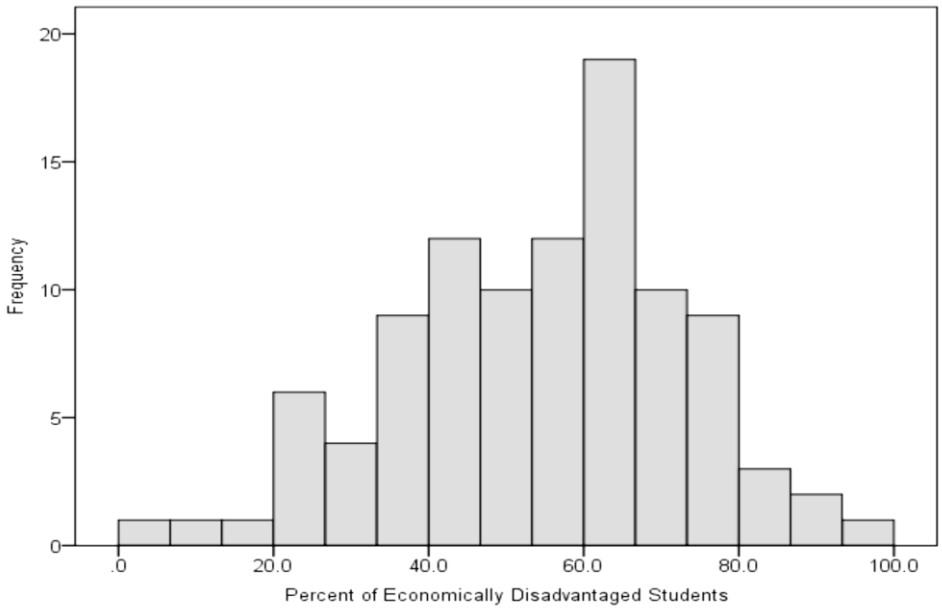

Figure 10. Economically Disadvantaged. Histogram showing frequency distribution of economically disadvantaged students.

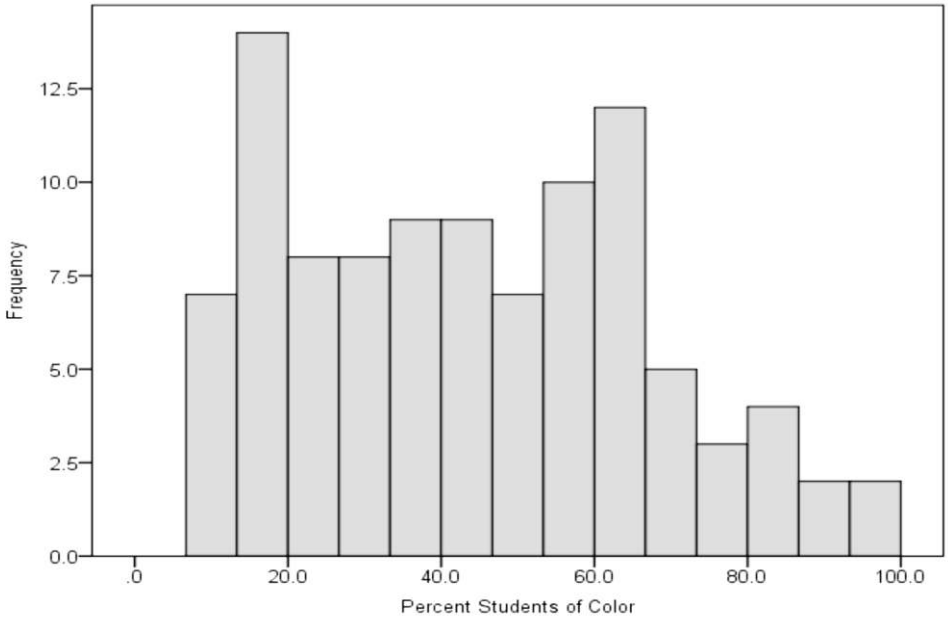

Figure 11. Students of Color. Histogram showing frequency distribution.

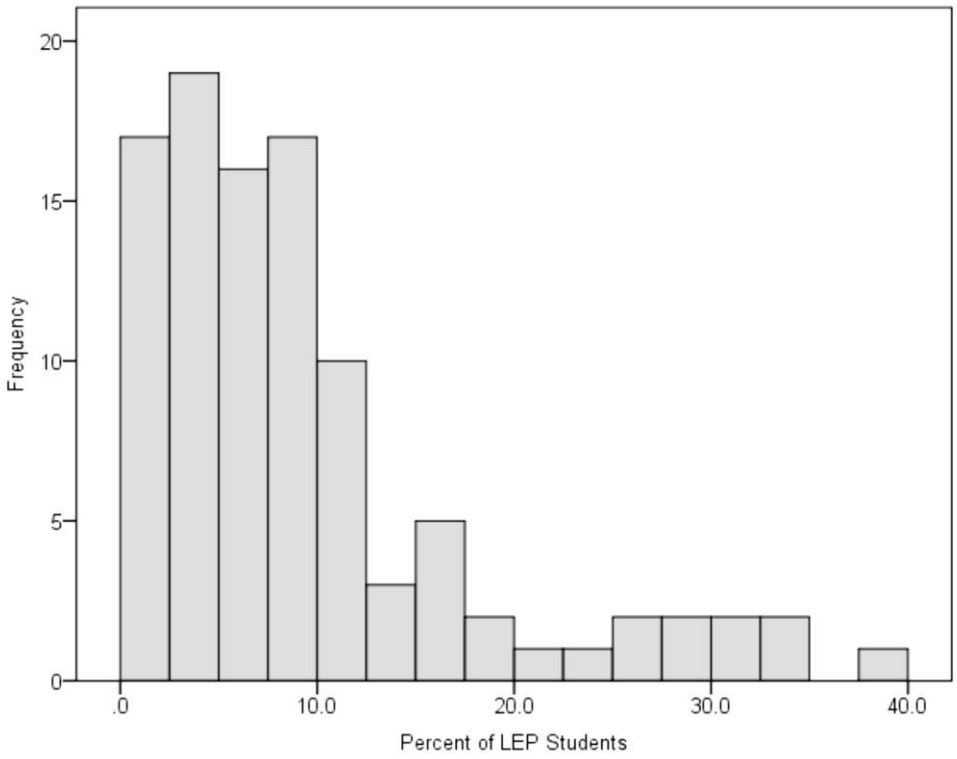

Figure 12. LEP Students. Histogram showing frequency distribution of LEP students.

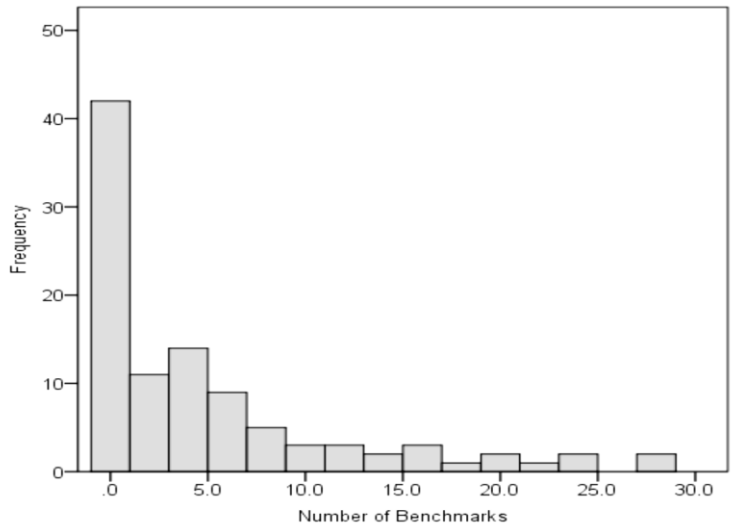

Figure 13. Benchmarks. Histogram showing frequency distribution of benchmarks.

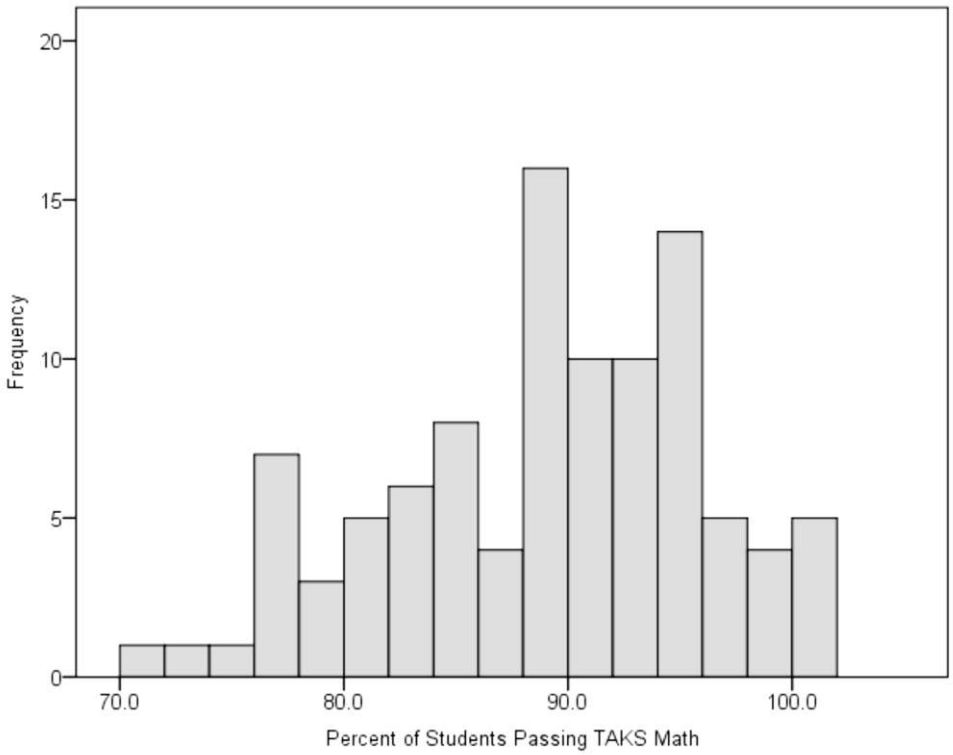

Figure 14. Students Passing Math TAKS. Histogram showing frequency distribution of students passing Math TAKS.

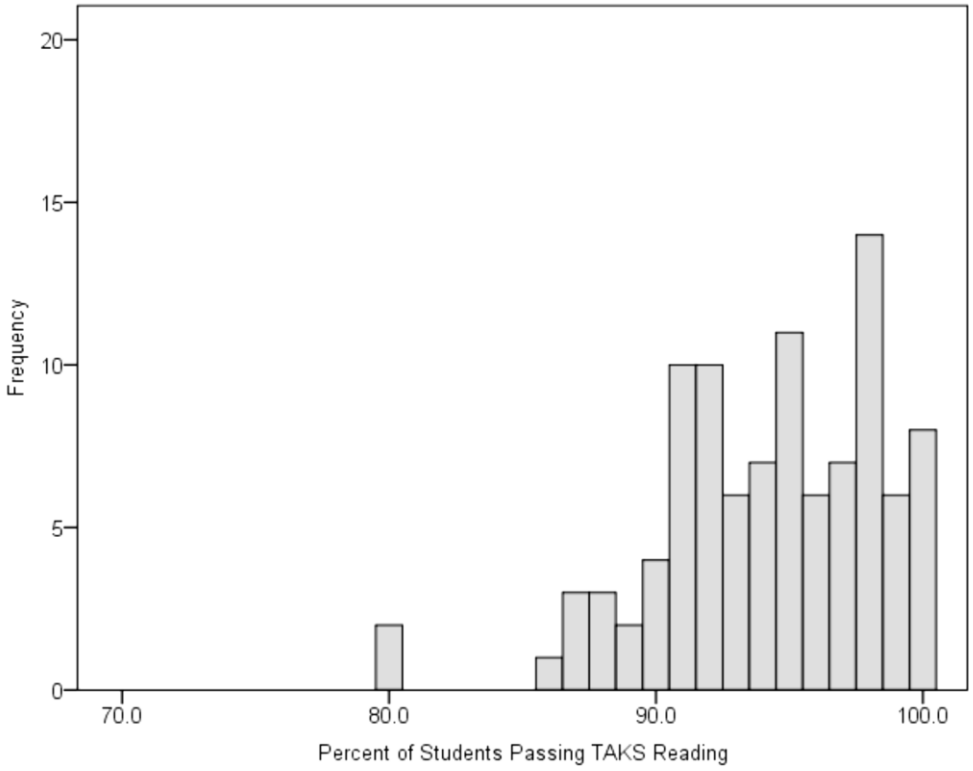

Figure 15. Students Passing Reading TAKS. Histogram showing frequency distribution.

Mediating Variables

A moderator is a variable that affects the strength and/or direction of the relation between an independent and dependent variable (i.e. moderating variables are based on categorical classification such as gender, ethnic background or whether study participants experience a particular treatment or program; Baron & Kenny, 1986). In contrast, a mediator variable accounts for the relation between the predictor and the outcome. Mediating variables are on a continuous scale of measurement (Baron & Kenny, 1986). A variable mediates the relationship between other variables when the relationship between the original variable, where the path originates, to the variable at the end of the causal chain (i.e. where the causal flow terminates) is reduced by an intervening (mediating)

variable being included in the equation (Jose, 2013). Mediating variables are used in the present study because of the level of measurement of the variables being on an interval or continuous level. For example, Percent of LEP students, Percent Students of Color, Percent of Economically Disadvantaged Students, and Number of Benchmarks are all mediating variables in this study. Figure 11 illustrates statistical mediation. In Figure 11, the variables X and Y are mediated by variable M. Error terms are associated with variables Y and M and these errors represent the part of the variables left unexplained due to random influences.

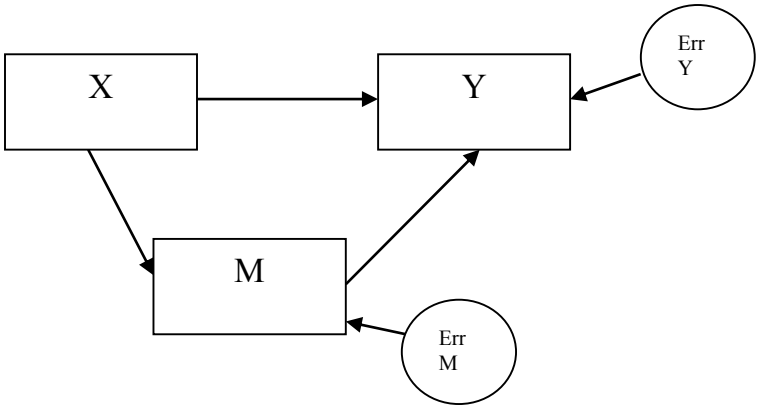

Figure 16. Sample Statistical Mediation. Sample statistical mediation. ErrY and ErrM = error terms.

Path Analysis

Regression analysis is a statistical method for estimating the relationships among variables. Multiple regression analysis includes multiple techniques for analyzing the relationship between a dependent variable and one or more independent variables (Whitley & Kite, 2012). It used to (a) understand which independent variable or variables are related to the dependent variable and (b) to develop regression and prediction equations. In multiple regression there is only one dependent variable (Keith, 2006).

Multiple regression analysis does not explain *how* variables relate to each other; only whether variables are related and to what degree. Explanation of the underlying the relationships in multiple regression analysis is guided by a theoretical framework. The theoretical framework can be new or existing. Path analysis is an extension of multiple regression used to determine if a multivariate, that is more than one dependent variable and more than one independent variable, set of nonexperimental data fits with a particular model (Pedhazur, 1982). Path analysis goes beyond regression by allowing for the analysis of complicated models designed to address complex questions about social systems (Streiner, 2005). Path analysis provides a graphic representation of the researcher's assumed theory and associated research questions, known as a path diagram. Path analysis can estimate the effect one variable has on another, specifically if the relationship is positive, negative, or not supported by the data (Keith, 2006). The correlation between variables can be partitioned into direct causal effects, indirect causal effects, and non-causal effects. Although the term "causal" is prevalent in the path analysis literature, the nature of causality is complex and statements about any causal effects depend on the research design used. For example, the present study is non-experimental, based on observational data already in existence and therefore is explanatory in nature. To this end, causal or cause and effect statements are not possible to make without further research into the possible mechanisms generating the patterns of relationships in the observed data.

In path analysis, a direct effect is the influence on one variable on another variable, with a single headed arrow representing that effect. In regression terminology, the variable where the arrow terminates is the Y (dependent or endogenous) variable. The

arrow where the arrow originates is the X (independent or exogenous) variable. The regression relationship is expressed as "the regression of Y on X" (see Y and X in Figure 11). An indirect, or mediating, effect is facilitated by at least one intervening variable. A double-headed arrow signifies a correlation between variables (Streiner, 2005). In the path analysis, rectangles represent observed, or manifest variables and a circle with an arrow pointing to a dependent variable is the error term (Keith, 2006). Path analysis allows for more complicated and realistic models than multiple regression and was the reason for its use in this study. An assumption not necessarily required to be included in the analyses. In the present model is that the variables of AYP Status, District Accountability Rating and District Type are allowed to correlate in both the reading and math analyses. The error terms 1, 2, and 3 also correlate. Figures 17 and 18 illustrate the path diagrams, one focusing on Math TAKS scores and the other on Reading TAKS scores, for this study. Next, the research questions that guide the study are presented.

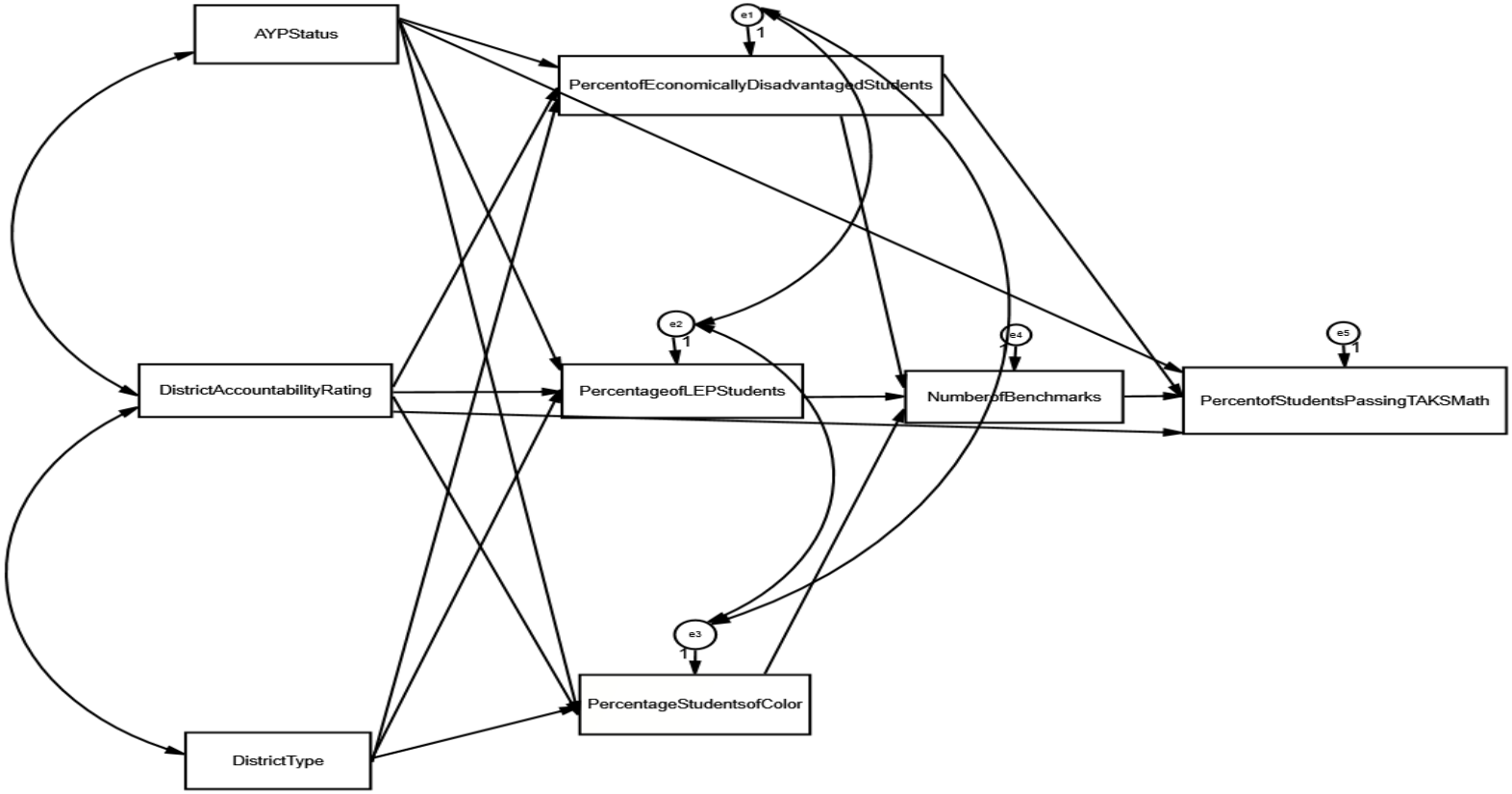

Figure 17. Math Conceptual Path Model. This conceptual path model depicts the structural equation model developed for the study for Math TAKS scores. Arrows drawn between variables specify mediating effects. The rectangles represent observed variables. Arrows point to the dependent variable on the right, the percent of students passing TAKS

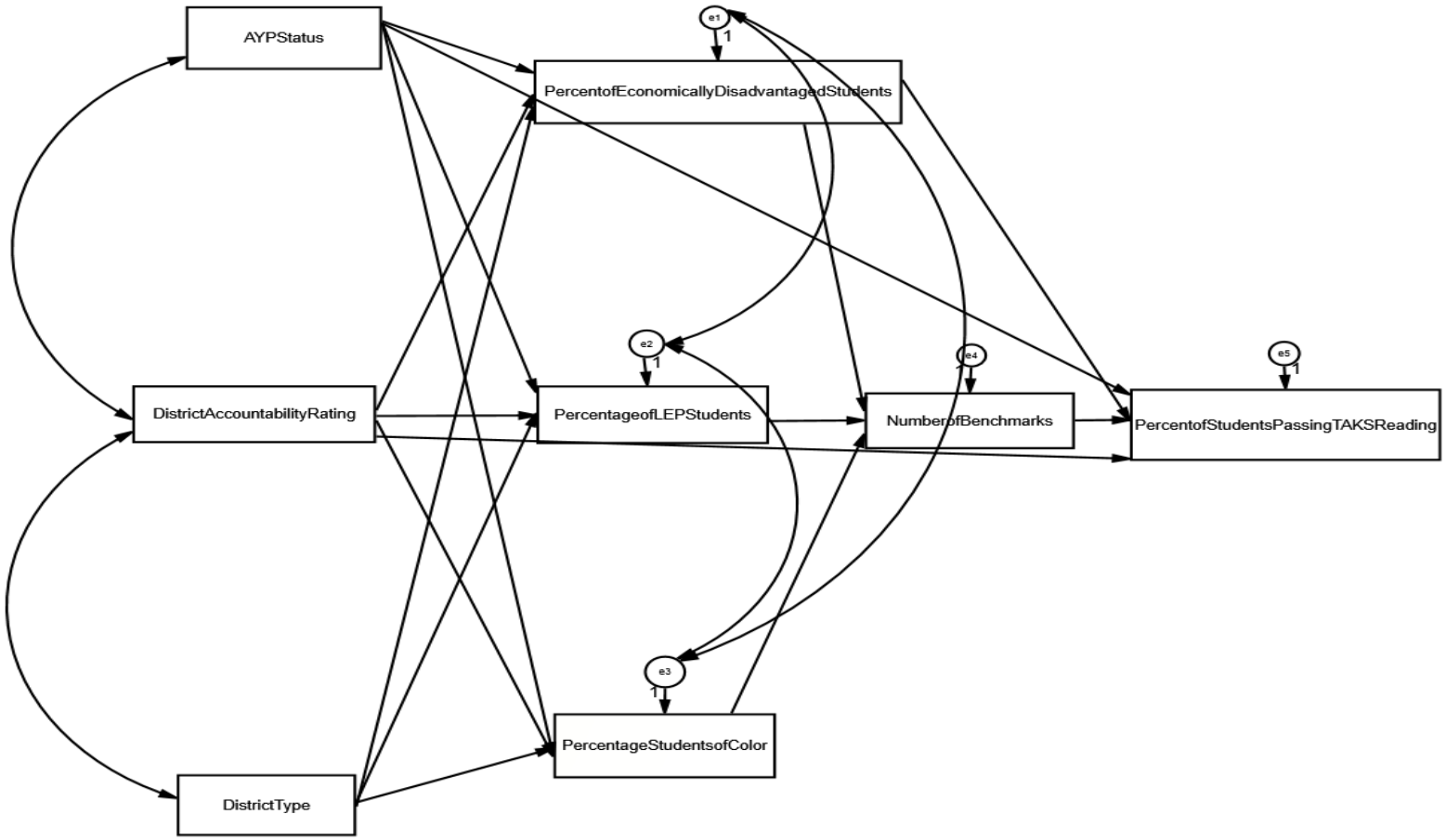

Figure 18. Reading Conceptual Path Model. This conceptual path model depicts the structural equation model developed for the study for Math TAKS scores. Arrows drawn between variables specify mediating effects. The rectangles represent observed variables. Arrows point to the dependent variable on the right, the percent of students passing TAKS.

