A NEW ERA OF EDUCATIONAL ASSESSMENT: THE USE OF STRATIFIED RANDOM SAMPLING IN HIGH STAKES TESTING

Stephanie N. Brown, B.A. M.S.

Dissertation Prepared for the Degree of

DOCTOR OF PHILOSPHY

UNIVERSITY OF NORTH TEXAS

December 2013

APPROVED:

Jimmy Byrd, Major Professor
Wendy Middlemiss, Minor Professor
John Brooks, Committee Member
Linda Stromberg, Committee Member
Nancy Nelson, Chair of the Department of Teacher Education and Administration
Jerry D. Thomas, Dean of the College of Education
Mark Wardell, Dean of the Toulouse Graduate School

Brown, Stephanie N. A new era of educational assessment: The use of stratified random sampling in high stakes testing. Doctor of Philosophy (Educational Administration), December 2013, 154 pp., 30 tables, 18 illustrations, references, 224 titles.

Although sampling techniques have been used effectively in education research and practice it is not clear how stratified random sampling techniques apply to high-stakes testing in the current educational environment. The present study focused on representative sampling as a possible means for reducing the quantity of state-administered tests in Texas public education. The purpose of this study was two-fold: (1) to determine if stratified random sampling is a viable option for reducing the number of students participating in Texas state assessments, and (2) to determine which sampling rate provides consistent estimates of the actual test results among the population of students. The study examined students' scaled scores, percent of students passing, and student growth over a three-year period on state-mandated assessments in reading, mathematics, science, and social studies. Four sampling rates were considered (10%, 15%, 20%, & 25%) when analyzing student performance across demographic variables, including population estimates by socioeconomic status, limited English proficiency, and placement in special education classes. The data set for this study included five school districts and 68,641 students.

Factorial ANOVAs were used initially to examine the effects of sampling rate on bias in reading and mathematics scores and bias in percentage of students passing these tests. Also 95% confidence intervals (CIs) and effect sizes for each model were examined to aid in the interpretation of the results. The results showed main effects for

sampling rate and campus as well as a two-way interaction between these variables. The results indicated that a 20% sampling rate would closely approximate the parameter values regarding the mean TAKS reading and mathematics scale scores and the percentage of students passing these assessments. However, as population size decreases, sampling rate may have to be increased. For example, in populations with 30 or fewer students in a subgroup it is recommended that all students be included in the testing program. This study situated in one state contributes to the growing body of research being conducted on an international basis in sample-based educational assessments.

ACKNOWLEDGMENTS

I would like to acknowledge that this work would not have been possible without the assistance and support of many individuals. First and foremost, I would like to express my unspeakable gratitude to my committee chair, Dr. Jimmy Byrd. Dr. Byrd strives for excellence and does not accept mediocrity. His methodological expertise and astute recommendations have fueled many of my intellectual endeavors. It has been a true honor and privilege to receive the mentorship of Dr. Byrd.

Additionally thank you to Drs. John Brooks, Linda Stromberg, and Wendy Middlemiss. They have given so generously their time, guidance, and support throughout the course of this research. They challenged me to think beyond my limits, they believed in me and my research, and their support allowed me to learn throughout this endeavor. I also wish to thank the faculty and staff of the Educational Leadership Department who provided timely advice and insight throughout my graduate coursework. Many thanks to my colleagues and friends; your care and support will never be forgotten.

I am forever appreciative of my loving family for supporting me on this long journey. My parents have always told me that I can do anything and everything I set my mind to accomplish, and I value their continued love and support. To my mother- your love is unconditional and you have always been my biggest advocate. To my dad- you graciously funded my college education and continuously supported me in my decision to keep furthering my education (P.S. - Dad I finally got my ticket!). To my brother David, thank you for always challenging me to greatness while providing humor for endurance. Finally to Grandma and Grandpa, thank you for your continued prayers and motivational wisdom as I advanced in this journey. Many thanks to you all.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Figure Page

CHAPTER 1

INTRODUCTION

Over the past 30 years, standards based accountability reform has taken on heightened importance in public education (Mayrowetz, 2009). This reform movement focuses on defined academic expectations, curricula standards, measureable assessments, and performance accountability (Hamilton, Stecher, Yuan, 2012). Many of the initiatives that have been adopted in response to the requirements of No Child Left Behind (NCLB, 2001) maintain origins in state and federal policy attributed to half a century of history from the 1960s to the 2000s. The beginning of educational reform in America and the first systemic push for testing in the United States commenced in 1958 with the national defense education act (NDEA). NDEA was passed as a response to the Soviet Union's early launch of the *Sputnik* satellite, which the United States viewed as a challenge to science and mathematics proficiency. Directly aligned with the efforts of standards-based reform, the legislative intent of NDEA was to strengthen the quality of American education (Flattau & Bracken, 2007; Jolly, 2009).

In 1983, increased accountability resulted from the publication of *A Nation at Risk* (National Commission on Excellence in Education, 1983). The report contended U.S. schools were performing inadequately in comparison to peer nations. Moreover, the findings recommended that schools and colleges set higher standards through increased student accountability. As a response, policy makers proposed improving the quality of American public education by concentrating on methods to raise expectations for teachers and students. This reform movement introduced the necessity to monitor student achievement in a systematic way (Wixson, Dutro, & Athan, 2003).

However, the most significant effort to implement reform in the American educational system started with the Elementary and Secondary Education Act (ESEA, 1965) during the presidency of Lyndon Johnson with a purpose to “provide financial resources to schools to enhance the learning experiences of underprivileged children” (Thomas & Brady, 2005, p. 51). Since its enactment, the law has been reauthorized eight times and has thus been a significant aspect of the American educational landscape (Anderson, 2005).

The 103rd Congress passed the Improving America’s Schools Act (IASA) in 1994. The intent of the IASA was to extend, for a five-year period, aspects of the Elementary and Secondary Education Act (ESEA) of 1965 as well as other educational functions, which first required states and school districts to identify schools in need of school improvement. The successor of the ESEA of 1994 is the No Child Left Behind Act (NCLB) of 2001, which is in place today. This landmark federal legislation requires states to develop assessments in basic skills in order to receive federal funding for schools. More concretely, the basic skills language of the NCLB legislation requires annual testing for students in Grades 3-8 and that all students be proficient in reading, writing, and math by 2014. Schools are required to report the scores of these annual tests to the public, disaggregating the data so that scores of minority students can be observed alongside the scores of the majority.

In an effort to close the achievement gap between low and high-performing schools, public education in the U.S. has focused attention on using assessment and accountability to ensure students are not just progressing through the educational system, but that they are learning core standards (Guskey, 2005). A major component

of standards based reform is the inclusion of assessments of student achievement with a growing emphasis on using tests to monitor progress and hold schools accountable. Public and professional attention to test scores has been growing since the establishment of National Assessment Educational Progress (NAEP) in the 1960s, and in the 1970s as tests started being linked with consequences for individual students (Koretz & Hamilton, 2006). Attributed to Popham (1987), the idea of "measurement-driven instruction" recognized that instruction is influenced by assessment, which led to research with innovative forms of assessment. This movement has further augmented growth in large-scale assessment, accompanied by initiatives to develop data systems to track student progress (Hamilton, Stetcher, & Yuan, 2012; USDO, 2010).

Today, schools are challenged to rethink assessment programs through the recent Race to the Top program authorized under the American Recovery and Reinvestment (RTTT) Act of 2009. The RTTT Assessment Program provides $350 million in competitive grants to support the development of a new generation of multi-state assessment systems. Unlike most existing state assessment systems, the vision expressed within the RTTT program features an integrated set of formative assessments for use by teachers within the flow of instruction, interim assessments to be given as progress checks throughout the year, and more focused summative accountability assessments (U.S. Department of Education, 2009). According the U.S. Department of Education (2010) this program seeks:

> …to develop assessments that are valid, support and inform instruction, provide accurate information about what students know and can do, and measure student achievement against standards designed to ensure that all students gain the knowledge and skills needed to succeed in college and the workplace. (¶ 1)

Pros and Cons of Standards Based Reform

Under the Bush administration, NCLB led supporters to believe that establishing high standards and accountability testing would reform American education (Amrein-Beardsley, 2009). In terms of the intended impact, research is mixed. A recent major U.S. review conducted by the National Academics of Sciences (Hout & Elliott, 2011) on the impact of incentives and test-based accountability in education reports:

> Test-based incentive programs…have not increased student achievement enough to bring the United States close to the levels of the highest achieving countries. When evaluated using relevant low-stakes tests, which are less likely to be inflated by incentives themselves, the overall effects on achievement tend to be small and are effective zero for a number of programs. (p. 3)

There are no consistent data to underscore whether high-stakes testing has had the intended effect of raising student achievement and/or closing the achievement gap (Nichols, Glass, & Berliner, 2012, Reardon, 2011; Timar & Maxwell-Jolly, 2012; Furthermore, schools serving larger populations of disadvantaged students are more likely to engage in narrowly targeted test preparation as a response to high-stakes testing (Nichols & Valenzuela, 2013).

A prominent concern regarding high-stakes testing is that it narrows curriculum and instruction to the content and format of the test (Au, 2007; Dee, Jacob, & Schwartz, 2013; Rothstein, Jacobsen, Wilder, 2008; Koretz, 2008). Specific narrowly targeted test preparation varies from over-emphasizing the tested content to drill and repetition of specific test items with similar format and even identical content to those on the high-stakes assessments. In response to the increased focus on high stakes testing, schools face negative consequences such as higher dropout rates, teaching to the test, and decreased student and teacher motivation, teacher exodus, and retention (Heilig &

Darling Hammond, 2008; Wang, Beckett, & Brown, 2006).

Anti-assessment driven reform argues high stakes assessments diminishes the student and teacher relationship and subsequently change the societal culture of the classroom (Nichols & Berliner, 2008). Supporters of NCLB, however, believe that NCLB initiatives will further democratize education through equality by establishing standards and providing resources to schools, irrespective of wealth, ethnicity, disabilities or language. Those in favor of such testing argue that the use of high-stakes assessments establishes a purposeful and explicit curriculum that allows all school personnel to understand what information should be taught (Herman & Haertel, 2005). Advocates of the accountability component of NCLB support standardized assessments as a measure to compare student competency across populations; NCLB desegregates data to highlight inequities among minority student populations, such as low socioeconomic (SES) students, English as a second language (ESL), and special education groups (TEA, n.d.).

In order for public schools to provide equal opportunity for all students our education system must change. Specifically, long standing achievement gaps must be addressed and closed with an eye towards the global economy of the future and America's shifting demographics (Baker, Sciarra, & Farrie, 2010). In Texas specifically, these reform efforts are evident in refocusing the landscape of public education on the individual child. A select group of school superintendents joined together in a visioning institute (Texas Association of School Administrators, TASA, 2008) to develop relevant core values to form new visions from which public education can emerge.

Evolved from the work of the TASA Public Education Visioning Institute, Senator Carona and Representative Strama established Texas High Performance Schools Consortium in SB 1557 (Texas High Performance Schools Consortium Act, 2011). This bill established a consortium of up to 20 districts and open-enrollment charter schools charged with developing methods for transforming public schools by improving student learning through the development of innovative, next-generation learning standards and assessment and accountability systems (TASA, 2008).

Visioning Institute envisions a comprehensive accountability system utilizing multiple measures with limited focus on high stakes assessments, in stark contrast to assigning accountability ratings based on student academic performance,. Furthermore, the institute emphasizes the importance of districts creating their own individualized accountability measures in accordance with state standards to meet the distinct needs of their children and ensure post-secondary success (TASA, 2008).

What is more evident is the intentional focus on reforming assessment measures. For example, the State of Texas was projected to spend more than $89 million on a new and revamped State of Texas Assessment of Academic Readiness (STAAR) testing program in 2012 (Cargile, 2012). This expenditure is almost 20% of the $468 million budgeted by the state for the testing program from 2010 to 2015. Comparable costs are reported among states nationwide as universal accountability mandates intensify (Obiakor & Beachum, 2005).

Even more startling is the understanding that these assessment measures lack predictive validity (Black, Bukrhardt, Daro, Jones, Lappan, Pead, & Stephens, 2012; Dobbelaer, 2010; Popham, 2001). Predictive validity is the degree of correlation

between the scores on a test and some other measure that the test is designed to predict (Messick, 1989). For example, the Scholastic Aptitude Test (SAT®) test is taken by high school students to predict their future performance in college (namely, their college grade point average (GPA). If students who scored at a high level on the SAT® tend to have high GPAs in college, then we can say that the SAT® has good predictive validity. But if there is no significant relation between SAT® scores and college GPA, then we would say the SAT® has low or poor predictive validity, because it did not predict what it was supposed to predict.

In the tough economic conditions that currently exist, the cost of developing academic assessments with little predictive validity (Black et al., 2012; Dobbelaer, 2010; Popham, 2001) is becoming cost prohibitive. As recommended by the Visioning Institute, perhaps a logical solution to diminishing the costs of developing statewide academic assessments is to examine how our peer countries model the ability to diagnose educational achievement through multi-stage sampling methods that accurately predict population results. Other developed high-performing nations do not administer individual student assessments annually (Savola, 2012). For example, Finland does not assess every student yearly and has not adopted the educational accountability framework that relies upon standardized testing (Sahlberg, 2007). Instead, Finland utilizes national sample-based assessments, meaning that each assessment administered is to a sample of students and the results are generalized to the larger population (Sahlberg, 2006).

With the current limited resources and the possibility of not receiving supplementary funding given future budget projections (Topol, Olson, Roeber, 2010;

National Center for Education Statistics, 2012), educators must be innovative and explore other assessment options to promote cost efficiency while improving instructional effectiveness. One plausible solution, as found in Finland, is to randomly select students to participate in state assessments. Randomly selecting students to participate in state-mandated assessments, while obtaining approximate parameter estimates of student mean scores has subsequently decreased the number of student participants and decreasing the overall costs associated with the nationwide testing program. Based on the work in Finland, the reduced student sample obtained from the population employing probability proportional to size (PPS) selection demonstrates that results can be achieved that are representative of the population parameter and substantially reduce the $1.7 billion expenditure on state assessments (Chingos, 2012). Visioning Institute has been tasked to remedy bureaucratic problems from a bottom-up approach beginning with the student as the centerpiece to education. In their principle for organizational transformation, it is clear that the theoretical approaches involve a multitude of perspectives and expertise strongly supporting a new vision for public education. Visioning Institute is currently exploring stratified random sampling methodology as an alternative to state assessments that test every child; ultimately transforming curricula from teaching to the test to teaching the child.

Statement of the Problem

A rising number of parents, school boards, teachers and civil rights organizations are beginning to question the fairness of the overreliance on standardized tests (Thomas, 2013). Recently over 300 groups, including the National Association for the

Advancement of Colored People (NAACP) legal defense fund, signed a Time Out from Testing (2012) petition to ask Congress to ban the use of such tests. The resolution calls on the U.S. Congress to overhaul the elementary and secondary education act, (NCLB, 2001), reduce the number of testing mandates, promote multiple forms of evidence of student learning and school quality in accountability, and not mandate any fixed role for the use of student test scores in evaluating educators. States such as New York and Florida have also followed in filing resolution petitions recognizing the overreliance on standardized testing. The Texas resolution states:

> The over reliance on standardized, high stakes testing as the only assessment of learning that really matters in the state and federal accountability systems is strangling our public schools and undermining any chance that educators have to transform a traditional system of schooling into a broad range of learning experiences that better prepares our students to live successfully and be competitive on a global stage. (TASA, 2013, ¶ 1)

While longitudinal research suggests SAT® and ACT® scores have a strong correlation in determining first year college success (Kobrin, Patterson, Barbuti, Shaw, Mattern, & Shaw, 2008; Patterson, Mattern, & Kobrin, 2009; Patterson & Mattern, 2011; Patterson & Mattern, 2012), there is no relevant research in terms of the predictive validity of state standardized high stakes tests and their relationship to college success. As Messick (1990) states “predictive validity indicates the extent to which an individual's future level on the criterion is predicted from prior test performance” (p. 7). Unfortunately, the evidence shows that such tests actually decrease student motivation and increase the proportion of students who leave school early (Darling-Hammond 2007a; Nichols et al., 2012). One might question the reliability and validity since test scores are not consistently improving on other data range measures of student achievement such as the NAEP and Program for International Assessment (PISA).

Reducing the number of students taking a test that has no established predictive validity in relation to college readiness even though it remains a state requirement will allow schools to focus on more substantive issues such as workforce and college and career readiness in the curriculum.

Purpose of the Study

While sampling techniques have been used effectively in education research and practice, it is not clear how stratified random sampling techniques apply to high stakes testing in the current educational environment in Texas. Perhaps a solution to diminishing the costs of developing statewide academic assessment is to examine how our peer countries model the ability to diagnose educational achievement through multi-stage sampling methods to ascertain population results. Therefore the purpose of this method study was two-fold: (1) to determine if stratified random sampling was a viable option for reducing the number of students participating in Texas state assessments, and (2) to determine which sampling rate provided consistent estimates of the actual test results among the subpopulations of students.

Research Questions

This study addressed these research questions:

1. How can stratified random sampling reduce the number of students taking state assessments in Texas school districts while accurately providing precise estimates of the mean scores, student growth, and the percentage of students passing high stakes assessments?

2. What is the recommended sampling rate among student subpopulations in Texas school districts to accurate, provide precise estimates of mean scores, student growth, and the percentage of students passing high stakes assessments among student subpopulations?

Theoretical Framework

The more bureaucratic a system, the less autonomy teachers have to alter instruction to meet the needs of students (Darling-Hammond, 2005). Not only is bureaucratic organization in opposition of innovation, it is substantially at odds with student learning (Darling-Hammond, 2005). In a concurring opinion, McNeil (2005) explains that our current education system is a bureaucratic hierarchy that is rigid in its implementation. In this system, rules are established at the top and there is no lee way given to schools or districts to choose the methods of implementation and evaluation that best suit their local needs (McNeil, 2005).

While assessment has become a primary feature of American school (Darling-Hammond, 2010; Ravitch, 2010) teachers consider the pressure of high-stakes assessment and no longer have the opportunity to employ creativity in classroom instruction. Instead educators' practice teacher led instruction coupled with test preparation, which consequentially narrows the curriculum (Musoleno, Malvern, White, 2010; 2010; Ravitch, 2010). According to Au (2011), "Knowledge learned for U.S. high-stakes tests is thus transformed into a collection of disconnected facts, operations, procedures, or data mainly needed for rote memorization in preparation for the tests" (p. 31). Dewey (1938) delineated four foundations of knowledge a teacher is recommended

to consider when lesson planning. They include knowing the child, individualizing curricula, understanding the social nature of learning, and preparation for life. Likewise, Vygotsky (1978) supported the teacher's role as a facilitator instead of the disseminator of all knowledge. The potential cost of this approach in regard to student learning is evidenced across educational theory and may be best understood by examining the educational approaches stemming from the work of Dewey and Vygotsky.

The Association of Experiential Education (AEE, 1994) defines experiential learning as "a process through which a learner constructs knowledge, skill, and value from direct experience" (p. 1). Bicknell-Holmes and Hoffman (2000) describe the main characteristics of discovery learning as 1) exploring and problem solving to produce, integrate, and generalize knowledge, 2) student driven, interest-based activities in which the student governs the sequence and frequency, and 3) activities to encourage integration of new knowledge into the learner's existing knowledge (schema) base. With origins in Dewey's conceptions of authenticity in instructional activities and Vygotsky's philosophies of social learning, experiential learning associates knowledge development to interaction and environmental experiences (Dewey, 1916; Dewey, 1938; Kolb, 1984; Vygotsky, 1978).

This study is guided by John Dewey's democratic theory. According to Dewey (1900), public schooling should nurture individual differences among learners in a common learning community. To serve American democracy, institutions should embrace and build bridges among learners' cultural and personal differences rather than create routinized training (Levin, 1991). Dewey (1916) utilized the term, active learner, to describe the learner who accessed sensory input to construct meaning. The

mind is an active element in the learning process of discovery learning. By actively pursuing new knowledge, learning is not defined as absorbing what is being said or read. Rather than the learner engaging in passive reception of information through lecture or drill and practice, students form deep applications for skills through problem solving. Dewey proposed the notion that learning did not occur through passive acceptance of information presented by the teacher but that learning involved the learner's engagement with the world. This passive rote memorization should be limited in learning (Jones, Jones, & Hargroves, 2003).

Dewey went on to recommend that learning was a social activity. Moreover, he suggested that student learning was enriched through connections, discussion and interaction with other students and the teacher (Edwards & Mercer, 1987). According to Dewey such reflective activity, engaging the student's mind, was crucial for the construction of knowledge and the learning process (Kolb, 1984). Dewey (1933) defined reflective thinking as "active, persistent, and careful consideration of any belief or supposed form of knowledge in light of the groups that support it and the future conclusions to which is tends" (p. 9).

Discovery learning promotes and increases student achievement when the students are learning skills rather than facts. Loyens, Rikers, and Schmidt (2007) reinforced this style of learning when they stated, "An important restriction of education is that teachers cannot simply transmit knowledge to students, but students need to actively construct knowledge in their own minds. That is they discover and transform information, check new information in comparison to old, and revise rules when they no longer apply" (p. 180). Dewey's (1910) democratic classroom environment supported

schools teaching students how to think instead of exactly what to think.

Applying Dewey's pivotal work as a foundation, Vygotsky (1978) suggested a necessary social aspect to constructivism. He theorized that individual students experienced two types of concepts in the classroom, spontaneous concepts and scientific concepts. Students readily integrated spontaneous concepts with previous knowledge and experiences (schema). Scientific concepts were more formal and abstract and had to be acclimated into the student's consciousness as a means to provide conceptual resources for spontaneous knowledge. Vygotsky (1978) termed the intersection of spontaneous and scientific concepts as the "zone of proximal development" (ZPD, p. 34). Additionally, he suggested that each learner articulated a different set of spontaneous concepts to the classroom, therefore the ZPD varied from one learner to the next. Vygotsky (1978) claimed that student learning in the ZPD would be enhanced "through collaboration with more able peers" (p. 86). Consequently, Vygotsky (1978) recognized that both the teacher and student peers developed a learning community, which influenced the student's intellectual development.

John Dewey's democratic theory of education suggests that, if the learning process engages the active mind and pursues a curriculum based on discovery learning, then student achievement will improve (Bahm, 2009). In a 2009 study, Bahm tested student's academic achievement, perception of inquiry learning skills, and retention of knowledge during a science model lesson. Using a pre- and post- test he compared discovery versus traditional learning techniques. Student achievement was significantly better in the group that utilized discovery-learning methods**.**

In this proposed authentic learning environment, students will be enabled to

make schematic connections from prior learning as recommended by the works of both Dewey and Vygotsky. By expanding the curriculum to involve higher order thinking skills such as synthesis and evaluation, students will be better equipped to retain knowledge for later learning. In addition, by expanding the curriculum and not teaching to the test, higher order thinking skills can differentiate instruction to meet the needs of all learners based on a multitude of varied intelligences. By incorporating the theoretical foundations of Dewey, Vygotsky, Bloom, and Gardner into an education program, students will participate in meaningful activities; thus their individual needs will be satisfied, resulting in a productive learning experience and improved student achievement.

As an extension of Dewey's work, individual differences transformed with the work of Gardener's (1993) theory of multiple intelligences. Gardener's work holds that there are many varieties of intelligence encompassing talents that are most often ignored by tests (Nelson & Eddy, 2008). While Dewey's theory redefines the purpose of education, the incorporation of Bloom's taxonomy aids a hierarchy for measuring the success of instruction.

Significance of the Study

Sampling or testing a portion of the students versus testing the entire population has several advantages including cost reduction, conservation of time, and reduction of the number of assessments administered. Instructional expenditures among school districts in the U.S. were approximately $610.1 billion for public elementary and secondary education in 2008–09, (USDOE, 2012; National Center for Education Statistics, 2012). Of this total, comprehensive evidence by the Brown Center on

Education Policy indicated roughly $1.7 billion was expended on annual student performance assessments alone (Chingos, 2012). Likewise, following the passage of NCLB (2001), annual state spending on standardized tests rose from $423 million to almost $1.1 billion in 2008. According to the Pew Center on the States (Vu, 2008), this figure represents a 160% increase compared to a 19.22% increase in inflation over the same period. Although costs are important, of greater concern is the persistent gap in academic achievement between children in the U.S. and their counterparts in other countries, which in economic terms, deprived the U.S. economy of as much as $2.3 trillion in economic output in 2008 (McKinsey & Company, 2009).

According to ACT®, approximately 28% of all 2012 ACT-tested high school graduates did not meet any of the ACT® college readiness benchmarks; meaning they were not prepared academically for first-year college courses in English composition, college algebra, biology, and social sciences (ACT®, 2013). Similarly, the SAT® report on college and career readiness revealed only 43% of SAT® takers in the class of 2012 graduated from high school with the level of academic preparedness associated with a high likelihood of college success (College Board, 2012, p. 152). One of the reasons to shift focus towards college and career readiness is state test results have increased while college entrance exams have remained rather static (ACT®, 2013; College Board, 2012). Meanwhile, researchers confirm that high school grade point average (HSGPA) has increased over time, whereas standardized test scores (ACT® or SAT® scores) remained stable (ACT® 2013).