Primary research questions:

1. Do significant relationships (i.e regression weights) exist between a district's descriptive factors and benchmark testing practices?
2. Does a significant relationship (i.e regression weight) exist between the number of benchmark tests a district requires and the percentage of students passing the TAKS test?

Supporting questions:

1. Is the effect of AYP Status on the number of benchmark tests mediated by the percentage of economically disadvantaged students in the district?
2. Is the effect of AYP Status on the number of benchmark tests mediated by the percentage of students of color in the district?
3. Is the effect of AYP Status on the number of benchmark tests mediated by the percentage of Limited English Proficient (LEP) students in the district?
4. Is the effect of a district's TEA state accountability rating on the number of benchmark tests given mediated by the percentage of economically disadvantaged students in the district?
5. Is the effect of a district's TEA state accountability rating on the number of benchmark tests mediated by the percentage of students of color in the district?
6. Is the effect of a district's TEA state accountability rating on the number of benchmark tests mediated by the percentage of Limited English Proficient (LEP) students in the district?
7. Is the effect of TEA district type on the number of benchmark tests mediated by the percentage of economically disadvantaged students in the district?

8. Is the effect of TEA district type on the number of benchmark tests mediated by the percentage of students of color in the district?
9. Is the effect of TEA district type on the number of benchmark tests mediated by the percentage of Limited English Proficient (LEP) students in the district?
10. Is the effect of TEA district type on the percent of students passing the Math TAKS test mediated by the percentage of economically disadvantaged students, percentage of LEP students, percentage of students of color, and number of benchmarks?
11. Is the effect of TEA district type on the percent of students passing the Reading TAKS test mediated by the percentage of economically disadvantaged students, percentage of LEP students, percentage of students of color, and number of benchmarks?
12. Is the effect of AYP Status on the percent of students passing the Math TAKS test mediated by the percentage of economically disadvantaged students, percentage of LEP students, percentage of students of color, and number of benchmarks?
13. Is the effect of AYP Status on the percent of students passing the Reading TAKS test mediated by the percentage of economically disadvantaged students, percentage of LEP students, percentage of students of color, and number of benchmarks?
14. Is the effect of a district's TEA state accountability rating on the percent of students passing the Math TAKS test mediated by the percentage of economically

disadvantaged students, percentage of LEP students, percentage of students of color, and number of benchmarks?

15. Is the effect of a district's TEA state accountability rating on the percent of students passing the Reading TAKS test mediated by the percentage of economically disadvantaged students, percentage of LEP students, percentage of students of color, and number of benchmarks?
16. What is the relationship between number of benchmarks a district requires and the percent of students passing Math TAKS?
17. What is the relationship between the number of benchmarks a district requires and the percent of students passing Reading TAKS?
18. How does the organizational theory of isomorphism help to explain the use of benchmark testing in school districts and is organizational theory congruent with the path model results observed in this study?

Assessing Overall Model Fit

The aim of structural equation modeling (SEM) is to posit a hypothetical model that produces an accurate fit to a set of observed data. Path analysis is a restricted case of SEM where only observed (manifest) variables are used; there are no unobserved or latent variables. Evaluating the efficacy of the model-data fit is central to structural equation modeling and path analysis. Table 6 provides a set of descriptive measures of fit, also known as goodness-of-fit measures, that assess discrepancies between the covariance matrix $(\mathbf{S})$ generated based on the observed data compared to the predicted matrix $(\hat{\mathbf{\Sigma}})$ implied by the path model. In the present study analyses, both models used an 8 x 8 variable multivariate covariance matrix based on the variables in Table 5. Due to

the skewed nature of some of the variables in this analysis, bootstrapping was performed to avoid estimation bias of the regression weights in the path analysis inference (Efron & Tibshirani, 1993). In the bootstrapping technique, the variables are allowed to be nonnormally distributed so the distributional shape for a variable can be of any type inference (Efron & Tibshirani, 1993). Descriptive measures of fit serve as monitors for the researcher to inform his or her evaluation of model-data fit or discrepancy. A model should not be accepted or rejected based on a single measure. Instead, a set of fit measures provide a means of viewing the model from multiple perspectives to assess model-data fit relative to accepted tolerances (Bollen, 1989; Garson, 2009).

Table 6

Model Fit Indices

	χ^2	*df*	*p*	CMIN/*df*	CFI	RMSEA	NPAR	AIC	BCC
Math	10.17	7	.18	1.45	.99	.07	29	68.17	73.97
Reading	7.60	7	.37	1.09	.99	.03	29	65.59	71.39

χ^2, df, p, and CMIN/df

The first four statistics include the chi-square statistic and related measures and are evaluated as a group. Generally, a chi-square statistic is observed as significant ($p <. 05$) indicates that a model-data fit is poor and the model should be rejected such as when the observed data matrix is not adequately able to be reproduced by the hypothesized path model matrix. In this case, $\chi^2 = 7.59$ with $df = 7$ and $p = .37$ (Reading) and $\chi^2 = 10.17$ with $df = 7$ and $p = .18$ (Math). The chi-square test was not significant in either the reading analysis ($p = .37$) or math ($p = .20$) analysis indicating an excellent fit of the model to the observed data; the observed data matrix is able to be accurately reproduced

by the hypothesized path model. Typically, the chi-square statistic is problematic in 1) highly complex models; 2) very large sample sizes; or 3) in cases where the assumption of multivariate normality is violated. In the present case, the model is of concern in both the first and third conditions. However, since the chi-square was not rejected the usual concern regarding the sensitivity of the test to be rejected too often is not applicable. The fourth column, CMIN/*df*, is referred to as relative chi-square, this measurement is computed as χ^2/df. Researchers vary in their opinions of the use of CMIN/*df* criteria. Byrne (2001) and Carmines and McIver (1981) recommend cut-off points ranging from two to five for acceptable models (Marsh & Hocevar, 1985; Wheaton, Muthén, Alwin & Summers, 1977). This model produces a relative chi-square of 1.09 for Reading and 1.45 for Math, both values are in acceptable ranges.

CFI

The Comparative Fit Index (CFI) is among those least affected by sample size. The CFI is derived by statistically comparing the hypothesized model to the independence (i.e. the completely uncorrelated) model. CFI values vary from zero to one with values approaching one suggestive of a better fit (Byrne, 2001). The recommended criteria for model acceptance are for 90% of the covariance data to be accounted for by the model. This is represented with a CFI statistic greater than or equal to .9 (Garson, 2009). This model has a CFI of .99 for Reading and .99 for Math, indicating a good fit for both models.

RMSEA

The Root Mean Square Error of Approximation (RMSEA) is one of the most informative measures of fit in structural equation modeling. RMSEA is a population-based measure

of fit that demonstrates the amount of error between the proposed model relative to the theoretical model (Price, Tulsky, Mills, & Weiss, 2002). The goal in path analysis is to evaluate how closely the implied or theoretical model is approximated by the sample data. The RMSEA is a measure that indexes the *closeness* of model-data fit in the population from which the data are drawn. To this end, the RMSEA indexes the amount of error or *misfit* between the model and the data. The index is sensitive to model complexity and is shown per degree of freedom in the proposed model. RMSEA yields a bias toward complex models (Arbuckle, 2008; Garson, 2009). Researchers agree that an RMSEA value less than .05 indicate a good fit, values between .05 and .08 indicate a reasonable fit, values of .08 to .10 indicate mediocre fit, and greater than .10 indicates a poor fit (Byrne, 2009; Hooper, Coughlan & Mullen, 2008). RMSEA value of 0.0 indicates a perfect model fit relative to the population (Byrne, 2009). The RMSEA value for this model is .03 for Reading, showing a good fit and .07 for Math, demonstrating a reasonable fit.

NPAR and AIC

The number of unique parameters to be estimated is represented as NPAR. The present model includes 30 hypothesized distinct parameters consisting of 21 regression weights, 5 co-variances, and 0 intercepts. Kline (2005) explains that each of these items correspond to assumed relationships among observed variables.

Another measure of model-data fit is the Akaike Information Criterion (AIC), based on the χ^2 statistic and incorporates the number of parameters to be estimated. AIC is computed as χ^2 + 2(NPAR). Values of AIC between the hypothesized, saturated, and

independence models represent the best fit (Byrne, 2010; Schumaker & Lomax, 2004). For this model, AIC is 65.60 for Reading and 68.17 for Math.

BCC

The Browne-Cudeck Criterion (BCC) is similar in function to the AIC, but takes sample size and degrees of freedom into account. It is computed as χ^2/n +2k/n-v-2 where n = sample size, v = number of variables, and k = (.5v(v + 1)) – df (Garson, 2009). The BCC statistic is more sensitive to model complexity than AIC because is includes the χ^2 as well as sample size and the number of variables. In this model BCC is 71.39 for Reading and 73.97 for Math. The best fitting model is shown with the smallest number between hypothesized, saturated, and independence models.

Regression Weights

Linear regression allows a researcher to analyze the relationship between a dependent variable and one or more independent variables (Hanneman, Kposowa, & Riddle, 2012; Randolph & Myers, 2013). Single-headed arrows in the path represent linear dependencies expressed as the regression of Y on X. The equation for describing the relationship of Y on X is $Y = \beta_0 + \beta_1 X$, where β_0 and β_1 are parameters. Multiple regression adds more independent variables to the relationship with a single dependent variable (Hanneman, Kposowa, & Riddle, 2012). A positive regression weight demonstrates that for each unit increase in X, the value of Y increases in direct proportion to the size of the regression weight ; a negative regression weight shows a relationship where X decreases when Y is increased (Randolph & Myers, 2013). *p*-values for regression weights tell whether or not a regression weight is statistically significant – it is not the actual weight or if the predictive or regressed relationship is significant. A *p*-value

of greater than 0.05 indicates that a relationship is not statistically significant, *p*-values of 0.05 or slightly less are marginally significant, and *p*-values of less than 0.01 are highly significant (Fisher, 1971). Tables 7 and 8 shows the standardized regression weights and significance levels for both models.

Table 7

Standardized Regression Weights for Math TAKS Model

	Estimate	*p*	Effect Size
Regression Weights			
Pct. LEP < --- AYP Status	-0.22	0.03	small
Pct. St. of Color< --- AYP Status	-0.35	< 0.01	medium
Pct. LEP < --- Dist. Acc. Rating	-0.16	0.11	small
Pct. St. of Color< --- Dist. Acc. Rating	-0.30	< 0.01	small
Pct. LEP < --- Dist. Type	0.38	< 0.01	medium
Pct. St. of Color< --- Dist. Type	0.31	< 0.01	small
Pct. Econ. Dis. < --- AYP Status	-0.41	< 0.01	medium
Pct. Econ. Dis. < --- Dist. Acc. Rating	-0.29	< 0.01	small
Pct. Econ. Dis. < --- Dist. Type	-0.09	0.27	small
No. Benchmarks< --- Pct. Econ. Dis.	-0.41	< 0.01	medium
No. Benchmarks < --- Pct. LEP	0.32	< 0.01	small
No. Benchmarks < --- Pct. St. of Color	0.27	0.05	small
Pct. Pass Math < --- No. Benchmarks	< -0.01	0.98	small
Pct. Pass Math < --- AYP Status	-0.13	< 0.01	small
Pct. Pass Math < --- Dist. Acc. Rating	0.44	< 0.01	medium
Pct. Pass Math < --- Pct. Econ. Dis.	-0.38	0.19	medium

Note. Effect size is based on the **partial correlation** based on the mediated path model and are interpreted as small=.10 (Cohen's *d*=.2 standard deviation units); medium=.3 (Cohen's *d*=.5 standard deviation units); large=.5 (Cohen's *d*=.8 standard deviation units), (Cohen, 1988, p. 80-81). Partial correlation coefficients are the standardized effect after the relationship among all other variables in the model are controlled or accounted for.

Table 8

Standardized Regression Weights for Reading TAKS Model

	Estimate	*p*	Effect Size
Regression Weights			
Pct. St. of Color< --- AYP Status	-0.22	0.31	small
Pct. LEP < --- Dist. Acc. Rating	-0.35	< 0.01	medium
Pct. St. of Color< --- Dist. Acc. Rating	-0.16	0.11	small
Pct. LEP < --- Dist. Type	-0.30	< 0.01	small
Pct. St. of Color< --- Dist. Type	0.38	< 0.01	medium
Pct. Econ. Dis. < --- AYP Status	0.31	< 0.01	small
Pct. Econ. Dis. < --- Dist. Acc. Rating	-0.41	< 0.01	medium
Pct. Econ. Dis. < --- Dist. Type	-0.29	< 0.01	small
No. Benchmarks< --- Pct. Econ. Dis.	-0.09	0.27	small
No. Benchmarks < --- Pct. LEP	-0.41	< 0.01	medium
No. Benchmarks < --- Pct. St. of Color	0.32	0.01	small
Pct. Pass Read < --- No. Benchmarks	0.27	0.05	small
Pct. Pass Read < --- AYP Status	-0.09	0.29	small
Pct. Pass Read < --- Dist. Acc. Rating	-0.03	0.73	small
Pct. Pass Read < --- Pct. Econ. Dis.	0.46	< 0.01	medium
Pct. St. of Color< --- AYP Status	-0.25	0.01	small

Note. Effect size is based on the **partial correlation** based on the mediated path model and are interpreted as small=.10 (Cohen's *d*=.2 standard deviation units); medium=.3 (Cohen's *d*=.5 standard deviation units); large=.5 (Cohen's *d*=.8 standard deviation units), (Cohen, 1988, p. 80-81). Partial correlation coefficients are the standardized effect after the relationship among all other variables in the model are controlled or accounted for.

Standardized Direct, Indirect, and Total Effects

Interpretation of the effects of causal variables within path analysis involves the differentiation of direct, indirect, and total effects (Cheong & MacKinnon, 2012; Alwin & Hauser, 1975). According to Cohen, Cohen, and West (2003), effect size is "a measure of the magnitude of a relationship" (p. 5). The standardized form of the coefficients allows meaningful judgments about effects of each variable on the dependent outcome to be made (Loehlin, 2007). Effect sizes can be described as small, medium, or large. In regression analyses, an effect size of 0.2 is small, 0.5 is medium, and > 0.8 is a large effect (Cohen, 1988). Effect sizes can be either positive or negative with a range of between -1.0 to +1.0; higher values are indicative of stronger effects. A positive correlation predicts that high values of the first variable will be found with high values of the second variable and low values of the first variable will be found with low values of the second variable (Randolph & Myers, 2012). A negative correlation predicts that high values of the first variable will be found with low values of the second variable and low values of the first variable will be found with high variables of the second variable (Randolph & Meyers, 2012). The statistical significance, *p*-value, determines if the effect size is significant; *p* -values range from 0 to 1. The lower the p-value, the greater the significance. A *p*-value of < 0.05 is considered significant and < 0.01 is highly significant (Boslough, 2012).

The direct effect is the influence of causal variables on other variables (Cheong & MacKinnon, 2012; Sobel, 1987). The present study is not concerned with the direct effects. The indirect effect is the effect of one variable on another intervened by at least one additional variable (Cheong & MacKinnon, 2012; Jose, 2013). The strength of the

indirect effect describes how strong of a moderation effect was achieved (Jose, 2013). The total effect is the sum of the direct and indirect effects.

Bootstrap Methods

The bootstrap is a data-based simulation method for statistical inference (Efron & Tibshirani, 1993) with broad applications for studying the underlying distribution for a variable or set of variables. The bootstrap method is a statistical method that simulates artificial data sets by resampling from the original data set (Chernick, 2008; Godfrey, 2009). Bootstrap tests are derived by repeatedly drawing random samples from the population data used for the study (Godfrey, 2009; Zieffler & Long, 2011).Bootstrap sample distribution has approximately the same shape and spread of the original data. In this study, the bootstrap method is particularly useful because of the non-normal distribution for several variables. For example, the using the bootstrap, the distributions are allowed to retain their original shape and statistical tests of significance are not expected to conform to the assumption of normality. The bootstrap method allows the marginal distribution to change with each replicate data set (Zieffler & Long, 2011). Bootstrap methods allow the researcher to measure the accuracy or validity of an estimator parameter (Ross, 2009). Bootstrapping methods are very useful in studying the estimates of indirect effects in mediation models (Shrout & Bolger, 2002). The use of bootstrap methods provides a high level of precision of confidence intervals, regardless of sample size or effect size (MacKinnon, Lockwood, & Williams, 2004; Mallinckrodt, Abraham, Wei, & Russell, 2006). The *p*-value of significance for indirect (mediation) effects can be estimated using bootstrapping. To this end, this study used boostrapping to estimate the two-tailed *p*-values and confidence intervals for the indirect effects in both

analyses. One-thousand (1,000) bootstrapped samples were generated for the math and reading analyses respectively. Tables 6 and 7 summarize the standardized total and indirect effects for each model along with point estimates for regression weights, *p*-values and confidence intervals.

Table 9

Effects for Math Model

	Standardized Total Effects			**Standardized Indirect Effects**		
	Weight (Sig.)	Lower 95%	Upper 95%	Weight (Sig.)	Lower 95%	Upper 95%
No. Benchmarks <--Dist. Type	0.24 (< 0.01)	0.13	0.38	0.24 (< 0.01)	0.13	0.38
Pct. Pass Math<--Dist. Type	0.03 (0.36)	-0.04	0.12	0.03 (0.36)	-0.04	0.12
No. Benchmarks	-0.02 (0.64)	0.13	0.10	-0.02 (0.64)	-0.13	0.10
Pct. Pass Math<--Dist	0.55 (< 0.01)	0.40	0.70	0.11 (0.02)	0.02	0.21
No. Benchmarks <--AYP	< 0.01 (0.91)	-0.11	0.12	< 0.01 (0.91)	-0.11	0.12
Pct. Pass Math<--AYP Status	0.03 (0.81)	-0.17	0.21	0.16 (< 0.01)	0.07	0.28
Pct. Pass Math <--- No.	< -0.01 (0.93)	-0.11	0.12	0.00 (0.91)	0.00	0.00

Table 10

Effects for Reading Model

	Standardized Total Effects			Standardized Indirect Effects		
	Weight (Sig.)	Lower 95%	Upper 95%	Weight (Sig.)	Lower 95%	Upper 95%
No. Benchmarks <--Dist. Type	0.24 (< 0.01)	0.13	0.38	0.24 (< 0.01)	0.13	0.38
Pct. Pass Reading<--Dist. Type	< 0.01 (0.36)	-0.06	0.08	< 0.01 (0.80)	-0.06	0.08
No. Benchmarks <--Dist. Acc. Rating	-0.02 (0.64)	-0.13	0.10	-0.02 (0.64)	-0.13	0.10
Pct. Pass Reading<--Dist. Acc. Rating	0.53 (< 0.01)	0.37	0.68	0.07 (0.02)	0.00	0.16
No. Benchmarks <--AYP Status	< 0.01 (0.91)	-0.11	0.12	< 0.01 (0.91)	-0.11	0.12
Pct. Pass Reading<--AYP Status	0.07 (0.81)	-0.16	0.32	0.10 (0.02)	< 0.01	0.25
Pct. Pass Reading <--- No. Benchmarks	-0.09 (0.93)	-0.26	0.11	0.00 (0.91)	0.00	0.00

Evaluation of the Research Questions

Due to the large number of similarities between the Math TAKS and Reading TAKS models, the evaluation of the research questions will be combined.

Primary Research Question 1

1. Do significant relationships (i.e regression weights) exist between a district's descriptive factors and benchmark testing practices?