While longitudinal research suggests SAT® and ACT® scores have a strong correlation in determining first year college success (Kobrin et al., 2008; Patterson,

Mattern, & Kobrin, 2009; Patterson & Mattern, 2011; Patterson & Mattern, 2012); there is no relevant research in terms of predictive validity of state standardized high stakes tests and the connection to college success, yet curriculum is based on standards directly associated with these tests. As Messick (1990) states “predictive validity indicates the extent to which an individual's future level on the criterion is predicted from prior test performance” (p. 7). Unfortunately, the evidence shows that such tests actually decrease student motivation and increase the proportion of students who leave school early (Darling-Hammond 2007a; Nichols et al., 2012). One might question the reliability and validity since test scores are not consistently improving on other data range measures of student achievement.

Definition of Terms

- Economically disadvantaged - The percent of economically disadvantaged students is calculated as the sum of the students coded as eligible for free or reduced-price lunch or eligible for other public assistance, divided by the total number of students: number of students coded as eligible for free or reduced-price lunch or other public assistance divided by total number of students (Texas Education Agency, n.d.).
- Explicit stratification - Explicit stratification consists of building separate sampling frames, according to the set of explicit stratification variables under consideration; used for categorical variables (National Center for Education Statistics, 2013).
- High stakes assessment - “A test is high-stakes when its results are used to

make important decisions that affect students, teachers, administrators, communities, schools, and districts" (Au, 2007, p. 258).

- Implicit stratification - A method of achieving the benefits of stratification often used in conjunction with systematic sampling. The sampling frame is sorted with respect to one or more stratification variables but is not explicitly separated into distinct strata (National Center for Education Statistics, 2013).
- Limited English proficient (LEP). These are students identified as LEP by the Language Proficiency Assessment Committee (LPAC) according to criteria established in the Texas Administrative Code. Not all students identified as LEP receive bilingual or English as a second language instruction, although most do. In the Profile section of the reports, the percentage of LEP students is calculated by dividing the number of LEP students by the total number of students in the school or district (Texas Education Agency, n.d.).
- No Child Left Behind Act of 2001 (NCLB) - NCLB is a federal legislation that enacts the theories of standards-based education reform. Pursuant to 20 USCS § 6301, NCLB ensures that all children have a fair, equal, and significant opportunity to obtain a high-quality education and reach, at a minimum, proficiency on challenging state academic achievement standards and state academic assessments.
- School type - For purposes of creating the campus groups, schools are placed into one of four classifications based on the lowest and highest grades in which students are enrolled at the school (i.e. in membership): elementary, middle (including junior high school), secondary, and both elementary

\secondary (K-12). Generally speaking, elementary schools are pre K-5 or pre K-6, middle schools are 6-8, and secondary schools are 9-12. Schools with grade spans that do not exactly match these are grouped with the school type most similar to their grade span (Texas Education Agency, n.d.).

- Special education - This refers to the population served by programs for students with disabilities. An Admission, Review, and Dismissal (ARD) committee makes assessment decisions for students in special education programs. The ARD committee is made up of the parent(s) or guardian, teacher, administrator, and other concerned parties. In the 2010-11 school year, a student in special education may have been administered the TAKS, TAKS (Accommodated), TAKS-Modified, or TAKS-Alternate (Texas Education Agency, n.d.).
- State of Texas Assessments of Academic Readiness (STAAR) – STAAR replaced the TAKS program beginning in spring 2012. The STAAR program at Grades 3-8 assess the same grades and subjects as was assessed on TAKS. For high school, general subject-area TAKS tests were replaced with twelve STAAR end-of-course (EOC) assessments (Texas Education Agency, n.d.).
- Texas Assessment of Knowledge and Skills (TAKS) - TAKS is a comprehensive testing program for public school students in Grades 3-11. The TAKS is designed to measure to what extent a student has learned, understood, and is able to apply the important concepts and skills expected at each tested grade level. All TAKS tests in Grades 3-6 are available in either

English or Spanish. The Academic Excellence Indicator System (AEIS) reports show performance on these language groups separately (Texas Education Agency, n.d.).

Limitations

Utilizing stratified random sampling (SRS) requires strata to be carefully defined. The strata in this case were delineated based on subgroups as defined by NCLB (2001) and Texas Education Agency (TEA, 2012). Due to the limitations of subgroup sizes, populations of limited English proficient (LEP) and special education (SPED) students could not be customized by ethnicity due to low sample sizes, which would automatically qualify for the prior recommendation to test all students in a sample less than 30. In Texas, LEP and SPED populations are not further reported by ethnicity from the state. These data points, however, would be ascertainable at the local school level from the student information system.

Employing SRS to augment testing every student, to testing a sample of students creates a cross-sectional, single point in time data and therefore does not enable longitudinal analysis for comparisons over time at the student level. While this study only sampled students in 7^{th} grade it is recommended that future studies sample a wider grade span to determine model reliability of the sampling program across grade levels. The student demographic variables as disaggregated by TEA are limited to the five school districts included in this study. While student demographic variables from the five school districts in the North Texas region were representative of the 68,462 students sampled in those districts, a larger statewide study is recommended to encompass the

ethnically diverse demographics of the state population. This study was further limited to TAKS scores as first year STAAR test results were not published until June 2012.

Organization of the Study

This report is organized into five chapters. Chapter 1 provides an introduction, statement of the problem, purpose of the study, research questions, definition of terms, assumptions and limitations of the study, significance of the study, and organization of the study. Chapter 2 is a review of the relevant literature, and Chapter 3 explains the methods used in the research. Chapter 4 provides a presentation of the results and analysis of the data. Chapter 5 explains the results of the study and provides recommendations for future research.

CHAPTER 2

LITERATURE REVIEW

The literature review focuses on accountability in schools as discussed in Chapter 1 and the research related to it. Chapter 2 is presented in three main sections that examine various dimensions of educational assessment. The first section reviews the theoretical foundation and history of high-stakes testing while providing a contextual background to understand the current high-stakes testing environment in public schools. In addition this section includes relevant research related to the achievement gap in terms of accountability measures. Understanding and identifying the factors influencing the achievement gap can assist educators, researchers, and policy makers in identifying effective and appropriate methods to help close the gap. The second section reviews how high-stakes testing and the accountability movement have negatively impacted the educational experiences and opportunities for students in public schools. Incorporated into this section are educational measurement issues of validity, reliability, cut scores and test inflation. The third section presents international assessment programs that employ stratified random sampling as an alternate form of accountability. Chapter 2 concludes with a summary of the three major sections that collectively attest to the relevance of this study.

In 1983, the National Commission on Excellence in Education conducted an analysis on the condition of education in America, which resulted in *A Nation at Risk: The Imperative for Educational Reform* that said America and "its educational institutions seem to have lost sight of the basic purposes of schooling, and of the high expectations and disciplined effort needed to attain them" (p. 1). This report led to the

implementation of standards based education in many states and the resulting Improving America's Schools Act of 1994 mandated state academic content standards and tests (USDOE, 2008). The current high-stakes testing environment in the United States (U.S.) was related to concerns regarding America's ability to educate its masses and compete in the world market.

The reauthorization of the Elementary and Secondary Education Act (ESEA) established the foundation for increased accountability through no child left behind (NCLB) (Schraw, 2010). One primary goal of NCLB was to increase student achievement for all students (Forte, 2010). The accountability aspect of NCLB focused on realizing predetermined proficiency rates and/or showing adequate yearly progress (AYP) for all students and subgroups (Education Commission of the States, 2004). The primary accountability objective was to raise student achievement by requiring states to institute performance standards and consequences associated with accountability measures (Kress, Zechmann, & Schmitten, 2011). One key requirement of NCLB was for every student to meet state proficiency thresholds by year 2014 (Burke, 2012).

The legislation of NCLB (2001) required states accepting federal funding to measure and report on results in terms of standards and accountability (NCLB, 2001). Moe (2002) explains, the rationale behind systems of high-stakes accountability:

> The movement for school accountability is essentially a movement for more effective top-down control of the schools. The idea is that, if public authorities want to promote student achievement, they need to adopt organizational control mechanisms-tests, school report cards, rewards and sanctions, and the like-designed to get district officials, principals, teachers, and students to change their behavior.... Virtually all organizations need to engage in top- down control, because the people at the top have goals they want the people at the bottom to pursue, and something has to be done to bring about the desired behaviors; the public school system is just like other organizations in this respect. (p. 81)

Today, NCLB continues to have a tremendous impact on school and district accountability at both the state and national levels. Consequently, the evaluation of student learning through various assessments has become increasingly important (Ercikan, 2006). Today, results of high-stakes testing are used to identify school performance labels and determine how schools and districts are recognized as low and high performing (Anderman, Anderman, Yough, & Gimbert, 2010).

Theoretical Foundation and History of Testing

The relationship concerning intelligence and academic success has consistently been strong (Bartels, Rietveld, Van Baal, & Boomsma, 2002; Jencks, 1979; Mackintosh; 1998; Spearman, 1904). One of the central purposes of testing, dating back to Binet in the 1910s and 1920s, was to measure intelligence and predict educational achievement (Binet & Simon, 1916; Spring, 2008). Binet, however, argued analyzing more complex abilities, such as judgment, memory and language, could only assess mental functioning.

Goddard (1913) was the first to translate and circulate Binet's scales in America for use. Psychologist, Lewis Terman, revised the test, developed an intelligence quotient (IQ) scale, and produced the Standford-Binet test (Spring, 2008). Terman's revised test marked the beginning of large-scale testing in the U.S. and coincided with the explosive growth of public school populations between 1890 and 1915, as well as an influx of immigrant children and the demise of the one-room schoolhouse. Terman's test appeared to answer the schools' need to classify children according to their skill levels, and it used intelligence testing for placement of children in the appropriate grade level (Zimmerman & Schunk, 2003).

The first noteworthy use of standardized assessment as part of a selection process transpired as a result of World War I. Desiring to determine which draftees were best-suited for various tasks, the U.S. Army developed and administered an intelligence test, called the army alpha, to each potential soldier (Haney, 1981). At about the same time the work of Carl Brigham culminated in the development of the Scholastic Aptitude Test (SAT®) (Spring, 2008). These tests, along with other achievement tests, were intended to assess the abilities of individual students and the effectiveness of particular curricular programs (Hamilton, Dunbar, & Linn, 2003).

While Binet, Goddard, Terman, and Brigham developed widely used assessments, the first theoretical concept of intelligence was attributed to Spearman (1904) and Thorndike (1913) (as cited in Zimmerman & Schunk, 2003). Continuing the original work of Spearman and Thorndike, more closely analyzed individual learner differences through statistical measures of correlation of stimulus and response connections. Through correlation methods, Thorndike concluded knowledge is experience and created the theory of intelligence otherwise known as sampling theory. Today his theory, although antiquated, serves as the basis for understanding the processes of knowledge and the overall ability to learn (Zimmerman & Schunk, 2003). Between 1908 and 1916, Thorndike (1918) and his students at Columbia University developed standardized achievement tests in arithmetic, handwriting, spelling, drawing, reading, and language ability. By 1918, there were well over 100 standardized tests, developed by different researchers to measure achievement in the principal elementary and secondary school subjects. In his famous quote Thorndike (1918) stated "whatever exists at all exists in some amount" (p. 16) conveyed his accomplishments in

quantitative educational measurement. Derived from his quantitative ideology, Thorndike (1942) created the CAVD college entrance assessment. In this test, Thorndike was able to differentiate between readiness (experience) and achievement (content exposure). In opposition, Dewey (1922) attested, “our mechanical, industrialized civilization is concerned with averages, with percent’s; the mental habit, which reflects this social scene subordinates education and social arrangements based on averaged gross inferiorities and superiorities” (as cited in Sokal, 1987, p. 73).

Two test-related developments in particular are significant to the rise of standardized tests: the introduction of the IQ test and the invention of the multiple-choice format. Frederick J. Kelly developed the first educational test using the multiple-choice format in 1915 (Samuelson, 1987). Since then, multiple choice has become the dominant format of standardized achievement tests. Upon its advent, many more items could be administered in a short period, and tests could be scored quickly and objectively. The multiple-choice design was implemented by the test publishing industry that emerged in the 1920s and evolved into a billion dollar industry (Clarke, Madaus, Horn, & Ramos, 2000). According to Wardrop (1976), there were three main reasons for the growth of the industry: new statistical procedures for analyzing and improving tests, faster ways of scoring the tests and reporting the results, and the “institutionalization of testing in American society” (p. 14).

By 1932, 75% of 150 large city school systems in the USA used group intelligence tests to divide students into ability groups, and, likewise, colleges also used tests to rationalize admissions procedures (Haney 1984). While in the late 1960s there was increasing concern for the population of students not attending college; many felt

too much attention had been attributed to “college-bound” students. To remedy this situation, special arrangements were employed with the Division of Employment Security for students intended for the workforce population. The result was the General Aptitude Test Battery (Camper, 1978).

Five waves of educational reform occurring between the 1950s and the 1990s included the role of tests in tracking and selection emphasized in the 1950s, the use of tests for program accountability in the 1960s, minimum competency testing programs of the 1970s, school and district accountability of the 1980s, and the standards-based accountability systems of the 1990s. According to Haney, Madaus, and Lyons (1993), there were five major social and political influences on the growth of educational testing: the launch of Sputnik in the 1950s, the Civil Rights Movement of the 1960s, the decline of SAT® scores in the 1970s, the emergence of the education reform movement of the 1980s, and national education reform proposals in the 1990s.

Achievement Gap

Many social scientists and psychologists of the 20th century were writing and speaking about the intellectual inferiority of different races and hence began documenting the legacy of educational inequities. Sir Frances Galton was among the first to study individual differences in mental ability and evaluated people based on their awards and accomplishments. This initial research established evidence that intelligence was hereditary and compelled further studies which involved evaluating individual differences in reaction time, which have since been verified to correlate with academic success (Zimmerman & Schunk, 2003). Most notably, psychologist Arthur

Jensen suggested that IQ differences between African Americans and Whites were due to genetic factors (Jensen, 1969; Jensen, 1992). Further, Army data acknowledged that members of immigrant groups scored lower than native-born Americans. Most recent immigrants from Southern and Eastern Europe scored lower than those from Northern and Western Europe. African American recruits, however, scored lowest of all (Snyderman & Rothman, 1988).

The presence of the achievement gap is not a new phenomenon in education. The need to close the gap is presently the driving force behind current influences of educational reform. Equality in education has been a long sought after principle since before the significant case of *Brown v. Board of Education* in 1954. Educational reform has been applied in various forms throughout the years with the ambition of achieving equity. The key focus of the No Child Left Behind Act of 2001 was to improve public education for all students in the U. S., with an emphasis on closing the achievement gap between advantaged and disadvantaged students (Kantor & Lowe, 2006). The literature has documented that non-white students continue to score at lower levels than their peers on standardized tests in reading and mathematics (National Center for Education Statistics, 2012). The achievement gap is defined as “the difference between how well low-income and minority children perform on standardized tests as compared to their white, more advantaged peers. For many years, low-income and minority children have been falling behind their white peers in terms of academic achievement” (USDOE, 2011).

NCLB (2001) defines four types of student subgroups: students from major racial groups (American Indian, Asian, Hispanic, African American, and White), students of

limited English proficiency, students with disabilities, and students who are socioeconomically disadvantaged (typically defined by eligibility for free or reduced-price lunch). Literature exits to correlate the relationship between race and student achievement (O'Conner, Hill, & Robinson, 2009). The rapid increase of the Hispanic population, not only in Texas but also in the United States, now represents the fastest-growing ethnicity inherent to the population. According to the U.S. Census Bureau data, the Hispanic population increased by about 58%, from 22 million in 1990 to 35 million in 2000, compared with an increase of about 13% for the total U.S. population (Hemphill & Vanneman, 2010). The increase of over 15 million Hispanic students from 2000 to 2010 accounted for more than half of the total population increase in the U.S. during the last decade (Humes, Jones, & Ramirez, 2011). To represent this shift, studies on the long-term trends of student performance in mathematics indicate the 2009 gap between White students and Hispanic students in 8th grade was 26 points; which was not statistically significant in comparison to either Grade 4 or Grade 8 results from the 1990 to 2009 span (National Assessment Educational Progress [NAEP], 2009). Similar results were identified in comparisons between White and African American students (NAEP, 2009).

Similar research has recognized that students who have limited English language proficiency (LEP) have greater difficulty in reading achievement than students who are not LEP (Allington & McGill-Franzen, 2003). The relationship between academic achievement relative to ethnic minorities and LEP student status has received increased interest from both a policy and research perspective (Lay & Stokes-Brown, 2009; Kim & Sunderman, 2005; Ravitch, 2011; Rothstein, Jacobsen, & Wilder, 2009).

A meta-analysis by White (1982) indicated that socioeconomic status (SES) measured at the level of the individual correlated modestly but significantly with academic achievement (r = 0.22). A more recent analysis reported a fairly similar correlation for the effect of individual SES on academic achievement (Johnson, McGue, & Lacono, 2007). Similarly, research that controls SES demonstrates that the condition of poverty is more significant than race in determining the achievement gap (Saenz, 2010; Yeung & Conley, 2008). In a related study by the Annie E. Casey Foundation (2011), researchers found that the gap for achievement test scores between rich and poor have grown by almost 60% since the 1960s and is almost twice as large as the gap between White students and other ethnicities. The significance of which lies in understanding the importance of college degrees for determining success in life; and only nine percent of low-income children will obtain those degrees (Bailey & Dynarski, 2011).

A primary purpose of NCLB (2001) was to close the achievement gap. NCLB (2001) failed to significantly increase average academic performance or to significantly narrow achievement gaps, as measured NAEP and Program for International Assessment (PISA) (Guisbond, 2012; Berliner, in press). The PISA data analyzed the percentage of disadvantaged students with the ability to realize acceptable scores and reported over 80% in Hong Kong, over 50% in Korea, and over 40% in Finland as compared to only 30% in the U. S. that achieved equal scores (OCED, 2010). Based on this data, Berliner (2014) reports, “the USA appears to have social and educational policies and practices that end up limiting the numbers of poor youth who can excel on tests of academic ability” (p. 13).

Finland is regarded as a model for public school education worldwide, despite the fact that this nation keeps testing, homework, and classroom hours to a minimum compared to other nations. In addition, grade retention for failing children is shunned-with only 2% of students held back and special education teachers assigned to work intensively with students who fall behind. On a national level, there is also a minimum of centralized government interference with local school policies. In Finland, the accountability within the school system relies on the judgment and professionalism of its teachers, rather than high stakes resting. It is also notable that the Finnish government's social policies have resulted in a children's poverty rate of well under 5%, compared with 20% in the U. S. (when poverty is defined as living in families with incomes <50% of the national median). In this context, the Finns have seen improvement in the last three PISA administrations (Berliner, in press).

United States Accountability System

The introduction of testing and reliance on assessments to hold schools accountable for student academic performance led to the prevalence of test-based accountability in the U. S. Large-scale, standardized tests of student achievement have long been a feature of K-12 education, as every K-12 student in the U.S. takes multiple state and district tests yearly (Hamilton, Stecher, & Klein, 2002). As legislative accountability mandates including the No Child Left Behind (NCLB) Act of 2001 require each state to measure student progress in reading and math in Grades 3 through 8 and at least once during Grades 10 through 12 (United States Department of Education, 2004). In addition, each state must meet the requirements of the previous law

reauthorizing ESEA (the Improving America's Schools Act of 1994) for assessments in reading and math in three grade spans (3-5; 6-9; and 10-12). States must also administer science assessments at least once during Grades 3-5; Grades 6-9; and Grades 10-12. Further, states must ensure that districts administer tests of English proficiency to measure oral language, reading and writing skills in English to all English language learner (ELL) students. Moreover, about half of U.S. states use large-scale assessments to evaluate students' qualifications for graduation or promotion from one grade to another. After high school graduation exams were implemented, 67% of the states posted a decrease in the rate by which students were graduated from high school (Amrein & Berliner, 2002).

Under the current regimen of NCLB, every K-12 student in the U.S. takes numerous state and district tests annually (Hamilton, Stecher, & Klein, 2002). In contrast, other developed nations do not administer individual student assessments annually (Savola, 2012). Finland, for example, does not assess every student and has not followed the global educational accountability movement utilizing standardized tests to hold schools and teachers accountable for student performance (Sahlberg, 2007). The educational system of Finland implements national sample-based assessments along with ongoing formative assessments develop. U.S. ranks twenty-eighth of forty countries in mathematics (Darling-Hammond, 2007a). The deficiency in academics among U.S. students compared to other industrialized nations weakens the global competitiveness of American students (Business Roundtable, 2005; Educational Testing Service, 2006; Rising above the Gathering Storm Committee, 2010). The Department of Education disseminates U.S. curriculum to exclude higher order thinking

skills that are not incorporated on multiple-choice assessments. The U.S. is not preparing students for the demands of the twenty first century and therefore, serving the nation a future disservice in its abilities to be globally competitive (Berliner, in press; Sahlberg, 2010).

According to Nelson and Eddy (2008), state standardized testing is not an effective way to measure school and individual success, nor guide instruction to improve learning. These types of tests, they argue, do not target skill specific areas nor do they provide educators the information necessary to design interventions to improve deficiencies. Furthermore, students with different learning styles have limited opportunities to demonstrate their knowledge with a single high stakes test (Nelson & Eddy, 2008).

Adobe™ (2013) released a research study that reveals the state of creativity in education. The study emphasizes the importance of preparing students to be innovators and how testing and government mandates are stifling creativity in the classroom. The author of the study concluded that, one of the top barriers to teaching creativity is that the educational system is too reliant on testing. An important step to promote and foster creativity in education is to reduce the number of tests (Adobe™, 2013). Sahlberg (2010) agrees by asserting that “test-based accountability [and] public ranking of schools based on those tests, and related rewards and sanctions are not contributing to ongoing efforts to sustainable improvement of the quality of public education” (p. 58).

Most districts have increased time for elementary school English language Arts (average 43%) or math (average 32%) and substantially decreased time for other subjects including social studies, science, art, music, and physical education (Center for

Education Policy, 2008). For example, Berliner (2009) states:

> American schools never allowed much time for individual or group work with high cognitive demands, but now even the teachers that made some use of problem based or project-based learning, forms of instruction that could ignite students' interests through a more personally tailored curriculum, are not allowed to do so. (p. 28)

According to Hollingworth (2007), teachers feel compelled to abandon what they know to be the best ways to instruct students and to resort, instead, to test preparation programs in efforts to raise test scores. There has been a shift from teaching for learning, or knowledge purposes, to teaching solely for high stakes testing, which has resulted in a narrowing of the curriculum, loss of instructional time, and loss of teacher autonomy (Higgins, Miller & Wegmann, 2006).

While there is no disagreement that measures of student performance are needed, the quantity and frequency of testing has changed the classroom focus from assessments to inform learning to test preparation (Au, 2007; Berliner, 2011; Holcombe, Jennings, & Koretz, 2013). Critics suggest NCLB narrows the curriculum in many schools focusing attention on the limited skills standardized tests measure (Lee & Lee, 2012). These negative effects fell most severely on classrooms serving low-income and minority children (Guisbond 2012, & National Center for Fair & Open Testing, 2012). As Guisbond (2012) explains:

> In fact, because of its misguided reliance on one-size-fits-all testing, labeling and sanctioning schools, it has undermined many education reform efforts. Many schools, particularly those serving low-income students, have become little more than test-preparation programs…Policymakers must abandon their faith-based embrace of test-and-punish strategies and, instead, pursue proven alternatives to guide and support the nation's neediest schools and student. (p. 1)

Conversely, experts stipulate while curriculum narrowing is harmful to all students, gifted students are affected severely (Peine & Coleman, 2010; Jolly & Kettler, 2008).

Gifted students require depth and exploration (Pandina, Callahan, & Urquhart, 2008; McCallister & Plourde, 2008), and yet curriculum narrowing does not allow for consistent depth and challenge of learning as it is often difficult to create learning experiences gifted students need (Scot, Callahan, & Urquhart, 2009). Since NCLB is an unfunded mandate, resources such as time and money must be re-allocated in order to support the requirements of NCLB (McCallister & Plourde, 2008). Reallocation might include cutting funds for gifted programs because those programs are not directly tied to the tenets of NCLB (Beisser, 2008; Hargrove, 2012).

There is rising concern that the testing required to fulfill the accountability requirements of NCLB (2001) are consequentially resulting in teacher burnout (Chang, 2009). The pressures of NCLB (2001) have indeed contributed negatively to the morale of both elementary and secondary teachers (Byrd-Blake, Afolayan, Hunt, Fabunmi, Pryor, & Leander, 2010). As Baker (2010) claims, “tying teacher evaluation and sanctions to test score results can discourage teachers from wanting to work in schools with the neediest students; while the large, unpredictable variation in the results and their perceived unfairness can undermine teacher morale” (p. 4).

Measurement Concerns

Multiple choice assessments are designed within a psychometric framework to produce “economically tractable and defensible reliability indices for ranking and norming purposes” (Sloane & Kelly, 2003, ¶ 7). While constructed along a bell curve, distribution, high stakes assessments are more appropriate to scaling students than providing valuable information capable of enhancing student learning (Ellison, 2012). To

generate the bell curve test makers usually eliminate questions that students with low overall scores might get right but those with high overall scores get wrong; therefore, most questions which favor minority groups are eliminated (Guisbond, 2012).

A test is able to measure only a fragment of a student's knowledge on a particular topic or subject. Studies and researchers question the use of a single measurement cut score with the analysis of school effectiveness (Darling-Hammond, 2007a; Harris, 2007; Hout, Elliot, & Frueh, 2012; Ravitch, 2010). Also, there is a consensus among researchers with regard to the reliability issues associated with the use of a single measurement to evaluate school effectiveness (Darling-Hammond, 2007a; Harris, 2007; & Ravitch, 2010).