In the present study, the standardized regression weight of percent of economically disadvantaged students on the number of benchmarks is -0.41, a medium effect size, with a significance of $p < 0.01$, for both models. This indicates a negative relationship that is highly statistically significant. The negative relationship means the greater the percent of economically disadvantaged students in a district, the fewer benchmarks the district requires.

There is a significant relationship between percent of LEP students and number of benchmarks with a standardized regression weight of 0.32, a small effect size, and a significance of $p < 0.01$ for both Math and Reading TAKS models. This indicates the greater the percent of LEP students in a district, the more benchmark tests the district requires.

The standardized regression weight of percent of students of color on the number of benchmarks 0.27, a small effect size, for Math and Reading with a significance of $p = 0.05$, marginally significant. This positive standardized regression weight indicates the greater the percent students of color in a district, the greater number of benchmarks the district requires.

Primary Research Question 2

2. Does a significant relationship (i.e regression weight) exist between the number of benchmark tests a district requires and the percentage of students passing the TAKS test?

In both models there is a negative correlation between the Number of Benchmarks and the Percent of Students Passing TAKS. The standardized regression weight of number of benchmarks on percent of students passing Math TAKS is < -0.01, a small effect size, with a significance of $p = 0.98$. The standardized regression weight of number of benchmarks on percent students passing Reading TAKS is -0.09, a small effect size, with a significance of $p = 0.29$. The p-values for both models are not statistically significant. Standardized regression weights in both models are negative, which indicates the greater the number of benchmarks, the lower the percentage of students passing Math or Reading TAKS.

Supporting Questions 1-3

1. Is the effect of AYP Status on the number of benchmark tests mediated by the percentage of economically disadvantaged students in the district?
2. Is the effect of AYP Status on the number of benchmark tests mediated by the percentage of students of color in the district?
3. Is the effect of AYP Status on the number of benchmark tests mediated by the percentage of Limited English Proficient (LEP) students in the district?

The paths leading from AYP statuses to number of benchmarks are mediated in three ways (see Figures 17 and 18). First, the effect of AYP status on number of benchmarks is mediated (i.e. the standardized path weights expressed as indirect effects)

by percent of economically disadvantaged students. Second, the effect of AYP status on number of benchmarks is mediated by percent LEP students. Third, the effect of AYP status on number of benchmarks is mediated by percent students of color. The standardized total effect of AYP status on number of benchmarks is < 0.01, a small effect size, with a significance of $p = 0.91$ for both math and reading models. The standardized total effect is small and not significant. The standardized indirect effect of AYP status on number of benchmarks is small at < 0.01 with a significance of $p = 0.91$. This is small and not statistically significant so AYP status does not help explain the number of benchmarks given in a district. The paths related to questions 1-3 are highlighted in bold.

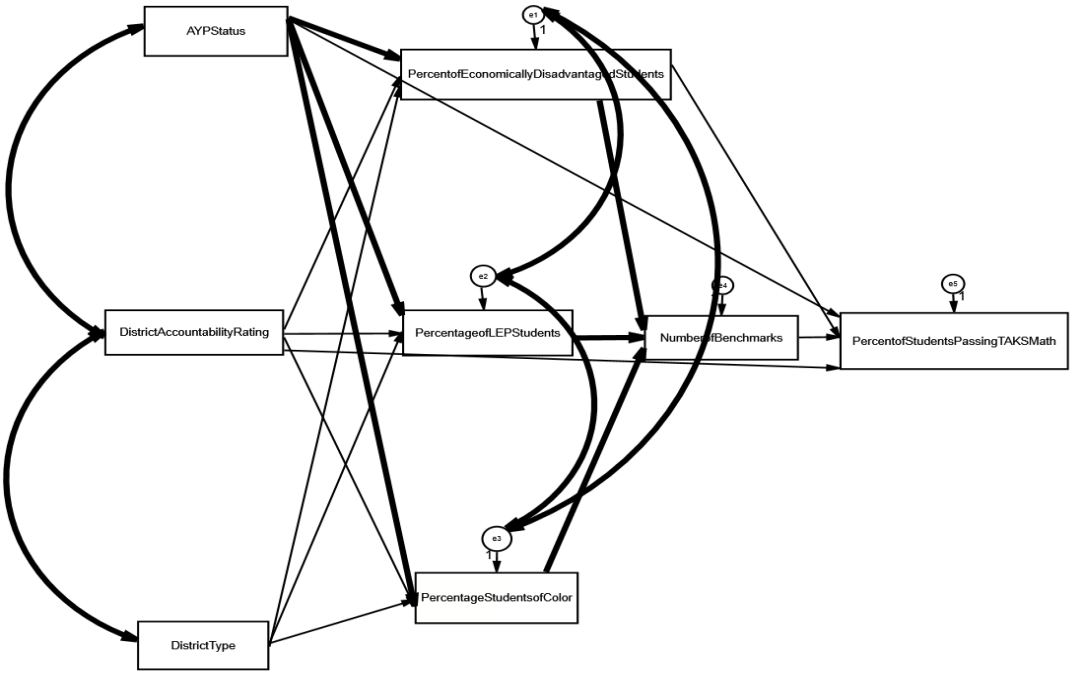

Figure 19. Math Conceptual Model. This conceptual path model depicts the structural equation model developed for the study for Math TAKS scores. The bold paths relate to supporting questions 1-3.

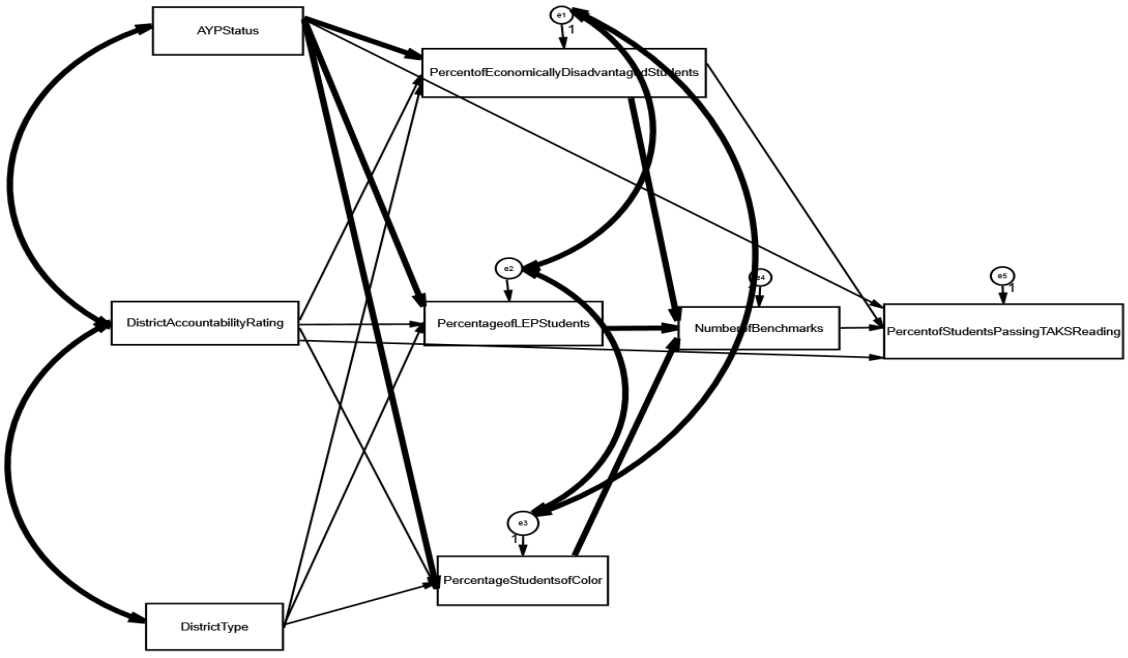

Figure 20. Reading Conceptual Model. This conceptual path model depicts the structural equation model developed for the study for Reading TAKS scores. The bold paths relate to supporting questions 1-3.

Supporting Questions 4-6

4. Is the effect of a district's TEA state accountability rating on the number of benchmark tests given mediated by the percentage of economically disadvantaged students in the district?
5. Is the effect of a district's TEA state accountability rating on the number of benchmark tests mediated by the percentage of students of color in the district?
6. Is the effect of a district's TEA state accountability rating on the number of benchmark tests mediated by the percentage of Limited English Proficient (LEP) students in the district?

All paths leading from state accountability rating to number of benchmark tests are mediated in three ways (see Figures 19 and 20). First, the effect of state accountability rating on number of benchmarks is mediated by percent of economically disadvantaged students. Second, the effect of state accountability rating on number of benchmarks is mediated by percent LEP students. Third, the effect of state accountability rating on number of benchmarks is mediated by percent students of color. The standardized total effect of state accountability rating on number of benchmarks is -0.02, a small effect size, for both models. The significance of the standardized total effects is $p = 0.64$ for both models, a result that is not statistically significant. The standardized indirect effect of state accountability rating is -0.02 with a significance of $p = 0.64$ for both models. This is a small and not significant effect. The effect of TEA state accountability rating does not help account for the number of benchmarks in a district. Figures 21 and 22 highlight the paths related to supporting questions 4-6.

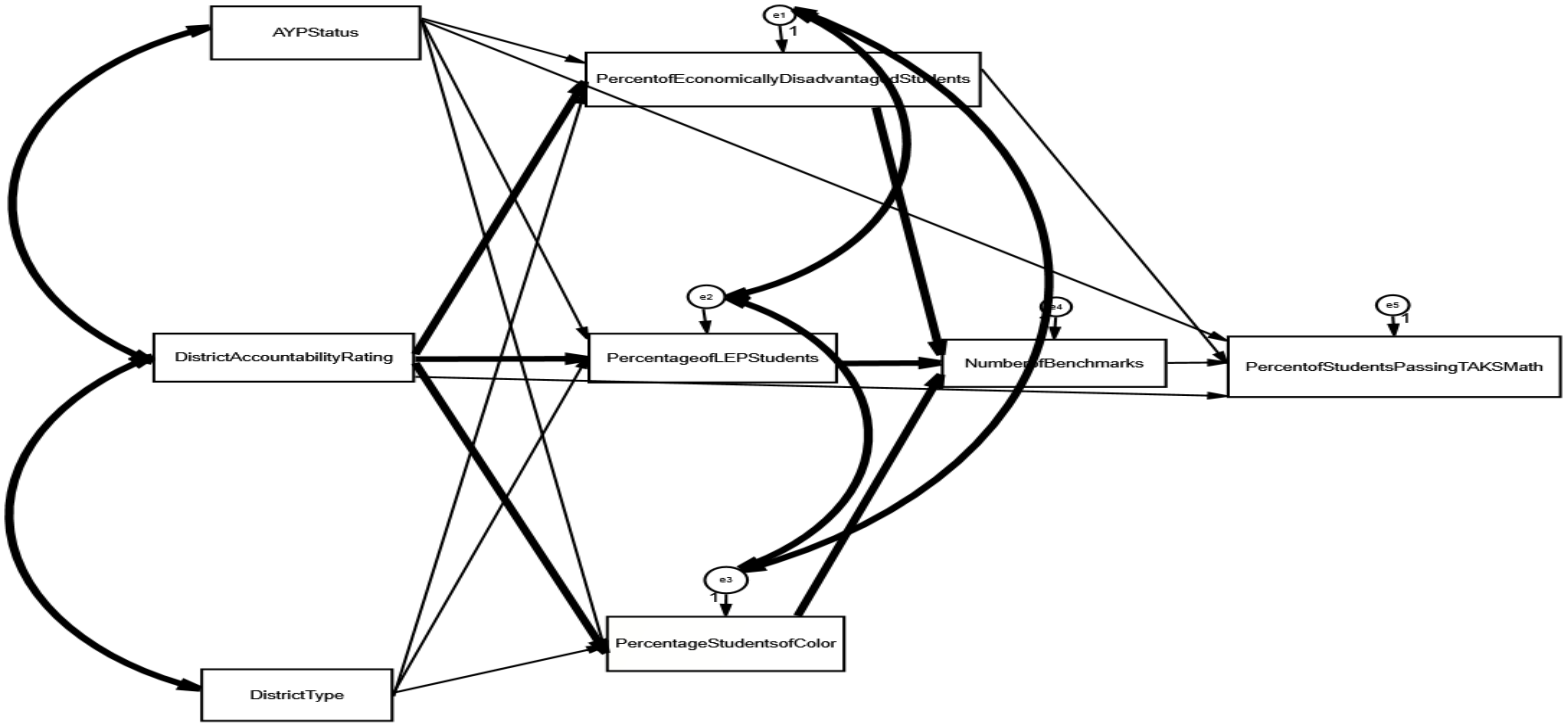

Figure 21. Math Conceptual Path Model. This conceptual path model depicts the structural equation model developed for the study for Math TAKS scores. The bold paths relate to supporting questions 4-6.

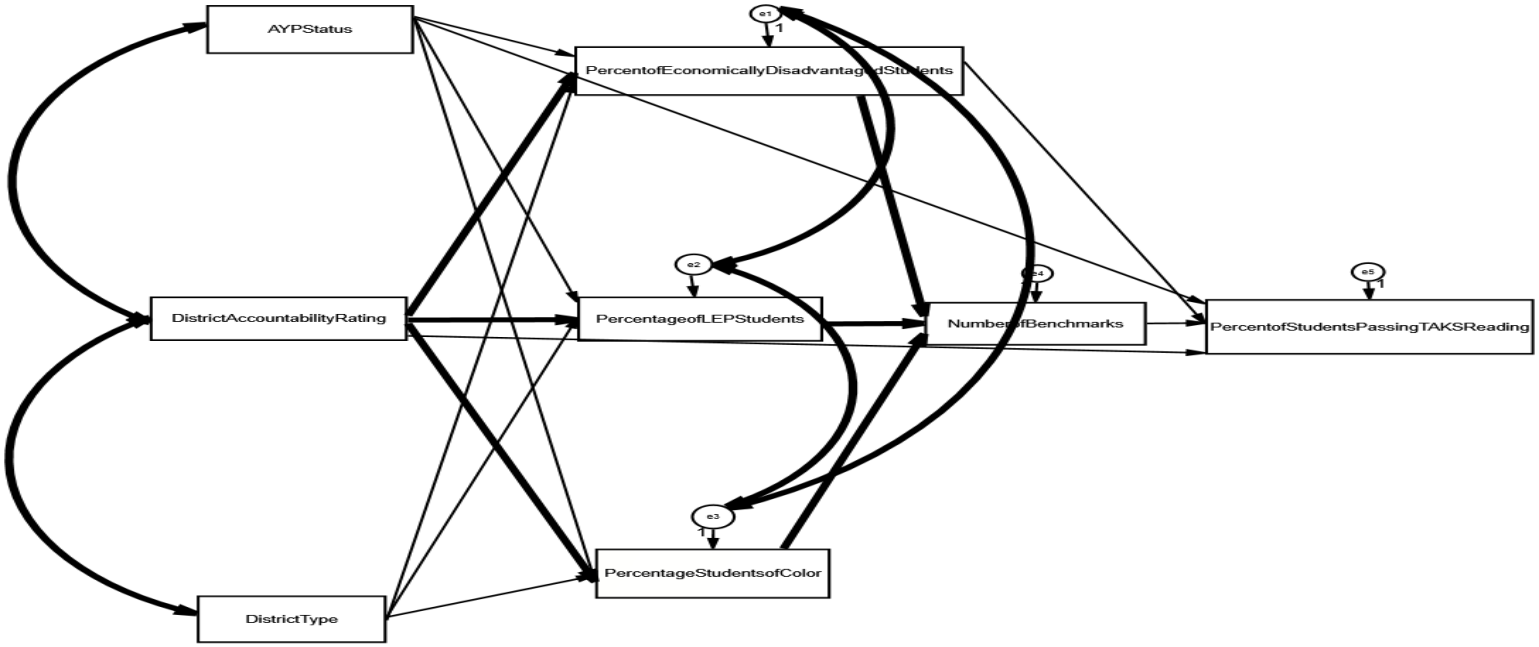

Figure 22. Reading Conceptual Path Model. This conceptual path model depicts the structural equation model developed for the study for Reading TAKS scores. The bold paths relate to supporting questions 4-6.

Supporting Questions 7-9

7. Is the effect of TEA district type on the number of benchmark tests mediated by the percentage of economically disadvantaged students in the district?
8. Is the effect of TEA district type on the number of benchmark tests mediated by the percentage of students of color in the district?
9. Is the effect of TEA district type on the number of benchmark tests mediated by the percentage of Limited English Proficient (LEP) students in the district?

The paths leading from TEA district types to number of benchmarks are mediated in three ways (see Figures 23 and 24). First, the effect of district type on number of benchmarks is mediated by percent of economically disadvantaged students. Second, the effect of district type on number of benchmarks is mediated by percent LEP students. Third, the effect of district type on number of benchmarks is mediated by percent students of color. The mediation effects are derived in consideration of the total effects in the path model. The standardized total effect of district type on number of benchmarks is 0.24, a small effect size, with a significance of $p < 0.01$ for math and reading models. The standardized indirect effect of district type on number of benchmarks is small at 0.24 both math and reading models. The significance of the standardized indirect effect is $p < 0.01$ for both models. The standardized total and indirect effect size is small and significant for both models. Because the *p*-value is significant, TEA district type does help explain the number of benchmarks required by a district. In Figures 23 and 24 the bold paths highlight supporting questions 7-9.

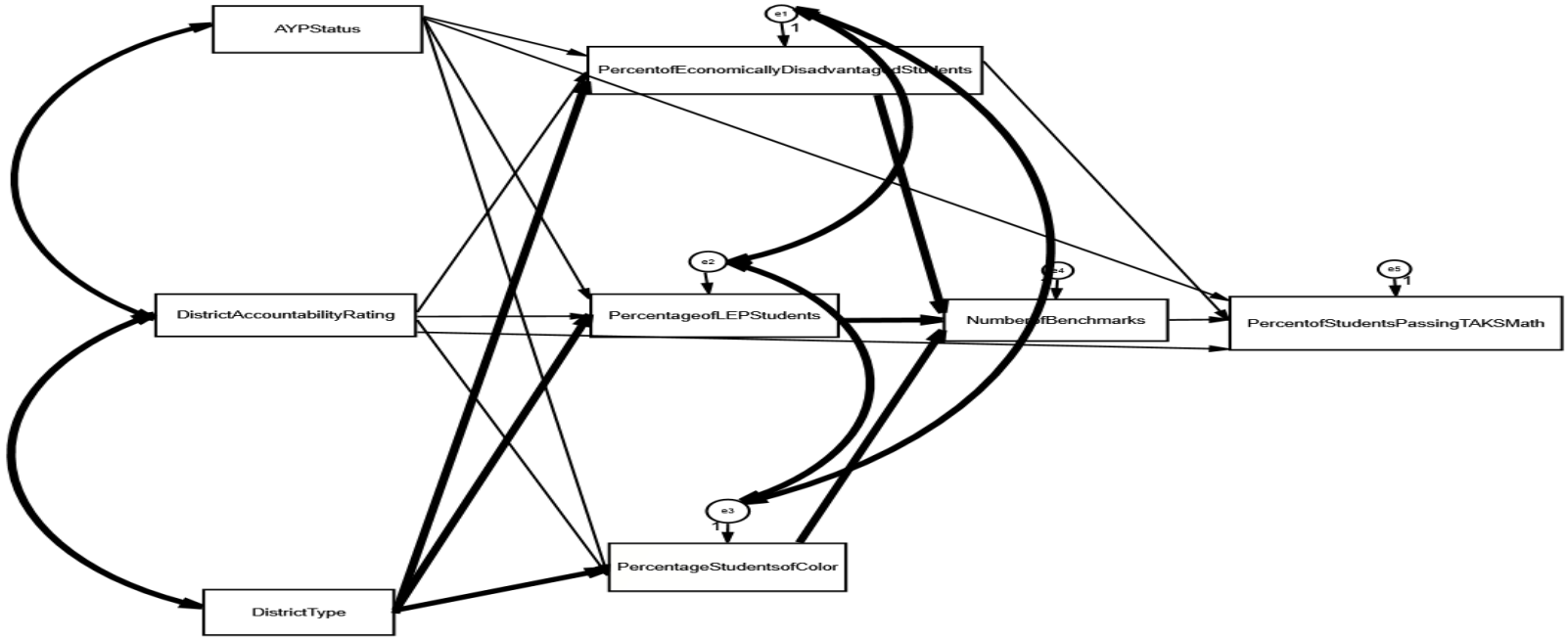

Figure 23. Math Conceptual Path Model. This conceptual path model depicts the structural equation model developed for the study for Math TAKS scores. The bold paths relate to supporting questions 7-9.

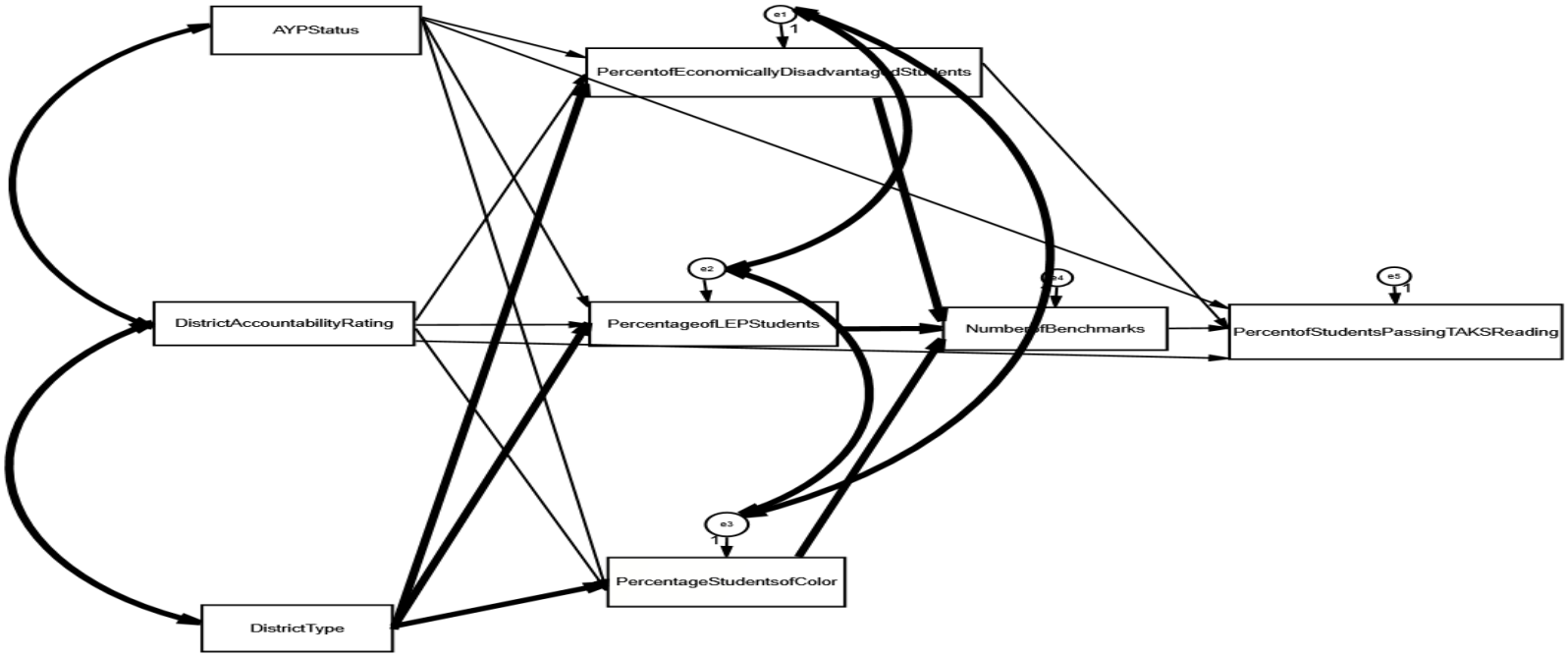

Figure 24. Reading Conceptual Path Model. This conceptual path model depicts the structural equation model developed for the study for Reading TAKS scores. The bold paths relate to supporting questions 7-9.