Cut score manipulation that states might implement can provide the appearance that gains have occurred without any improvement in the conditions of school programs or classroom instruction (Darling-Hammond, 2007a; Harris, 2007; Ravitch, 2010). High-stakes testing demonstrates no evidence of improved scores on the NAEP (Nichols & Berliner, 2008). Some analysts contend that NCLB established a biased incentive for states to set low cut-points when defining percent proficient, given the Act's mandate of universal proficiency by 2014 (Fuller, Wright, Gesicki, & Kang 2007). Mean annual gains in reading, reported from state test scores, continued to climb after NCLB's inception in 2002, but NAEP proficiency levels hit a flat plateau or even declined in some instances (Fuller et. al., 2007). Still, annual mean score gains in math tended to range higher for state test results, compared to the gradual pace of progress revealed by NAEP scores. Meanwhile, "Texas reported that 83% of its 4th graders met the state proficiency standard in reading, whereas the most recent round of testing by the NAEP showed that

only 23% of those 4th graders met the proficiency level as defined by NAEP" (Cawelti, 2006, p. 66). If the accountability system of NCLB continues, student tests scores may increase; however, student learning may not (Amrein-Beardsley, 2009).

Comparatively, research has shown that incentives can encourage teachers to teach to the test by narrowing their focus to the material most likely to appear on the test (Palmer & Rangel, 2011; Center for Education Policy, 2008; Hollingworth, 2007; Berliner, 2011). As a result, their students' scores may be artificially inflated because the score reflects only partial knowledge of the material the students should know about the subject. Test preparation practices are perhaps artificially inflating gains in student learning and academic achievement (Amerin-Beardsley, 2009; Koretz, 2010). To judge true score gains, students are frequently tested utilizing alternative forms of assessments and more authentic forms of assessments in the equivalent content areas. These authentic forms of assessment engage more open-ended questions and problem solving questions (Sloane & Kelly, 2003). If students are able to score well on both the standardized test and the authentic assessment instrument, then a true gain is achieved (Amrein-Beardsley, 2009).

A further implication associated with high stakes assessment is the concern of test validity. Messick (1990) defines validity as "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores or other modes of assessment" (p. 1). In terms of validity, certain questions deserve consideration in test design:

- Inferences: Can the users make valid inferences about the student's capabilities in the subject from the test results achieved by that student?
- Evaluation and decision: Can users evaluate the results and use them in making decisions, including pedagogical ones, with confidence that the results are a dependable basis, free of bias effects and reflecting a comprehensive interpretation of the subject's aims?
- Range and variety: Does the variety of tasks in the test match the range of educational and performance aims – as set out in, for example, Common Core State Standards (CCSS) or other specification of the aims of the intended curriculum?
- Extrapolation: Does the breadth and balance of the domain actually assessed justify inferences about the full domain, if the former is a subset of the latter?
- Effects the test has on what happens in classrooms: Both commonsense and data from observations show that, where there are high-stakes tests, the task types in the test dominate the pattern of learning activities in most classrooms. Does this influence represent the educational aims in a balanced way? Is this a "test worth teaching to"? Given this inevitable effect on classrooms, this question summarizes a very important criterion of validity (Black, Bukrhardt, Daro, Jones, Lappan, Pead, & Stephens, 2012).

While the *Standards for Educational and Psychological Testing* (1999) recognizes test validity as a the most fundamental consideration in developing and evaluating tests (American Educational Research Association, 1999) the literature suggests most high stakes tests do not score well on the above mentioned criteria (Black et al., 2012).

Koertz (2013) recognizes the ethical obligation to evaluate the effects of testing, as such effects can decrease the extent to which inferences based on scores are accurate and justified. According to Koch and DeLuca (2012) the process of test validation is critical to the assurance of the appropriate interpretation of the results. Lai and Waltman (2008) also contend the validity of test scores are questioned when instruction is concentrated only on the knowledge and skills represented on the test. Koch and DeLuca (2012) argue, "…an important aspect of the process of validation, which is that one assessment purpose can be interpreted in different way by various stakeholders and assessment users. This is especially true for the purposes and uses of assessments, which represent complex constructs such as accountability and student achievement" (p. 101). Another implication influencing test validity is the concern of multiple secondary uses of high stakes assessment for student grades and teacher effectiveness (Koch & DeLuca, 2012).

Sampling and Assessments

Most business and government agencies consider it impractical and cost prohibitive to review all available data (U.S. Department of the Treasury, 1998). It would also take a colossal amount of time to collect data from the entire population. Sampling, with accurate results, is used to contain cost and save time (World Bank, 2007). According to Wright and Farmer (2000), sampling methods are widely used throughout the world by groups, organizations, and individuals, as well as by all levels of the government. Wright and Farmer (2000) further maintain that sampling methods are utilized successfully in virtually every field, including but not limited to, financial

institutions, businesses, healthcare, pharmaceuticals, demographics, transportation, economics, manufacturing industries, and even in education to monitor progress. One of the longest users of sampling is The United States Census Bureau. The Bureau started using sampling to estimate unemployment in 1937 and implemented decennial census in 1940 (U.S. Census Bureau, n.d.) and has been using sampling ever since. The U.S. Census Bureau (n.d.) contends “sampling made it possible to ask additional detailed questions of the population without unduly increasing cost or respondent burden” (¶ 5). Over the years, the census bureau has expanded the use of sampling to virtually all their data collection efforts. Moreover, they use sampling techniques in planning, development and implementation of their statistical programs and operations to ensure they are of high quality and efficient.

Review of Sampling Strategies among National and International Studies

In education, sampling is utilized mainly in national and international large-scale assessments and in conducting educational surveys. Sampling on state summative assessments is mainly used in the field-testing of items to be included in state assessments. In Texas, the Texas Education Agency (TEA) applies various sampling strategies in their test development and research activities. By contrast, educational policies in most highly regarded Nordic countries do not dictate annual individual student testing for all students and or test-based accountability. Finland’s intelligent accountability is a more trust-based accountability system valuing teacher professionalism and does not mandate annual individual student testing (Savola, 2012; Sahlberg, 2006). Rather, it advocates for professionalism, valid and reliable measures

that do not undermine the purpose of education, and measures that encourage the development of the whole child, emphasizing student self-evaluation (Cowie & Croxford, 2007).

In comparison to the U.S. and many other industrialized nations, the Finns have implemented a profoundly different model of educational reform based on a balanced curriculum and professionalization, not testing. Not only do Finnish educational authorities allocate more recess to students than their U.S. counterparts - 75 minutes a day in Finnish elementary schools versus an average of 27 minutes in the U.S; they also promote arts and crafts, more learning by doing, rigorous standards for teacher certification, higher teacher compensation, and appealing working conditions (Sahlberg & Hargreaves, 2011). This is a far cry from the U.S. concentration on testing in reading and math since the enactment of NCLB (2001), which has led school districts across the country to significantly narrow their curricula.(Center for Education Policy, 2009).

By contrast, assessments in Finland are intended to guide and support learning for students to receive frequent and varied written and oral feedback throughout the year (Savola, 2012). Primary schools are considered “testing-free zones” and are reserved for learning and for sustaining children’s natural curiosity (Sahlberg & Hargreaves, 2011). The Finnish students do not receive numerical grades through fifth grade and normally they are not compared with other students on their performance on any standardized assessments administered (Cowie & Croxford, 2007). Teachers spend more time planning their curriculum and focus on teaching and learning (Sahlberg & Hargreaves, 2007). Teachers give students feedback in a variety of ways, which include, but are not limited to, diagnostic, formative, performance, and summative

evaluations (Sahlberg, 2011). Hence, the evaluation of student outcomes is the responsibility of the teachers and schools (Sahlberg, 2007).

Annual high-stakes tests are not required of students in Finland (Sahlberg, 2007). The only high-stakes test all students are required to participate in is the Matriculation Examination at the culmination of secondary school. The National Board of Education in Finland conducts studies to monitor the state of education in the country; however, the studies have no consequences on individual students, teachers, or schools. These studies rely on sample-based assessments (Sahlberg, 2007), which normally sample 10% of the students (Savola, 2012). These sample-based assessments are developed locally by teachers to evaluate its national standards at only two grade levels (Darling-Hammond, 2010). Sahlberg (2006) points out that "sample-based assessments … together with continuous teacher-made classroom assessments provide well-founded and immediate feedback that promote insight into performance and support planning and decision making about what works and what should be improved" (p. 22).

Likewise, countries such as New Zealand, Scotland and Australia employ various sampling methods that test 10-20% of the entire student population (Australian Curriculum, Assessment and Reporting Authority 2008; National Education Monitoring Project, 2010; Scotland, 2012). Australia's National Assessment Program (NAP), for example, is based on a two-stage cluster sample. Schools are selected with a probability proportional to size, and disproportionate sampling ratios among strata, from a national stratified sampling frame. Students were selected using a simple random sample that included 15 students from within each sampled school. Weights were

applied to the sample in order to estimate population parameters and confidence intervals associated with each estimate were computed using replication methods (The Australian Council for Educational Research, 2008). Furthermore, the Scottish Survey of Literacy and Numeracy (SSLN, 2011) utilizes simple random sampling with a 50/50 gender split in the pupil sample from each school (2011). The overall target pupil sample size of about 4,000 pupils per stage, selected at random, was based on two pupils per stage in primary schools and 12 pupils per stage in secondary schools. Weighting was then applied to the data to account for the fact that sampled pupils were representing schools of varying size (SSLN, 2011).

Continuing in this area of research, assessment frameworks in Sweden, Britain, Wales, Ireland, Queensland, Australia, and Hong Kong rely exclusively on local assessments that emphasize inquiry, application of knowledge, and are developed and scored by teachers (Darling-Hammond, 2010). As an example, for the past 42 years, Queensland, Finland, and Sweden have been using school-based assessments developed and scored by teachers "in relation to the national curriculum guidelines and state syllabi (also developed by teachers), and moderated by panels that include teachers from other schools and professors from the university system" (Darling-Hammond, 2010, p. 290).

The term accountability is not a common phrase in Finnish educational policy (Sahlberg, 2010). Ninety-nine percent of all Finnish students' complete basic school and 90% graduate from high school (Sahlberg, 2006). Student performance on the international assessments (Savola, 2012) is also high. According to the 2003 Programme International Student Assessment (PISA) data, the percentage of 15-year-

olds who scored at the lowest proficiency level of 1 or 0 in math was 6.8% for Finnish students contrasted with 25.7% of their counterparts in the U.S. The average for the participating Organization for Economic Cooperation Development (OCED) countries was 21.4% (OCED, 2004). Further, 77% of the Finnish students obtained a proficiency level of three or higher suggesting these students had acquired the literacy skills to thrive in today's knowledge societies (OCED, 2004).

Sampling Designs

Sampling or testing a portion of the students versus testing the entire population has several advantages which include: cost reduction, conservation of time, reduction of the number of assessments administered, manageability, and feasibility when it is impossible to reach the target population. More specifically, the use of stratified random sampling (SRS) has been shown to yield a representative sample that leads to more accurate estimation of the parameters under investigation (Texas Education Agency [TEA], 2008). In SRS, the population is divided into strata based on population characteristics. According to Cochran (1977), "If each stratum is homogeneous in that the measurement vary little from one unit to another, a precise estimate of any stratum mean can be obtained from a small sample in that stratum; these estimates can be combined into a precise estimate for the whole population" (p. 90). The purpose of sampling theory is to make sampling more efficient. It attempts to develop methods of sample selection and estimation that provide, at the lowest possible cost, estimates that are precise. Equally significant, SRS ensures that the drawn sample will be distributed in the same manner as the target population (Bryman, 2008).

Stratification is the categorization of sampling units in the frame according to a characteristic prior to drawing the sample. Stratification is mostly used to improve the efficiency of the sample design and to apply various sample designs to individual groups while ensuring adequate representation in the sample of specific groups from the target population (Joncas & Foy, 2011). There are essentially two types of stratification, explicit and implicit. Explicit stratification consists of building separate sampling frames according to the stratification variables under consideration. Examples of explicit stratification variables could be states or regions of a country. Meanwhile, implicit stratification sorts schools distinctively within each explicit stratum by a set of selected implicit stratification variables. Examples of implicit stratification variables could be type of school or minority composition. This type of stratification is a way of ensuring a strictly proportional sample allocation of schools across all implicit strata (U.S. Department of Education, 2013; Institute of Education Sciences, 2011; National Center for Education Statistics, 2011).

There are many sampling techniques that can be utilized to make inferences. The choice of which sampling technique to use depends on the objective of the study (U.S. Department of the Treasury, 1998). Two widely used techniques are probability and non-probability sampling. In probability sampling, each student in the target population has a known non-zero probability of being selected (Murphy & Schulz, 2006). Some of the probability sampling techniques include simple random sampling, systematic sampling, stratified random sampling, and cluster random sampling. By contrast, in non-probability sampling, not all students have an equal chance of being selected. Murphy and Schulz (2006) caution against using non-probability sampling

techniques. I note that: “the use of non-probability sampling methods may lead to controversy and ultimately criticism of [the sampling design]” (Murphy & Schulz, 2006, p. 4). Likewise, the US Department of Treasury (1998) cautions against inferences made based on non-statistical or judgmentally derived samples. The World Bank favors using stratified random sampling (SRS) to ensure that specific groups that might ordinarily be missed are included in the sample (World Bank, 2007).

Large scale assessments such as the NAEP, Trends in International Mathematics and Science Studies (TIMSS), Programme for International Student Assessment (PISA), Progress in International Reading Literacy Study (PIRLS), and National Assessment of Education Progress Transcript Studies use sophisticated sampling designs that incorporate stratification and multistage sampling to assess and collect data from samples of students rather than testing the entire student population (National Center for Education Statistics, 2011). These studies apply “rigorous school and classroom sampling techniques so that achievement in the student population as a whole may be estimated accurately by assessing just a sample of students from a sample of schools” (Joncas & Foy, 2011, p. 1). Using complex sample designs enables one to draw on a small sample which in most cases encompasses a few hundred to a couple thousand subjects to make inference to a target population that may be 10 times larger than the sample (Yang, 2008).

The NAEP studies utilize a probability sample design in which schools and students have a known probability of being selected to assess a representative sample of students rather than the entire student population (National Center for Education Statistics, 2011). For example, the 2007 sample design for the NAEP aimed to attain a

nationally representative sample of students in the target population at the time of assessment. The NAEP study adopted a 2-stage sample design. In the first stage, samples of schools were selected with probability proportional to the grade level enrollment, and in the second stage, students were sampled within the selected schools (NCES, 2007). Correspondingly, the research design for the 2009 NAEP state assessments also utilized a 2-stage sample design to model probability samples of schools and students to represent the diverse United States population.

The PISA studies also utilize two-stage stratified random sampling to evaluate education systems all over the world every three years to assess 15-year-olds' proficiencies in reading, math and science (OECD, 2004). In 2008, the National Center for Education Statistics implemented a two-stage sampling design to identify students to be included in the 2009 PISA. In their sampling design, schools were stratified by census region (northeast, midwest, south and west) and school type. Within each stratum, schools were sorted by school characteristics and a systematic sample was randomly selected. The sample size for each school was proportionate to the school size. The target sample was 35 students from each school; however, the study sampled 42 students to account for students who might not attend school on the day of testing (NCES, 2011).

The Trends in International Mathematics and Science Study (TIMSS) and the Progress in International Reading and Literacy Study (PIRLS) also use this 2-stage random sample design in which the schools are selected on the first stage and one or more intact classes are selected on the second stage (Joncas & Foy, 2011; NCES, 2009). For example, the United States national sample from the 2007 TIMSS two-stage

sampling processes with the schools being selected in the first stage and a sample of classrooms within the sampled schools in the second stage (NCES, 2009). The TIMSS and PIRLS studies opt to select classes in the second stage because their studies attempt to determine curricular experiences, which are typically organized on a classroom basis. Moreover, sampling intact classes causes less school disruption to the school's day to day operations than individual student sampling (Joncas & Foy, 2011). It should be noted that these national and international assessments require small sample sizes that provide precise parameters. For example, TIMSS and PISA studies in Australia require 4,500 students nationwide to generate meaningful estimates with 95% confidence (Murphy & Schulz, 2006). Equally, in the 2007 U.S. NAEP studies, the NCES targeted samples of 6,500 and 9,750 fourth and eighth graders respectively in each participating jurisdiction. In the 2007 TIMSS study, the U.S. fourth grade stratified systematic sample included 257 schools and 7,896 grade students. The 8th grade sample included 239 schools and 7,377 students (NCES, 2009). The 2011 PIRLS studies proposed to sample 150 schools and a student sample of 4,000 per each target grade per country (Joncas & Foy, 2011). These sample sizes were meant to yield aggregate estimates with equal precision for all participating jurisdictions. Further, these studies are meant to provide valid, reliable and timely data on achievement of U.S. students compared to their counterparts in other countries, while minimizing the burden on schools, teachers and students (Joncas & Foy, 2011). All the sampling designs the national and international studies apply have yielded precise population parameters that are used to make inferences on the status of education in the United States and in other countries, respectively (NCES, 2009).

In addition to sampling students to be assessed, the TIMSS studies apply matrix-scaling techniques whereby assessment item pools are divided so that each sampled student responds to a portion of the test items (Martin, Mullis, & Chrostowski, 2003). Matrix-scaling accords NCES an opportunity to test a wide range of items while keeping the response burden on individual students to minimum. In 2003, for example, the TIMSS fourth grade math and science tests had 313 items, which were allocated to 28 matrices, which resulted in the development of 12 test forms. Similarly, the Grade 5 test had 383 math and science items that were divided into 28 matrices.

Another area where sampling has been used in education is in the administration of educational surveys. For example, in conducting educational surveys in Australia, Murphy and Schulz (2006) used a multiple-stage cluster sampling approach. In the first stage, a group of schools were sampled and then, within the sampled schools, students were randomly sampled based on their grade level and demographics. Murphy and Schulz (2006) used cluster sampling because it was cost effective and reduced the burden of surveying entire populations of interest. They pointed out that the disadvantage of cluster sampling is that it requires a larger sample to attain the same precision as the simple random sampling.

Prior to sampling, Murphy and Schulz (2006) recommend ensuring that the lists of schools and students to sample from are comprehensive, accurate, and current. Further, they recommend stratifying the sampling frame by key characteristics related to the outcomes of interest, which include, but are not limited to, location, socio-economic status (SES), school size, and school level. Stratification improves the precision of the estimates, helps identify sample size desired for each stratum, ensures all specific

groups of the target population are adequately represented, and ensures smaller schools are not overburdened (Joncas & Foy, 2011, Murphy & Schulz, 2006).

Stratifying on excessive criteria, however, proves counterproductive. The requirements imposed by fine stratification often increase the sample size. Likewise, the number of cases identified in the wrong stratum may increase with the number of strata, especially those that are based on more volatile or less reliable data; for instance, the number of staff members or student enrollment. Stratification can increase both statistical and overall efficiency, moderating the size and cost of the sample while sustaining the level of reliability (Greaney & Kellaghan, 2012). Stratification is also significant in regards to skewed populations. Stratification also prevents drawing an unusual sample. In SRS, sample selection is determined entirely by chance. Stratified sampling attempts to restrict potentially extreme samples by taking steps to ensure that certain categories of the student population are included in the sample (Greaney & Kellaghan, 2012).

The sample size needed to make precise estimates depends on the population of interest, and the confidence and precision level desired (World Bank, 2007). Krejcie and Morgan (1970) recommended sampling about a third of the total population when a population of interest is less than a thousand. For larger populations, such as a population of a million, they recommended sampling about 10% (World Bank, 2007). Whenever possible, the World Bank (2007) recommends sampling a larger group to compensate for the possibility of a less than perfect response rate. Joncas and Foy (2011) note that nonresponse and/or nonparticipation can lead to sample bias and misleading results. They maintain that TIMSS and PIRLS studies aim for 100%

participation; however, they recognize that some degree of non-participation may be unavoidable. With this in mind, they have a criterion for accepting samples with less than perfect participation. For a sample to be accepted, it must have either:

- Eighty-five percent of the originally sampled schools participating; AND
- Ninety-five percent of the originally sampled classroom participation; AND
- Eighty-five percent of the originally sampled student participation from sampled schools; or
- Seventy-five percent of the originally sampled combined school, classroom and student participation.

Classrooms with less than 50% participation rates are not included in the overall aggregates (Joncas & Foy, 2011).

Although sampling is subject to some error irrespective of the method used (World Bank, 2007), the precision of the estimates of the sampled data to the target population is of paramount importance (Joncas & Foy 2011). Higher sample sizes decrease the sampling error (Bryman, 2008). Simply put, decisions on acceptable confidence intervals and margins of error have to be resolved. The less sampling error one is willing to endure, the larger the sample size must be (Bryman, 2008). The standard confidence level is 95%. Lower confidence levels, say 90%, require smaller sample sizes. Conversely, higher confidence levels e.g. 99% require larger sample sizes (World Bank, 2007). The TIMSS and PIRLS studies apply 95% confidence intervals to their studies. To meet this confidence interval, TIMSS and PIRLS sampling standards provide for a standard error no greater than .035 standard deviation units for the country's mean achievement (Joncas & Foy, 2011).

Conclusions

This chapter reviewed the relevant literature for the present study. Articles discussing the historical development and various limitations of assessments and their use were included. The use of sampling designs and stratification were discussed. Stratified random sampling was evaluated as demonstrated in Finland and sampling rates were reviewed. Finally, comparisons of international assessments that employ stratified random sampling were examined.

CHAPTER 3

METHODOLOGY

Research Design

While sampling techniques have been used effectively in education research and practice, it is not clear how stratified random sampling techniques apply to high stakes testing in the current educational environment in Texas. Therefore the purpose of this sampling study was two-fold: 1) to determine if stratified random sampling was a viable option for reducing the number of students participating in Texas state assessments, and 2) to determine which sampling rate provided consistent estimates of the actual test results among the subpopulations of students.

This study examined how stratified random sampling can be used as a method of reducing the quantity of state-administered tests in order to afford more educational funding and time to curricula that are more closely related to long-term student academic and career success. Specifically, this study examined the utility of stratified random sampling in providing accurate estimates of population scores on math and reading assessments and the percentage of students passing those assessments. Further, this study examined how the effects of stratified random sampling bias in population estimates differ by socioeconomic status, English proficiency, and placement in special education classes. Lastly, this study examined the effects of stratified random sampling on bias in estimates of student growth in assessments over time. The study examined scale scores, percent passing, and student growth over a three-year period on state-mandated assessments in reading, mathematics, science, and social studies. Four sampling rates were considered (10%, 15%, 20%, & 25%) when analyzing student

performance across demographic variables within and across each participating district.

Participants

The participants in this study included students in Grade 5 districts across the state of Texas. Table 1 displays demographic data for each of the five districts. More specifically, three districts were identified as mid-size districts that included 15 elementary campuses, 5 middle schools, and 6 high schools, while two districts were identified as small districts that included one elementary campus, one middle school campus, and one high school.

Table 1

Enrollment and School Data for the Five School Districts

District	District Size	Total Enrollment	Number of Elementary Schools	Number of Middle Schools	Number of High Schools
1	Mid-Size	55,836	42	13	8
2	Mid-Size	4,229	4	1	1
3	Mid-Size	8,139	6	4	2
4	Small	155	1	1	1
5	Small	102	1	1	1

Table 2 describes student demographics for each district. Regarding descriptive measures among the campuses examined in this study the students enrolled in District 1 (*N* = 55,836), 42.8% were classified as White, 21.9% Hispanic, 11% African-American, while 20.5% are identified as Asian. In this study, further student demographic variables representative of District 1 indicated 25.9% low socioeconomic status (SES), 10.1% special education (SPED), while 12.0% are limited English

proficient (LEP).

Table 2

Student Demographics for the Five School Districts

District	African American	Anglo	Asian	Hispanic	Native American	Low SES	SPED	LEP
1	11%	42.80%	20.50%	21.90%	0%	25.90%	10.10%	12%
2	32.80%	36.20%	1.70%	28.10%	0.40%	75.80%	14.70%	19.20%
3	68.70%	10.70%	0.60%	18.50%	0%	63.80%	8.80%	5.50%
4	0.60%	72.30%	1.30%	24.50%	0%	48.40%	11.60%	0%
5	0%	78.40%	0%	21.60%	0%	32.40%	9.80%	7.80%

Variable Examined

Dependent Variables

Dependent variables for the study included bias in mean reading and math scores on the State of Texas Assessment of Academic Readiness (STAAR) and bias in the percentage of students passing these assessments. Additionally, growth in average reading and math scores were assessed from Grades 3 to 5 in a subset of schools, where growth was defined as changes in students' *z*-scores on the tests over time. As of spring 2012, the STAAR test has replaced the Texas Assessment of Knowledge and Skills (TAKS). The STAAR program at Grades 3–8 assesses the same subjects and grades that are currently assessed on TAKS. At the high school level, however, grade-

specific assessments have been replaced with five end-of-course (EOC) assessments: Algebra I, Biology, English I, English II, and U.S. History (TEA, n.d.).

Independent Variables

For the ANOVA analyses, sampling rates were treated as categorical variables (10% = 1, 15% = 2, 20% = 3, and 25% = 4). Campuses were also classified as categorical variables, and were assigned three-digit numbers based on district (1 through 5), campus type (1 = elementary, 2 = middle school, and 3 = high school), and campus number within the district by campus type. Campuses were ordered alphabetically and were assigned numbers based on the total number of campuses by type. In other words, the first number in the campus label represented the district in which the campus is located, the second number represented the campus type, and the third number represented the campus number within the district based on campus type. For example, the label 321 indicates that the campus is from District 3, the campus is a middle school campus, and the campus is the first middle school campus (based on alphabetizing of campus names). The first student demographic variable was ethnicity, where African American = 1, Anglo = 2, Asian = 3, Hispanic = 4, Native American = 5, and 6 = Other. Second, participants identified as receiving special education services were coded as 1, whereas students not receiving special education services were coded as 0. Lastly, limited English proficient (LEP) students were coded as 1, whereas non-LEP students were coded as 0.

Procedure

Initially, student-level data was obtained from each of the five participating districts. Data was then obtained from the student information system database with the permission of each district's superintendent. The data was then formatted and entered into the Statistical Analysis System (SAS©, www.sas.com) database, and screened for erroneous entries. Subsequently, data labels were defined and variables recoded for use in the study. Next, descriptive statistics including frequencies and percentages for categorical variables (i.e. gender, ethnicity, LEP, SPED), as well as means and standard deviations for continuous variables (i.e., average test scores and percent of students passing tests) were first calculated. Additionally, histograms were created to examine variable distribution and help detect errors in data entry.

Method

Stratified sampling is a method for utilizing auxiliary information about the population to partition the population into regions or strata, where samples are selected by design from each stratum (Thompson, 2012). Essential to the performance of a stratified sampling scheme is within-stratum homogeneity, where more precise estimates of population parameters are obtained if subjects in each stratum are more similar with regard to the variable of interest. Each stratum is considered an independent sub-population from which a sample is drawn independently and parameters are estimated for that sub-population. Estimates from all sub-populations are finally combined to arrive at estimates of population parameters. With homogeneous strata, stratified sampling provides more precise estimates of population parameters,

compared to simple random sampling of the same size (Thompson, 2012; Lehtonen & Pahkinen, 2004).

There are several possible strategies to allocate the total sample size to different strata. If the sample sizes are proportional to the population size within each stratum, proportional allocation will be achieved, which results in equal sampling rates across strata. This technique usually demonstrates more precise estimates for characteristics of the whole population.

Since a sample is drawn independently from each stratum, sample size determination can also be done independently for each with regard to stratum characteristics (size, variance and practical concerns). Proportional allocation is an approach where a proportion of the total sample size is allocated to each stratum based solely on the stratum's size relative to the population. Variance of population characteristics can then be calculated using the following formula:

$$V(\hat{Y}) \approx (1-f)\sum_{i=1}^{m} W_i V(\hat{Y}_i);$$

where, $V(\hat{Y})$ is the variance of the population estimator, $f = \frac{n}{N}$ is the sampling rate (percentage of population in the sample), $W_i = \frac{N_i}{N}$ is relative size of the stratum, and $V(\hat{Y}_i)$ is the within-stratum variance. The summation in this formula runs through all strata from 1 to m. Given a desired precision, sampling rate (f) can be calculated using the same formula (Rao, 2000). Alternative to the mathematical approach discussed above, the current study sought to examine the sensitivity of the stratified sampling scheme by comparing the precision of four sampling rates for the population of students, using resampling techniques.

Sampling Procedure

Statistical analysis software (SAS®) was utilized to first identify the general population of interest and subsequently identify student scores within the population based on ethnicity, placement in regular or special education (SPED) classes, and classification as limited English proficient (LEP). Once the specific student population was identified, SAS® was used to obtain stratified random samples of students in each middle school campus based on probability proportional to size. Specifically, 1,000 re-samplings were conducted for each of the four different sampling rates (10%, 15%, 20%, and 25%) to determine the minimum rate that would provide accurate estimates of population parameters.

In order to obtain the stratified random samples, the target population was divided into subpopulations (*strata*), and samples were drawn independently from each stratum. Stratified random sampling was considered preferable to simple random sampling in the current study, because it generally produces more precise estimates of the total population, and allows for the use of various sampling rates in different strata to control the precision of sub-group estimates.

Precision of estimates is assessed using confidence intervals and margins of error. A confidence interval is an interval estimate for a parameter consisting of a range of values, with which a specific level of confidence is accompanied (Gravetter and Wallnau, 2004). In contrast to a point estimate, an interval estimate specifies a range of values within which the population value is expected to lie. The wider such an interval, at a given confidence level, the less precise the estimate is.

Moreover, the margin of error reflects the uncertainty associated with a

population estimate when using a random sampling design. In other words, since only a proportion of the entire population (i.e. the sample) is being examined, any generalization to the population based on this sample involves uncertainty and error. This uncertainty or sampling error is reflected in the standard error associated with each estimate, as well as the confidence interval calculated for each. The smaller the standard error or equivalently the shorter the width of the confidence interval, the more precise our estimates are. Margin of error for each estimate is defined as half the width of the confidence interval associated with that estimate. Hence at 95% confidence, the margin of error based on the normal approximation for every estimate would be:

$$Margin\ of\ Error = 1.96 \times SE;$$

Where *SE* is the standard error associated with that estimate, which depends on sample size, population characteristics (variability in the population), and the sampling design. In the case of stratified random sampling the standard error for an estimate of the population mean can be calculated using the following formula:

$$SE = \frac{1}{N}\sqrt{\sum_{h=1}^{H} N_h^2 \left(\frac{N_h - n_h}{N_h}\right)\frac{s_h^2}{n_h}};$$

In which N is the total population size, N_h is the size of stratum h, s_h^2 is the estimated variance in stratum h, and n_h is the sample size for stratum h. The summation runs through all the strata 1 to H. If strata variances are known, the known variance for each stratum σ_h^2 can be replaced in the formula for the estimate s_h^2 (Thompson, 2012).

The general rule relative to acceptable margins of error in educational and social

research is as follows: For categorical data, 5% margin of error is acceptable, and, for continuous data, 3% margin of error is acceptable (Krejcie & Morgan, 1970).

While the precision of estimates for the subpopulations mainly depends on the sample size and the spread of scores within these subpopulations, varying sampling rates across strata might be optimal for enhancing the precision of estimates for special groups and thus oversampling may be required. To study how sample size and spread among scores impact the precision of parameter estimates in varied sample sizes, four different sampling rates were considered (10%, 15%, 20%, and 25%). For example, assume the total population of one campus included 1,000 regular education students and was comprised of 500 Anglo students, 400 Hispanic students, 40 Asian students, and 60 students classified as other. In this example, 100 students would be randomly selected (at the 10% sampling rate), and the sample would consist of 50 Anglo students, 40 Hispanic students, 4 Asian students, and 6 students identified as other.

In addition to precision, the performance of the sampling rate is assessed by looking at the bias of calculated estimates. A sample statistic is considered to be biased if on the average it overestimates or underestimates the corresponding population parameter (Gravetter & Wallnau, 2004). To determine bias in the dependent variables, 1,000 sampling replications were conducted and bias for the mean reading and math STAAR scores and the percentage of students passing these assessments for each sampling rate in each district was calculated. Specifically, bias was calculated as: ((true population value – calculated value)/ true population value)*100.

Effects of Sampling Rate, Ethnicity, and Bias in Sampling Rate

In order to examine bias in sampling rate, student and campus-level demographics in average state-mandated reading and math scores and bias in percentage of students passing these tests, a series of factorial analyses of variance (ANOVAs) models were conducted. Specifically, these Factorial ANOVAs assessed how bias in test scores and percentage of students passing differed across sampling rate, ethnicity, and campus. Additionally, two-way interactions (i.e., sampling rate by campus, sampling rate by ethnicity, ethnicity by campus) were examined. In order to assist with parsimonious interpretations of the models, the three-way interactions were not considered. Lastly, 95% confidence intervals (CIs) and effect sizes for each model were examined to aid in the interpretation of the models.

Effects of Sampling Rate and Campus on Bias in Economically Disadvantaged Students

Factorial ANOVAs were conducted to examine the effects of sampling rate and campus on bias in average reading and math scores and bias in percentage of students passing these tests in economically disadvantaged, SPED, and LEP students. These ANOVAs examined main effects of sampling rate and campus, as well as the two-way interaction between these variables. Lastly, 95% confidence intervals (CIs) and effect sizes for each model were examined to aid in the interpretation of the models.

Effects of Sampling Rate and Campus on Bias in Limited English Proficient (LEP) Students

Lastly, ANOVAs were utilized to examine the role of sampling rate, ethnicity, and campus on bias in mean growth on reading and math scores. These ANOVAs assessed

how average growth in test scores (changes in student *z*-scores over time) differed across sampling rate, ethnicity, and campus. Additionally, two-way interactions (i.e., sampling rate by campus, sampling rate by ethnicity, ethnicity by campus) were examined. The three-way interaction was not considered. Lastly, 95% confidence intervals (CIs) and effect sizes for each model were examined to aid in the interpretation of the models.

CHAPTER 4

RESULTS

This sampling study sought to determine if stratified random sampling was a viable option for reducing the number of students participating in Texas state assessments, and to determine which sampling rate provided consistent estimates of the actual test results among the subpopulations of students. The study examined scale scores, percent passing, and student growth over a three-year period on state-mandated assessments in reading, mathematics, science, and social studies. Four sampling rates were considered (10%, 15%, 20%, & 25%) when analyzing student performance across demographic variables within and across each participating district. The districts participating in the study included five districts in the North Texas region.

The sampling design for this study employed stratified sampling which divides the target population into subpopulations called strata, and then samples are drawn independently from each stratum. For purposes of the current study the subpopulations are defined by demographic variables present in the target population at the individual campus level and divided into ethnic subpopulations. There are several possible advantages of allocating the total sample size to the different strata. If the sample sizes are proportional to the population size within each stratum, proportional allocation is achieved which results in equal sampling rates across strata. This technique usually demonstrates more precise estimates for characteristics of the whole population. While the precision of estimates for the subpopulations mainly depends on the sample size within these subpopulations, varying sampling rates across strata might be reasonable to enhance the precision of estimates for special groups and thus oversampling may

transpire. Therefore, stratified sampling demonstrates two main advantages; it generally conducts more precise estimates of the total population. In addition, this sampling method allows using various sampling rates in different strata to control precision of sub-group estimates as a result allowing true sample estimates to be realized (Thompson, 2012).

Average TAKS Reading and Math Scores

Table 3 displays the descriptive measures of the TAKS reading and math mean scores comparing the sample results to the population values by ethnicity. Regarding the sample mean scores for reading and math, the reading mean scores among Asian students ranged from 799.55 (*SD* = 100.53) (95% CI = 767L - 832U) for the 10% sampling rate to 822.16 (*SD* = 45.95) (95% CI = 807L - 837U) for the 25% sampling rate. The mean population reading score among all Asian students was 820.58 (*SD* = 121.51). Similarly, the sample mean reading scores among White students ranged from 811.14 (*SD* = 22.09) (95% CI = 804L - 818U) for the 15% sampling rate to 820.44 (*SD* = 18.97) (95% CI = 814L - 829U) for the 10% sampling rate. The mean population reading score among White students was 818.28 (*SD* = 82.30.). Similar results were reported for African American and Hispanic students. Note in each subgroup, the 95% CI captured the true population value. Regarding the Black and Hispanic students, the mean reading scores among the sampling rates were similar to the population values with the 95%CIs capturing the true population values. As expected, as the sampling rate increases the standard errors decreased overall. For example, the standard errors were greater for the 10% and 15% sampling rates compared to the 20% and 25% sample

rates.

Concerning the accuracy of the samples replicating the true population values, the sample results were very similar for each sampling rate among ethnic groups for reading. However, there were greater disparities between the sample and population results related to TAKS math. A plausible explanation for the increased variation in TAKS math scores among samples is the larger standard deviations associated with TAKS math scores. With the exception of Asians, the TAKS math standard deviations are almost twice as large as the TAKS reading standard deviations. It is important to have some knowledge of the standard deviation of the true population prior to conducting a stratified random sample due to the fact that more disperse populations require a larger sample size in order to attain the same level of precision for sampling estimates (Thompson, 2012). In some instances, it may be necessary to increase the sampling rate among strata with increased standard deviations to obtain results that closely replicate the true population values.

Table 3

Descriptive Measures-Comparing the Average Scale Score on the TAKS Reading and Math Assessments among Regular Education Students in Seventh Grade by Ethnicity

		Sample						Population				
		Reading			Math			Reading		Math		
Ethnicity	Sampling Rate	Mean	*SE*	95%CI	Mean	*SE*	95%CI	Mean	*SD*	Mean	*SD*	Number of Students
A	10	799.55	16.10	767-832	852.86	20.89	811-895	820.58	121.51	885.91	165.34	746
	15	818.36	10.75	796-840	872.10	8.54	855-889	820.58	121.51	885.91	165.34	746
	20	810.33	9.65	791-830	882.11	13.96	853-910	820.58	121.51	885.91	165.34	746
	25	822.16	7.35	807-837	876.54	10.63	855-898	820.58	121.51	885.91	165.34	746
B	10	767.61	7.15	753-782	700.26	18.66	663-738	765.18	88.65	686.56	213.55	385
	15	771.27	6.84	757-785	693.56	19.10	655-732	765.18	88.65	686.56	213.55	385
	20	760.95	5.35	750-771	687.00	14.82	657-717	765.18	88.65	686.56	213.55	385
	25	761.20	5.91	749-773	682.16	14.37	653-711	765.18	88.65	686.56	213.55	385
H	10	763.61	8.26	747-780	716.32	16.67	683-750	764.54	107.81	705.10	213.94	759
	15	767.10	8.40	750-784	725.21	15.86	693-757	764.54	107.81	705.10	213.94	759
	20	768.04	6.27	755-781	712.14	13.32	685-739	764.54	107.81	705.10	213.94	759
	25	771.34	6.31	759-784	726.37	14.05	698-754	764.54	107.81	705.10	213.94	759
W	10	820.44	3.04	814-826	812.21	9.18	794-831	818.28	82.30	813.75	180.74	1740
	15	811.14	3.54	804-818	806.23	9,13	788-825	818.28	82.30	813.75	180.74	1740
	20	816.55	3.20	810-823	808.01	8.11	792-824	818.28	82.30	813.75	180.74	1740
	25	815.79	2.50	811-821	803.97	7.11	790-820	818.28	82.30	813.75	180.74	1740

The results displayed in Table 4 revealed no statistically significant differences in TAKS reading score bias among the main effects considered in the study (sampling rate, ethnicity, and campus). However, when examining the interaction terms, TAKS reading score bias differed across the interaction of sampling rate by campus (F = 1.45, df = 36,623, p = .040) and ethnicity by campus (F = 1.91, df = 36, 623, p = .001). The effect size associated with sampling rate by campus was approximately five percent (R^2 = .048), indicating that sampling rate by campus explained approximately five percent of the variance in bias, while the effect size associated with the interaction terms that included ethnicity by campus was approximately 7% (R^2 = .074, further indicating that sampling rate by campus accounted for approximately seven percent of the variance in TAKS reading score bias.

Table 4

TAKS Reading Average Scale Score Bias by Sampling Rate, Ethnicity, and Campus

Source	Type III Sum of Squares	*df*	Mean Square	*F*	Sig.
Corrected Model	.380[a]	99	.004	1.545	.001
Intercept	.000	1	.000	.104	.747
Sampling Rate	.004	3	.001	.561	.641
Ethnicity	.007	3	.002	.985	.399
Campus	.041	12	.003	1.379	.172
Sampling Rate X Ethnicity	.024	9	.003	1.095	.365
Sampling Rate X Campus	.132	36	.004	1.474	.040
Ethnicity X Campus	.171	36	.005	1.914	.001
Error	1.301	524	.002		
Total	1.681	624			
Corrected Total	1.680	623			

a. R Squared = .226 (Adjusted R Squared = .080)

The results displayed in Table 5 revealed no statistically significant differences in TAKS reading score bias among the main effects considered in the study (sampling rate, ethnicity, and campus). However, when examining the interaction terms, TAKS reading score bias differed across the interaction of sampling rate by campus (F = 1.45, df = 36,623, p = .040) and ethnicity by campus (F = 1.91, df = 36, 623, p = .001). The effect size associated with sampling rate by campus was approximately five percent (R^2 = .048), indicating that sampling rate by campus explained approximately five percent of the variance in bias, while the effect size associated with the interaction terms that included ethnicity by campus was approximately seven percent (R^2 = .074, further indicating that sampling rate by campus accounted for approximately 7% of the variance in TAKS reading score bias.

Table 5

TAKS Reading Average Scale Score Bias by Sampling Rate, Ethnicity, and Campus

Source	Type III Sum of Squares	*df*	Mean Square	*F*	Sig.
Corrected Model	.380[a]	99	.004	1.545	.001
Intercept	.000	1	.000	.104	.747
Sampling Rate	.004	3	.001	.561	.641
Ethnicity	.007	3	.002	.985	.399
Campus	.041	12	.003	1.379	.172
Sampling Rate X Ethnicity	.024	9	.003	1.095	.365
Sampling Rate X Campus	.132	36	.004	1.474	.040
Ethnicity X Campus	.171	36	.005	1.914	.001
Error	1.301	524	.002		
Total	1.681	624			
Corrected Total	1.680	623			

a. R Squared = .226 (Adjusted R Squared = .080)

To gain further insight into the statistically significant results, post-hoc confidence intervals comparing sampling rate by ethnicity were examined across campuses. The results displayed in Figure 1 show that bias among the Asian students varied greatly in Campus 3 at the 10% and 15 % sampling rates. In addition, there was a statistically significant difference in TAKS reading score bias between Campus 10 and Campuses 2, 4, and 9. When the sampling rate increased to 20% and 25% the results stabilized with confidence intervals among all campuses crossing the horizontal reference line. Note confidence intervals crossing the horizontal reference line indicate that TAKS reading score bias was not statistically significantly different from zero. In summary, it appears that a 25% sampling rate is adequate to replicate the population values.

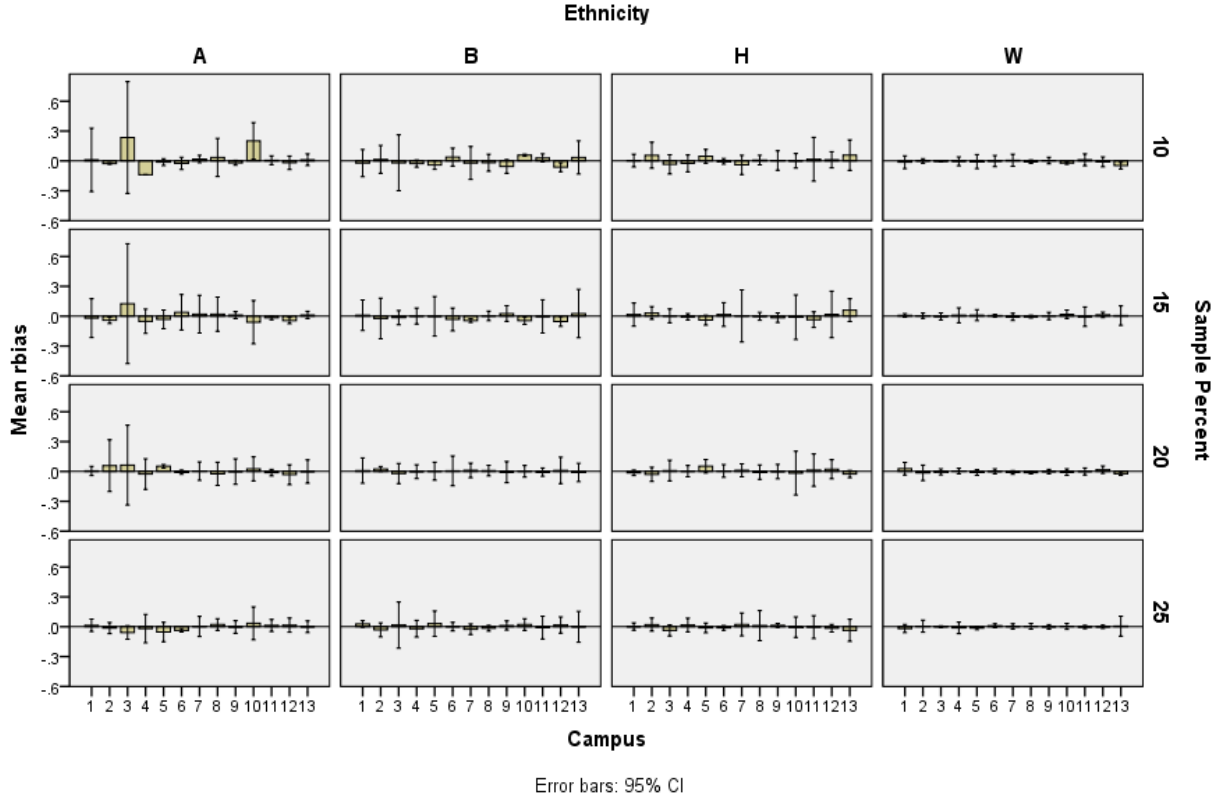

Figure 1. Comparison of average TAKS reading score bias among special education students by sampling rate and campus.

The results displayed in Table 6 revealed no statistically significant differences in TAKS math score bias among the main effects and the related interaction terms included in the model. More specifically, bias in the TAKS math scores was similar across each sampling rate, campus and ethnic group examined in the study.

Table 6

TAKS Math Average Scale Score Bias by Sampling Rate, Ethnicity, and Campus

Source	Type III Sum of Squares	*df*	Mean Square	F	Sig.
Corrected Model	.937[a]	99	.009	.978	.544
Intercept	.005	1	.005	.518	.472
Sampling Rate	.002	3	.001	.059	.981
Ethnicity	.003	3	.001	.094	.964
Campus	.059	12	.005	.507	.911
Sampling Rate X Ethnicity	.053	9	.006	.604	.794
Sampling Rate X Campus	.468	36	.013	1.342	.092
Ethnicity X Campus	.353	36	.010	1.014	.450
Error	5.074	524	.010		
Total	6.016	624			
Corrected Total	6.011	623			

a. *R* Squared = .156 (Adjusted *R* Squared = -.004)

Percent Passing TAKS Reading and Math

Table 7 displays the descriptive measures of the percentage of regular education seventh grade students passing TAKS reading and math assessments. The sample results are compared to the overall population of seventh grade students by ethnicity

within the district. Regarding the sample mean percent passing for reading, the mean percent passing and the standard errors associated with each sampling rate appear to be similar across each ethnic group. However, there are apparent differences noted in the percent of students passing TAKS math. As the sampling rate increases, the standard errors decrease. This is especially true when comparing the10% and 15% sampling rates to the 20% and 25% sampling rates.

Concerning the accuracy of the sample replicating the true population values, the sample results were similar across sampling rates among ethnic groups for both reading and mathematics. This was especially interesting given the increased standard deviation related to TAKS math in comparison to TAKS reading. Prior research has indicated that greater spread in population scores (i.e., increased standard deviations) lead to greater inaccuracy in replicating the populations values, thus requiring a larger sampling rate (increasing sample size) to obtain the true population parameter.

Table 7

Descriptive Measures-Comparing the Percentage of Students Passing TAKS Reading and Math Assessments by Ethnicity among Regular Education Students in Seventh Grade District-wide

Ethnicity	Sampling Rate	Sample						Population				
		Reading			Math			Reading		Math		
		Mean	SE	95%CI	Mean	SE	95%CI	Mean	SD	Mean	SD	Number of Students
A	10	92.95	2.61	88-98	87.86	2.75	82-93	95.37	21.03	90.28	29.65	648
	15	94.60	1.66	91-98	90.57	2.20	86-95	95.37	21.03	90.28	29.65	648
	20	93.58	1.45	91-97	89.20	2.43	84-94	95.37	21.03	90.28	29.65	648
	25	95.99	1.05	94-98	89.02	1.97	85-93	95.37	21.03	90.28	29.65	648
B	10	92.31	2.40	86-97	68.95	4.23	60-77	86.43	34.29	54.29	49.88	420
	15	95.79	1.42	93-98	62.69	3.97	54-70	86.43	34.29	54.29	49.88	420
	20	87.99	2.21	84-92	55.09	3.08	49-61	86.43	34.29	54.29	49.88	420
	25	90.84	1.86	86-94	58.48	3.26	51-65	86.43	34.29	54.29	49.88	420
H	10	88.78	2.32	84-94	69.79	3.57	63-77	86.40	34.30	61.19	48.76	809
	15	88.17	2.27	84-93	63.22	3.76	56-70	86.40	34.30	61.19	48.76	809
	20	87.37	1.82	84-91	63.88	3.08	58-70	86.40	34.30	61.19	48.76	809
	25	88.78	1.81	85-92	65.29	3.14	59-71	86.40	34.30	61.19	48.76	809
W	10	97.67	.730	96-99	81.82	1.86	78-85	96.21	19.11	77.93	41.48	1767
	15	96.27	.741	95-98	78.50	2.08	75-83	96.21	19.11	77.93	41.48	1767
	20	98.03	.526	96-99	77.65	1.83	74-81	96.21	19.11	77.93	41.48	1767
	25	97.49	.474	96-98	78.38	1.77	75-82	96.21	19.11	77.93	41.48	1767

Reading Pass Bias

Table 8 displays the factorial ANOVA results examining bias in the percentage of students passing the TAKS reading assessment across sampling percent, student ethnicity, and seventh grade campuses. The results indicated that ethnicity was the only statistically significant term in the model. Further insight revealed that bias among Black students was statistically significantly different form Hispanic students when the sampling rate was less than 20%. However, when the sampling percent was increased to 20% and 25% respectively, no statistically significant differences were noted. The effect size associated with ethnicity was .008, indicating that ethnicity explained less than one percent of the variance in bias related to the percent of students passing TAKS reading.