Supporting Question 10

10. Is the effect of TEA district type on the percent of students passing the Math TAKS test mediated by the percentage of economically disadvantaged students, percentage of LEP students, percentage of students of color, and number of benchmarks?

The paths leading from district type to percent of students passing Math TAKS is mediated in three ways (see Figure 25). First, the effect of district type on percent of students passing Math TAKS is mediated by percent economically disadvantaged students and number of benchmarks across districts. Second, the effect of district type on percent of students passing Math TAKS is mediated by percent LEP students and number of benchmarks across districts. Third, the effect of district type on percent of students passing Math TAKS is mediated by percent students of color and number of benchmarks across districts. The mediation effects are derived in consideration of the total effects in the path model. Both the standardized total effect size and standardized indirect effect size of district type on percent students passing Math TAKS is 0.03, a small effect size, with a significance level of $p = 0.36$. The standardized total and standardized indirect effects of district type on percent of students passing Math TAKS is small and not significant. TEA district type does not relate to the percent of students passing Math TAKS. Figure 25 highlights paths related to question 10.

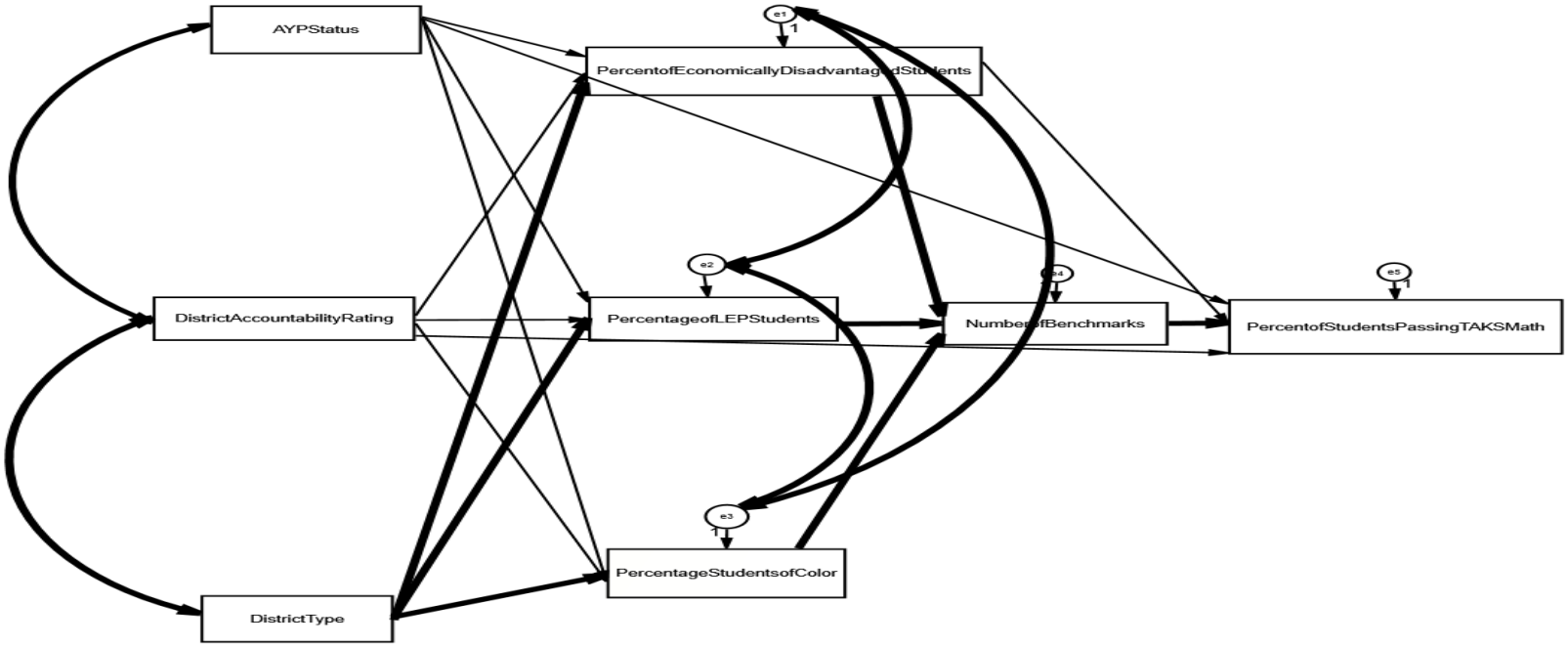

Figure 25. Math Conceptual Path Model. This conceptual path model depicts the structural equation model developed for the study for Math TAKS scores. The bold paths relate to supporting question 10.

Supporting Question 11

11. Is the effect of TEA district type on the percent of students passing the Reading TAKS test mediated by the percentage of economically disadvantaged students, percentage of LEP students, percentage of students of color, and number of benchmarks?

The paths leading from district type to percent of students passing Reading TAKS are mediated in three ways (see Figure 26). First, the effect of district type on percent of students passing Reading TAKS is mediated by percent economically disadvantaged students and number of benchmarks. Second, the effect of district type on percent students passing Reading TAKS is mediated by percent LEP students and number of benchmarks. Third, the effect of district type on percent of students passing Reading TAKS is mediated by percent students of color and number of benchmarks. The standardized total effect of district type on percent of students passing Reading TAKS is small at < 0.01 with a significance of $p = 0.36$. The standardized total effect size is small and not statistically significant. The standardized indirect effect size of district type on percent students passing Reading TAKS is < 0.01, a small effect size, with a significance of $p = 0.80$. The standardized indirect effect is also small and not significant. TEA district type does not help explain the percent of students passing Reading TAKS in a district. Figure 26 highlights paths related to question 11.

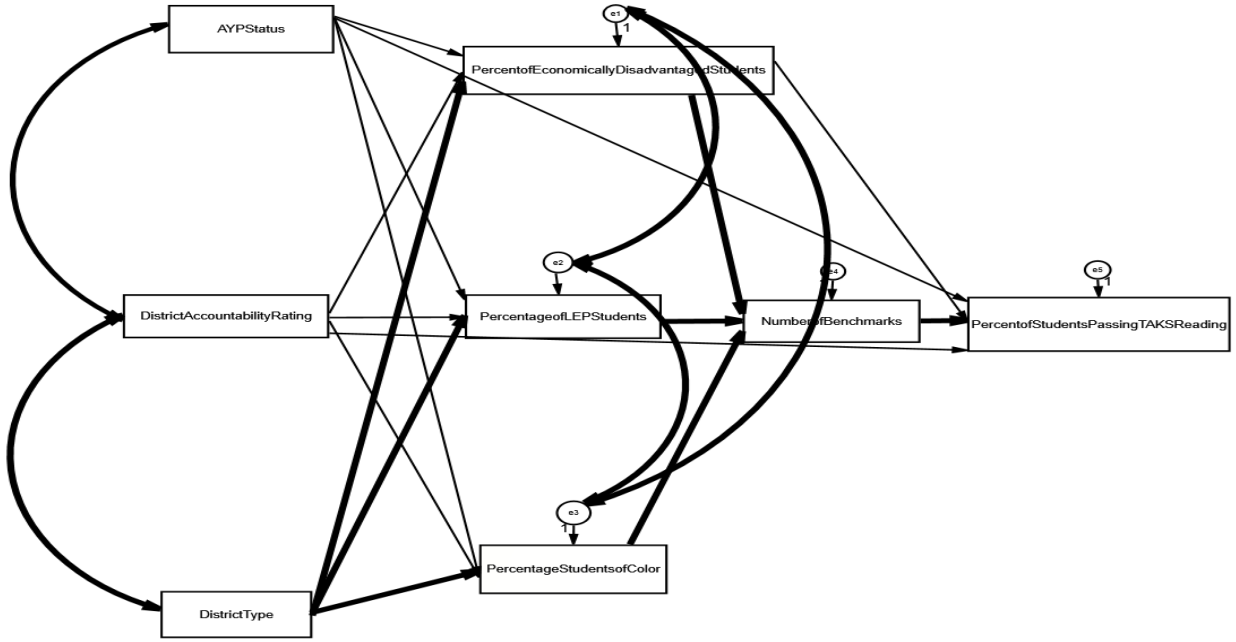

Figure 26. Reading Conceptual Path Model. This conceptual path model depicts the structural equation model developed for the study for Reading TAKS scores. The bold paths relate to Supporting Question 11.

Supporting Question 12

12. Is the effect of AYP Status on the percent of students passing the Math TAKS test mediated by the percentage of economically disadvantaged students, percentage of LEP students, percentage of students of color, and number of benchmarks?

All paths leading from AYP status to percent of students passing Math TAKS are mediated in three ways (see Figure 27). First, the effect of AYP status on percent of students passing Math TAKS is mediated by percent economically disadvantaged students and number of benchmarks. Second, the effect of AYP status on percent of students passing Math TAKS is mediated by percent LEP students and number of benchmarks. Third, the effect of AYP status on percent of students passing Math TAKS is mediated by percent students of color and number of benchmarks. The standardized total effect of AYP status on percent passing Math TAKS is 0.03, a small effect size, with a significance level of $p = 0.91$. This shows a small effect size that is not statistically significant. The standardized indirect effect of AYP status on percent students passing Math TAKS is small at 0.16. The significance of the standardized indirect effect is $p < 0.01$. This small, statistically significant effect indicates that a district's AYP status does help explain the number of students passing Math TAKS. Paths related to question 12 are highlighted in Figure 27.

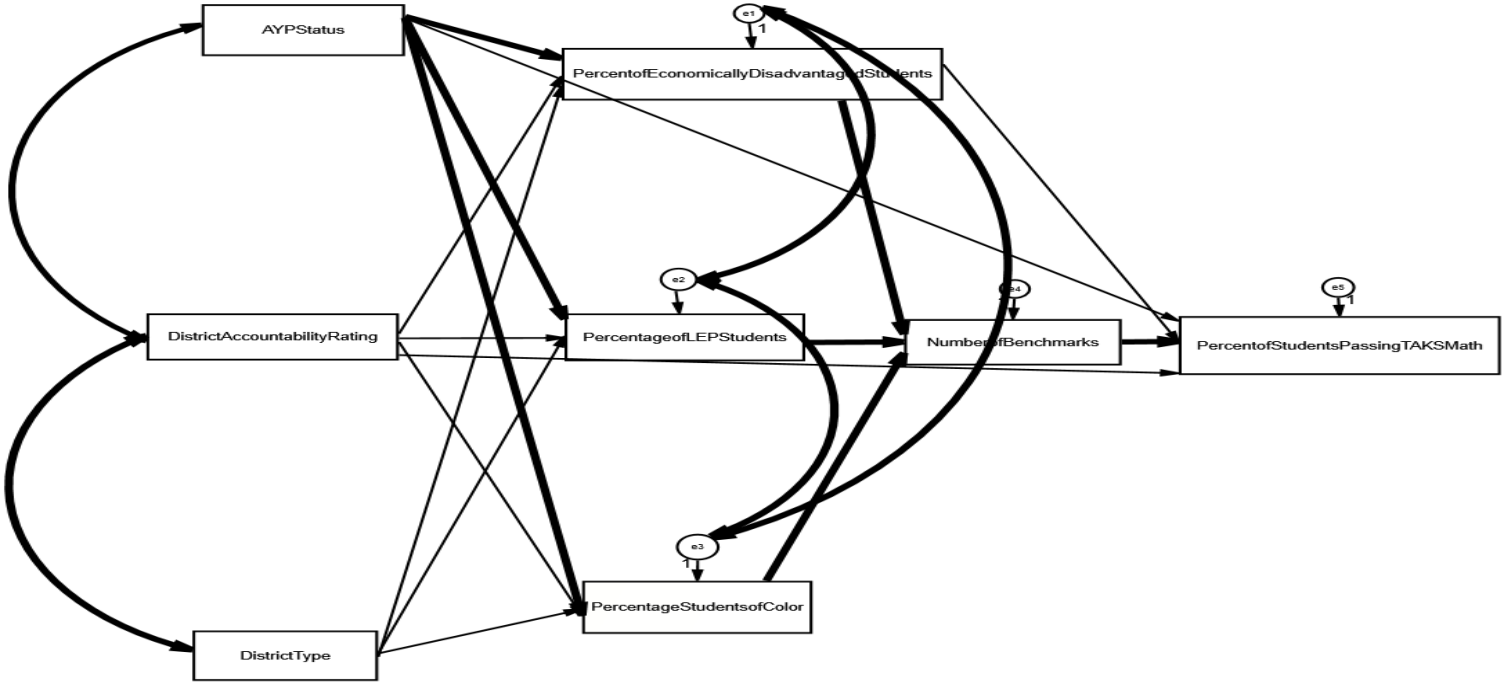

Figure 27. Math Conceptual Path Model. This conceptual path model depicts the structural equation model developed for the study for Math TAKS scores. The bold paths relate to Supporting Question 12.

Supporting Question 13

13. Is the effect of AYP Status on the percent of students passing the Reading TAKS test mediated by the percentage of economically disadvantaged students, percentage of LEP students, percentage of students of color, and number of benchmarks?

The paths leading from AYP status to percent students passing Reading TAKS are mediated in three ways. First, the effect of AYP status on percent students passing Reading TAKS is mediated by percent economically disadvantaged students and number of benchmarks. Second, the effect of AYP status on percent students passing Reading TAKS is mediated by percent LEP students and number of benchmarks. Third, the effect of AYP status on percent students passing Reading TAKS is mediated by percent students of color and number of benchmarks. The standardized total effects for the effect of AYP status on percent students passing Reading TAKS is 0.07, a small effect size. The significance level of the standardized total effect is $p = 0.81$, not statistically significant. The standardized indirect effect of AYP status on percent students passing Reading TAKS is 0.10, a small effect size, with a significance of $p = 0.02$. This is a small effect size that is statistically significant. AYP status does relate to the percent of students passing Reading TAKS. Paths related to question 13 are highlighted in Figure 28.

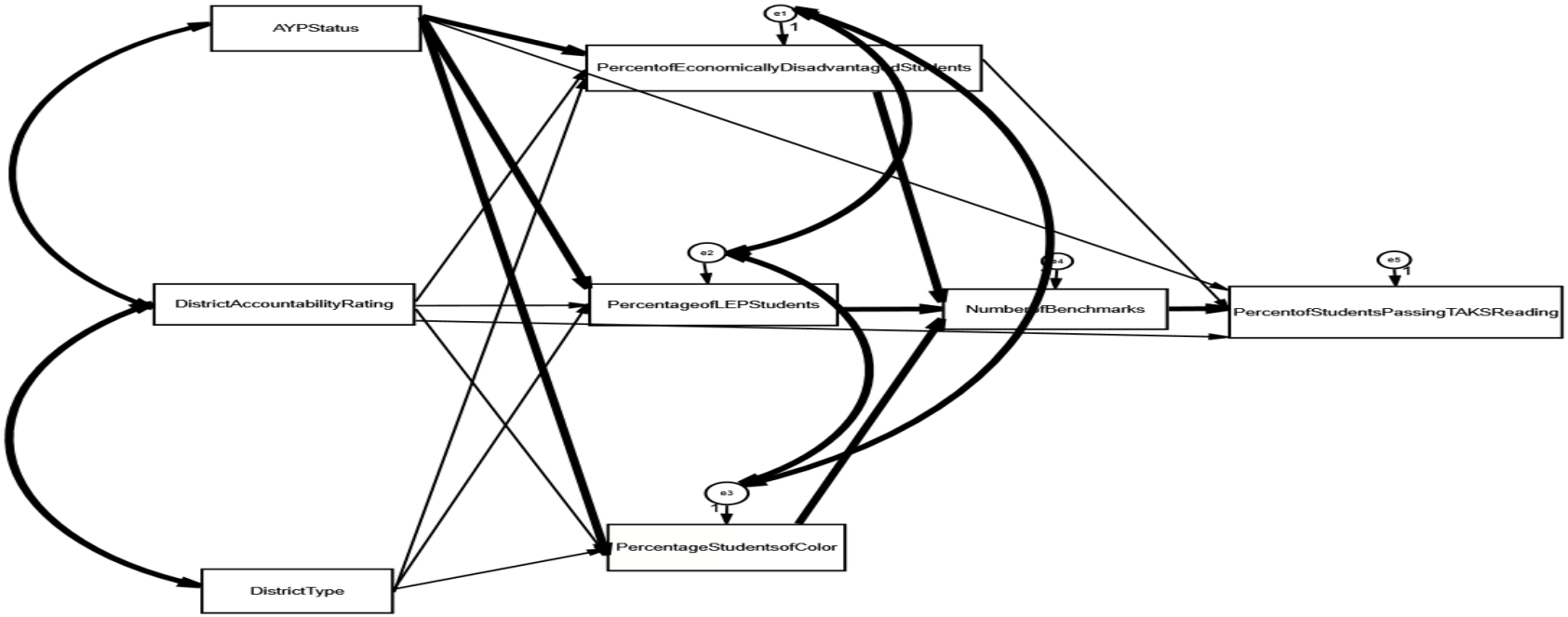

Figure 28. Reading Conceptual Path Model. This conceptual path model depicts the structural equation model developed for the study for Reading TAKS scores. The bold paths relate to Supporting Question 13.

Supporting Question 14

14. Is the effect of a district's TEA state accountability rating on the percent of students passing the Math TAKS test mediated by the percentage of economically disadvantaged students, percentage of LEP students, percentage of students of color, and number of benchmarks?

Paths leading from district accountability rating to percent students passing Math TAKS are mediated in three ways (see Figure 29). First, the effect of district accountability rating on percent passing Math TAKS is mediated by percent economically disadvantaged students and number of benchmarks. Second, the effect of district accountability rating on percent passing Math TAKS is mediated by percent LEP students and number of benchmarks. Third, the effect of district accountability rating on percent passing Math TAKS is mediated by percent students of color and number of benchmarks. The standardized total effect of district accountability rating on percent of students passing Math TAKS is 0.55, a medium effect size. The significance level of the standardized total effect size of district accountability rating on percent passing Math TAKS is $p < 0.01$, highly significant. The standardized indirect effect of district accountability rating on percent passing Math TAKS is small at 0.11 with a significance of $p = 0.02$. The standardized indirect effect is small and not significant. TEA state accountability rating does not help explain the percent of students passing Math TAKS. Paths related to question 14 are highlighted in Figure 29.

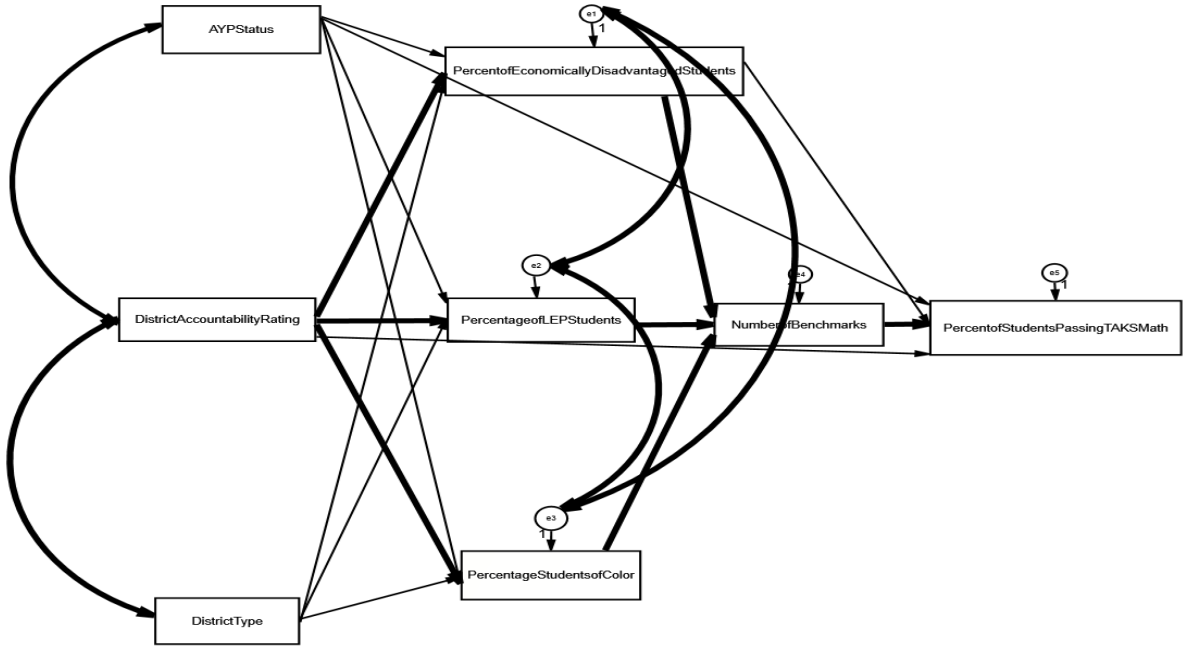

Figure 29. Math Conceptual Path Model. This conceptual path model depicts the structural equation model developed for the study for Math TAKS scores. The bold paths relate to Supporting Question 14.