Table 8

Bias in the Percentage of Students Passing TAKS Reading by Sampling Rate, Ethnicity, and Campus

Source	Type III Sum of Squares	*df*	Mean Square	*F*	Sig.
Corrected Model	1.245[a]	99	.013	1.048	.368
Intercept	.005	1	.005	.425	.515
Sampling Rate	.042	3	.014	1.159	.325
Ethnicity	.141	3	.047	3.926	.009
Campus	.144	12	.012	1.003	.445
Sampling Rate X Ethnicity	.158	9	.018	1.463	.158
Sampling Rate X Campus	.383	36	.011	.887	.660
Ethnicity X Campus	.376	36	.010	.871	.686
Error	6.288	524	.012		
Total	7.538	624			
Corrected Total	7.532	623			

a. *R* Squared = .165 (Adjusted *R* Squared = .008)

Math Pass Bias

The results reported in Table 9 reveal a statistically significant difference in bias among the percentage of students passing TAKS math across sampling rates (Sampling Rate), while no statistically significant differences in bias was associated with the remaining variables. Although the sampling rates were statistically significant, the effect size was .002, suggesting that the sampling rate explained less than one percent of the variance in bias associated with the percentage of students passing TAKS math.

Table 9

Bias in the Percentage of Students Passing TAKS Math by Sampling Rate, Ethnicity, and Campus

Source	Type III Sum of Squares	*df*	Mean Square	*F*	Sig.
Corrected Model	5.489[a]	99	.055	.988	.516
Intercept	.088	1	.088	1.571	.211
Sampling Rate	.868	3	.289	5.157	.002
Ethnicity	.252	3	.084	1.495	.215
Campus	.136	12	.011	.201	.998
Sampling Rate X Ethnicity	.911	9	.101	1.804	.065
Sampling Rate X Campus	1.773	36	.049	.878	.674
Ethnicity X Campus	1.601	36	.044	.793	.802
Error	28.948	516	.056		
Total	34.505	616			
Corrected Total	34.437	615			

a. *R* Squared = .159 (Adjusted *R* Squared = -.002)

To provide further insight into how TAKS math bias differed by sampling rates and ethnicity, 95% confidence intervals (CI) were examined. The results displayed in Figure 2 shows a statistically significant difference among Blacks between the 10% and 20% sampling rates; with bias associated with the 10% sampling rate well below the horizontal reference line (horizontal reference line indicates zero bias). Concerning the remaining ethnic groups, the 95%CIs cross the horizontal reference, indicating that bias was not statistically significantly different from zero. Note how the 95%CIs narrow as the sampling rate increases, indicating increased accuracy in the results. At the 25% sampling rate, all 95%CIs crossed the horizontal reference line among all ethnic groups, which provided further evidence that a 25% sampling rate was adequate to provide results that closely approximate the population values.

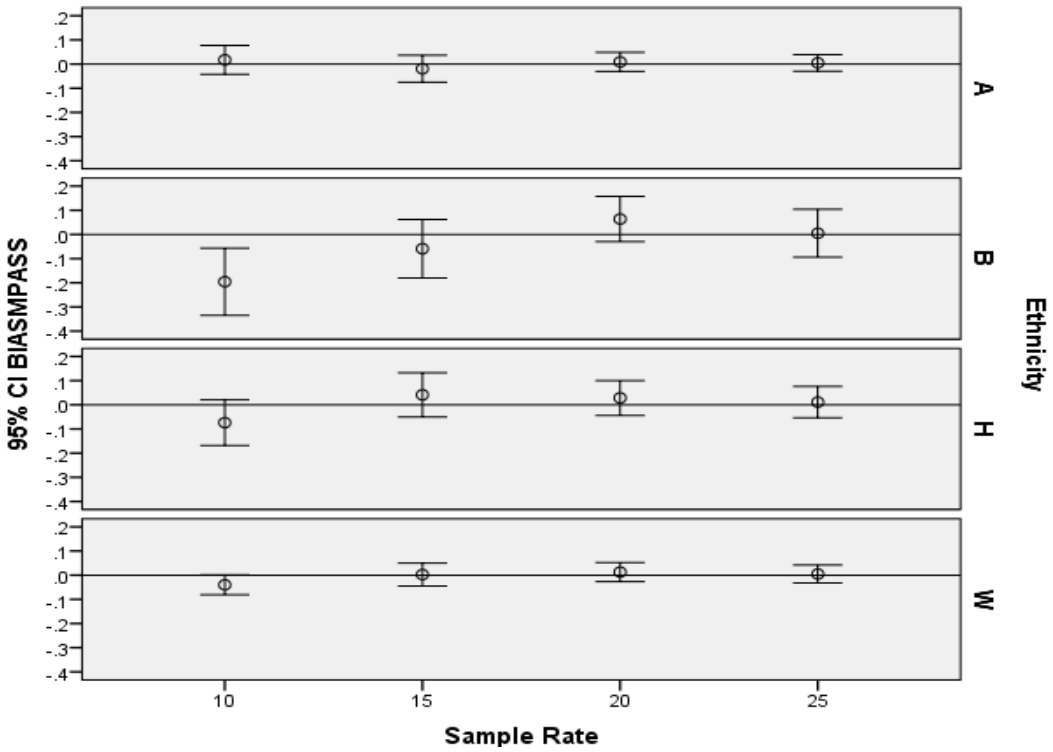

Figure 2. Ninety-five percent confidence intervals comparing TAKS math score bias by sampling rate and ethnicity.

The distribution curves displayed in Figure 3 compares TAKS math score bias by sampling rate. The results show that, as expected, as the sample rate increased, bias became less spread and approximated a normal distribution. Note the vertical reference line is associated with zero bias. The results in Figure 4 coupled with the findings reported from Figure 3 further indicate that a 25% sampling rate would be adequate to obtain accurate estimates of the population parameter.

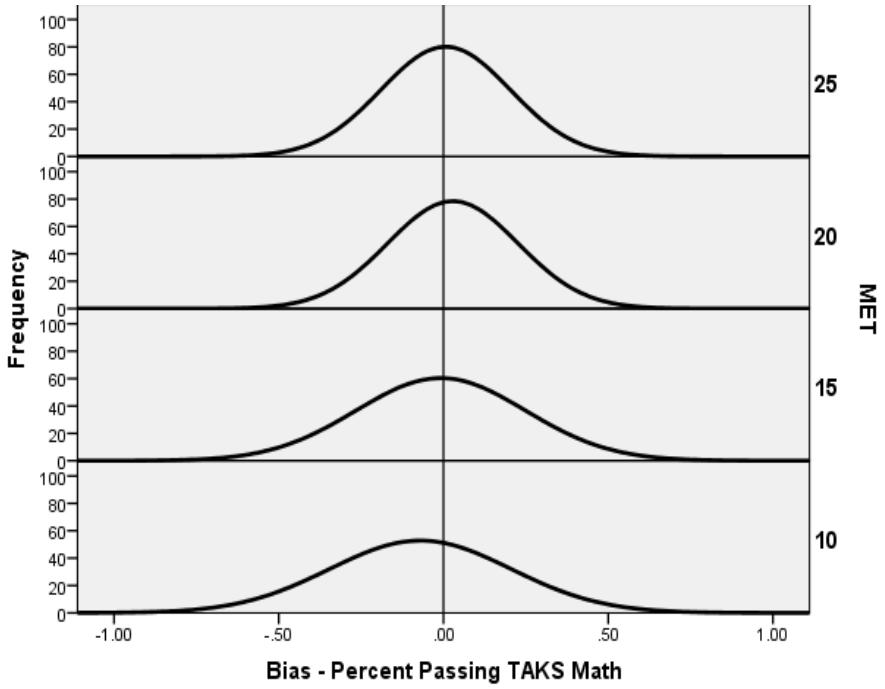

Figure 3. TAKS math score bias distribution by sample rate.

The distribution curves displayed in Figure 5 compares TAKS math score bias by sampling rate. The results show that, as expected, as the sample rate increased, bias became less spread and approximated a normal distribution. Note the vertical reference line is associated with zero bias. The results in Figure 4 coupled with the findings reported from Figure 3 further indicate that a 25% sampling rate would be adequate to obtain accurate estimates of the population parameter.

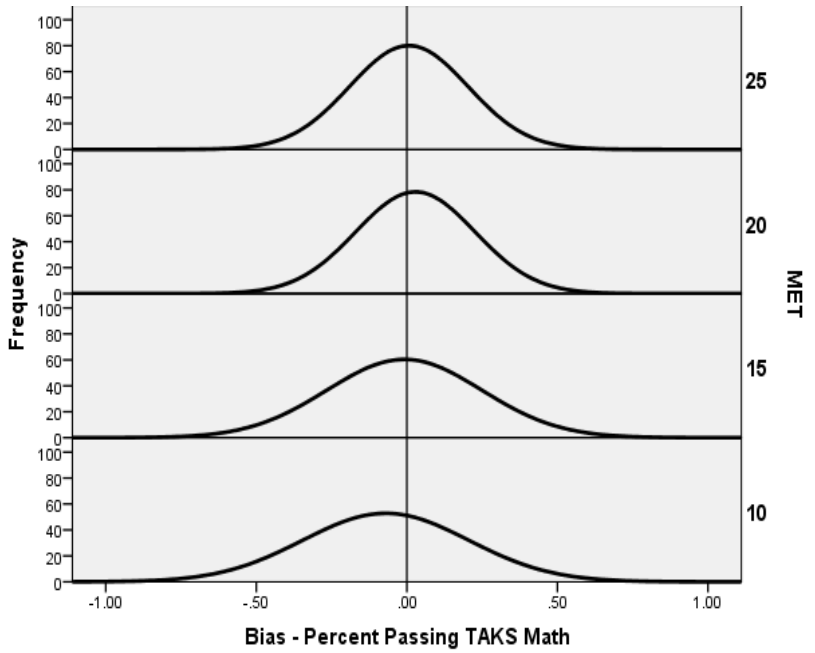

Figure 4. TAKS math score bias distribution by sample rate.

TAKS Reading and Math Score by Sampling Rate among Economically Disadvantaged Students

The results displayed in Table 10 reports similar outcomes across sampling rates for the reading assessment, with average scores ranging from 744.61 (*SE* = 44.72) (95% CI = 736_L - 752_U) at the 10% sampling rate to 751.59 (*SE* = 42.14) (95% CI = 744_L - 758_U) at the 15% sampling rate. To serve as a reference point, the mean population value of the TAKS reading score among economically disadvantaged students was 752.65 (*SD* = 103.11).

Regarding the sample results for TAKS math, average math scores ranged from 675.78 (*SE* = 103.51) (95% CI = 657_L - 693_U) for the 10% sampling rate to 706.89 (*SE* = 88.41) (95% CI = 691_L - 722_U) for the 15% sampling rate. The standard error associated with the sample scores for the 10% and 15% sampling rates were much greater than the standard errors reported for scores related to the 20% and 25% sampling rates. The

standard errors are important because as decreased standard errors lead to greater the accuracy of the results (i.e., less spread in the results). Based on these findings, it appears that sampling rates of 20% to 25% will return similar results.

Comparing the average TAKS reading score for each sampling rate to the mean population score revealed that the sample mean for each sampling rate closely approximated the true score. While the sample reading results closely approximated the population mean, there was a slightly larger difference between the sample math mean score and the true population value (average difference between sample and population math scores = 26.38; average difference between sample and population reading scores = 9.67). A plausible explanation for increased difference in math versus reading scores is the greater variability in the math scores within the population from which the samples were selected. The standard deviation associated with the population math scores was more than twice that of reading scores (*SD*-reading = 103.11; *SD*-math = 213.15).

Table 10

Descriptive Measures-Comparing the Average Scale Score on the TAKS Reading and Math Assessments among Economically Disadvantaged Regular Education Students in Seventh Grade

	Sample						Population				
	Reading			Math			Reading		Math		
Sampling Rate	Mean	*SE*	95%CI	Mean	*SE*	95%CI	Mean	*SD*	Mean	*SD*	Number of Students
10	744.61	3.92	736-752	675.78	9.08	657-693	752.65	103.11	678.09	213.15	937
15	751.59	3.70	744-758	706.89	7.75	691-722	752.65	103.11	678.09	213.15	937
20	747.61	3.23	741-754	683.73	6.20	671-696	752.65	103.11	678.09	213.15	937
25	746.34	3.05	740-753	686.53	6.37	673-699	752.65	103.11	678.09	213.15	937

The results presented in Table 11 show that bias in the TAKS reading score did not differ across sampling rate, campus in which students were enrolled or the interaction term that included sampling rate by campus. A plausible explanation for the statistically insignificant results is that the TAKS reading scores were similar among economically disadvantaged students across participant campuses. Note the more similar scores are among participants; reduced sampling rates tend to provide estimates that closely approximate the true population values (i.e., less biased results).

Table 11

Reading Average Scale Score Bias by Sampling Rate and Campus among Economically Disadvantaged Students

Source	Type III Sum of Squares	*df*	Mean Square	*F*	Sig.
Corrected Model	.102[a]	51	.002	.798	.840
Intercept	6.727E-006	1	6.727E-006	.003	.959
Sampling Rate	.006	3	.002	.779	.506
Campus	.021	12	.002	.707	.745
Sampling Rate X Campus	.075	36	.002	.829	.750
Error	1.178	468	.003		
Total	1.280	520			
Corrected Total	1.280	519			

a. *R* Squared = .080 (Adjusted *R* Squared = -.020)

Regarding the average TAKS math score bias among economically disadvantaged students, the results illustrated in Table 12 show that the sampling rate was the only statistically significant variable in the overall model (F= 4.526, df = 3, 5.401, p = .004). Based on the descriptive measures reported in Table 7, a plausible

explanation for the statistically significant findings could be the increased standard errors associated with the 10% and 15% sampling rates compared to the 20% and 25% sampling rates. Nonetheless, while sampling rate was statistically significant, the effect size was less than one percent (R^2 = .004) suggesting that the sampling rate explained less than one percent of the variance in bias related to the TAKS math average scale score. To provide further insight into the statistically significant results, 95% CIs were examined that compared bias in TAKS math scale score averages across the sampling rates examined in the current study.

Table 12

TAKS Math Average Scale Score Bias by Sampling Rate and Campus among Economically Disadvantaged Students

Source	Type III Sum of Squares	*df*	Mean Square	*F*	Sig.
Corrected Model	.550[a]	51	.011	1.039	.404
Intercept	.002	1	.002	.200	.655
Sampling Rate	.141	3	.047	4.526	.004
Campus	.063	12	.005	.510	.909
Sampling Rate X Campus	.345	36	.010	.925	.596
Error	4.852	468	.010		
Total	5.403	520			
Corrected Total	5.401	519			

a. *R* Squared = .102 (Adjusted *R* Squared = .004)

The resulting 95% CIs displayed in Figure 5 show that bias similar for the 10, 20 and 25 sampling rates with overlapping confidence intervals noted. However, the results indicated that the TAKS math score was biased downward in the 15% sampling rate.

The results show that there was indeed a statistically significant difference in TAKS math score bias between the 15% sampling rate and the remaining sampling rates. Bias in the TAKS math scores stabilized at the 20% and 25% sampling rates, indicating that a 25% sampling rate would be adequate to obtain results that closely approximated the population parameter.

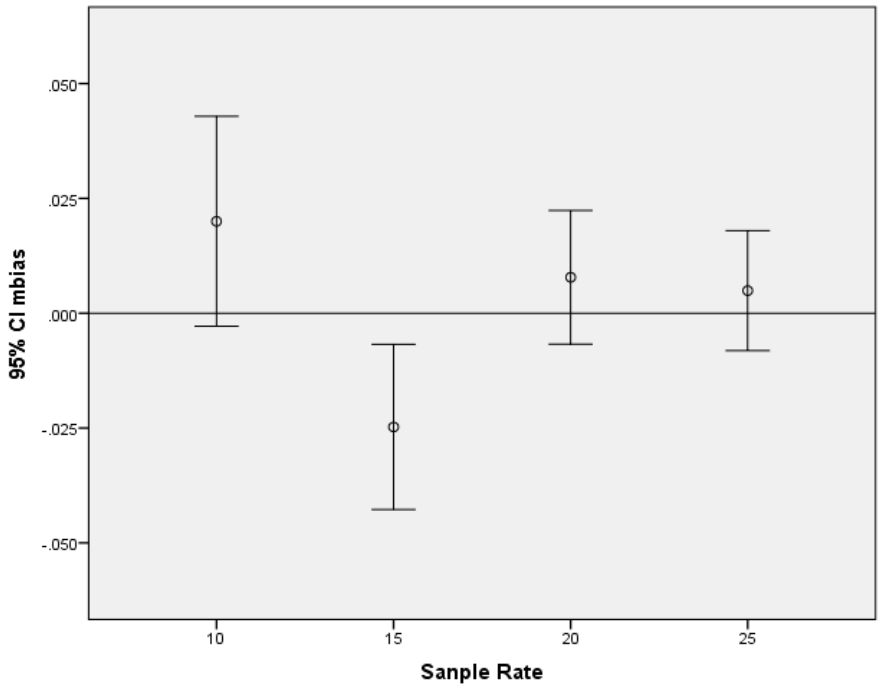

Figure 5. Confidence intervals comparing average TAKS math scale score bias by sample rate.

Figure 6 displays the distribution of the bias in the TAKS math score by sampling rate. The results show that as the sampling rate increased, bias became less spread, which is related to the decreased standard errors accompanying the increased sampling rates reported in Table 7. The results displayed in Figures 6 combined with Figure 5 further underscore the recommended 25% sampling rate as a benchmark to obtain accurate estimates of the average TAKS math score among the population.

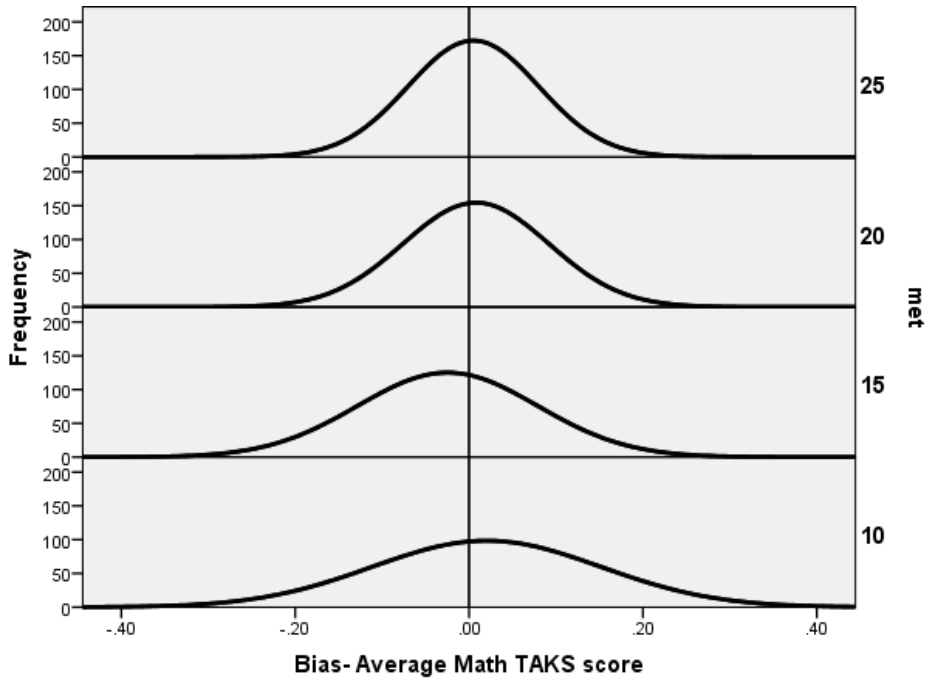

Figure 6. TAKS math score bias distribution by sample rate.

Economically Disadvantaged

Students Passing TAKS Reading and Math

The comparison of the percentage of economically disadvantaged students passing the TAKS reading and math assessments between the samples and the true population value is displayed in Table 13. Regarding TAKS reading, the percent passing is similar across each of the sampling rates considered. Average passing rates ranged from 83.72 (*SE* = 10.89) (95% CI = 82_L - 85_U) for the 25% sampling rate to 85.53 (*SD* = 13.84) (95% CI = 83_L - 88_U) for the 15% sampling rate. The population-passing rate was 83.02 (*SD* = 37.57). As expected, while the mean scores were similar, standard errors decreased as sampling rate increased. Regarding the percentage of students passing the TAKS math assessment among sampled students, the average passing rates ranged from 57.71 (*SE* = 17.51) (95% CI = 54_L - 62_U) for the 20% sampling rate to 61.66

(*SE* = 22.76) (95% CI = 57_L - 65_U) for the 10% sampling rate. The population percent passing TAKS math was 54.51 (*SD* = 49.82). Although the percent passing TAKS math was similar across sampling rates, the passing rates fluctuated less for TAKS reading compared to TAKS math across sampling rates more than likely due to the decreased standard deviation associated with the population percent passing the TAKS reading assessment.

Table 13

Descriptive Measures-Comparing the Percentage of Students Passing the TAKS Reading and Math Assessments among Economically Disadvantaged Regular Education Students in Seventh Grade District-wide

	Sample						Population				
	Reading			Math			Reading		Math		
Sampling Rate	Mean	SE	95%CI	Mean	SE	95%CI	Mean	SD	Mean	SD	Number of Students
10	85.50	1.45	83-88	61.66	2.03	57-65	83.02	37.57	54.51	49.82	954
15	85.53	1.21	83-88	59.80	1.78	56-63	83.02	37.57	54.51	49.82	954
20	84.07	1.06	82-86	57.71	1.54	54-62	83.02	37.57	54.51	49.82	954
25	83.72	.955	82-85	58.38	1.36	54-61	83.02	37.57	54.51	49.82	954

TAKS Reading Pass Rate

Bias related to the percentage of economically disadvantaged students passing the TAKS reading assessment was not statistically significant different among the variables examined in the model. The results for bias among the percent passing the TAKS reading assessment reported in Table 14 paralleled the findings reported for bias among the TAKS reading average scale scores. A likely explanation for the statistically

insignificant results is that the reading scores were similar among economically disadvantaged students across all campuses considered in the study.

Table 14

Bias in the Percentage of Students Passing TAKS Reading by Sampling Rate and Campus among Economically Disadvantaged Students

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	1.073[a]	51	.021	1.018	.444
Intercept	.011	1	.011	.531	.467
Sampling rate	.108	3	.036	1.745	.157
Campus	.285	12	.024	1.149	.318
Sampling Rate X Campus	.680	36	.019	.914	.615
Error	9.674	468	.021		
Total	10.758	520			
Corrected Total	10.747	519			

a. *R* Squared = .100 (Adjusted *R* Squared = .002)

TAKS Math Pass Rate

Comparing the campus and sampling rates among economically disadvantaged students, bias among the percentage of students passing the TAKS math assessment differed across sampling rates. The results displayed in Table 15 revealed that the sampling rate was the only statistically significant term in the model ($F = 3.307$, $df = 3$, 35.908, $p = .020$). However, while the sampling rate was statistically significant, the effect size was less than one percent ($R^2 = .008$), suggesting that the sampling rate explained less than one percent of the variance in TAKS math score bias.

Table 15

Bias in the Percentage of Students Passing TAKS Math by Sampling Rate and Campus among Economically Disadvantaged Students

Source	Type III Sum of Squares	*df*	Mean Square	*F*	Sig.
Corrected Model	3.802[a]	51	.075	1.084	.327
Intercept	.170	1	.170	2.465	.117
Sampling Rate	.682	3	.227	3.307	.020
Campus	.445	12	.037	.539	.889
Sampling Rate X Campus	2.715	36	.075	1.097	.325
Error	32.107	467	.069		
Total	36.068	519			
Corrected Total	35.908	518			

a. *R* Squared = .106 (Adjusted *R* Squared = .008)

To provide insight into the statistically significant results, post-hoc 95% CIs were examined. The resulting 95% CIs displayed in Figure 7 compared the bias in the percent passing TAKS reading by sample rate. The results show the TAKS math scores were negatively biased for the 10% sampling rate, with a statistically significant difference noted bet the 105 and 15% sampling rates. The results began to stabilize at the 20% sampling rate with similar results noted for the 25% sampling rate.

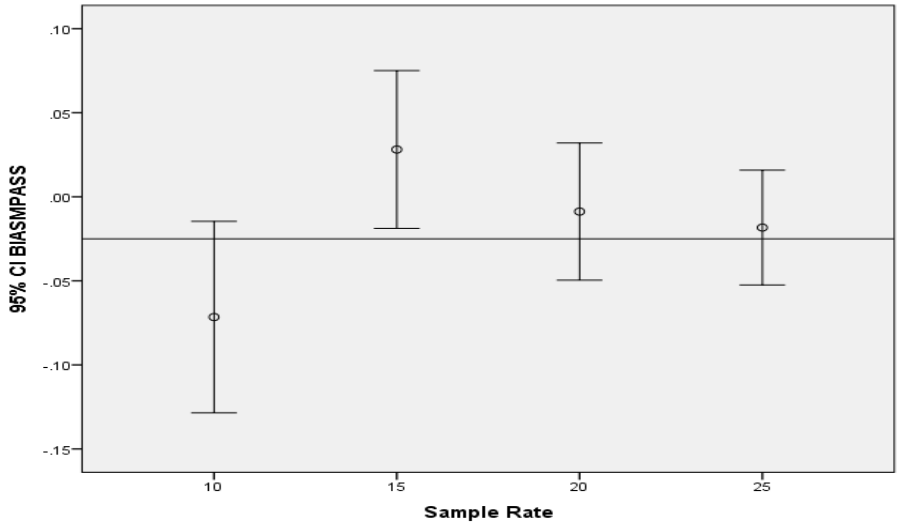

Figure 7. Confidence intervals comparing bias in the percentage of economically disadvantaged students passing TAKS math.