Supporting Question 15

15. Is the effect of a district's TEA state accountability rating on the percent of students passing the Reading TAKS test mediated by the percentage of economically disadvantaged students, percentage of LEP students, percentage of students of color, and number of benchmarks?

The paths leading from district accountability rating to percent students passing Reading TAKS are mediated in three ways (see Figure 30). First, the effect of district accountability rating on percent passing Reading TAKS are mediated by percent economically disadvantaged students and number of benchmarks. Second, the effect of district accountability rating on percent passing Reading TAKS is mediated by percent LEP students and number of benchmarks. Third, the effect of district accountability rating on percent passing Reading TAKS is mediated by percent students of color and number of benchmarks. The standardized total effect of district accountability rating on percent of students passing Reading TAKS is 0.53, a medium effect size. The significance level of the standardized total effect of district accountability rating on percent passing Reading TAKS is $p < 0.01$, highly significant. The standardized indirect effect of district accountability rating on percent passing Reading TAKS is 0.07, a small effect size, with a $p = 0.02$. This is small and not significant so TEA district type does not help explain the percent of students passing Reading TAKS. Paths related to question 15 are highlighted in Figure 30.

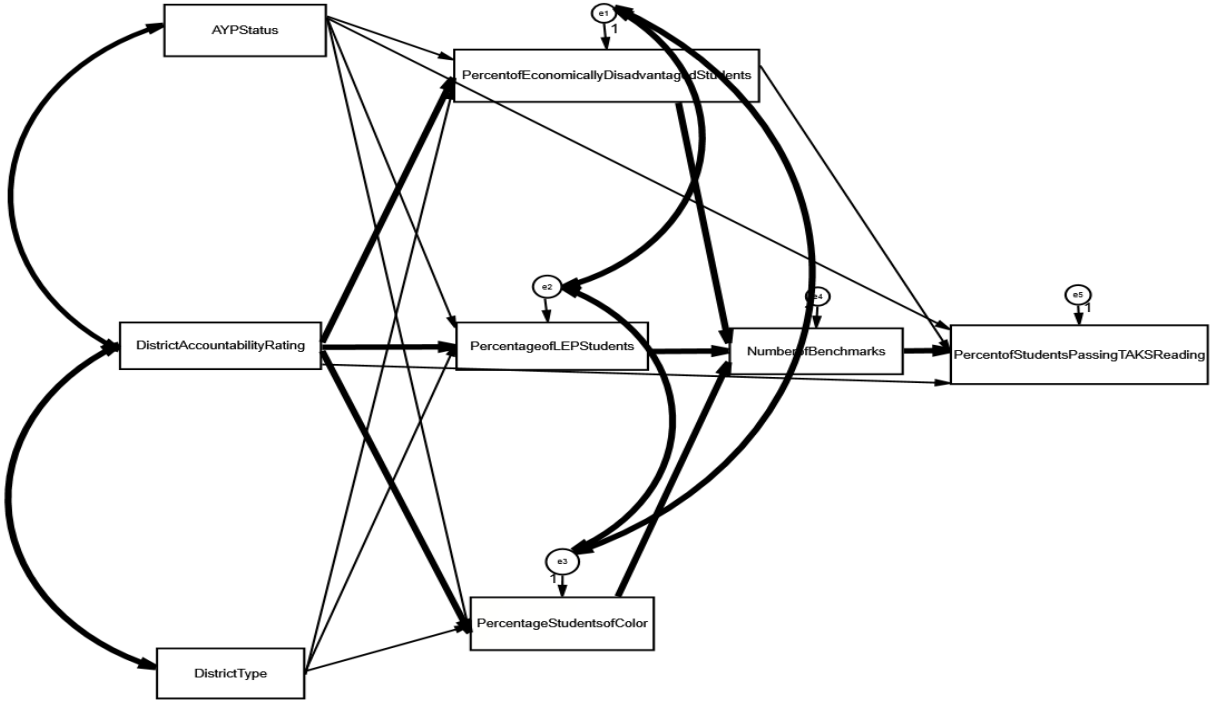

Figure 30. Reading Conceptual Path Model. This conceptual path model depicts the structural equation model developed for the study for Math TAKS scores. The bold paths relate to Supporting Question 15.

Supporting Question 16

16. What is the relationship between number of benchmarks a district requires and the percent of students passing Math TAKS?

The path from number of benchmarks to percent of students passing Math TAKS has a standardized total effect size of < -0.01, a small effect size, with a significance level of $p = 0.93$. This is a small effect size that is not statistically significant. The standardized indirect effect of number of benchmarks on percent of students passing Math TAKS is small, 0.00. The significance of the standardized indirect effect is $p = 0.91$, not significant. Because the *p*-value is not significant, the number of benchmarks a district requires does not relate to the percent of students passing Math TAKS.

Supporting Question 17

17. What is the relationship between the number of benchmarks a district requires and the percent of students passing Reading TAKS?

The path from number of benchmarks to percent of students passing Reading TAKS has a standardized total effect size of -0.09, a small effect size, with a significance level of $p = 0.93$. This standardized total effect size is small and not statistically significant. The standardized indirect effect of number of benchmarks on percent students passing Reading TAKS is 0.00, a small effect. The significance of the standardized indirect effect is $p = 0.91$, not statistically significant. The large *p*-value indicates that the number of benchmarks a district requires does not relate to the percent of students passing Reading TAKS.

Limitations of the Study

The present study relied upon districts self-reporting benchmark testing data. Districts that did not respond to the request for benchmark information were not included in the study. It is possible that this self-reporting affected the results. The number of benchmarks given in a district varied from 0 to 28 benchmark tests, with a mean of 4.83 benchmark tests. This data was skewed with a Z-score of 7.28 that was significant at $p < 0.05$. In the present study, 40% of the districts reported zero benchmark tests required per school year. It is possible that districts requiring multiple benchmark tests did not respond to the researcher's request for information.

The present study requested school districts' testing calendar of required benchmark or practice assessments. The literature shows the terms "practice test" and "benchmark test" are used synonymously in the literature since NCLB (Haertel, 1999; Linn, 2000; Hamilton, 2003; Trimble, Gay & Matthews, 2005). This definition in literature may differ from the practical language used by school districts. There may be assessments required for particular schools or populations of students within a district that do not appear on the district-wide calendar of tests. Some school districts in Texas require assessments throughout the year to gauge student progress with names such as "mid-year progress" or "short-cycle" assessments. These may be in addition to "benchmark tests." The discrepancy in naming of required assessments across districts could influence the numbers of assessments self-reported by districts. Reliance on district self-reporting data may not provide a representative description of benchmark testing practices in all Texas public school districts.

Summary

Chapter four provides the results of the path analyses conducted in this study. This chapter examines describes the mediated effects of relationships among observed variables on Texas school district's use of benchmark tests and the percent of students passing the Math and Reading TAKS tests. Adequacy of the model fit is supported by the chi-square fit statistic, CFI and RMSEA indices. Table 11 summarizes the findings for each research question.

Table 11

Summary Table

Question	Statistical result	Interpretation
Primary research questions		
1. Do significant relationships (i.e regression weights) exist between a district's descriptive factors and benchmark testing practices?	Pct. Econ. Dis. -0.41 $p < 0.01$	There is a highly significant inverse relationship between pct. econ. dis. students and number of benchmarks, with a medium effect size.
	Pct. LEP 0.32 $p < 0.01$	There is a highly significant positive relationship between pct. LEP students and number of benchmarks, with a small effect size.
	Pct. St. of Color 0.27 $p = 0.05$	There is a moderately significant positive relationship between pct. st. of color in a district and number of benchmarks, with a small effect size.

Table 11 (continued)

Question	Statistical result	Interpretation
2. Does a significant relationship (i.e regression weight) exist between the number of benchmark tests a district requires and the percentage of students passing the TAKS test?	Math < -0.01 $p = 0.98$ Reading -0.09 $p = 0.29$	There is a negative relationship between number of benchmarks and pct. of st. passing Math TAKS that is not significant, with a small effect size, There is a negative relationship between number of benchmarks and pct. of st. passing Reading TAKS that is not significant, with a small effect size.
Supporting research questions		
1. Is the effect of AYP Status on the number of benchmark tests mediated by the percentage of economically disadvantaged students in the district?	Math and Reading < 0.01 $p = 0.91$	The mediated effect of AYP status on number of benchmarks is positive and not significant, with a small effect size
2. Is the effect of AYP Status on the number of benchmark tests mediated by the percentage of students of color in the district?		

Table 11 (continued)

Question	Statistical result	Interpretation
3. Is the effect of AYP Status on the number of benchmark tests mediated by the percentage of Limited English Proficient (LEP) students in the district?		
4. Is the effect of a district's TEA state accountability rating on the number of benchmark tests given mediated by the percentage of economically disadvantaged students in the district?	Math and Reading -0.02 $p = 0.64$	The mediated effect of a district's accountability rating on number of benchmarks is negative and not significant, with a small effect size.
5. Is the effect of a district's TEA state accountability rating on the number of benchmark tests mediated by the percentage of students of color in the district?		

Table 11 (continued)

Question	Statistical result	Interpretation
6. Is the effect of a district's TEA state accountability rating on the number of benchmark tests mediated by the percentage of Limited English Proficient (LEP) students in the district?		
7. Is the effect of TEA district type on the number of benchmark tests mediated by the percentage of economically disadvantaged students in the district?		
8. Is the effect of TEA district type on the number of benchmark tests mediated by the percentage of students of color in the district?	Math and Reading 0.24 $p < 0.01$	The mediated effect of TEA district type on number of benchmarks is significant, with a small effect size.

Table 11 (continued)

Question	Statistical result	Interpretation
9. Is the effect of TEA district type on the number of benchmark tests mediated by the percentage of Limited English Proficient (LEP) students in the district?		
10. Is the effect of TEA district type on the percent of students passing the Math TAKS test mediated by the percentage of economically disadvantaged students, percentage of LEP students, percentage of students of color, and number of benchmarks?	Math 0.03 $p = 0.36$	The mediated effect of number of TEA district type pct. of st. passing Math TAKS is not significant with a small effect size.
11. Is the effect of TEA district type on the percent of students passing the Reading TAKS test mediated by the percentage of economically disadvantaged students, percentage of LEP students, percentage of students of color, and number of benchmarks?	Reading < 0.01 $p = 0.36$	The mediated effect of TEA district type on pct. of st. passing Reading TAKS is not significant, with a small effect size.

Table 11 (continued)

Question	Statistical result	Interpretation
12. Is the effect of AYP Status on the percent of students passing the Math TAKS test mediated by the percentage of economically disadvantaged students, percentage of LEP students, percentage of students of color, and number of benchmarks?	Math 0.03 $p = 0.91$	The mediated effect of AYP status on pct. of st. passing Math TAKS not significant with a small effect size.
13. Is the effect of AYP Status on the percent of students passing the Reading TAKS test mediated by the percentage of economically disadvantaged students, percentage of LEP students, percentage of students of color, and number of benchmarks?	Reading 0.07 $p = 0.81$	The mediated effect of AYP status on the pct. of st. passing Reading TAKS is small and not significant.

Table 11 (continued)

Question	Statistical result	Interpretation
14. Is the effect of a district's TEA state accountability rating on the percent of students passing the Math TAKS test mediated by the percentage of economically disadvantaged students, percentage of LEP students, percentage of students of color, and number of benchmarks?	Math 0.55 $p < 0.01$	The mediated effect of TEA district type on pct. of st. passing Math TAKS is medium and highly significant.
15. Is the effect of a district's TEA state accountability rating on the percent of students passing the Reading TAKS test mediated by the percentage of economically disadvantaged students, percentage of LEP students, percentage of students of color, and number of benchmarks?	Reading 0.53 $p < 0.01$	The mediated effect of TEA district type on pct. of st. passing Reading TAKS is medium and highly significant.

Table 11 (continued)

Question	Statistical result	Interpretation
16. What is the relationship between number of benchmarks a district requires and the percent of students passing Math TAKS?	Math < -0.01 $p = 0.93$	The mediated effect between number of benchmarks and pct. of st. passing Math TAKS is negative, very small and not significant.
17. What is the relationship between the number of benchmarks a district requires and the percent of students passing Reading TAKS?	Reading -0.09 $p = 0.93$	The mediated effect of the number of benchmarks on the pct. of st. passing Reading TAKS is negative, small, and not significant.

Note. pct.= percent, st.= students.

This chapter provided the results of the analyses conducted in this study. Several finds are noteworthy and are highlighted next. The effect of district type on the number of benchmarks tests required by a district was small, 0.24, and highly significant at $p < 0.01$. The standardized total effect of a district's TEA state accountability rating on the use of benchmark tests was 0.55 (Math) and 0.53 (Reading), a medium effect, with a value of $p < 0.01$, a highly significant. The standardized total effect number of benchmarks on percent of students passing Math TAKS was < -0.01 and on percent of students passing Reading TAKS was -0.09. Both are small and negative effects. The negative effects show an inverse relationship where an increased number of benchmarks relate to a lower

percentage of students passing TAKS. The level of statistical significance for the number of benchmarks on percent of students passing TAKS was $p = 0.93$ (Math and Reading). This is not a statistically significant relationship.

This analyses show some important relationships between a district's descriptive factors, benchmark testing practices, and percent of students passing TAKS. The highly significant ($p < 0.01$) relationship between percent economically disadvantaged students and the number of benchmarks indicates the greater the percent of economically disadvantaged students in a district, the fewer benchmarks the district requires. The positive and highly significant at $p < 0.01$ relationship between percent LEP students on the number of benchmarks shows the greater the percent of LEP students in a district, the greater the number of benchmarks the district requires. The positive and moderately significant ($p = 0.05$) relationship between percent students of color and the number of benchmarks given in a district indicate the greater the percent of students of color, the greater the number of benchmarks. The relationship between the number of benchmarks and the percent of students passing Math TAKS is not significant ($p = 0.93$), indicating the number of benchmarks in a district does not help explain the percent of students passing Math TAKS. The relationship between the number of benchmarks and the percent of students passing Reading TAKS is not statistically significant with $p = 0.93$, showing that the number of benchmarks does not help explain the percent of students passing Reading TAKS. TEA district type helps explain the number of benchmark tests a district requires with a highly significant $p < 0.01$. A district's TEA state accountability rating also helps explain the number of benchmarks given in a district, with a highly significant $p < 0.01$.Chapter 5 will discuss possible theories for these relationships

between a district's descriptive factors, number of benchmarks, and percent of students passing TAKS.

CHAPTER 5

DISCUSSION AND CONCLUSIONS

In this chapter, the results of the study, which explored the relationships between school district descriptive factors, their benchmark testing practices, and the percent of students passing a mandatory state standardized test, are examined through the lens of isomorphism. These results are then discussed in relationship to the scholarly literature along with implications for practice, policy and future research.

Review of the Findings in Relation to the Scholarly Literature

The sanctions embedded within NCLB (2002) have created high-stakes consequences for students and educators alike. These sanctions include students being retained at grade level, educators being reassigned or losing their jobs, and entire schools being reconstituted (McGhee & Nelson, 2005). Because the stakes are so high, much attention is given to trying to ensure student success on accountability measures. Benchmark tests are one tool many districts use to prepare students for accountability tests and to identify students who may need intensive intervention to be successful.

Although the intent of using benchmark tests is to assist students, the scholarly literature suggests there are negative consequences to this practice. For example, the use of benchmark tests and other forms of test practice detracts from learning by disrupting the flow and content of instruction. Educators often narrow the curriculum to focus on subject areas that will be tested (Popham, 2001), and then further narrow the curriculum within the tested subjects. For example, if the state mandated math test centers around low-level math skills, many educators emphasize low-level math skills throughout the

year's curriculum (Popham, 2001). Further, because assessments measure learning rather than advance learning, time spent practicing for accountability tests decreases the amount of time spent actually teaching (McNeil, 2000). This lost instructional time is not insignificant. One study found that teachers in Texas spent an average of eight to ten hours of instructional time each week coaching students for the state test (Hoffman, et al. 2001). This equates to more than 45 instructional days or 25% of the typical 180-day school year devoted to test preparation. This test preparation is often little more than a steady regimen of worksheets that do not actually ensure that students have learned the skills or content. Rather, test preparation often leads to students who can follow established steps well enough to be declared proficient on the state test, but cannot answer similar questions when asked in a slightly different format or apply their knowledge to new settings (Shepard, 2010). Alarmingly, this practice of trading instructional time for test preparation is more likely to occur in schools serving large numbers of students of color and low socio-economic students (Causey-Bush, 2005; Darling-Hammond, 2011; McNeil, 2000; Sheppard, 2002). This means that students of color and low socio-economic students spend more time practicing for accountability tests, and consequently less time learning, than their white, more affluent peers.

The present study supports the findings of previous studies that suggest not all students are equally affected by test preparation practices. In this study, a positive and significant relationship was found between the percent of English language learners (ELL), or LEP students, in a school district and the number of benchmark tests given, meaning the greater the percent of ELL students, the greater the number of benchmark tests. A similar relationship was found between percent of students of color and the

number of benchmark tests: the greater the percent of students of color in a district, greater the number of benchmark tests. Because the methods employed in this study accounted for a variety of district descriptive factors, the results suggest LEP status and race/ethnicity were greater factors in determining whether a district used benchmark tests and how often they administered them than seemingly more related factors such as district accountability rating or the AYP status of the district. In other words, districts with large numbers of ELL students and students of color were more likely to utilize benchmarking regardless of how students were performing on the test.

What makes the finding that LEP status and race/ethnicity were deciding factors in determining which students must trade learning time for test practice all the more disturbing is the lack of empirical evidence to support the practice of benchmarking. Many vendors of benchmark tests suggest these tests are formative assessments and cite research on the value of formative assessments to student learning. However, many benchmark assessments do not have the characteristics of formative assessments (Black & Wiliam, 1998; Perie, Marion & Gong, 2009). Moreover, benchmark tests have not been show to actually improve test scores. A prior study revealed that while 63% of districts in a random sample reporting using benchmark tests, the use of benchmark tests appeared to have little to no influence on standardized test scores in those districts (Nelson, et al, 2007). The present study reveals a similar finding. In this study, a negative relationship was found between the number of benchmark tests given and the percentage of students passing accountability tests. In other words, the greater the number of benchmark tests given in a district, the fewer the number of students who passed the state test. Because this finding was not statistically significant, it cannot be argued with

any level of assurance that benchmark testing harms test performance. However, it can be argued that benchmark tests do not improve scores on accountability tests and *may* decrease them.

This finding raises important questions. Why are so many school districts willing to dedicate valuable instructional time to benchmark testing when there is little evidence that such tests actually improve test scores? More importantly, why are ELL students and students of color being more frequently subjected to this unproven practice when the stakes are so high? The theory of isomorphism helps to explain this phenomenon. District leaders utilize benchmark testing because they see the practice being utilized in in other districts and, therefore, assume it is an effective practice in spite of having little evidence to support this belief.

Isomorphism as a Lens for Understanding the Results

Strategic isomorphism is the resemblance of an organization's practices to the practices of other organizations in its industry (Heugens & Lander, 2007; Meyer & Rowan, 1977; Abrahamson & Hegeman, 1994; Deephouse, 1996). The theory of isomorphism suggests organizations tend to mimic other organizations that are perceived as successful (Haberberg, 2005; DiMaggio & Powell, 1983). This is particularly the case in times of high stress, such as when organizations are competing for legitimacy from government regulators or public opinion (Deephouse & Suchman, 2008; Aldrich & Fiol, 1994; Jepperson, 1991; Deephouse, 1996). The status of an organization rises when government regulators or public opinion recognizes an organization as legitimate, that is the organization is recognized as one that does what it is purported to do in an effective way. (Baum & Oliver, 1991; Deephouse, 1996; Meyer & Scott, 1983). Because the path

to legitimacy is uncertain (Abrahamson & Hegeman, 1994), organizational leaders often employ rational myths in decision-making. Rational myths are institutional rules that are perceived to be of value, although there may be no evidence of their effectiveness (Tsoudkas and Knudsen, 2003; Meyer & Rowan, 1977). Rational myths emerge from distorted or limited interpretations of events or phenomenon. They are passed from one member of an organization to another until they become a socially constructed reality that is accepted throughout the organization (Mizruchi & Fein, 1999).

Rational myths have powerful effects. First, rational myths reduce uncertainty within an organization because members of the organization assume the rational myth is based on evidence of effectiveness (Myer & Rowan, 1977). Secondly, rational myths increase organizational legitimacy by spreading from one organization to another until the rational myth is accepted as standard practice for an industry. Once a rational myth reaches this level of acceptance, it becomes a measure that influences how an organization is viewed by regulators and public opinion (Deephouse, 1996). At this point, rational myths become a constraint on the behavior of individuals in an organization as organizations conform to these rational myths to increase legitimacy and reduce uncertainty (Aldrich and Fiol, 1994; Meyer and Rowan, 1977; Miner, 2005).

The results of this study suggest that, as in other fields, isomorphism may be present in education. With the publication of *A Nation At Risk* in 1983, the legitimacy of public schools was called into question. In response, a complex system of accountably policies and regulations was developed (NCLB, 2002). The system is fraught with ambiguity including, in the case of Texas, state and national accountability measures which are not aligned and conflict with one another. This has created a high degree of

uncertainty and fear within schools. As the theory of isomorphism would suggest, this fear and uncertainty has led to the proliferation of rational myths, benchmarking evidently being one of them.

The Rational Myths of Educational Accountability

Perhaps it is no surprise that educators have turned to rational myths as a means of responding to the pressures of high-stakes accountability. The entire educational accountability system is predicated on rational myths. Beginning with the idea that public schools were failing in the first place, the accountability movement has been propelled by a series of rational myths. These ideas are not proven by facts, but are prevalent in academic and popular literature. They have been repeated so often they are generally accepted by not only by policymakers and the general public, but by educators themselves. The data used to support these rational myths is skewed or non-existent, but they have nonetheless dramatically altered the context of education.