The distribution of bias in the percent passing TAKS math by sample rate is displayed in Figure 7. The results show that as sample rate increases, the distribution in bias approximates a normal distribution. Examined together, the results displayed in Figures 7and 8 indicate that a sampling rate of 25% is a reasonable sample rate to provide accurate estimates of the population value.

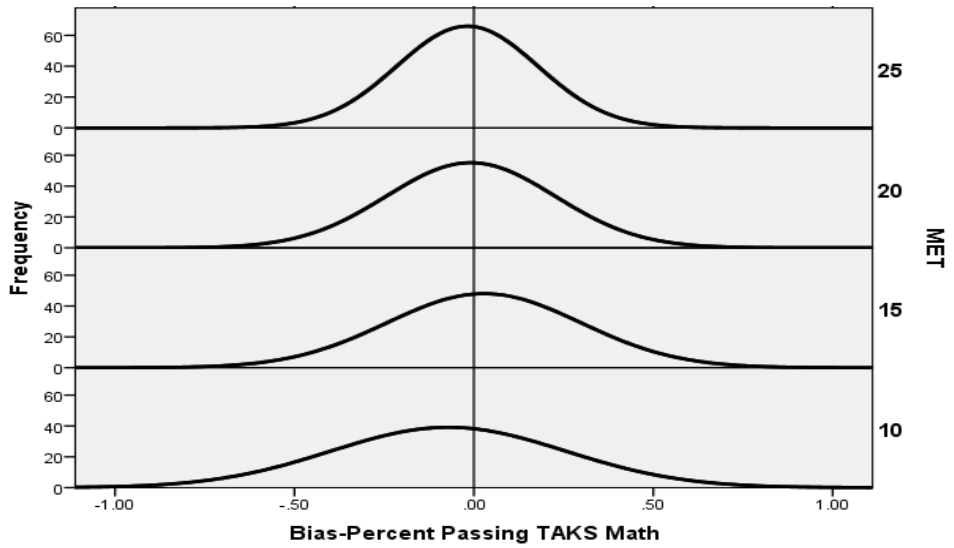

Figure 8. Distribution in bias in the percent of students passing TAKS math by sample rate.

Limited English Proficiency

TAKS Math by Sample Rate

Descriptive measures comparing bias in the average TAKS reading and math scale scores by sampling rate among LEP students regarding LEP students, the TAKS reading scores ranged from 662.02 (*SE* = 115.33) for the 15% sampling rate to 671.63 (*SE* = 140.55) for the 10% sampling rate. The overall population mean reading score was 670.81 (*SD* = 149.16). Concerning TAKS math, scores among the sampling rates ranged from 664.52 (*SE* = 124.57) for the 25% sampling rate to 677.72 (*SE* = 132.74) for the 20% sampling rate. The overall population mean for the TAKS math assessment among LEP students was 635.53 (*SD* = 232.09). While the results for the TAKS reading assessment among sampling rates compared reasonably well the true population value (average difference of 4.30 scale score points), the math results were not as close. On average, the difference between the TAKS math scores among sampling rates and the true population value was 35.18. A plausible explanation between the augmented difference between TAKS math scores among sampling rates and the true population is the larger standard deviation associated with the population TAKS math scores. The standard deviation for the TAKS math scores is approximately 35% larger than the standard deviation associated with the TAKS reading assessment. Note as the scores become more spread out (increased standard deviation), a larger sampling rate is required to replicate the true population value. The results are displayed in Table 16 below.

Table 16

Descriptive Measures-Comparing the Average Scale Score on the TAKS Reading and Math Assessments among LEP Students in Seventh Grade District-wide

	Sample						Population				
	Reading			Math			Reading		Math		
Sampling Rate	Mean	*SE*	95%CI	Mean	*SE*	95%CI	Mean	*SD*	Mean	*SD*	Number of Students
10	671.63	12.32	647-696	672.96	14.64	644-702	670.81	149.16	673.63	210.09	361
15	662.02	10.12	642-682	666.50	14.45	638-695	670.81	149.16	673.63	210.09	361
20	667.96	9.23	650-686	677.72	11.64	655-701	670.81	149.16	673.63	210.09	361
25	664.42	7.62	640-670	664.52	10.93	643-686	670.81	149.16	673.63	210.09	361

The results displayed in Tables 17 and 18 revealed no statistically significant differences in bias between sampling rates and campuses related to TAKS reading and math assessments.

Table 17

TAKS Reading Average Scale Score Bias by Sampling Rate and Campus among Limited English Proficient Students

Source	Type III Sum of Squares	*df*	Mean Square	*F*	Sig.
Corrected Model	1.262[a]	51	.025	1.036	.411
Intercept	.004	1	.004	.156	.693
Sampling Rate	.039	3	.013	.546	.651
Campus	.121	12	.010	.421	.955
Sampling Rate X Campus	1.102	36	.031	1.281	.132
Error	11.182	468	.024		
Total	12.447	520			
Corrected Total	12.444	519			

a. *R* Squared = .101 (Adjusted *R* Squared = .003)

Table 18

TAKS Math Average Scale Score Bias by Sampling Rate and Campus among Limited English Proficient Students

Source	Type III Sum of Squares	*df*	Mean Square	*F*	Sig.
Corrected Model	1.665 [a]	51	.033	.866	.732
Intercept	.127	1	.127	3.360	.067
Sampling Rate	.056	3	.019	.497	.685

(*table continues*)

Table 18 (*continued*).

Source	Type III Sum of Squares	*df*	Mean Square	*F*	Sig.
Campus	.271	12	.023	.599	.843
Sampling Rate X Campus	1.338	36	.037	.986	.496
Error	17.650	468	.038		
Total	19.443	520			
Corrected Total	19.316	519			

a. *R* Squared - .086 (Adjusted *R* Squared = -.013)

Table 19 compares the passing rates among LEP students for the TAKS reading and math assessments by sampling rates to the true population values. Regarding TAKS reading, as the sampling rate increased, the results trended towards the population value of 67.57%. The results for the 25% sampling rate appear to be adequate to gain insight into the percent of LEP students passing TAKS reading district-wide (difference of 1.68 percentage points between sample and true population value). Further, similar standard errors were reported for each sampling rate. As for the percentage of LEP students passing TAKS math, the results were not as accurate as the percentage of LEP students passing TAKS reading. On average, there was a difference of approximately 12 percentage points between the sampling rates and the true population value. A probable explanation for the discrepancy between the percent of LEP students passing the TAKS reading and math assessments is the wide variation in math scores compared to the reading scores reported in Table 17. Further

investigation indicated that a 30% sampling rate would yield results that closely approximated the true population value.

Table 19

Descriptive Measures-Comparing the Average Percentage of Students Passing the TAKS Reading and Math Assessments among Limited English Proficient Students in Seventh Grade District-wide

	Sampling						Population				
	Reading			Math			Reading		Math		
Sampling Rate	Mean	*SE*	95%CI	Mean	*SE*	95%CI	Mean	*SD*	Mean	*SD*	Number of Students
10	76.36	2.23	72-81	62.92	2.60	58-68	67.57	46.87	52.77	50.05	367
15	74.02	2.02	70-79	63.55	2.58	58-69	67.57	46.87	52.77	50.05	367
20	71.16	2.05	67-75	59.10	2.52	54-64	67.57	46.87	52.77	50.05	367
25	65.89	1.95	62-70	56.21	2.20	52-61	67.57	46.87	52.77	50.05	367

TAKS Reading Pass Rate

The results displayed in Table 20 revealed that bias among the percentage of students passing the TAKS reading assessment differed across sampling rates (F = 4.93, df = 3, 478, p = .002) and campuses in which LEP students were enrolled (F = 6.31, df = 12, 478, p = .001) . Note the study included each of the middle schools in a selected district. The effect size associated with the sampling rates was R^2 = .09, indicating that the sampling rate explained approximately 9% of the variance in bias, while the effect size related to campuses was R^2 = .13, indicating that the campuses in which the LEP students were enrolled accounted for approximately 13% of the bias in the percentage of LEP students passing the TAKS reading assessment. The interaction term that included sampling rate by campus was not statistically significant.

Table 20

Bias among Percentage of Students Passing TAKS Reading across Sampling Rates

Source	Type III Sum of Squares	*df*	Mean Square	*F*	Sig.
Corrected Model	11.246[a]	51	.221	2.387	$p < .01$
Intercept	3.006	1	3.006	32.536	$p < .01$
Sampling Rate	1.367	3	.456	4.934	.002
Campus	6.998	12	.583	6.312	.001
Sampling Rate X Campus	4.103	36	.114	1.234	.171
Error	39.449	427	.092		
Total	52.092	479			
Corrected Total	50.695	478			

a. *R* Squared = .222 (Adjusted *R* Squared = .129)

A descriptive bar chart with overlaid confidence intervals was examined by sampling rate and campus to gain further insight into how the statistically significant main effects impacted bias in the percentage of LEP students passing the TAKS reading assessment. The results displayed in Figure 9 show that bias differed considerably across campuses and sampling rates, especially among the 10%, 15%, and 20% sampling rates. However, at the 25% sampling rate, the results were stable across campuses with narrowing confidence intervals. Note at the 25% sampling rate, confidence intervals crossed zero indicating that bias in the percent of LEP students passing the TAKS reading assessment was not statistically significantly different from zero.

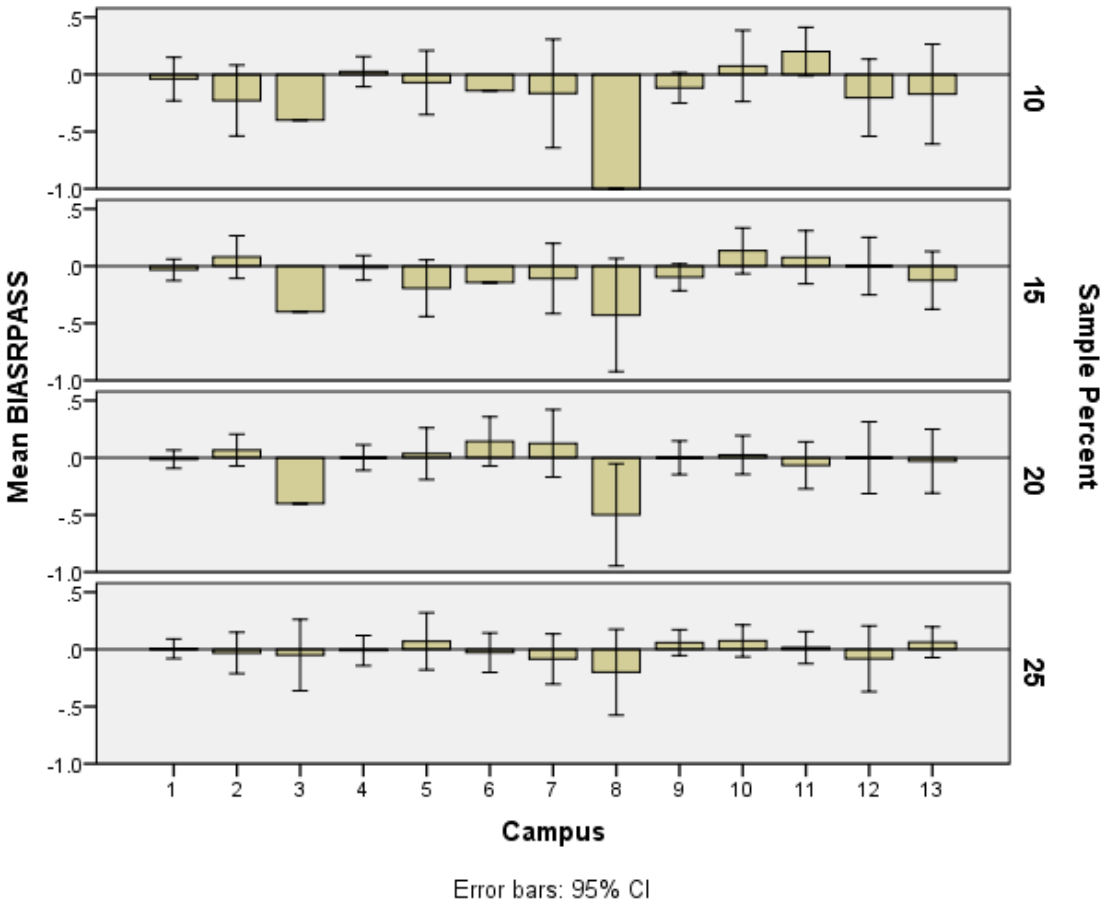

Figure 9. Comparison of bias among LEP students passing TAKS reading by sampling rate and campus.

Figure 10 shows that as the sampling rate increased, bias became less spread (decreased standard error) and approximated a normal distribution. Based on the results displayed in Figures 9 and 10, it appears that a 25% sampling rate will closely approximate the true population value.

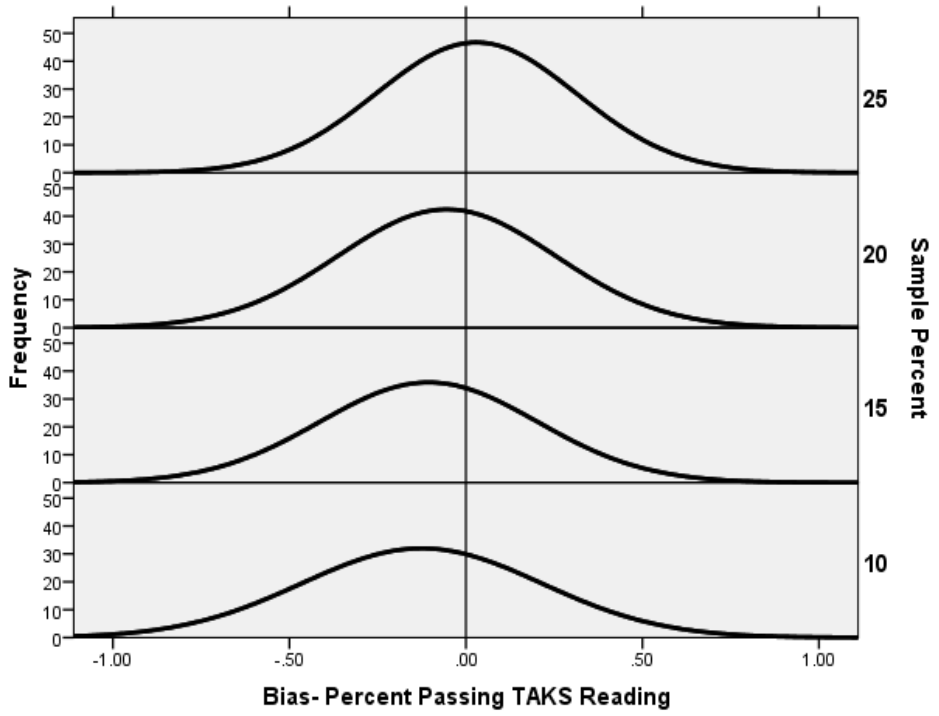

Figure 10. Bias Distribution in the Percent of LEP Students Passing the TAKS Reading Assessment by Sampling Rate.

TAKS Math Pass Rate

Similar to the results reported for bias in the percentage of LEP student passing TAKS reading, the results displayed in Table 21 indicate that bias among LEP students passing the TAKS math assessment differed across sampling rates (F = 8.62, df = 3, 469, p = p < .01) and campuses (F = 2.15, df = 12, 469, p = .013). The effect size associated with the sampling rate was R^2 = .08, indicating that the sampling rate accounted for approximately 8% of the variance in bias, while the effect size associated with campuses in which the participants were enrolled was R^2 = .09, demonstrating that campuses explained approximately 9% of the bias among LEP students passing the TAKS math assessment. The interaction term that included sampling rate by campus was not statistically significant.

Table 21

Bias in the Percentage of Limited English Proficient Students Passing TAKS Math by Sampling Rate and Campus

Source	Type III Sum of Squares	*df*	Mean Square	*F*	Sig.
Corrected Model	12.192[a]	51	.239	1.706	.003
Intercept	2.651	1	2.651	18.921	$p < .01$
Sampling Rate	3.622	3	1.207	8.617	$p < .01$
Campus	3.612	12	.301	2.148	.013
Sampling Rate X Campus	5.469	36	.152	1.084	.344
Error	58.569	418	.140		
Total	72.373	470			
Corrected Total	70.761	469			

a. *R* Squared = .172 (Adjusted *R* Squared = .071)

The results displayed in Figure 11 show considerable variation in bias across campuses for the ten percent sampling rate. However, as the sampling rate increases, the variance in bias decreases, especially for the 20% and 25%sampling rates. While the 25% sampling rate returned more stable results, there were statistically significant differences in bias among LEP students passing TAKS math between Campus 1 and Campus 11. Further investigation indicated that a 30% sampling rate provided stable results across all campuses that closely approximated the true population value with bias that was not statistically significantly different from zero.

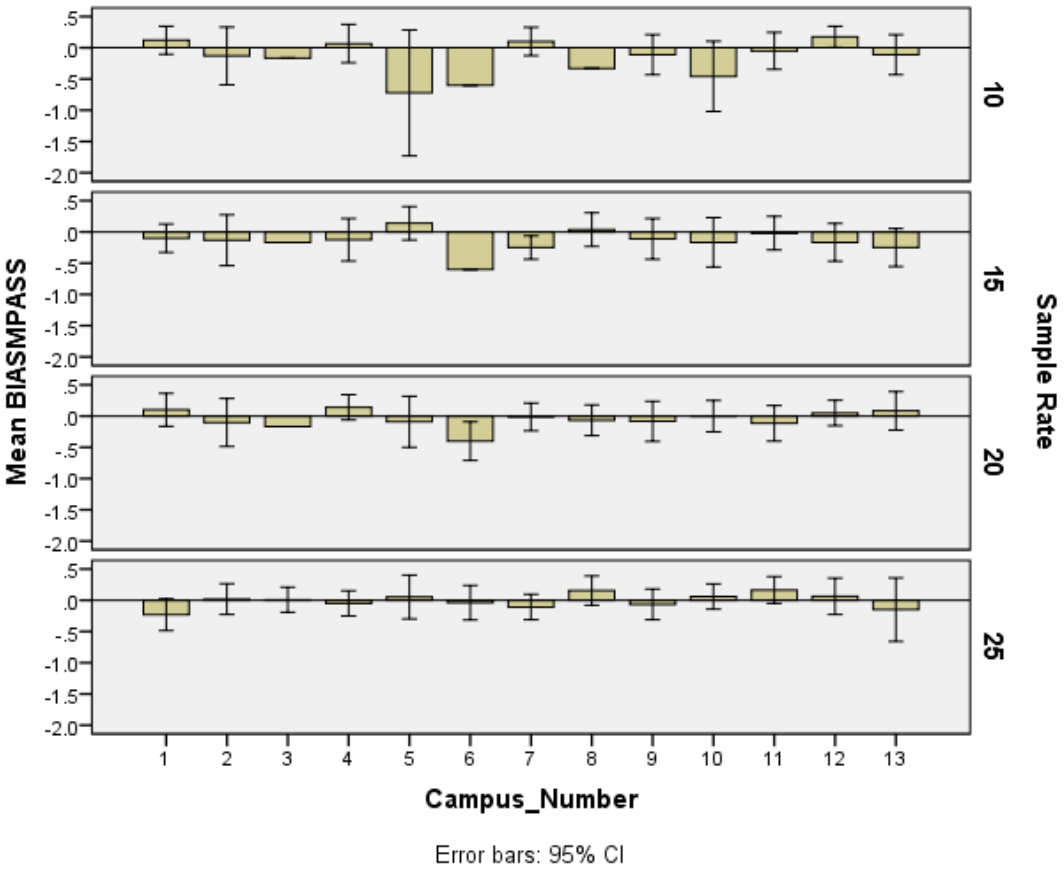

Figure 11. Comparison of bias among LEP students passing TAKS math by sampling rate and campus.

Figure 12 shows that as the sampling rate increased, bias in the percentage of LEP students passing TAKS math became less spread (decreased standard error) and approximated a normal distribution. However, the distribution of bias at the 25% sampling rate was similar to the 20% sampling rate. Further investigation indicated that a 30% sampling rate would closely approximate a normal distribution and provide acceptable results that closely approximated the true population value.

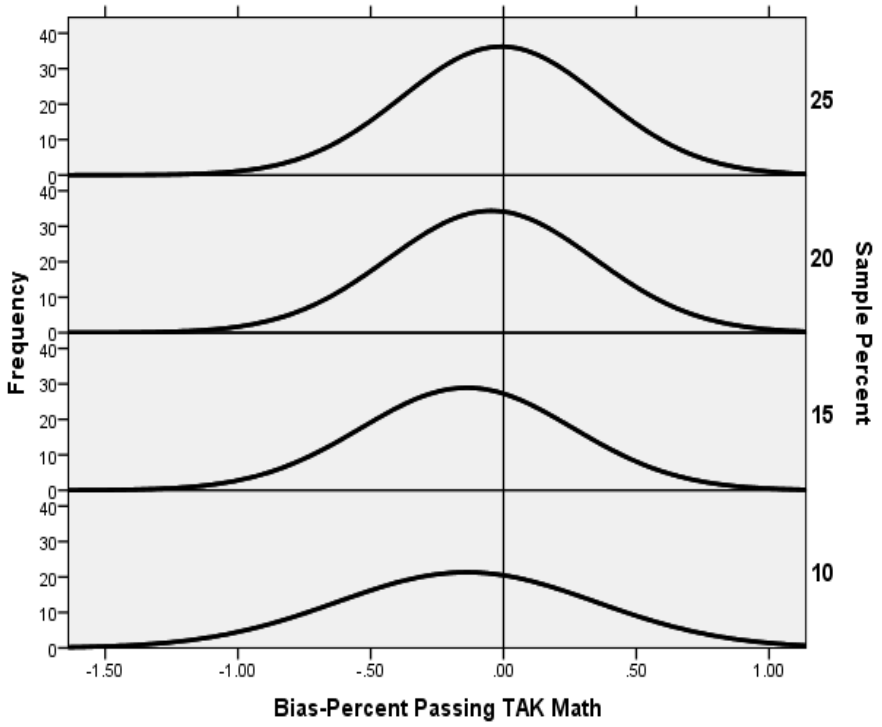

Figure 12. Bias distribution in the percent of LEP students passing the TAKS Math Assessment by Sampling Rate.

Special Education

Reading Average

Comparing the average TAKS reading score for each sampling rate to the mean population score in Table 22 revealed that the sample mean for each sampling rate closely approximated the true score. While the sample reading results closely approximated the population mean, there was a slightly larger difference between the sample math mean score and the true population value (average difference between sample and population math scores = 6.5; average difference between sample and population reading scores = 2.35). A plausible explanation for slight increased difference in math versus reading scores is the greater variability in the math scores

within the population from which the samples were selected. The standard deviation associated with the population math scores was almost twice that of reading scores (SD-reading = 106.73; SD-math = 209.95).

Table 22

Descriptive Measures-Comparing the Average Percentage of Students Passing the TAKS Reading and Math Assessments among Special Education Students in Seventh Grade District-wide

Sample							Population				
Reading				Math			Reading		Math		
Sampling Rate	Mean	*SE*	95%CI	Mean	*SE*	95%CI	Mean	*SD*	Mean	*SD*	Number of Students
10	710.29	2.18	708-715	613.94	4.34	605-623	713.26	106.73	624.12	209.95	259
15	712.86	1.74	710-716	617.03	3.55	610-624	713.26	106.73	624.12	209.95	259
20	712.31	1.50	709-715	619.99	3.12	614-626	713.26	106.73	624.12	209.95	259
25	710.71	1.27	708-713	621.88	2.80	616-627	713.26	106.73	624.12	209.95	259

The results presented in Table 23 show that TAKS reading score bias among special education students did not differ across sampling rates. However, there was a statistically significant differ across campuses in which students were enrolled (F = 48.28, df = 12, 5123, $p = p < .01$). The effect size associated with campus was $R^2 = .09$, indicating that the campus in which the participants were enrolled accounted for approximately 9% of the variance in TAKS reading score bias. The interaction term that included sampling rate by campus was not statistically significant.

Table 23

TAKS Reading Average Scale Score Bias by Sampling Rate and Campus among Special Education Students

Source	Type III Sum of Squares	*df*	Mean Square	*F*	Sig.
Corrected Model	4.331[a]	51	.085	12.045	$p < .01$
Intercept	.215	1	.215	30.448	$p < .01$
Sampling Rate	.012	3	.004	.552	.647
Campus	4.084	12	.340	48.278	$p < .01$
Sampling Rate X Campus	.240	36	.007	.944	.565
Error	35.759	5072	.007		
Total	40.319	5124			
Corrected Total	40.090	5123			

a. *R* Squared = .108 (Adjusted *R* Squared = .099)

The post-hoc results examining TAKS reading score bias among special education students by sampling rate and campus is displayed in Figure 13. The results

revealed that bias was most prevalent at Campus 3 for each sampling rate considered. While there was variation in TAKS reading score bias for the 10%, 15%, and 20% sampling rates, the results stabilized at the 25% sampling rate. With the exception of Campus 3, bias was similar across campuses with each confidence interval cross zero, indicating that the bias in TAKS reading scores was not statistically significantly different from zero.

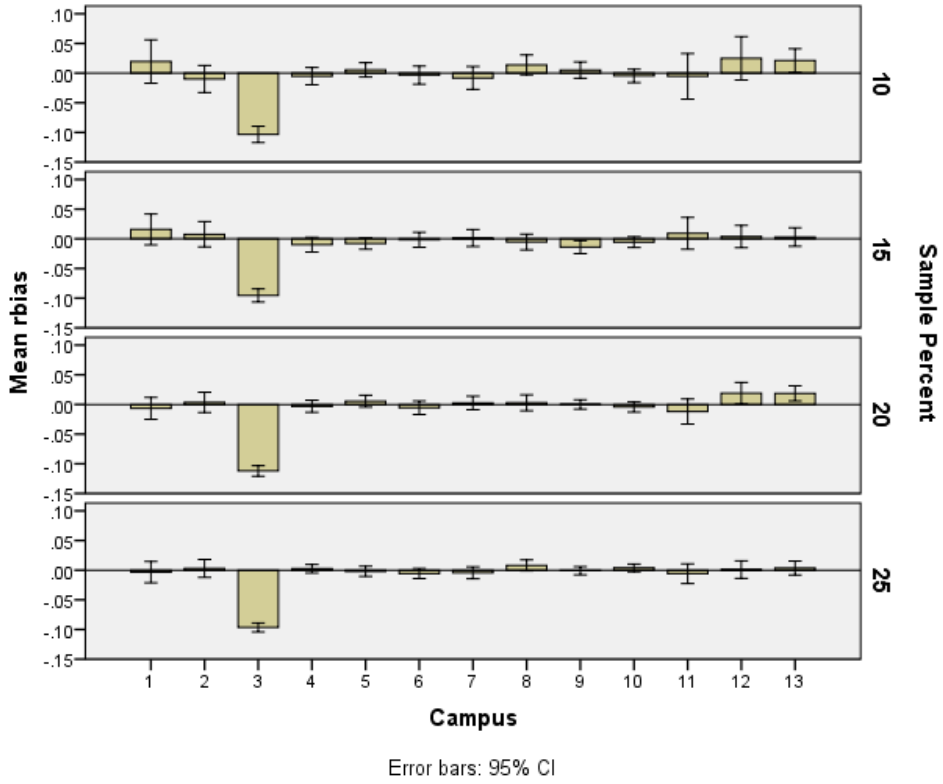

Figure 13. Comparison of average TAKS reading score bias among special education students by sampling rate and campus.

Similar to the results reported in Tables 23 and 24 shows that TAKS math score bias among special education students did not differ across sampling rates. However, there was a statistically significant difference across campuses in which students were enrolled (F = 48.4.61, df = 12, 5123, p = p < .01). The effect size associated with

campus was R^2 = .10, indicating that the campus in which the participants were enrolled accounted for approximately 10% of the variance in TAKS math score bias. The interaction term that included sampling rate by campus was not statistically significant.

Table 24

TAKS Math Average Scale Score Bias by Sampling Rate and Campus among Special Education Students

Source	Type III Sum of Squares	*df*	Mean Square	*F*	Sig.
Corrected Model	3.566[a]	51	.070	2.050	$p < .01$
Intercept	.093	1	.093	2.724	.099
Sampling Rate	.076	3	.025	.743	.526
Campus	1.887	12	.157	4.609	$p < .01$
Sampling Rate X Campus	1.604	36	.045	1.306	.105
Error	173.007	5072	.034		
Total	176.665	5124			
Corrected Total	176.572	5123			

a. *R* Squared = .020 (Adjusted *R* Squared = .010)

The post-hoc analysis results displayed in Figure 14 compare TAKS math score bias among special education students by sampling rate and campus. Bias varied considerably across campuses at the 10%ten and 15% sampling rates but appeared to stabilize at the 20% sampling rate. Similar to the TAKS reading results, bias was most prevalent at Campus 3 at each sampling rate. With the exception of Campus 3, bias was similar across campuses with each confidence interval cross zero at the 25% sampling rate. Thus, indicating that the bias in TAKS math scores was not statistically significantly different from zero.

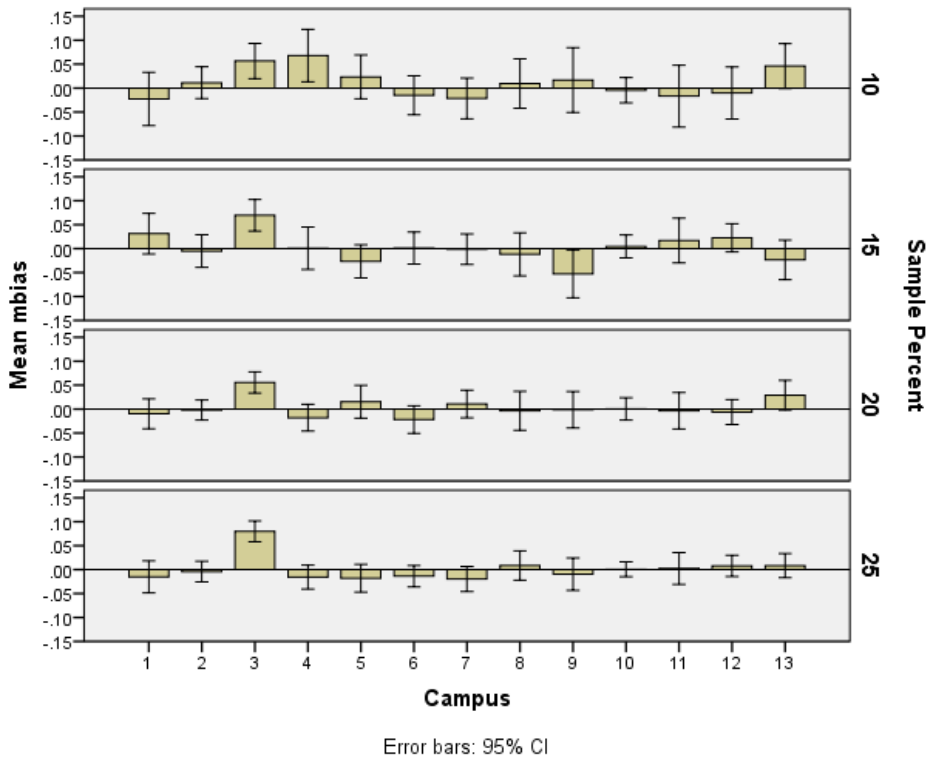

Figure 14. Comparison of average TAKS math score bias among special education students by sampling rate and campus.

Students Passing TAKS Reading and Math

Table 25 compares the passing rates among special education students for the TAKS reading and math assessments by sampling rates to the true population values. Regarding TAKS reading, as the sampling rate increased, the results trended towards the population value of 78.77%. The results for the 25% sampling rate appear to be adequate to gain insight into the percent of special education students passing TAKS reading district-wide (difference of .10 percentage points between sample and true population value). Further, similar standard errors were reported for each sampling rate. As for the percentage of special education students passing TAKS math, the results were similar to the TAKS reading results. On average, there was a difference of

approximately 4.75 percentage points between the sampling rates and the true population value. Based on the results reported in Table 25, a 25% sampling rate seems reasonable to replicate the true population values.

Table 25

Descriptive Measures-Comparing the Average Percentage of Students Passing the TAKS Reading and Math Assessments among Special Education Students in Seventh Grade District-wide

	Sample						Population				
	Reading			Math			Reading		Math		
Sampling Rate	Mean	SE	95%CI	Mean	SE	95%CI	Mean	SD	Mean	SD	Number of Students
10	81.50	.659	80-83	64.22	.770	62-66	78.77	40.97	51.71	50.06	292
15	80.43	.608	79-82	56.57	.655	55-57	78.77	40.97	51.71	50.06	292
20	79.16	.549	78-80	54.12	.610	53-55	78.77	40.97	51.71	50.06	292
25	78.61	.494	77-79	53.40	.544	51-54	78.77	40.97	51.71	50.06	292

Regarding bias in the percentage of special education students passing TAKS reading, the results displayed in Table 26 revealed that bias differed across sampling rates ($F = 6.86$, $df = 3, 5129$, $p = p < .01$), campus in which the participants were enrolled ($F = 5.03$, $df = 12, 5129$, $p = p < .01$), and the interaction term that included sampling rate by campus ($F = 2.16$, $df = 36, 5129$, $p = p < .01$). The effect size associated with sampling rate was $R^2 = .04$, suggesting that sampling rate explain approximately four percent of the variation in TAKS reading score bias. In addition, the

effect size associated with campus on which the student enrolled was R^2 = .06, indicating that the campus on which the participants were enrolled accounted for six percent of the variance, while the interaction term (sampling rate by campus) accounted for approximately eight percent of the variance in percent of students passing TAKS reading (R^2 = .08).

Table 26

Bias in the Percentage of Special Education Students Passing TAKS Reading by Sample Percent and Campus

Source	Type III Sum of Squares	*df*	Mean Square	*F*	Sig.
Corrected Model	9.120[a]	51	.179	2.878	$p < .01$
Intercept	.728	1	.728	11.724	.001
Sampling Rate	1.274	3	.425	6.834	$p < .01$
Campus	3.748	12	.312	5.026	$p < .01$
Sampling Rate X Campus	4.840	36	.134	2.164	$p < .01$
Error	315.544	5078	.062		
Total	325.220	5130			
Corrected Total	324.664	5129			

a. *R* Squared = .028 (Adjusted *R* Squared = .018)

The results displayed in Figure 15 show considerable variation in bias related to the percentage of special education students passing TAKS reading. The variation was most prevalent among Campuses 7 and 8 for the 10% and 15% sampling rate. However, as the sampling rate increased to 20%, the results stabilized with no statistically significant differences noted across campuses. Further, confidence intervals for each campus crossed the horizontal reference line, indicating that bias in the

percentage of special education students passing TAKS reading across campuses was not statistically significantly different from zero. Similar results were reported for the 25% sampling rate.

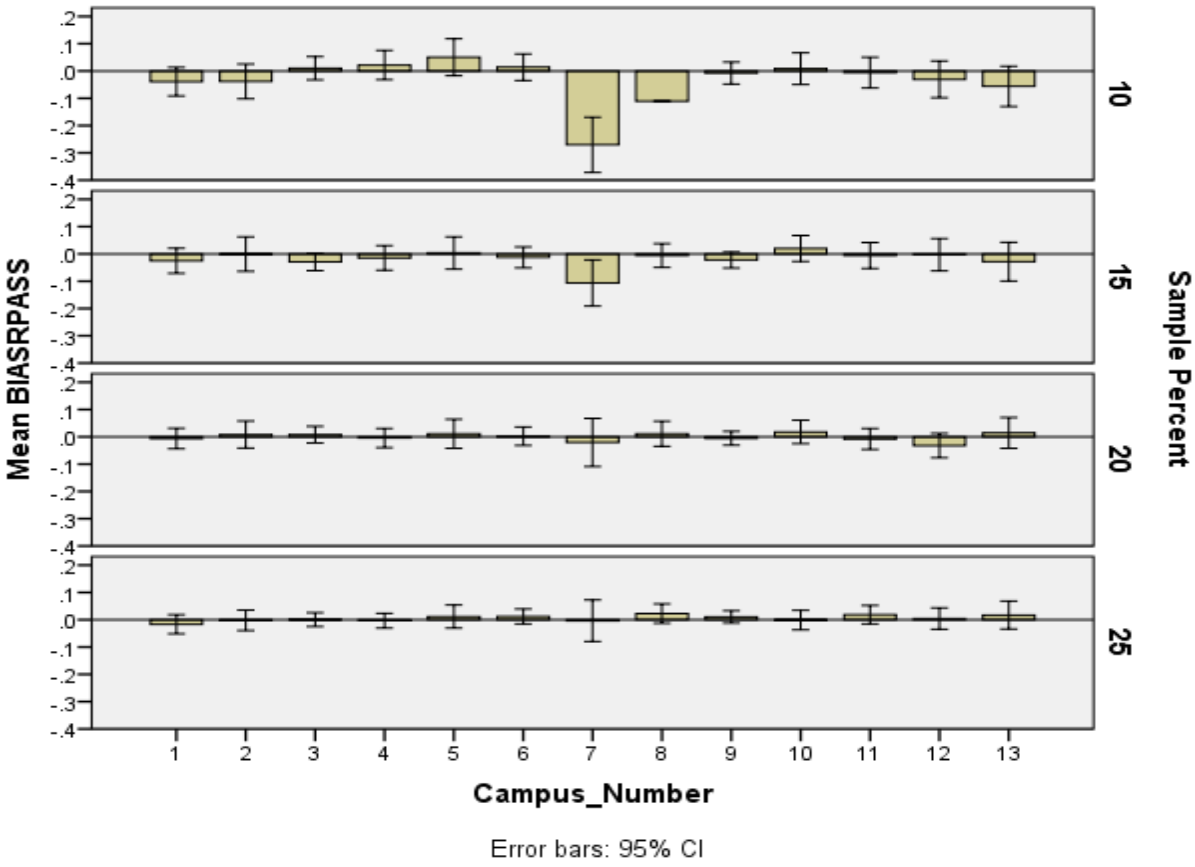

Figure 15. Comparison of bias among special education students passing TAKS reading by sampling rate and campus.

The distribution of bias in the percent passing TAKS reading by sample rate is displayed in Figure 15. The results show that as sample rate increases, the distribution in bias approximates a normal distribution. The findings in Figure 14 coupled with the results displayed in Figure 15 above indicate that a sampling rate of 25% is a reasonable sample rate to provide accurate estimates of the population value.

TAKS Math Pass Rate

The results reported in Table 27 revealed that bias in the percentage of special education students passing TAKS math differed across sampling rates (F = 80.88, *df* = 3, 4706, *p* = *p* < .01), campus in which the participants were enrolled (*F* = 19.31, *df* = 12, 4706, *p* = *p* < .01), and the interaction term that included sampling rate by campus (*F* = 5.70, *df* = 36, 4706, *p* = *p* < .01). The effect size associated with sampling rate was R^2 = .04, suggesting that sampling rate explains approximately four percent of the variation in TAKS math score bias. In addition, the effect size associated with campus on which the student enrolled was R^2 = .03, indicating that the campus on which the participants were enrolled accounted for three percent of the variance, while the interaction term (sampling rate by campus) accounted for approximately 2.8% of the variance in percent of students passing TAKS math (R^2 = .028).

Table 27

Bias in the Percentage of Special Education Students Passing TAKS Math by Sample Percent and Campus

Source	Type III Sum of Squares	*df*	Mean Square	*F*	Sig.
Corrected Model	87.747[a]	51	1.721	10.835	*p* < .01
Intercept	56.576	1	56.576	356.277	*p* < .01
Sampling Rate	38.526	3	12.842	80.870	p < .01
Campus	36.802	12	3.067	19.313	p < .01
Sampling Rate X Campus	32.528	36	.904	5.690	p < .01
Error	739.204	4655	.159		
Total	865.932	4707			
Corrected Total	826.951	4706			

a. *R* Squared = .106 (Adjusted *R* Squared = .098)

The results displayed in Figure 16 show considerable variation in bias related to the percentage of special education students passing TAKS math. The variation was most prevalent at the ten, fifteen, and twenty percent sampling rate. However, as the sampling rate increased to 25%, the results stabilized with no statistically significant differences noted across campuses. Further, confidence intervals for each campus crossed the horizontal reference line, indicating that bias in the percentage of special education students passing TAKS math across campuses was not statistically significantly different from zero.

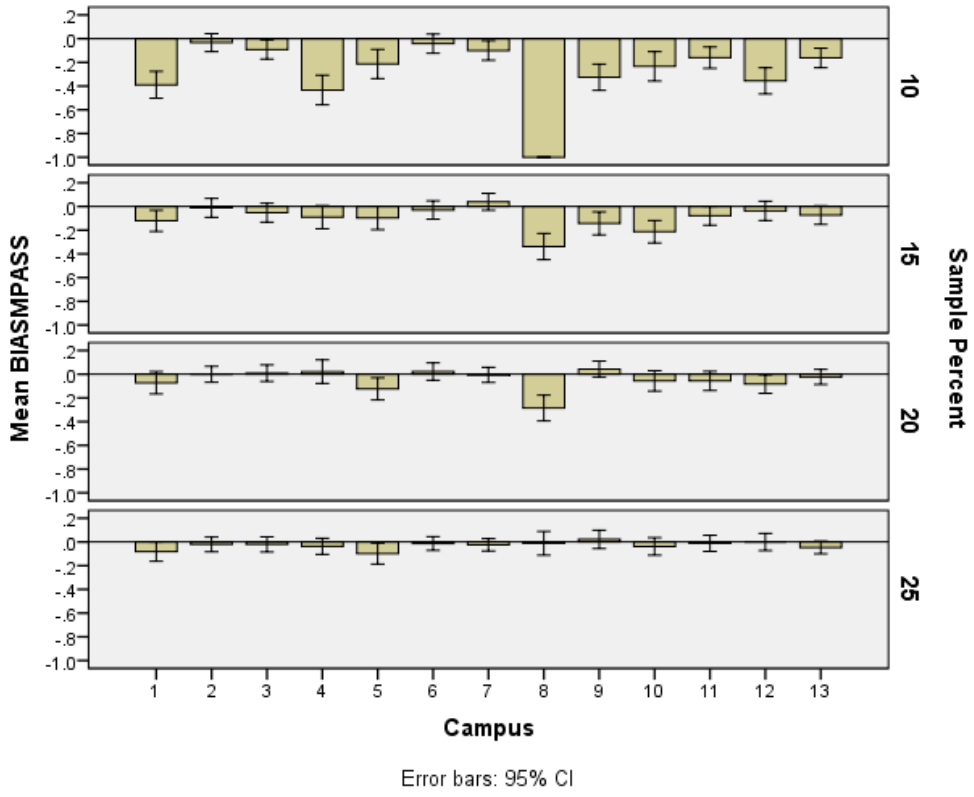

Figure 16. Comparison of bias among special education students passing TAKS math by sampling rate and campus.

The distribution of bias in the percent passing TAKS math by sample rate is displayed in Figure 17. The results show that as sample rate increases, the distribution

in bias approximated a normal distribution. The findings in Figure 17 coupled with the results displayed in Figure 16 above indicate that a sampling rate of 25% is a reasonable sampling rate to provide accurate estimates of the true population value related to the percent of special education students passing the TAKS math assessment.

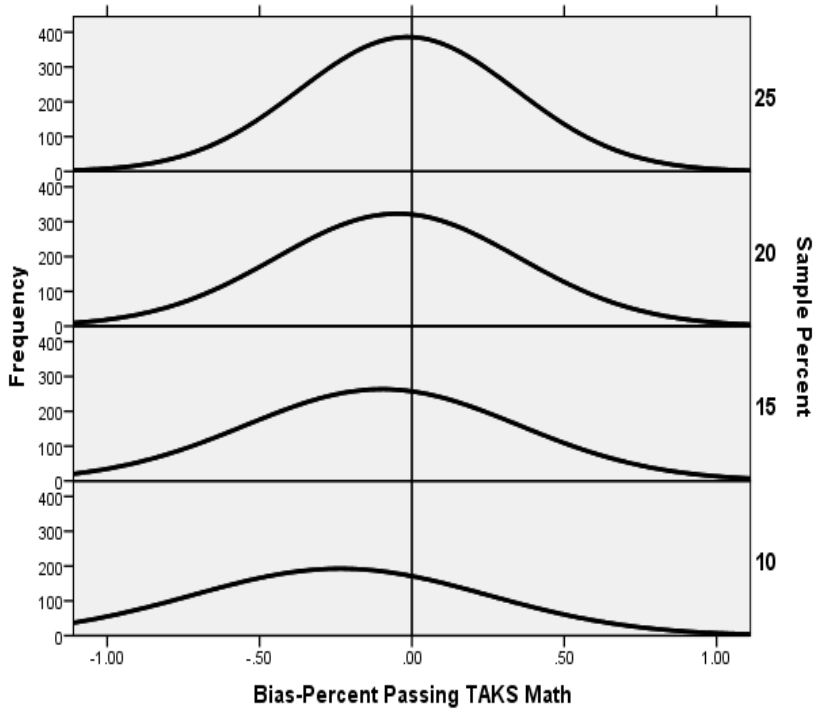

Figure 17. Bias in TAKS math scores among special education students by sampling rate and campus.

Growth

This section of the study examined the impact of stratified random sampling on student academic growth. Table 28 displays the true population academic growth values among regular education students who participated in the TAKS math assessment over a three-year timeframe (Grades 3-5). Due to the TAKS math

assessment not being vertically linked, the exam cannot be used to measure growth without making adjustments. In this study, the TAKS math scores were converted to standard scores (z-scores) at the individual student level for each year examined. Although not an ideal measure of growth, the standard scores provide an improved method of monitoring growth than using raw scale scores.

The results displayed in Table 28 show similar gains or growth across all ethnic groups among the campuses examined. The greatest amount of growth was noted at campus “B” while students at campus “A” exhibited the least amount of growth overall. Average growth ranged from .99 (*SD* = .86) to 1.02 (*SD* = 1.11) in Campus “A” compared to Campus “B” where average growth ranged from 1.02 (*SD* = 1.21) to 1.59 (*SD* = 1.56).

Table 28

Comparison of Growth among Fifth Grade Students by Ethnicity and Campus for Total Population

							Students Enrollment by Campus		
Ethnicity	Mean	*SD*	Mean	*SD*	Mean	*SD*	A	B	C
Black	.99	.86	1.59	1.56	1.36	1.03	45	50	35
Hispanic	1.26	.99	1.41	1.66	1.24	1.12	92	76	55
White	1.02	1.11	1.02	1.21	.95	0.93	168	129	210

Table 29 displays the average growth by ethnicity and campus for each of the sampling rates considered in the current study. Similar growth scores were noted across each student subgroup and campus for each sample rate. Note at the 15% sample rate, scores mean score increased slightly among Hispanic students at

Campuses “A” and “B”. Finally, as expected, the standard errors decreased as the sampling rate increased. In each scenario, the 95% CI captured the true population parameter.

Table 29

Comparison of Growth among Fifth Grade Students by Ethnicity, Sampling Rate, and Campus

		Campus								
		A			B			C		
Sampling Rate	Ethnicity	Mean	*SE*	95%CI	Mean	*SE*	95%CI	Mean	*SE*	95%CI
	Black	.99	.041	.91-1.07	1.61	.070	1.47-1.75	1.40	.056	1.28-1.50
10	Hispanic	1.30	.047	1.21-1.40	1.46	.066	1.62-1.59	1.20	.047	1.11-1.29
	White	1.03	.032	.91-1.10	1.07	.041	.99-1.16	.94	.025	.89-.99
	Black	1.01	.034	.95-1.08	1.50	.055	1.39-1.61	1.35	.045	1.26-1.44
15	Hispanic	1.27	.033	1.20-1.34	1.38	.058	1.26-1.50	1.28	.041	1.20-1.36
	White	1.01	.027	.96-1.06	1.06	.030	1.00-1.12	.92	.019	.88-.96
	Black	.98	.026	.93-1.03	1.50	.043	1.41-1.58	1.34	.039	1.24-1.42
20	Hispanic	1.24	.026	1.19-1.29	1.42	.045	1.33-1.51	1.24	.030	1.18-1.30
	White	.98	.020	.94-1.02	.99	.025	.94-1.03	.93	.016	.90-.97
	Black	.99	.024	.93-1.04	1.56	.041	1.50-1.65	1.33	.033	1.26-1.40
25	Hispanic	1.24	.022	1.20-1.29	1.39	.035	1.32-1.46	1.21	.030	1.15-1.27
	White	1.01	.017	.97-1.04	1.00	.021	.96-1.04	.95	.012	.93-.97

To determine if bias in the mean growth rate differed across sampling rate, ethnicity, and campus, a factorial ANOVA was conducted. The results displayed in Table 30 revealed that bias in the mean growth score differed across sampling rates (F = 7.695, df = 5, 5399, p < .01). The effect size associated with the sampling rate was R^2 = .04, indicating that the sampling rate explained approximately four percent of the variation in mean growth score bias. The remaining variables (both main effects and interaction terms) were not statistically significant).

Table 30

Bias in Growth on the TAKS Math Assessment among Regular Education Students by Sample

Source	Type III Sum of Squares	*df*	Mean Square	*F*	Sig.
Corrected Model	4.221[a]	33	.128	1.703	.007
Intercept	.001	1	.001	.012	.914
Sampling Rate	2.890	5	.578	7.695	p < .01
Ethnicity	.057	2	.028	.377	.686
Campus	.017	2	.009	.116	.890
Sampling Rate X Ethnicity	.446	10	.045	.594	.820
Sampling Rate X Campus	.384	10	.038	.511	.884
Ethnicity X Campus	.426	4	.107	1.419	.225
Error	403.095	5366	.075		
Total	407.317	5400			
Corrected Total	407.316	5399			

a. R Squared = .010 (Adjusted R Squared = .004)

To provide insight into the statistically significant results reported in Table 25, post-hoc 95% CIs were examined. The resulting 95% CIs displayed in Figure 18 compared the bias in the growth scores by sample rate. The results show the TAKS math growth scores were positively biased for the 10% sampling rate and negatively biased for the 15% sampling rate. A statistically significant difference noted bet the 10% and 15% sampling rates. However, the results began to stabilize at the 20% sampling rate with similar results noted for the 25% sampling rate. Bases on these results and further in-depth analysis, it is recommended that a 20% sampling rate be employed to capture the true population growth values on high stakes assessments.