Rational Myth: Schools are Failing

In the 1970s, the U.S. economy began to deteriorate from the post-WWII boom (Cantor & Land, 1985; Akard, 1992; Hodrick & Prescott, 1997). Although there were complex reasons for this deterioration, there was widespread belief that much of the blame rested with schools that were inadequately preparing students for international competition in the workforce (Vinoviskis, 1999). This concern led to the creation of the National Commission on Excellence in Education, which issued *A Nation at Risk* (1983), a scathing report about the state of U.S. education. In spite of national test data that illustrated students were performing at the same or better levels than a decade before (Vinoviskis, 1999), *A Nation at Risk* (1983) announced the failings of schools in the

United States. The report proclaimed severe deficiencies of the United States educational system that would lead to dire consequences for the country. These failings included declining test scores and the idea that youth would not have the academic skills of previous generations (Alexander & Pallas, 1984; Davies, 2007). The report also stated that the poor quality of education in the United States would lead the country to lose international economic standing (Alexander & Pallas, 1984; Rothstein, 2008).

This view of schools as failing has been widely disputed since *A Nation at Risk* was published. Berliner and Biddle (1996) questioned the foundation for the claims made in report. They noted that many of statistics in the report lacked citations and even where data were presented with citations, the findings suggested a skewed interpretation meant to highlight failings that do not actually exist (Berliner & Biddle, 1996). More evidence that *A Nation at Risk* (1983) was flawed came with the publication of what has come to be known as the *Sandia Report* (1993). In 1990, then U.S. Secretary of Energy, Admiral James Watkins, commissioned a study by the nation's top research laboratory to document the decline of U.S. schools as reported in *A Nation at Risk*. Rather than verify the decline of education, the Sandia study called into question the findings of *A Nation at Risk*. In the official report of the Sandia study (Carson, Huelskamp, & Woodall, 1992, the authors stated, "To our surprise, on nearly every measure [of educational quality] we found steady or slightly improving trends, (p. 259). While the Sandia report did not completely vindicate public education, the report provided clear evidence that the state of public education was not as dire as it had been made out to be. Nonetheless, the myth that public schools needed fixing persisted.

Rational Myth: Business Models Can Be Used to Improve Education

Out of the myth that schools were failing grew another rational myth: business models can be used to improved education. This myth emerged out of the assumption that unproven educational theories had created the problem and, therefore, educators could not be trusted to fix the problem.. Instead, because the state of education had been linked to the decline of the U.S. economy (Vinoviskis, 1999), business leaders were called on to come up with a strategy for improving education.

In the 1980s, business leaders in the United States believed improving education would improve productivity of industry (McDonnell & Fuhrman, 1985). Members of the business community believed that systems of mandatory standardized tests would help guarantee that high-school graduates would possess basic skills needed to ensure success of American businesses (Sipple, 1999; Popham, 2001; Brooks-Buck, 2008). The popular mantra in education of "data-based decision making" comes from a business model of continual improvement (Deming, 1986; Spillane, Halverson & Diamond, 2001). The idea of using data on an ongoing basis seems intuitive and works with automobiles, but there is little evidence it works with student learning (Shepard, 2010). With business leaders heading the charge for educational reform, it is not surprising that successful business models were believed to be the basis for successful schools.

In order to ensure these business models were embedded in the education system, business leaders took a dominant role in the construction of federal education policy. A direct link between U.S. business interests and NCLB (2001) is evident. Business groups including the National Alliance of Business, Achieve, the Business Roundtable, and the Business Coalition for the Excellence in Education (BCEE) publicly supported and

actively lobbied for the annual testing of students in grades 3-8 in the areas of reading and math (Hoff, 2006; DeBray-Pelot & McGuinn, 2009). BCEE, an alliance of more than 70 business groups and individual companies, was formed for the purpose of ensuring NCLB (2001) and its requirements for testing students in grades 3-8 in the areas of math and reading became law (Hoff, 2006) Business leaders and groups including the U.S. Chamber of Commerce and the Business Roundtable, well-known organizations representing business owners and chief executives of large corporations, were prominent supporters of the reauthorization of NCLB in 2007 (Hoff, 2006; Brooks-Buck; 2008).

Although NCLB (2001) is often thought of as a testing program, NCLB is actually a comprehensive policy that encompasses almost every aspect of public education. At the heart of this policy are two prevalent business concepts: competition improves efficiency and negative consequences spur positive change. Supporters of the business model have argued public education is an inefficient monopoly that has no motivation to change (Hoff, 2006). For this reason, NCLB (2001) contains extensive provisions to support school choice and to sanction schools that do not show improvement. Supporters of school choice and competition believe the market will drive innovation and greater efficiency in education as it has in business (Ravitch, 2010). However, policies to motivate schools through competition and sanctions are based on faulty assumptions of motivation (Sheldon & Biddle, 1998; Natriello & Pallas, 2001). Moreover, there is little research to support the long-term successes of treating education as a business (Ravitch, 2010).

Rational Myth: Test Scores Demonstrate Learning

Even if competitive business models worked in education, there is still the issue of how best to measure quality in schools. Because teaching and learning is complex, developing instruments or tests to measure the effects of teaching and the level of learning is difficult and requires careful attention psychometric standards (AERA, APA & NCME, 1999). While a well-developed measurement is a useful tool, even the most carefully constructed test is limited in terms of what it can measure and under what conditions. Likewise, the results of any single measure are only useful in the context for which the instrument was designed (AERA, APA & NCME, 1999; Linn & Gronlund, 2000). Psychometricians often caution against using the results of any single test to make important decisions (AERA, APA & NCME, 1999; Gipps, 1994; Linn & Gronlund, 2000).

This advice has gone largely unheeded in the development of educational accountability policy. Policymakers have built an entire system around the rational myth that standardized test scores equal learning. According to Popham (1999), "if a school's standardized test scores are high, people think the school's staff is effective" (p. 8). There exists the assumption that assessment scores are a significant indicator of the educational possibilities in a school (Delandshere, 2001; Watson & Robbins, 2008; Ravitch, 2010). The American public supports testing because it believes that test scores are valid indicators of a child's learning (Haladyna & Haas, 1998; Ravitch, 2010). Newspapers publish district and school scores on high-stakes standardized tests because there is a public belief that these scores reflect instructional quality (Popham, 2001). Often lawmakers use standardized tests for purposes for which they were not designed, leading

public expectations to surpass the capacity of the tests (Heubert & Hauser, 1999; Wilson, 2007). Policymakers and the public seem to view assessment as an unfaultable method for making decisions about student and program success (Delandshere, 2001; Valenzuela, 2005). "The critical question of 'How do we teach Tracy the things she needs to know?' is forced aside by this far less important one: 'How do we improve Tracy's scores on the high-stakes test she will be taking?'" (Popham, 2001, p. 16).

The Elementary and Secondary Education Act of 1965 was an early example of policymakers using assessment as a measure of school success as it tied accountability to funding (Linn, 2001). During the 1970s and 1980s in the United States, the index for measuring students' success in schools was their scores on standardized tests (Stiggins, 1991). There is a vast amount of knowledge and skills a student is expected to learn at each grade level, and standardized tests assess a small collection of that information (Popham, 1999). If test-focused, unexciting drill activities focused on test content replace a diverse curriculum and actually raise students' test scores, it is almost certain the test is inappropriate and measures low-level skills (Popham, 2001). Even in 1959 Daly cautioned against using standardized test scores as the only measure of a student's success. "A test score can help the teacher or administrator who is working with a pupil if the score is given full consideration in light of all other findings" (Daly, 1959, p. 46).

Scores on standardized tests do not provide a complete measure of educational achievement (Bracey, 1991; Harris, Smith & Harris, 2011; Koretz, 2002; Gutierrez, 2008; Kohn, 2001; Popham, 2001; Natriello & Pallas, 2001). The multiple-choice format of many standardized tests does not simulate real-world situations (Sacks, 1999). Standardized tests measure a small part of what constitutes intelligence, leaving an

incomplete view of what a student knows (Sternberg, 1996; Gardner, 1995; Baker & Linn, 2004). Concentrating on student test scores is based on the idea that a school is a stable unit that can be credited with changing student performance (Baker & Linn, 2004). Changes in student population and school staff year to year make using the school as a unit of comparison invalid (Baker & Linn, 2004). Popham (2001), states that some schools labeled "failing" or "in need of improvement may have exemplary staff but the students are being assessed by an inaccurate assessment. There is also the problem of believing that a school with high test scores is full of successful educators because inaccurate assessments sometimes focus on the knowledge students *bring* to school, not what students *learn* in school (Popham, 2001). By focusing on scores on standardized tests, schools are looking at narrow view a student's knowledge and abilities.

The Rational Myth of Benchmark Testing

Given that the entire accountability system seems to be based on a series of rational myths, perhaps it is no surprise that school districts adhere to yet another rational myth in an effort to prepare students for accountability tests. Benchmark testing is a practice perceived to increase students' scores on standardized tests, although there is little evidence to suggest this is actually the case (Herman & Baker, 2005; McNeil, 2000; Kulik, Kulik & Bangert, 1984; Popham, 2001; Nabors-Olah, Lawrence & Riggan, 2010). School districts view benchmarks as an important link between district policies classroom teachers' practice (Buckley, Christman, Goertz & Lawrence, 2010). Part of the rational myth of benchmark testing is based on the literature supporting formative assessments. Formative assessments can be useful for guiding educators to adjust ongoing classroom instruction (Black & Wiliam, 1998; Herman & Baker, 2001). Many developers of

benchmark assessments flaunt the power of benchmarks as formative assessment to increase student achievement. However, little attention is given to the actual formative assessment research on which these claims are based (Shepard, 2009). Most benchmark tests are not used in the way formative assessments are described in the literature (CITE), so the references used to support them are not valid.

Benchmarks as Formative Assessment

The seminal report by Black and Wiliam (1998) brought together diverse research around formative assessment. Formative assessment is a systematic process of continuously gathering evidence of student performance and providing feedback while instruction is under way (Black & Wiliam, 1998; Heritage, et al, 2009). They examined studies that analyzed students' self-perception and motivation, teachers' assessment methods, features of assessments, and feedback teachers provided to students (Black & Wiliam, 1998). Effective feedback from teachers to students is required for students to improve performance (Black & Wiliam, 1998). Formative assessments are only useful when the information collected from the assessments is used as a piece of a system of coordinated assessment and instruction (Black & Wiliam, 1998).

How educators use assessment results is one breakdown between formative assessments as studied by Black and Wiliam (1998) and modern benchmark tests (Heritage, et al, 2009; Shepard, 2009). Studies focused on formative assessment have shown that how teachers provide feedback to students is important. Improved student learning is more likely to occur when feedback from teachers contains clear guidance of how to improve (Nichols, Meyers & Burling, 2009; Shepard, 2009). Students who develop "mindfulness" and reflective thinking on their own strateiges are more likely to

show improvement (Bangert-Downs, Kulik, Kulik & Morgan, 1991). Simply correcting student errors is not effective in improving students' performance (Shepard, 2009).

Teacher feedback must be part of a system that includes interpreting student results and effectively modifying instruction (Perie, Marion & Gong, 2009). Research has shown that the task of using assessment information to plan subsequent instruction for students is difficult for teachers (Heritage, et al, 2009). A teacher's process for interpreting benchmark assessment data is influenced by a variety of factors including knowledge and perceptions of the content and a student's background and past performance (Nabors Olah, Lawrence & Riggan, 2010). This, then, determines a teacher's decision of what and how to reteach. Often teachers focus their reteach, or change in instruction, on students who were not successful on the benchmark assessment. Little time is spent addressing the needs of students who were successful on the benchmark, even though they may have areas of misconception that need further instruction (Nabors Olah, Lawrence & Riggan, 2010). Teachers' limitations in their analysis of benchmark data leads to a superficial approach to reteaching and altering instruction (Nabors Olah, Lawerence & Riggan, 2010). Therefore, benchmarks as they are commonly utilized should not be considered formative assessments. They are a mechanism for preparing student for state tests.

Benchmarks as Test Preparation

The present study found that there is no indication that benchmark tests improve students' scores on TAKS, the state mandated, high-stakes test. The present study found that the more benchmark tests a district gave, the fewer students passed the state standardized test. The relationship between number of benchmark assessments and

percent of students passing TAKS was not significant ($p = 0.93$ for both Math and Reading). The relationship between number of benchmarks and percent of students passing TAKS was negative (< - 0.01, Math and – 0.09, Reading). This negative relationship indicates the greater the number of benchmark tests, the lower the percent of students passing TAKS. In this regard, benchmarks are not an effective means of preparing students for accountability tests.

Why the Myth of Benchmarking Persists

When members in a field of related organizations give a practice merit beyond what can be objectively ascertained, and it becomes a driving force in their behavior, isomorphism is at play (Schelling, 1978; Jones, 2009). Organizational legitimacy can be status conferred on an organization by groups such as policymakers and government regulators (Baum & Oliver, 1991; Scott, 1983). Public opinion is another factor that influence's an organization's legitimacy (Deephouse, 1996). Public school districts seek legitimacy from many of these groups.

School districts in Texas seek organizational legitimacy from TEA's rating system, which influences public opinion of a district's success (Popham, 2001). The basis of the ratings from TEA is student performance on mandatory high-stakes tests. Student performance on these tests can lead to sanctions, such as a loss of funding for districts or the closing of schools (NCLB, 2002). School districts look to each other for strategies believed to increase student performance on the high-stakes tests.

The present study found 60% of school districts in the sample use benchmark tests. Due to districts self-reporting the number of benchmark tests data and the variety of names for practice tests used in Texas school districts, there could be more than 60% of

districts using benchmark tests. Previous studies have shown that benchmark assessments are an integral part of the instructional plan in many school districts in the United States (Boyd & Christman, 2003; Porter, Chester & Schlesinger, 2004; Blanc, et al, 2010; Buckley, et al. 2010). The present study found a highly significant relationship between TEA district type and the number of benchmark tests given. A significant relationship was also found between a district's state accountability rating and the number of benchmark tests. A school district's use of benchmark testing was found to be related to the district's size, population, and accountability rating. These findings support the idea of isomorphism, which is the similarity of an organization's strategies to other organizations in its industry (Meyer & Rowan, 1977; Abrahamson & Hegeman, 1994; Deephouse, 1996).

Texas public school districts' use of benchmark testing practices when they do not seem to correlate with student success on TAKS could relate to strategic isomorphism. A central idea to isomorphism is that organizations conform to rational myths about what constitutes a successful organization in a particular field (Boxenbaum & Jonsson, 2008). The theory of isomorphism has been applied to public schools in the past. In 1970, Lamon and Scott discussed the structure of elementary school mathematics programs as a result of isomorphism. Rowan (1982) examined the expansion of administrative positions in public school districts as an isomorphic practice. Sweeping changes to the structure of New York City's Department of Education were examined through the lens of isomorphism (Carolan, 2008). Using isomorphism to understand organization and structure of school districts as organizations has occurred in the past. Directing the lens of

isomorphism on the particular strategy of benchmark testing has not been explored in depth in the literature.

School districts are looking for legitimacy from TEA ratings. They fear sanctions from NCLB such as schools closing or funding reduced. This need for legitimacy and fear of sanctions can lead organizations to isomorphic practices. Why benchmark testing is used by a majority of school districts in Texas although it is ineffective is a topic for further research. Benchmark tests seem to have become a rational myth in school districts that influence practice with little data to support their connection to student success on high-stakes tests.

Implications for Practice

Curriculum directors are responsible for ensuring that district policies match state and federal requirements for curriculum and standardized testing (Honig & Coburn, 2008). Benchmark tests are supposed to measure students' progress throughout the school year to provide educators with information on how to prepare students for the high-stakes standardized tests (Bancroft, 2010; Nese, et al, 2011; Olson, 2005). This study has shown that an increased number of benchmark tests corresponds to fewer students passing the state assessment.

NCLB (2002) requires that programs are "scientifically based" and "data-driven." The present study examined the relationship (regression weight) of the effect of the number of benchmarks required by a district and the percent of students passing the state standardized test, TAKS. This study revealed the relationship between the number of benchmarks and the percent of students passing TAKS is small at < -0.01 (Math) and -0.09 (Reading) and not significant, with $p = 0.93$ for Math and Reading state tests. This

indicates the number of benchmarks a district administers does not help explain the percent of students passing Math or Reading TAKS. This information calls into question the practice of benchmark testing. Is the practice of benchmark testing "scientifically based" and "data-driven" as required by NCLB (2002)? Curriculum leaders and other school district leaders need to examine the practice of benchmark testing in their districts. Is this practice worth time, money, and energy from both students and educators, that is required? District and school leaders need to examine how benchmarks are being utilized. Are they actually being used to change instructional practices? How are teachers analyzing the data? Many teachers do not have deep knowledge and understanding in how to scrutinize benchmark data from a variety of viewpoints and then use it to change instruction. Are teachers simply reteaching material in the same way it was taught the first time, or are they actually changing they way they deliver instruction?

Professional development and training for teachers can help solve some of the dilemmas of utilizing benchmark test data effectively. Districts and schools should think about the amount of time teachers have for data analysis and planning new ways to teach content. Are forty-five minute planning periods and afterschool meetings providing enough time to delve into benchmark data? Districts should look at the number of benchmark tests they are implementing and determine how much time teachers need to analyze and then create new plans of instruction based on the analysis. This time should be structured in the school's yearly calendar. Perhaps schools could provide substitutes to allow teachers half-days to plan and prepare. Teacher in-service days could be devoted to using benchmark data to change instructional practices.

Increased time to analyze data and plan will only be beneficial if teachers understand the variety of ways to examine data. Perhaps in-depth training with a few teacher leaders on each campus could provide resource and structure for the data analysis meetings. This training needs to be ongoing; a one-time, training cannot meet the needs of all teachers. As teachers implement strategies in using benchmark data they will need a place to ask questions and reflect on the effectiveness of their practice.

Benchmark test practices do not affect all students equally. Researchers have found that students in less affluent schools or low-performing schools spend more time practicing for the high-stakes state test than students in affluent schools (McNeil, 2000; Sheppard, 2002; Valenzuela, 2005). Educators working with students of color spend more time preparing for state tests, than educators working with white students (Darling-Hammond, 2011). The present study supports some of the previous findings. School district leaders need to examine their benchmark testing practices to see if all student groups are affected equitably. Districts need to look at their own benchmark data. Does their use of benchmark tests and student success on the state test mirror what was found in this study? If so, what is their justification for using a practice that does not improve student success? And how to they justify using an ineffective practice with a higher rate for students of color and LEP students?

Implications for Policy

The original intent of accountability policies was to ensure all students were receiving the same access to educational opportunities and that students were successfully learning the content taught to them. These are benevolent intents, but evidence shows there are many negative effects. The loss of high-quality diverse

instruction, being replaced with test preparation, has been shown in multiple studies. In light of this evidence, some policymakers have protested that the increase in test preparation is not their doing.

In Texas, lawmakers are now regulating the amount of time schools can spend preparing students for tests. In the 2013 legislative session, Texas state policymakers passed a bill limiting the use of benchmark tests. House Bill 5 (HB 5), Section 37 defines "benchmark test instrument" to mean "a district-required assessment instrument designed to prepare students for a corresponding state-administered assessment instrument" (section 39.0263). HB 5 goes on to restrict the number of benchmark tests a Texas school district can administer in a school year. "A school district may not administer to any student more than two benchmark assessment instruments to prepare the student for a corresponding state-administered assessment instrument" (HB 5, section 39.0263).

Due to the discrepancies in the naming of assessments in Texas school districts, it is possible that districts will continue to require "mid-year progress" or "short-cycle" assessments in addition to the "benchmark tests" as defined by HB 5. School districts that wish to continue administering large number of practice assessments could simply rename their assessments to fall within the regulations of the law.

In the present study, the mean number of benchmark tests reported was 4.83 per district. This number is below the number allowed by HB 5. HB 5 allows not more than two benchmark assessments for each state-administered assessment. Using the definition in HB 5, an eighth grade student could take six benchmark assessments per year – two for the math assessment, two for the reading assessment, and two for the science assessment. The six allowed benchmark tests is more than mean reported by districts in the present

study. Texas HB 5 could increase the number of benchmark tests some school districts administer.

Another piece of HB 5 is a new index rating system for public schools and districts. Schools will be rated on student achievement, student progress, closing performance gaps, and postsecondary readiness. Beginning in 2013, schools and districts in Texas will receive ratings of Met Standard, Met Alternative Standard, or Improvement Required. In 2014 the accountability system for schools and districts will expand to include a rating of A-F for schools and districts. A rating of A, B, or C will indicate acceptable performance; a rating of D or F will reflect unacceptable performance.

Though HB 5 changes the rating system for schools and districts in Texas, there are two important points to note. One is that there is still a rating system to publicly rank and compare schools. This is related to the rational myth of business models and competition helping education. The other point to note is that the four indices the Texas rating system will use are based on student scores on the standardized state assessments. This relates to the rational myth of student test scores indicating learning and student success. The Texas system may be changing, but its dependence on rational myths in education is unshakable.

NCLB (2001) created the system that led to the prevalence of high-stakes standardized testing. NCLB (2001) is legislation, created by politicians not educators. The high-stakes sanctions politicians put in place with NCLB created a culture of dependence on test scores. The dependence on test scores led to wide-spread benchmark testing. Now legislators and creating more policies to limit the amount of benchmark

testing caused by legislators with NCLB. Politicians, created a system of fear based on test scores and now they are moving to further regulate this system.

Implications for Future Research

Previous research (Darling-Hammond, 2011; Valenzuela, 2005) that found students from low socio-economic backgrounds spend more time preparing for high-stakes tests than their more affluent peers. The present study did not support that finding. This study found a negative and highly significant ($p < 0.01$) relationship between percent of economically disadvantaged students and number of benchmarks. This study indicates the greater the percent of economically disadvantaged students, the fewer benchmarks given in a district. This counter information may be related to the limitations of the data collection previously. Futher research is needed to examine the relationship between socio-economic status and school districts' benchmark testing practices.

The present study focused on quantitative data related to benchmark testing practices. Future research could examine qualitative questions related to the data. Interviewing curriculum directors and district leaders about their thoughts, opinions, and reasons for benchmark testing could provide valuable information. It may also be helpful to examine how benchmark test scores are used by teachers and administrators. Are the scores on benchmark tests used to alter instruction to meet the needs of all learners?

Previous research by Darling-Hammond (2011) and Valenzuela (2005) found students from low socio-economic backgrounds participated in more benchmark testing that their more affluent peers. The present study did not support that finding. This study found a significant negative relationship between economically disadvantaged students and the number of benchmark tests. Findings of this study indicate the greater the percent

of economically disadvantaged students in a district, the fewer the number of benchmark tests given.

Summary

This research examines an important topic in elementary and secondary public schools. Students lose high-quality instruction and it is replaced with benchmark testing. The practice of benchmark testing does not affect all students equitably. The present study showed LEP students and students of color take more benchmark tests than other students. This study indicated benchmark testing does not improve students' performance on the high-stakes state test. Texas public school districts seem to be under the influence of isomorphism when making benchmark testing decisions. The pressures on districts form NCLB, the state, and public opinion to improve test scores affect instructional practices. Sixty percent of school districts using benchmark assessments, even though they do not relate to higher test scores, indicates isomorphism may be influencing school leaders' decisions. These findings should force school district leaders to examine their benchmark testing practices. Are benchmark tests really worth the time, money, and energy expended by educators and students?

REFERENCES

(1983). A Nation at Risk: The Imperative for Education Reform. N. C. o. E. i. Education. Washington, D.C., U.S. Government Printing Office. (2001). NCLB. U. S. D. o. Education.

Abrahamson, E. & Hegeman, R. (1994). Strategic conformity: An institutional theory explanation. In *The annual meeting of the Academy of Management, Dallas, TX.*

Akard, P. J. (1992). Corporate mobilization and political power: The transformation of the U.S. economic policy in the 1970s. *American Sociological Review 57*(5), 597-615.

Aldrich, H. E. & Fiol, C. M. (1994). Fools rush in? The institutional context of industry creation. *Academy of Management Review, 19*(4), 645-670.

Alexander, K. L. and A. M. Pallas (1984). "Curriculum Reform and School Performance: An Evaluation of the "New Basics"." *American Journal of Education 92*(4), 391-420.

Alwin, D. F. & Hauser, R. M. (1975). The decomposition of effects in path analysis. *American Sociological Review, 40*(1), 37-47.

Arbuckle, J. (2008). *Amos 17.0 user's guide.* SPSS Incorporated.

Archbald, D. A. & Porter, A. C. (1990). Reforming the curriculum: Will empowerment policies replace control? *Journal of Education Policy, 5*(5), 11-36.

Argote, L. & Greve, H. R. A behavioral theory of the firm: 40 years and counting: Introduction and Impact. *Organization Science, 18*(3), 337-349.

Argote, L. (1999). Organizational learning: Creating, retaining, and transferring knowledge. Boston: Kluwer Academic.

Ashforth, B. E. & Gibbs, B. W. (1990). The double-edge of organizational legitimation. *Organization Science, 1*(2), 177-194.

Au, W. (2007). High-stakes testing and curricular control: A qualitative metasynthesis. *Educational Researcher, 36*(5), 258-267.

Au, W. (2010). The idiocy of policy: The anti-democratic curriculum of high-stakes testing. *Critical Education, 1*(1).

Babcock, C. D. (1965). The emerging role of the curriculum leader. The role of the supervisor and curriculum director in a climate of change. R. R. Leeper. Washington, D.C, ASCD.

Bancroft, K. (2010). Implementing the mandate: The limitations of benchmark tests. *Educational assessment, evaluation, and accountability.*22(1): 53-72.

Bangert-Downs, R. L., Kulik, J. A., & Morgan, M. T. (1991). The instructional effect of feedback in test-like events. *Review of Educational Research 61*, 213-238.

Barnard, A. F. (1937). Who does your thinking? *Secondary Education, 6*, 200-204.

Baron, R. M. and Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology, 51*(6), 1173-1182.

Bauer, D. J. (2003). Estimating multilevel linear models as structural equation models. *Journal of Educational and Behavioral Statistics, 28,* 135-167.

Baum, J. A. C. & Oliver, C. (1991). Institutional linkages and organizational mortality. *Administrative Science Quarterly, 36*(2), 187-218.

Behrens, J. T. (1997). Principles and procedures of exploratory data analysis. *Psychological Methods 2*(2), 131-160.

Bellack, A. A. (1972). History of Curriculum Thought and Practice. Current Readings in Improvements in Curriculum. Arlington, VA, College Readings Inc.

Berliner, D. and B. Biddle (1995). The Manufactured Crisis: Myths, Fraud, and the Attack on America's Public Schools. Reading, M.A., Addison-Wesley.

Black, P. and D. Wiliam (1998). "Assessment and classroom learning." Assessment in education. *5*(1), 7-74.

Blanc, S., Christman, J. B., Liu, R., Mitchell, C., Travers, E., and Buckley, K. E. (2010). Learning to learn from data: Benchmarks and instructional communities. *Peabody Journal of Education 85*(2), 205-225.

Bloom, B. S. (1971) Mastery learning. In J. H. Block(Ed.), *Theory and Practice.* New York: Holt Rinehart & Winston.

Bobbitt, J. F. (1912). Education: A first book. *Elementary school teacher, 13*(3), 154-155.

Bobbitt, J. F. (1918). *The curriculum.* New York: Houghton Mifflin Company.

Bobbitt, J. F. (1924). *How to make a curriculum.* New York: Houghton Mifflin Company.

Bohrnstedt, G. W., & Knoke, D. (1994). *Statistics for social data analysis* (3rd ed). Peacock Publishing.

Bollen, K. A. (1989). *Structural equations with latent variables.* Wiley.

Booher-Jennings, J. (2005). Below the bubble: 'Educational triage' and the Texas accountability system. *American educational research journal, 42*(2), 231-268.

Borman, G. D. (2005). National efforts to bring reform to scale in high-poverty schools: Outcomes and implications. *Review of research in education, 29,* 1-27.

Boslaugh, S. (2012). *Statistics in a nutshell* (2nd ed.). Sebastopol, CA: O'Reilly Media. Bracey, G. W. (1991). Why can't they be like we were? *Phi Delta Kappan, 73*(2), 104-117.

Boxenbaum, E., & Jonsson, S. (2008). Isomorphism, diffusion and decoupling. In R. Greenwood, C. Oliver, K. Sahlin, & R. Suddaby (Eds.), *The SAGE handbook of organizational institutionalism* (pp. 78-98). London: Sage Publications.

Boyd, W. L., & Christman, J. B. (2003). A tall order for Philadelphia's new approach to school governance: Heal the political rifts, close the budget gap, and improve schools. In L. Cuban, M. D. Usdan, & E. L. Hale (Eds.), *Powerful reforms with shallow roots: Improving America's urban schools* (pp. 96-124). New York: Teachers College Press.

Brooks-Buck, J. (2008). Schools as markets: Bilking the young and powerless. In R. Hopson, C. Camp-Yeakey, & F. Boakari (Eds.), *Advances in education in diverse communities: Research, policy and praxis, Vol. 6* (pp. 117-147).

Broom, C. A. (2011). Community in the early twentieth century schools, a case study: 1920s-1940s. *Historical Studies in Education, 23*(2).

Buckley, K. E., Christman, J. B., Goertz, M. E., Lawrence, N. R. (2010). Building with benchmarks: The role of the district in Philadelphia's benchmark assessment system. *Peabody Journal of Education 85*(2), 186-204.

Bulkley, K. E., Nabors Olah, L., & Blanc, S. (2010). Introduction to the special issue on benchmarks for success? Interim assessments as a strategy for educational improvement. *Peabody journal of education, 85*(2), 115-124.

Burch, P. (2007). Educational policy and practice from the perspective of institutional theory: Crafting a wider lens. *Educational Researcher, 36*(2), 84-95.

Byrne, B. M. (2001). Structural equation modeling with AMOS, EQS, and LISREL: Comparative approaches to testing for the factorial validity of a measuring instrument. *International journal of testing, 1*(1), 55-86.

Byrne, B. M. (2009). *Structural equation modeling with AMOS: Basic concepts, applications, and programming* (2nd ed.). New York: Routledge.

Byrne, B. M. (2010). *Structural equation modeling with AMOS: Basic concepts, applications, and programming* (2nd ed.). New York: Routledge.

Cantor, D. and Land, K. C. (1985). Unemployment and crime rates in the post-World War II United States: A theoretical and empirical analysis. *American Sociological Review 50*(3), 317-332.

Carnoy, M. and Loeb, S.(2003). Does external accountability affect student outcomes? A cross-state analysis. *Educational Evaluation and Policy Analysis 24* (4), 305–31.

Carolan, B. V. (2008). Instituitonal pressures and isomorphic change: The case of New York City's Department of Education. *Education and Urban Society, 40*(4), 428-451.

Carson, C. C., Huelskamp, R. M., and Woodall, T. D. (1992). Perspectives on education in America: An annotated briefing. *Journal of Educational Research 86*(5), 259-310.

Caswell, H. L. (1966). Emergence of the Curriculum as a Field of Professional Work and Study. Precedents and Promises in the Curriculum Field. H. F. Robison. New York, Teachers College Press**:** 1-11.

Caswell, H. L. (1988). Emergence of the Curriculum as a Field of Professional Work and Study. Curriculum: An Introduction to the Field. J. R. Gress. Berkeley, CA, McCutchan Publishing Corporation**:** 21-31.

Causey-Bush, T. (2005). Keep your eye on Texas and California: A look at testing, school reform, no child left behind, and implications for students of color. *The Journal of Negro Education 74*(4), 332-343.

Chapman, P. (1988). Schools as Sorters. New York, New York University Press.

Cheong, J. & MacKinnon, D. P. (2012). Mediation/indirect effects in structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (417 – 435). New York: Guilford Press.

Chernick, M. R. (2008). *Bootstrap methods: A guide for practitioners and researchers.* Hoboken, N.J.: Wiley.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.) Hillsdale, NJ: Erlbaum.

Cohen, J. (1990). Things I have learned (so far). *American Pyschologist, 45*(12), 1304-1312.

Cohen, J., & Cohen, P. (1984). *Applied multiple regression/correlation analysis for the behavioral sciences.* Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Cohen, J., & Cohen, P., and West, S. G. (2003). *Applied multiple regression/Correlation analysis.* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Cremin, L. A. (1975). Public education and the education of the public. *Teachers college record, 77,* 1-12.

Crocker, L. (2003). Teaching for the test: Validity, fairness, and moral action. Educational *Measurement, Issues and Practice*. *22*(3), 5.

Croninger, R. G., L. Valli, et al. (2003). Mapping the Policy Environment for High-Quality Teaching. Can we get there from here? American Educational Research Association, Chicago, IL.

Cyret, R. M., & March, J. G. (1963). A behavioral theory of the firm. *University of Illinois at Urbana-Champaign's Academy for Entrepreneurial Leadership Historical Research Reference in Entrepreneurship.*

Daft, R. L. (2008). *Organization theory and design.* Mason, OH: Cengage Learning.

Daly, W. C. (1959). Test scores: Fragment of a picture. *The Elementary School Journal, 60*(1), 43-46.

Darling-Hammond, L. (2011). *Testing, No Child Left Behind, and educational equity* in Diversity in American Higher Education. L.M. Stulberg and S.L. Weinberg (eds). New York, NY: Routledge.

Davies, G. (2007). *See government grow: Education politics from Johnson to Reagan.* Lawernce, KN: University Press of Kansas.

DeBray-Pelot, E. and McGuinn, P. (2009). The new politics of education: Analyzing the federal education policy landscape in the post-NCLB era. *Educational Policy 23*(1), 15-42.

Deephouse, D. L. (1996). Does isomorphism legitimate? *Academy of management journal, 39*(4), 1024-1039.

Deephouse, D. L. and Suchman, M. (2008). Legitimacy in organizational institutionalism. *The Sage handbook of organizational institutionalism,* 49-77.

Delandshere, G. (2001). Implicit Theories, Unexamined Assumptions and the Status Quo of Educational Assessment. Assessment in Education: Principles, Policy & Practice, Routledge. *8***,** 113-133.

Deming, W. E. (1986). *Out of the crisis.* Cambridge, MA: Massachusetts Institute of Technology, Center of Advanced Engineering Study.

Dimaggio, P., & Powell, W. W. (1983). The iron cage revisited: Institutional isomorphism and collective rationality in organizational fields. *American sociological review, 48,* 147-160.

Domino, G. and Domino, M. L. (2006). *Psychological testing: An introduction.* New York, NY: Cambridge University Press.

Duckworth, A. L., Quinn, P. D., Tsukayama, E. (2012). What No Child Left Behind leaves behind: The roles of IQ and self-control in predicting standardized achievement test scores and report card grades. *Journal of Educational Psychology, 104*(2), 439-451.

Edwards, V. B. (2006). Quality counts at 10: A decade of standards-based education. *Education Week, 25*(17), 17.

Efron, B. & Tibshirani, R. (1993). *An introduction to the bootstrap* (Vol. 57). New York, NY: Chapman & Hall/CRC.

Elbourne, E. T. (1914) *Factory administration and accounts with contributions on the general problem of industrial work design.* Longmans, Green and Company.

English, F. W. (1975). *School organization and management.* Worthington, OH: Charles A. Jones Publishing Company.

Erickson, H. L. (2002). *Concept-based curriculum and instruction: Teaching beyond the facts.* Thousand Oaks, CA: Corwin Press, Inc.

Eye, G. G., L. Netzer, et al. (1971). Supervision of instruction. New York, Haper & Row.

Fiedler, F. E. (1967). *A theory of leadership effectiveness.* New York: McGraw-Hill.

Field, A. (2005). *Discovering statistics using SPSS.* London: Sage Publications Inc.

Fisher, R. A. (1971). *The design of experiments* (9th ed.) New York:Macmillan.

Freire, P. (1970). *Pedagogy of the Oppressed.* New York: Continuum International Publishing Group Inc.

Frey, B. B. and V. L. Schmitt (2007). Coming to Terms With Classroom Assessment. *Journal of Advanced Academics*, Prufrock Press. *18***,** 402-423.

Galaskiewicz, J. (1985). Professional networks and the institutionalism of a single mindset. *American sociological review, 50*(5), 639-658.

Gallagher, C. J. (2003). Reconciling a Tradition of Testing with a New Learning Paradigm. *Educational Psychology Review 15,* 83-99.

Gardner, H. (1995). Reflections on multiple intelligences: Myths and messages. *Phi Delta Kappan, 77*, 206-209.

Garman, N. (2006). Curriculum Leaders as Public Intellectuals in an Impoverished Landscape. *Journal of Curriculum & Pedagogy 3*(1), 73-78.

Garson, G. D. (2009). Computerized simulation in the social sciences: A survey and evaluation. *Simulation and Gaming, 40*(2), 267-279.

Gergen, M., & Gergen, K. J. (2003). *Social construction*. Thousand Oaks, CA: Sage.

Gill, B. P. and S. L. Schlossman (2004). Villain or Savior? The American Discourse on Homework, 1850-2003. *Theory into Practice*, Lawrence Erlbaum Associates. *43***,** 174-181.

Gipps, C. V. (1994). *Beyond testing: Towards a theory of educational assessment.* London: Falmer Press.

Giroux, H. A. (1978). Developing educational programs: Overcoming the hidden curriculum. *The Clearing House, 52*(4), 148-151.

Glatthorn, A. A., Boschee, F. and Whitehead, B. M. (2005). *Curriculum Leadership: Development and Implementation.* Thousand Oaks, CA: SAGE Publications.

Godfrey, L. (2009). *Bootstrap tests for regression models.* Basingstoke, Hampshire:Palgrave Macmillan.

Goetz, M. E. (2001). Redefining government roles in an era of standards-based reform.*Phi Delta Kappan 83,* 62-66.

Goldberg, M. and J. Harvey (1983). A Nation at Risk: The Report of the National Commission on Excellence in Education. *The Phi Delta Kappan 65*(1), 14-18.

Goodson, I. F. (2005). *Learning, curriculum and life politics: The selected works of Ivor F. Goodson.* Abingdon, UK: Routledge.

Gress, J. R. and D. E. Purpel, Eds. (1988). *Curriculum: An introduction to the field*. Berkeley: McCutchan Publishing Corporation.

Gunzenhauser, M. G. (2003). High-stakes testing and the default philosophy of education. *Theory into Practice 42*(1), 51-58.

Gutierrez, R. (2008). A "gap-gazing" fetish in mathematics education? Problematitzing research on the achievement gap. *Journal for Research in Mathematics Education,* 357-364.

Haberberg, A. B. (2005). Isomorphism in strategic decsion-making. (Unpublished thesis). Retrieved from Electronic Theses Online Service: British Library. (uk.bl.ethos.418937).

Haertel, E.H. (1999). Performance assessment and education reform. *Phi Delta Kappan 80*, 662-666.

Haladyna, T., N. Haas, N., Allison, J. (1998). Continuing tensions in standardized testing. *Childhood Education 74*(5), 262.

Hall, J. D. (2005). The long civil rights movement and the political uses of the past. *The Journal of American History, 91*(4), 1233-1263.

Hamilton, L. S. (2003). Assessment as a policy tool. *Review of research in education 27*, 25-68.

Hamilton, L. S., B. M. Stecher, et al., Eds. (2002*). Making Sense of Test-Based Accountability in Education*. Santa Monica, CA, RAND.

Hamm, J. D. (1993). A Qualitative Study of the Work and the Contextual Factors That Affect the Work of Exemplary Public School Curriculum Directors in Washington State: A Preliminary Analysis. *Annual Meeting of the Americal Educational Research Association.* Atlanta, GA.

Hanneman, R. A., Kposowa, A. J., and Riddle, M. (2012). *Basic statistics for social research.* San Francisco: Jossey-Bass.

Harris, D. and Herrington, C. (2004). Accountability and the achievement gap: Evidence from NAEP. Unpublished manuscript, Department of Educational Leadership and Policy Studies, Florida State University.

Harris, D. and Herrington, C. (2006). Accountability, standards, and the growing achievement gap: Lessons from the past half-century. *American Journal of Education 112*(2), 209-238.

Harris, P., Smith, B. M., & Harris, J. (2011). *The myths of standardized tests: Why they don't tell you what you think they do.* Lanham, MD: Rowman & Littlefield Publishers.

Harrison-Jones, L. (2007). No child left behind and implications for black students. *The Journal of Negro Education,* 76(3), 346-356.

Hawes, G. R. (1964). *Educational Testing for the Millions: What Tests Really Mean for Your Child*. New York, McGraw-Hill Book Company.

Heilig, J. V. and Darling-Hammond, L. (2008). Accountability Texas style: The progress and learning of urban minority students in a high-stakes testing context. *Educational Evaluation and Polity Analysis 30*(2), 75-110.

Heritage, M. (2007). Formative Assessment: What Do Teachers Need to Know and Do? *Phi Delta Kappan 89*(2), 140-145.

Herman, J. L. and E. L. Baker (2005). Making benchmark testing work. *Educational Leadership 63*(3), 48-54.

Herman, J. L., Baker, E. L. & Linn, R. L. (2004). Accountability systems in support of students learning: Moving to the next generation. CRESST Line, Spring 2004. Center for Research on Evaluation Standards and Student Testing (CRESST). ED483383

Heubert, J. and R. Hauser, Eds. (1999). *High Stakes: Testing for Tracking, Promotion, and Graduation*. Washington, D.C., National Academic Press.

Heugens, P. M. and Lander M. (2007). Testing the strength of the iron cage: A meta-analysis of neo-institutional theory. Retrieved from ERIM Report Series Reference No. ERS-2007-007-ORG. http://ssm.com/abstract=962252.

Hlebowitsh, P. S. (2005). General ideas in curriculum: A historical triangulation. *Curriculum inquiry, 35*(1), 73-87.

Hodrick, R. J. and Prescott, E. C. (1997). Postwar U.S. business cycles: An empirical investigation. *Journal of Money, Credit, and Banking 29*(1), 1-16.

Hoff, D. J. (2006). Big business going to bat for NCLB: Competitiveness is cited as reason to retain law. *Education week 26*(8), 1-24.

Hoffman, J. V., L. C. Assaf, et al. (2001). The effects of standardized testing on teaching and schools. *Educational Measurement: Issues and Practice 12*(1), 20-25, 41-42.

Honig, M. I. (2006). Street-level bureaucracy revisited: Frontline district central-office administrators as boundary spanners in education policy implementation. *Educational Evaluation and Policy Analysis, 28*(4), 357-383.

Honig, M. I. and Coburn, C. (2008). Evidence-based decision making in school district central offices: Toward a policy and research agenda. *Educational Policy*. *22*(4), 578-608.

Hooper, D. Coughlan, J. and Mullen, M. (2008). Structrual equation modeling guidelines for determining model fit. *Electronic Journal of Business Research Methods, 6*(1), 53-60. Retrieved from http://arrow.dit.ie/buschmanart

Hunter, M. C. (1977). How can I plan more effective lessons. *Instructor, 87*(2), 74-75.

Hunter, M. C. (1979). Teaching is decision making. *Educational leadership, 37*(1), 62-64.

Hunter, M. C. (1982). *Mastery teaching: Increasing instructional effectiveness in elementary and secondary schools, colleges, and universities.* Thousand Oaks, CA: Corwin Press.

Hunter, M. C. (1989). Madeline Hunter in the English classroom. *The English journal, 78*(5), 16-18.

Hunter, M. C. (1991). Hunter lesson design helps achieve the goals of science instruction. *Educational leadership, 48*(4), 79-81.

Hunter, M. C. (1994). *Enhancing teaching.* New York: Macmillan College Publishing Company.

Hunter, M. C. (1994). *Mastery teaching.* Thousand Oaks, CA: Corwin Press, Inc.

Hursh, D. (2005). The growth of high-stakes testing in the USA: Accountability, markets and the decline in educational equality. *British Educational Research Journal, 31*(5).

IBM Corp. Released 2012. IBM SPSS Statistics for Windows, Version 21.0. Armonk, N.Y.: IBM Corp.

Jacob, B. A. (2005). Accountability, incentives and behavior: The impact of high-stakes testing in the Chicago public schools. *Journal of public economics, 89*(5/6), 761-796.

Jacobs, H. H. (1997). *Mapping the big picture: Integrating curriculum and assessment K-12.* Alexandria, VA: Association for Supervision and Curriculum Development.

Janesick, V. J. (2003). *Curriculum Trends: A reference handbook*. Santa Barbara, CA, ABC-CLIO.

Jennings, J. F. (1998). *Why National Standards? Politics and the Quest for Better Schools*. Thousand Oaks, CA: Sage Publications.

Jepperson, R. L. (1991). Institutions, institutional effects, and institutionalism. In W. W. Powell, P. J. DiMaggio (Eds.), *The new institutionalism in organizational analysis* (pp. 143-163). Chicago: University of Chicago Press.

Jeynes, W. (2007). *American educational history: School, society, and the common good.* Thousand Oaks, CA: Sage Publications.

Joint Committee on Standards for Educational and Psychological Testing of the AERA, APA, and NCME. (1999). *Standards for educational and psychological testing*. Washington DC: American Educational Research Association.

Jones, G. R. (2009). *Organizational theory, design, and change, 6th edition.* Old Tappan, NJ: Prentice Hall.

Jose, P. E. (2013). *Doing statistical mediation and moderation.* New York: Guilford Press.

Kamphaus, R. W., Winsor, A. P., Rowe, E. W., and Kim, S. (2005). A history of intelligence test interpretation. In D. P. Flanagan (Ed.) *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 23-38). New York: Guilford Press.

Keith, T. (2006). *Multiple regression and beyond.* Boston: Pearson Education.

Kelting-Gibson, L. M. (2005). Comparison of curriculum development practices. *Educational Research Quarterly 29*, 26-36.

Kenny, D. A. (2008). Refelctions on mediation. *Organizational Research Methods, 11*(2), 353-358.

Kimball, D. (1913). *Principles of industrial organization.* McGraw.

King, B. M. & Minium, E. W. (2003). *Statistical reasoning in psychology and education* (4th ed). Hoboken, NJ: John Wiley & Sons, Inc.

Kite, M. E. & Whitley, B. E. (2012). Ethnic and nationality stereotypes in everyday language. *Teaching of psychology, 39*(1), 54-56.

Kliebard, H. M. (1968). The Curriculum Field in Retrospect. *Technology and the Curriculum*. P. W. F. Witt. New York, Teachers College Press**:** 69-84.

Kliebard, H. M. (1995). *The Struggle for the American Curriculum*. New York, Routledge.

Kline, R. B. (2005). *Principles and practice of structural equation modeling* (2nd ed). New York: The Guilford Press.

Kohn, A. (2001). Fighting the tests: A practical guide to rescuing our schools. *Phi Delta Kappan, 82*(5), 348-357.

Koretz, D. M. (2002). Limitations in the use of achievement tests as measures of educators' productivity. *The Journal of Human Resources, 37*(4), 752-777.

Kulik, J. A., Kulik, C. C., & Banger, R. L. (1984). Effects of practice on aptitiude and achievement test scores. *American educational research journal, 21*(2), 435-447.

Labaree, D. F. (2005). Progressivism, schools and schools of education: An American romance. *Paedagogica historica, 41*(1-2), 275-288.

Ladson-Billings, G. (2006). From the achievement gap to the education debt: Understanding achievement in US schools. *Educational Researcher, 35*(7), 3-12.

Lammers, J. C. and Barbour, J. B. (2006). An institutional theory of organizational communication. *Communication Theory, 16,* 356-377.

Lamon, W. E. and Scott, L. F. (1970). An investigation of structure in elementary school mathematics: Isomorphism. *Educational Studies in Mathematics, 3*(1), 95-110.

Levitt, B., & March, J. G. (1988). Organizational learning. *Annual review of sociology, 14*, 319-340.

Lieberman, A. (2005). Introduction: The growth of educational change as a field of study: Understanding its roots and branches. *The Roots of Educational Change,* 1-8.

Linn, R. L. (2000). Assessments and accountability. *Educational Researcher 29*(2), 4-16.

Linn, R. L. (2001). A Century of Standardized Testing: Controversies and Pendulum Swings. *Educational Assessment*, Lawrence Erlbaum Associates. **7:** 29-38.

Linn, R. L. and Gronlund, N. E. (2000). *Measurement and assessment in teaching, 8th ed.* Upper Saddle River, N.J.: Prentice Hall.

Loehlin, J. C. (2004). *Latent variable models: An introduction to factor, path, and structural equation analysis* (4th ed). Mahwah, NJ: Lawrence Erlbaum Associates.

Long, M., Wood, C., Littleton, K., Passenger, T., and Sheehy, K. (2010). *The Psychology of Education: The Evidence Base for Teaching and Learning.* New York: Routledge.

MacKinnon, D. P., Lockwood, C. M. and Williams, J. (2004). Confidence limits for the indirect effect: Distribution of the product and resampling methods. *Multivariate behavioral research, 39*, 99-128.

Madaus, G. F., Russell, M. K., and Higgins, J. (2009). *The paradoxes of high stakes testing: How they affect students, their parents, teachers, principals, schools, and society.* Charlotte, NC: Information Age Publishing Incorporated.

Magnuson, K. A. and Valdfogel, J. (2008). *Steady gains and stalled progress: Inequality and the black-white test score gap.* New York: Russell Sage Foundation.

Mallinckrodt, B., Abraham, W. T., Wei, M., and Russell, D. W. (2006). Advances in the testing the statistical significance of mediation effects. *Journal of counseling psychology 53*(3), 372-378.

Manicas, P. T. (2006). *A realist philosophy of social science: Explanation and understanding.* United Kingdom: Cambridge University Press.

March, J. G., & Simon, H. A. (1958). *Organizations.* Oxford England: Wiley.

Marsh, H. W., & Hocevar, D. (1985). Application of confirmatory factor analysis to the study of self-concept: First- and higher order factor models and their invariance across groups. *Psychological bulletin, 97*, 562-582.

Martins, L. L. (2005). A model of the effects of reputational rankings on organizational change. *Organization Science, 16*(6), 701-720.

McDonnell, L. M., Fuhrman, S., & National Association of State Boards of Education, A. A. (1985). Meeting education policymakers' information needs: The role of the national organizations.

McGhee, M. W., & Nelson, S. W. (2005). Sacrificing leaders, villainizing leadership: How educational accountability policies impair school leadership. *Phi delta kappan, 86*(5), 367-372.

McIver, J. P., & Carmines, E. G. (1981). *Unidimensional scaling.* Beverly Hills, CA: Sage Publications.

McKenney, S., Nieveen, N., van den Akker, J. (2006). Design research from a curriculum perspective. *Educational design research,* 67-90.

McNeil, L. and A. Valenzuela (2001). The Harmful Impact of the TAAS System of Testing in Texas: Beneath the Accountability Rhetoric. *Raising Standards or Raising Barriers? Inequality and High-Stakes Testing in Public Education.* G. Orfield and M. L. Kornhaber. New York, Century Foundation Press**:** 127-150.

McNeil, L. M. (2000). *Contradictions of school reform: Educational costs of standardized testing*. New York: Routledge.

Merenda, P. F. (2005). Cross-cultural adaptation of educational and psychological testing. In R. K. Hambleton, P. F. Merenda, and C. D. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 321-341). Mahwah, NJ: Lawrence Erlbaum Associates, Inc., Publishers.

Meyer, J. W., & Rowan, B. (1977). Institutionalized organizations: Formal structure as myth and ceremony. *American journal of sociology, 83*(2), 340-363.

Meyer, J. W., & Scott, W. (1983). *Organizational environments: Ritual and rationality.* United States of America: Sage.

Miner, J. B. (2005). *Organizational behavior: Essential theories of motivation and leadership.* New York: M.E. Sharpe.

Mizruchi, M. S. and Fein, L. C. (1999). The social construction of organizational knowledge: A study of the uses of coercive, mimetic, and normative isomorphism. *Administrative Science Quarterly 44*(4), 653-683.

Monroe, W. S., DeVoss, J. C., & Kelly, F. J. (1917). *Educational tests and measurements.* Houghton.Nabors Olah, L., Lawrence, N. R., Riggan, M. (2010). Learning to learn from benchmark assessment data: How teachers analyze results. *Peabody Journal of Education 85*(2), 226-245.

Mooney, J. D., and Reiley, A. C. (1931). *Onward Industry.* New York: Harper & Row.

Naftali, T. (2007). *George H. W. Bush: The American presidents series: The 41st president, 1989-1993.* New York: Henry Holt and Company.

National Education Association of the United States. (2009). *Report of the Committee of Ten on Secondary School Studies Appointed at the Meeting of the National Educational Association July 9, 1892.*

Natriello, G. and A. M. Pallas (2001). The development and impact of high-stakes testing. *Raising Standards or Raising Barriers? Inequality and High-Stakes Testing in Public Education*. New York, Century Foundation Press**:** 19-38.

Nelson, S., M. McGhee, et al. (2007). Supplanting Teaching with Testing.

Nese, J.F, Park, B.J., Alonzo, J., Tindal, G. (2011). Applied curriculum-based measurement as a predictor of high-stakes assessment. *Elementary school journal*. 111(4): 608-624.

Newton, P. E. (2007). Clarifying the purposes of educational assessment. *Assessment in Education: Principles, Policy & Practice*,*14,* 149-170.

Nichols, P. D., Meyers, J. L. and Burling, K. S. (2009). A framework for evaluating and planning assessments intended to improve student achievement. *Educational measurement: Issues and practice 28*(3), 14-33.

Nichols, S. L. (2007). High stakes testing: Does it increase achievement? *Journal of Applied Psychology, 23*(2).

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

Noddings, N. (2005). What does it mean to educate the whole child? *Educational Leadership, 63*(1), 8.

Norman, G. R. & Streiner, D. L. (2003). *PDQ statistics* (3rd ed.). Hamilton, Ontario, Canada: BC Decker, Inc.

Novak, J and Fuller, B. (2003). *Penalizing diverse schools: Similar test scores but different students bring federal sanctions.* Berkley, CA: Policy Analysis for California Education.

Olson, A. (2007). Growth measures for systemic change. *School Administrator 64*(1): 10-14.

Olson, L. (2005). Benchmark assessments offer regular checkups on student achievement. *Education Week*. November 30, 2005. p. 13

Patton, M. Q. (2002). *Qualitative Research and Evaluation Methods.* Thousand Oaks, CA: Sage Publications.

Pedhazur, E. J. (1982). *Multiple regression in behavior research: Explanation and prediction* (2nd ed.). New York: Holt, Rinehart, and Winston.

Perie, M., Marion, S., & Gong, B. (2009). Moving toward a comprehensive assessment system: A framework for considering interim assessments. *Educational measurement: Issues and practice, 28*(3), 5-13.

Perie, M., Marion, S., Gong, B. & Wurtzel, J. (2007). *The role of interim assessments in a comprehensive assessment system: A policy brief.* Aspen, CA: The Aspen Institute.

Peterson, J. (1983). *The Iowa Testing Programs.* Iowa City, University of Iowa Press.

Pett, M. A. (1997). *Nonparametric statistics for health care research.* Thousand Oaks, CA: Sage Publications, Inc.

Pfeffer, J., & Salancik, G. R. (1978). Uncertainty, secrecy, and the choice of similar others. *Social psychology, 41*(3), 246-255.

Popa, A. B. (2009). Form follows function: A backward design to develop leadership ethics curriculum. *Volume 8, Number 1-Summer 2009,* 59.

Popham, J. (2006). Phony formative assessments: Buyer beware! *Educational Leadership 64*(3): 86-87.

Popham, W. J. (1999). Why standardized tests don't measure educational quality. *Educational Leadership 56*(6): 8.

Popham, W. J. (2001). *The Truth about Testing: An Educator's Call to Action.* Alexandria, VA: ASCD.

Porter, A. C., Chester, M. D., & Schlesinger, M. D. (2004). Framework for an effective assessment and accountability program: The Philadelphia example. *Teachers college record, 106*(6), 1358-1400.

Presthus, R. V. (1958). Toward a theory of organizational behavior. *Administration Science Quarterly, 3*, 48-72.

Price, L. R., Tulsky, D., Millis, S., & Weiss, L. (2002). Redefining the factor structure of the Wechsler Memory Scale-III: Confirmatory factor analysis with cross-validation. *Journal of clinical and experimental neuropsychology, 24*(5), 574-585.

Randolph, K. A. and Myers, L. L. (2013). *Basic statistics: Multivariate analysis.* New York: Oxford University Press.

Ravitch, D. (1996). The case for national standards. *The Clearing House 69*(3): 134-35.

Ravitch, D. (2010). *The death and life of the great American school system*. Philadelphia, PA: Basic Books.

Rogers, B. (1999). Conflicting approaches to curriculum: Recognizing how fundamental beliefs can sustain or sabotage school reform. *PJE. Peabody Journal of Education 74*(1), 29.

Ross, S. M. (2009). *Introduction to probability and statistics for engineers and scientists* (4th Ed). Saint Louis, M.O.: Academic Press.

Rothman, R. (1995). *Tests of Significance.* San Francisco: Jossey-Bass.

Rothstein, R. (2008). "A nation at risk" twenty five years later. *Cato Unbound.*

Rowan, B. (1982). Organizational structure and the institutional environment: The case of public schools. *Administrative Science Quarterly, 27*(2), 259-279.

Rowe, K. (2006). The measurement of composite variables from multiple indicators: Applications in quality assurance and accreditation systems – childcare. *Camberwell, Victoria: Australian council for educational research.*

Sacks, P. (1999). *Standardized Minds.* Cambridge, M.A.: Perseus Books.

Sadler, D. R. (1989). Formative assssment and the design of instructional systems. *Instructional Science 18*, 145-165.

Schelling, T. C. (1978). *Micromotives and macrobehavior.* Ipswich, MA: Norton.

Schmidt, J. N. (2009). *The impact of federal and state accountability policies on the breadth of education in low-income schools*. (master's thesis). Retrieved from Sacramento State Scholarworks.

Schumaker, R. E. & Lomax, R. G. (2004). *A beginner's guide to structural equation modeling* (3rd ed.). New York: Routledge.

Schwab, J. J. (1969). The practical: A language for curriculum. *The School Review 78*(1), 1-23.

Schwandt, T. A. (2005). The centrality of practice to evaluation. *American Journal of Evaluation 26*(1), 95-105.

Scott, W. G., & Mitchell, T. R. (1976). *Organization theory: A structural and behavioral analysis* (3rd ed.). Homewood, IL: Irwin.

Scott, W. R. (2007). *Institutions and organizations: Ideas and interests.* Thousand Oaks, CA: Sage Publications, Incorporated.

Scriven, M. (1967). The methodolgy of evaluation. *Perspectives on curriculum evaluation*. R. Tyler, R. Gagne and M. Scriven. Chicago, Rand McNally and Co.

Sears, L. (2007). Edward Lee Thorndike (1874-1949): A look at his contributions to learning and reading. *Shaping the Reading Field: The Impact of Early Reading Pioneers, Scientific Research, and Progressive Ideas,* 119.

Selznick, P. (1957). *Leadership in administration.* Evanston, IL: Row Peterson.

Senge, P. M. (1990). *The fifth discipline: The art and practice of the learning organization.* Michigan: Currency/Doubleday.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Belmont, CA: Wadsworth/Cengage Learning.

Sheldon, K. M., & Biddle, B. J. (1998). Standards, accountability, and school reform: Perils and pitfalls. *Teachers college record, 100*(1), 164-180.

Shepard, L. (2009). Commentary: Evaluating the validity of formative and interim assessment. *Educational measurement: Issues and Practice 28*(3), 32-37.

Shepard, L. A. (2002). The hazards of high-stakes testing. *Issues in science and technology. 19*(2), 53-58.

Shepard, L. A. (2010). What the marketplace has brought us: Item-by-item teaching with little instructional insight. *Peabody journal of education, 85*(2), 246-257.

Shook, C. L., Ketchen, D. J., Hult, G. T. M., & Kacmar, K. M. (2004). An assessment of the use of use of structural equation modeling in strategic management research. *Strategic Management Journal, 25*, 397-404.

Shrout, P. E. and Bolger, N. (2002). Mediation in experimental and nonexperiemental studies: New procedures and recommendations. *Psychological Methods, 7*(4), 422-445.

Sipple, J. W. (1999). Institutional contraints on business involvement in K-12 education policy. *American Educational Research Journal 36*(3), 447-488.

Sobel, M. E. (1987). Direct and indirect effects in structural equation models. *Sociological methods and research, 16*(1), 155-177.

Solley, B. A. (2007). On standardized testing: an ACEI position paper. *Childhood Education, 84*(1), 31-37.

Spillane, J. P., Halverson, R., and Diamond, J. B. (2001). Investigating school leadership practice: A distributed perspective. *Educational Researcher 30*(3), 23-28.

Spring, J. H. (2008). *The American school: From the puritans to No Child Left Behind.* McGraw-Hill.

Standards for educational and psychological testing. (2002). Washington, D. C.: American Educational Research Association (AERA), American Psychological Association (APA), National Council on Measurement in Education (NCME).

Sternberg, R. (1996). *Successful intelligence: How practical and creative intelligence determine success in life.* New York: Plume.

Stiggins, R. J. (1991). Facing Challenges of a New Era of Educational Assessment. *Applied Measurement in Education 4*(4): 263.

Stiggins, R. J. (2002). Assessment Crisis: The Absence Of Assessment FOR Learning. *Phi Delta Kappan 83*(10): 758.

Stiggins, R. J. and Chappuis, S. (2005). Putting testing in perspective: It's for learning. *Principal Leadership* 6(2): (Middle School Ed.)

Stiggins, R. J. and DuFour, R. (2009). Maximizing the power of formative assessments. *Phi Delta Kappan, 90*(9), 640-644.

Streiner, D. I. (2005). Finding our way: An introduction to path analysis. *Canadian journal of psychiatry, 50*(2), 115-122.

Sunderman, G. and Kim, J. (2004). *Inspiring vision, disappointing results: Four studies on implementing the No Child Left Behind Act.* Cambridge, MA: Harvard Civil Rights Project.

Taba, H. (1962). *Curriculum Development Theory and Practice*. New York: Harcourt, Brace, & World, Inc.

Taras, M. (2005). Assessment: Summative and formative and some theoretical reflections. *British Journal of Educational Studies, 53***:** 466-478.

Taras, M. (2008). Summative and formative assessment: Perceptions and realitites. *Active Learning in Higher Education 9*(2): 172-192.

Taubman, P. M. (2009). *Teaching by Numbers: Deconstructing the Discourse of Standards and Accountability in Education.* New York: Routledge.

Taylor, F. W. (1903). Shop management. *Transactions, 24,* 1337-1456.

Terman, L. (1919). *The Intelligence of School Children*. Cambridge, M.A., Riverside.

Thorndike, E. L. and E. O. Bergman (1934). *Measurement of Intelligence.* New York: Columbia University Press.

Tosi, H. L. (2009). It's about time!!!! *Journal of management inquiry, 18*(2), 175-178.

Tosi, H. L., & Hamner, W. C. (1974). Organizational behavior and management: A contingency approach. Chicago: St. Clair Press.

Trimble, S., Gay, A., Matthews, J. (2005). *Middle School Journal (36)*4, 26-32.

Tsoukas, H., & Knudsen, C. (2003). *Oxford handbook of organization theory.* New York: Oxford University Press.

Turkey, J. W. (1977). *Exploratory data analysis.* Reading, MA: Addison-Wesley.

Tyack, D. B. (1974). *The One Best System*. Cambridge, MA: Harvard University Press.

Tyler, R. W. (1950). *Basic Principles of Curriculum and Instruction.* Chicago, University of Chicago Press.

Urban, W. J. (2010). *More than science and Sputnik: The National Defense Education Act of 1958.* University of Alabama Press.

Valenzuela, A. (2005). *Leaving children behind: How "Texas-style" accountability fails Latino youth.* Albany: State of New York Press.

Valli, L., R. G. Croninger, R. G., Chambliss, M. H., Graeber, A. O., Buese, D. (2008). *Test Driven: High-Stakes Accountability in Elementary Schools.* New York, Teachers College Press.

Vinovskis, M. A. (1999). The road to Charlottesville: The 1989 education summit. *National Education Goals Panel.*

Walker, D. F. and J. F. Soltis (1986). *Curriculum and Aims*. New York, Teachers College Press.

Walsh, W. B. and N. Betz (1995). *Tests and assessment*. Englewood Cliff, N.J., Prentice-Hall.

Wang, L., G. H. Beckett, et al. (2006). Controversies of standardized assessment in school accountability reform: A critical synthesis of multidisciplinary research evidence. *Applied Measurement in Education 19*: 305-328.

Watson, D. and Robbins, J. (2008). Closing the chasm: Reconciling contemporary understandings of learning with the need to formally assess and accredit learners through the assessment of performance. *Research Papers in Education, 23*(3), 315-331

Weinfurt, K. P. (2000). Repeated measures analysis: ANOVA, MANOVA, and HLM. In L. G. Grimm, P. R.Yarnold (Eds.), *Reading and understanding multivariate statistics*, (pp. 317-361). Washington, DC: American Psychological Association.

Welkowitz, J., Cohen, B. H., and Lea, R. B. (2012). *Introductory statistics for the behavioral sciences, 7th ed.* Hoboken, N.J.: John Wiley & Sons.

Wheaton, B., Muthén B., Alwin, D. F., & Summers, G. F. (1977). Assessing reliability and stability in panel models. *Sociological methodology, 8*, 84-136.

Wiggins, G. and J. McTighe (2005). *Understanding by Design*. Alexandria, VA: ASCD.

Wilcox, R. R. (2009). *Basic statistics: Understanding conventional methods and modern insights.* New York: Oxford University Press.

Wilcox, R. R. (2009). Basic statistics: Understanding conventional methods and modern insights. Oxford: Oxford University Press.

Wiliam, D. (2006). Formative assessment: Getting the focus right. *Educational Assessment 11(3/4): 283-289.*

Williamson, D. (2008, March). Legislative history of alternative education: The policy context of continuation in high schools. Presented at AERA Annual Meeting. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.170.2361&rep=rep1&type=pdf

Wilson, L. D. (2007). High-stakes testing in mathematics. In F. Lester (Ed.), *Second handbook of research on mathematics teaching and learning* (pp. 1099-1110). Charlotte, NC: Information Age Publishing Incorporated.

Wraga, W. (2006). Curriculum theory and development and public policy making. *Journal of Curriculum & Pedagogy 3*(1): 83-87.

Wright, S. (1921). Correlation and causation. *Journal of Agricultural Research* 20: 557-585.

Yeager, M. (2007). Understanding NAEP: Inside the nation's education report card. *Education Sector.*

Zanderland, L. (1998). *Measuring Minds*. Cambridge, U.K., Cambridge University Press.

Zemelman, S., Daniels, H., Hyde, A. (2005). *Best Practice: Today's Standards for Teaching and Learning in America's Schools.* Portsmouth, NH, Heinemann.

Zieffler, A. H. and Long, J. D. (2011). *Comparing groups: Randomization and bootstrap methods using R.* Hoboken, N.J.: Wiley.