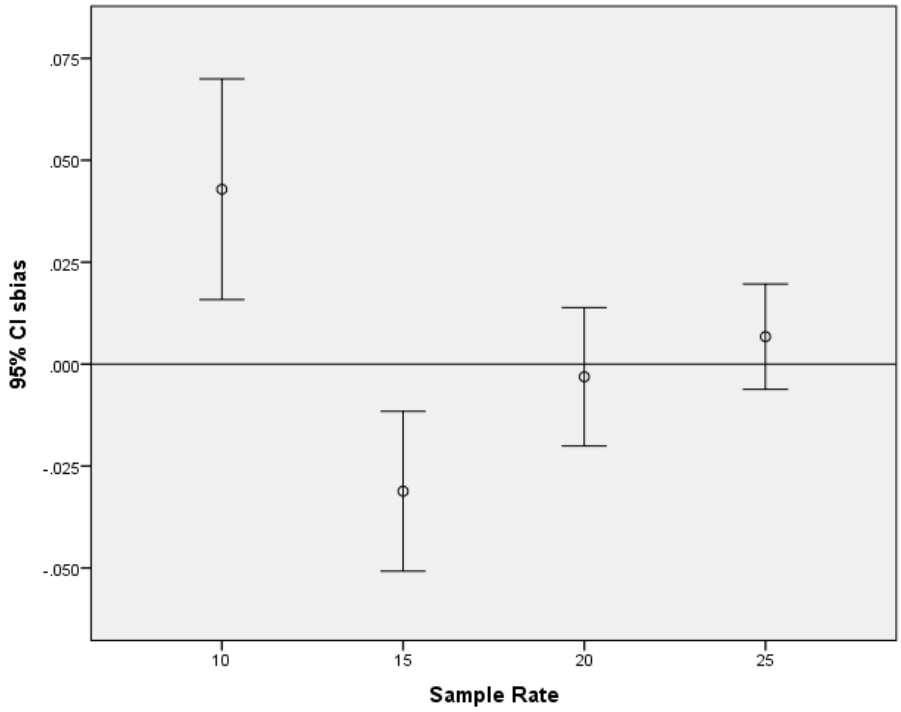

Figure 18. *Confidence intervals comparing bias in with scores on the TAKS math assessment.*

CHAPTER 5

DISCUSSION

This chapter summarizes the findings of the study organized by the research questions posed in Chapter 1. Relevant implications and conclusions are drawn based on these findings in terms of potential influence on research and practice. Finally, this chapter presents recommendations for future research.

Sampling techniques have been used effectively in education research and practice (Australian Curriculum, Assessment and Reporting Authority 2008; Darling-Hammond, 2010; Joncas & Foy, 2011; Martin, Mullis, & Chrostowski, 2003; Murphy & Schulz, 2006; NCES, 2011; NCES, 2009; NCES 2007; National Education Monitoring Project, 2010; OCED, 2004; Sahlberg, 2006; Sahlberg, 2007; Savola, 2012; Scotland, 2012; TEA, 2008; Thompson, 2012); it is not clear how stratified random sampling techniques apply to high stakes testing in the current educational environment in Texas. The purpose of this sampling study was to determine if stratified random sampling was a viable option for reducing the number of students participating in Texas state assessments, and to determine which sampling rate provided consistent estimates of the actual test results among the subpopulations of students.

A quantitative sampling research design was utilized that examined average scale scores, percentage of students passing, and student growth over a three-year period on state-mandated assessments in reading, mathematics, science, and social studies. Four sampling rates were considered (10%, 15%, 20%, & 25%) when analyzing student performance across demographic variables, including population estimates by socioeconomic status, limited English proficiency, and placement in special education

classes. Note the demographic variables considered in the study were based Academic Excellent Indicator System (AEIS), which is the Texas accountability system (TEA, 2013).

Results Related to Research Question 1

1. How can stratified random sampling reduce the number of students taking state assessments in Texas school districts while accurately providing precise estimates of the mean scores, student growth, and the percentage of students passing high stakes assessments?

The sample results were compared to the overall population of seventh grade students by ethnicity within the district. Regarding the sample mean percent passing for reading, the mean percent passing and the standard errors associated with each sampling rate appear to be similar across each ethnic group. However, there are apparent differences noted in the percent of students passing TAKS math. As the sampling rate increases, the standard errors decrease. This is especially true when comparing the 10% and 15% sampling rates to the twenty and 25% sampling rates. Regarding the sample mean scores for reading and math, the reading mean scores among Asian students ranged from 799.55 (*SD* = 100.53) (95% CI = 767L - 832U) for the 10% sampling rate to 822.16 (*SD*= 45.95) (95% CI = 807L - 837U) for the 25% sampling rate. The mean population reading score among all Asian students was 820.58(*SD* = 121.51). Similarly, the sample mean reading scores among White students ranged from 811.14 (*SD* = 22.09) (95% CI = 804L - 818U) for the 15% sampling rate to 820.44 (*SD* = 18.97) (95% CI = 814L - 829U) for the 10% sampling rate. The mean population reading score among White students was 818.28 (*SD* = 82.30.). Similar

results were reported for African American and Hispanic students. Note in each subgroup, the 95% CI captured the true population value. Regarding the Black and Hispanic students, the mean reading scores among the sampling rates were similar to the population values with the 95% CIs capturing the true population values. As expected, as the sampling rate increases the standard errors decreased overall. For example, the standard errors were greater for the 10% and 15% sampling rates compared to the 20% and 25% sample rates.

Concerning the accuracy of the samples replicating the true population values, the sample results were very similar for each sampling rate among ethnic groups for reading. However, there were greater disparities between the sample and population results related to TAKS math. A plausible explanation for the increased variation in TAKS math scores among samples is the larger standard deviations associated with TAKS math scores. With the exception of Asians, the TAKS math standard deviations are almost twice as large as the TAKS reading standard deviations. The standard deviation associated with the population math scores was more than twice that of reading scores (*SD*-reading = 103.11; *SD*-math = 213.15).It is important to have some knowledge of the standard deviation of the true population prior to conducting a stratified random sample due to the fact that more disperse populations require a larger sample size in order to attain the same level of precision for sampling estimates (Thompson, 2012). Prior research has indicated that greater spread in population scores (i.e., increased standard deviations) lead to greater inaccuracy in replicating the populations values, thus requiring a larger sampling rate (increasing sample size) to obtain the true population parameter (Thompson, 2012). In some instances, it may be

necessary to increase the sampling rate among strata with increased standard deviations to obtain results that closely replicate the true population values.

Results Related to Research Question 2

2. What is the recommended sampling rate among student subpopulations in Texas school districts to accurately provide precise estimates of mean scores, student growth, and the percentage of students passing high stakes assessments among student subpopulations?

Based on the results of this study, it appears that a 20% sampling rate would closely approximate the parameter values regarding the mean TAKS reading and mathematics scale scores and the percentage of students passing these assessments. Although the current study reported results related to ethnicity among regular education students, similar results were found among all student demographic variables across each district type examined.

Compared to the Finland studies, where they sample 10% of their students, the sampling rate found to be adequate to replicate the parameter values in the current study was slightly higher. A plausible explanation could be attributed to the fact that that Finnish study samples all students nationwide, which yields less variation in the scores compared to individual campuses within districts, which are more varied. Note that scores with greater spread require a greater sampling rate to derive parameter estimates that closely approximated the true population values.

The findings suggest that a 20% sampling rate is adequate in most cases. However, when there is increased variation in the data, a 25% sampling rate was

sufficient. For example, regarding LEP TAKS reading, as the sampling rate increased, the results trended towards the population value of 67.57%. The results for the 25% sampling rate appear to be adequate to gain insight into the percent of LEP students passing TAKS reading district-wide (difference of 1.68 percentage points between sample and true population value). Further, similar standard errors were reported for each sampling rate. As for the percentage of LEP students passing TAKS math, the results were not as accurate as the percentage of LEP students passing TAKS reading. On average, there was a difference of approximately 12 percentage points between the sampling rates and the true population value. A probable explanation for the discrepancy between the percent of LEP students passing the TAKS reading and math assessments is the wide variation in math scores compared to the reading scores. This is especially true among the special education subgroup and subgroups that have a minimum of 30 students. With 30 or fewer students in a subgroup it is recommended that all students be included in the testing program. Although not reported here due to space limitations, the results of the current study found it difficult to obtain stable results with less than 30 students in a subgroup at the sampling rates considered in the current study.

Discussion

The public at all levels have argued whether high stakes testing associated with the current accountability systems are working as advocates intended. As Au (2008) mentions standardized testing does not prepare their children for the intellectual rigors demanded within the globalized economy. Districts must find ways to foster innovation

and responsiveness without compromising equity, access, and the public purpose of schools to prepare citizens who can live, work, and contribute to a common democratic society (Darling-Hammond, 2010). This study addresses these questions by demonstrating that not all students necessitate testing to demonstrate student output in terms of academic achievement. This study was executed as a direct response to the national impetus for new accountability policies.

Originally, the purpose of NCLB (2001) asserted a meaningful purpose by focusing on every child, but has inadvertently increased the demand on students and teachers at the sacrifice of the child's educational experience to be focused on the test (Au, 2011; Dee, Jacob, & Schwartz, 2013; Darling-Hammond, 2010; Rothstein, Jacobsen & Wilder, 2008; Koretz, 2008; Musoleno, Malvern, & White, 2010; Nelson & Eddy, 2008; Ravitch, 2010) versus post-secondary success (ACT, 2013; Au 2008; College Board, 2012). Moreover, some argue that focusing on the state-mandated assessments narrows the curriculum by eliminating crucial concepts from the curriculum that are not covered on the state assessment (Au, 2011; Dee, Jacob, & Schwartz, 2013; Darling-Hammond, 2010; Rothstein, Jacobsen, Wilder, 2008; Koretz, 2008; Musoleno et al., 2010; Nelson & Eddy, 2008; Ravitch, 2010).

The foundation of high stakes testing was to ensure all children are well educated; however, the current accountability system forfeits higher levels of thinking by constraining students to sheer recall to pass the test. According to Bloom's (1969) hierarchy of learning, abilities of recall and application occur within level one otherwise referred to as the knowledge level of learning. Reducing the number of students taking the state-mandated assessments and allowing educators autonomy to teach with the

theoretical foundations of Dewey, Vygotsky, Bloom, and Gardner (as they do in Finland) would improve instructional effectiveness. By alleviating the anxiety associated with the test, teachers can allocate more time for instruction, which will benefit all students. More importantly, supplementary time and effort can be focused on the lowest students to progress higher.

As conveyed in popular press, political reasons exist to discount the use of SRS. In 2000, the federal court prohibited the US Census Bureau from utilizing statistical sampling as opposed to the traditional person-by-person headcount. The sampling plan was widely criticized for undercounting minorities and attributed to manipulating the allocation of federal capital for political gain. Unlike the US Census Bureau's sampling plan of 2000, this study purposefully disaggregated test data based on both TEA (2013) and NCLB (2001) student subgroup definitions. Each subgroup is sampled independently to ascertain true values relevant to the population of interest to improve the representation of the individual strata (subgroups) themselves.

There is a wealth of literature (Casbarro, 2005; Wong, 2013; Jacobsen & Young, 2013) that addresses the politics behind high-stakes testing. Ample evidence exists to suggest that proponents of high stakes testing are financially motivated to lobby for standardized testing via policy mandates. Private industry-including the three largest textbook companies- have lobbied extensively in Washington to promote an agenda of high stakes testing that mandates the use of the very types of test that they develop and publish (Jacobsen & Young, 2013). One Government Accountability Office (GAO) estimate produced for Congress calculated that states would need to spend between 1.9 and 5.3 billion dollars to produce more 433 tests in order to satisfy NCLB mandates

(GAO, 2003).

Although the Student Response System (SRS) is a logical solution to increase student achievement by allowing more time for learning with a focus on college and workforce readiness, it is foreseen that the testing industry will be in direct opposition. According to public records of the contract between Pearson Education Management and the Texas Education Agency, Texas is projected to spend more than $468 million between 2010-2015 on its new and revamped State of Texas Assessment of Academic Readiness (STAAR) assessment (House Bill 5, 2012). This new assessment system, to be fully implemented in the spring of 2014, includes annual testing in reading and mathematics for all students in Grades 3 - 11. Additionally, students in Grades 5 and 8-11 are assessed on science and social studies exams. Students in Grades 4, 7, 9, 10, and 11 take writing assessments. This includes end of course (EOC) assessments in secondary education. With resources constrained, the price of developing these assessments seems both cost prohibitive and economically inefficient. Yet, the state's new testing program has seen high failure rates, particularly among at-risk students. Approximately 61% of Texas students come from low-income homes and among those students, 47% has failed at least one of the standardized exams (TEA, 2012).

Texas is making progress in the area of the number of tests students take. In the area of assessment, House Bill 5 reduces the number of end-of-course exams (EOCs) from 15 to 5. The five EOCs will consist of English II (reading and writing), Algebra I, Biology and U.S. History. It also eliminates the requirement that EOCs count for 15% of a student's final course grade and the requirement for students to earn a certain cumulative score on the EOC. By reducing the number of EOC exams TEA had to

develop and administer, HB 5 would result in savings of $12.1 million annually, according to the fiscal note (House Research Organization, 2013).

Conclusion

Reduction in student-administered tests is a new ascertainable paradigm. Fewer tests will allow state and districts to direct limited funding to redirect dollars targeted for testing toward the use of deeper forms of questioning, psychometrics, new assessment technology, and project based assessments. In reducing the number of state assessments schools can allocate additional time to instruction with emphasis on college and career readiness.

Since ESEA was not reauthorized, a wavier system was developed for schools to opt out of NCLB (2001) accountability measures. To receive flexibility from NCLB (2001), states must adopt and have a strong plan to implement college- and career-ready standards. States must also create comprehensive systems of teacher and principal development; evaluation and support that include factors beyond test scores, such as principal observation, peer review, student work, or parent and student feedback. States receiving waivers must set new performance targets to improve student achievement and close achievement gaps.

States receiving flexibility are required to implement accountability systems that reward high-performing schools, while targeting interventions for the lowest-performing schools. In terms of the achievement gap, all schools will be required to develop and implement plans for improving educational outcomes for underperforming subgroups of students. The states with waivers must agree to accept the Common Core State

Standards, a national curriculum in mathematics and English language arts developed by nongovernment groups that has not undergone field-testing (Ravitch, 2012). In addition, states with waivers must agree to evaluate teachers based on student test scores. Teachers with populations of greatest need such as low SES, SPED, LEP, and at-risk students with likely experience the greatest consequences for insufficient scores. The sum of all these changes means that test scores will matter even more in the states with waivers than in the states oppressed by NCLB's heavy-handed regulations (Ravitch, 2012).

The House of Representative approved the Student Success Act (SSA) (2012) to reform the accountability mandates of NCLB. Significant changes of (SSA) include restoration of state authority for establishing performance ratings, the elimination of Adequate Yearly Progress (AYP), removal of the "highly qualified" teacher mandate, and provides funding flexibility while empowering states to design school improvement strategies (Burke, 2012). Under the provisions of the bill continued testing of students in grades 3-8 is required and the use of student test scores remains a significant aspect of teacher evaluation. With the recent legislative updates it seems testing remains a significant aspect of the educational system. However, it is surprising that educational reform has not analyzed how are peer countries address education they continue to outperform the United States on international assessments and rankings. The new proposed wavier system seems to be more of the same and lacks all the qualities of an intelligent accountability system as suggested by Sahlberg (2010):

> More intelligent accountability involves all stakeholders, including students and parents, in discussing and determining the extent that jointly set goals have been attained. It combines data from student assessments, external examinations, teacher-led classroom assessments, feedback from parents and school self-

> evaluations. Intelligent accountability draws on data from samples rather than census-based assessments that, by themselves, limit the stakes of student testing. (p. 58)

The total annual cost of national assessment in Finland is less than $5 million (Sahlberg, 2011). Regarding efficiency, Texas spends $44 billion per year on public education. Of that, almost $1 billion is spent on testing days. This amount is staggering. Based on employing a stratified random sample with verified methodology requiring only 20% of the student population to take assessments, there could be substantial savings while still obtaining an accurate measure of students' academic progress.

U.S. and Finnish education policies have appeared to be progressing in opposite directions. While U.S. public schools moved to standardized testing, Finnish schools avoided nationwide tests to evaluate teachers, students or schools, instead relying on sample-based testing. Based on the work in Finland, the reduced student sample obtained from the population employing probability proportional to size (PPS) selection, results can be achieved that are representative of the population parameter and substantially reduce expenditure on state assessments.

The use of assessments have increased to the point that the majority of children are tested on high stakes assessments with little predictive validity linked to a student's future success. Although countries such as Finland have resorted to stratified random sampling to reduce the number of students tested while maintaining accountability, it is not clear if such a program could work in the United States. While there are theoretical constructs that establish statistical sampling as a valid form of assessment, it is not clear that the theory will be supported in actual practice. Research suggests stratified random sampling is a suitable methodology in order to make proportionate, and

therefore meaningful, comparisons between sub-groups in the population (Gay, 1987). Robson (2002) notes sampling theory supports stratified random sampling because the means of the stratified samples are likely to be closer to the mean of the population overall. Further Leary (1995) recommends a stratified random sample will reflect the characteristics of the population as a whole. The need for testing every child may not be needed or necessary shown that unlike the U.S. where testing every student is prominent, Finland uses sampled-assessments (Sahlberg, 2007) whereby they sample 10% of their students (Savola, 2012) and only test every student in two grade levels (Darling-Hammond, 2010). Although Finland does not test every student, it remains one of the prominent educational systems in the world (Sahlberg, 2010).

This chapter is a comprehensive overview of the study. Chapter 5 summarized the study, the findings, identified limitations, and provided recommendations for future research. One primary goal of this study was to determine if stratified random sampling was a viable option for reducing the number of students participating in Texas state assessments, and to determine which sampling rate provides consistent estimates of the test results among the population of students. The study examined scale scores, percent passing, and student growth over a three-year period on state-mandated assessments in reading, mathematics, science, and social studies. Four sampling rates were considered (10%, 15%, 20%, & 25%) when analyzing student performance across demographic variables within and across each participating district. Based on the findings of this study a 20% sampling rate would closely approximate the parameter values regarding the mean TAKS reading and mathematics scale scores and the percentage of students passing these assessments.

This study attempted to statistically determine if stratified random sampling was a viable option for reducing the number of students participating in Texas state assessments, and determine which sampling rate provided consistent estimates of the actual test results among the subpopulations of students. The results of the study indicate that employing a stratified random sample is a viable option to reduce the number of students participating in a testing program while obtaining an accurate estimate of the scores for the population of students.

Future Research

Certain relevant facets of this study warrant further investigation. Though this study recommends the use of SRS to augment and redefine accountability systems the logistical challenges of implementation in a schoolhouse are considerable. As education in the U.S. remains a local function, the logistical tactics of implementation are beyond the scope of this study and remain at the discretion and expertise of legislatures and school administrators.

While this study focused on TAKS assessment, future research entails applying SRS to both STAAR assessments and curriculum benchmark assessments (CBAs) to determine if stratified random sampling is a viable option for reducing the number of students participating, and determines if the recommended 20% sampling rate provides consistent estimates of the test results among the population of students. In addition, the current study was based on five public school districts while future research should replicate this study using more districts with varied student demographics to include both low SES and at-risk student populations to ensure that the results of this study are

generalizable. It is recommended low SES populations be included, as research has found that state-level family socioeconomic status contributed most to the differences in average student achievement across states (Berliner, in press; Wei, 2012).

REFERENCES

ACT (2013). The Condition of College & Career Readiness. *The condition of college & career readiness 2013-National readiness report.* Retrieved from http://www.act.org/research/policymakers/cccr13/pdf/CCCR13-NationalReadinessRpt.pdf

Adobe Systems Inc. (2013). *Barriers to creativity in education: Educators and parents grade the system* Retrieved from http://wwwimages.adobe.com/www.adobe.com/content/dam/Adobe/en/education/pdfs/barriers-to-creativity-in-education-study.pdf.

Allington, R. L., & McGill-Franzen, A. (2003). The impact of summer setback on the reading achievement gap. *Phi Delta Kappan*, *85*(1), 68-75.

American Educational Research Association. (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Amrein-Beardsley, A. (2009). This is jeopardy. *Education Digest*, *74*(5), 14-18. Retrieved from http://www.eddigest.com/

Amrein, A. L., & Berliner, D. C. (2002). High-stakes testing & student learning. *Education Policy Analysis Archives*, *10*, 18.

Anderson, L. (April 4, 2005). The No Child Left Behind Act and the legacy of federal aid to education. *Education Policy Analysis Archives, 13*(24). Retrieved July 27, 2008 from http://epaa.asu.edu/epaa/v13.n4/

Anderman, E. M., Anderman, L. H., Yough, M. S., & Gimbert, B. G. (2010). Value-added models of assessment: Implications for motivation and accountability. *Educational Psychologist*, *45*(2), 123-137.

Annie E. Casey Foundation. (2011). 2011 Kids Count Data Book. *State profiles of child well-being.* Baltimore, MD: The Anne E. Casey Foundation.

Association for Experiential Education (AEE). (1994). *AEE definition of experiential education*. Boulder, CO: Association for Experiential Education

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

Au, W. (2008) Between education and the economy: High stakes testing and the contradictory location of the new middle class. *Journal of Education Policy*, *23*(5), 501–513.

Au, W. (2011). Teaching under the new Taylorism: High-stakes testing and the standardization of the 21st century curriculum. *Journal of Curriculum Studies*, *43*(1), 25-45.

Australian Curriculum, Assessment and Reporting Authority. (2008). Sydney, N.S.W: Australian Curriculum, Assessment and Reporting Authority.

Bahm, A. (2009). The effects of discovery learning on students' success and inquiry learning skills. *Eurasian Journal of Educational Research*, *35*, 1-20.

Bailey, M. J., & Dynarski, S. M. (2011). *Gains and gaps: Changing inequality in US college entry and completion* (No. w17633). National Bureau of Economic Research.

Baker, B. D., Sciarra, D. G., Farrie, D., & Education Law Center. (2010). Is school funding fair?: A national report card. Newark, NJ: Education Law Center.

Baker, E. L., & Economic Policy Institute. (2010). *Problems with the use of student test scores to evaluate teachers*. Washington, DC: Economic Policy Institute.

Bartels, M., Rietveld, M. J. H., Van Baal, G. C. M., & Boomsma, D. I. (2002). Genetic and environmental influences on intelligence. *Behavior Genetics*, *32*(4), 237–249.

Beisser, S. R. (2008). Analysis of Funding & Services for NCLB - Special Education and Gifted Children in USA. Research study conducted spring 2008.

Berliner, D. (in press). Effects of inequality and poverty vs. teachers and schooling on America's youth. *Teachers College Record*, *116*(1), Retrieved from http://www.tcrecord

Berliner, D. C. (June, 2009). *Rational response to high-stakes testing and the special case of narrowing the curriculum.* In International Conference on Redesigning Pedagogy, National Institute of Education, Nanyang Technological University, pp. 1-13).

Berliner, D. (January 01, 2011). Rational responses to high stakes testing: The case of curriculum narrowing and the harm that follows. *Cambridge Journal of Education*, *41*(3), 287-302.

Bicknell-Holmes, T., & Hoffman, P. S. (2000). Elicit, engage, experience, explore: Discovery learning in library instruction. *Reference Services Review*, *28*(4), 313-322.

Binet, A., & Simon, T. (1916). *The development of intelligence in children: The Binet-Simon scale* (No. 11). Baltimore, MD: Williams & Wilkins Company.

Black, P., Burkhardt, H., Daro, P., Jones, I., Lappan, G., Pead, D., & Stephens, M. (2012). High-stakes examinations to support policy: Design, development and implementation, *Journal of the International Society for Design and Development in Education*, *2*(5), 1-31.

Bloom, B. S. (1969). Some theoretical issues relating to educational evaluation. Educational evaluation: New roles, new means*. The 63rd yearbook of the National Society for the Study of Education*, (Part 2), 26-50.

Brown v. Board of Education, 347 U.S. 483 (1954).

Bryman, A. (2008). *Social research methods*. (3rd ed.). New York: Oxford University Press.

Burke, L. M. (2012). The student success Act: Reforming federal accountability requirements under No Child Left Behind. WebMemo. (No. 3461). Heritage Foundation.

Business Roundtable. (2005). *Tapping America's potential: The education for innovation initiative*. Washington, DC: Business Roundtable.

Byrd-Blake, M., Afolayan, M. O., Hunt, J. W., Fabunmi, M., Pryor, B. W., & Leander, R. (2010). Morale of teachers in high poverty schools: A post-NCLB mixed methods analysis, *Education & Urban Society*, *42*(4), 350-472.

California Education (Organization). (2006). *Is the No Child Left Behind Act working?: The reliability of how states track achievement*. Berkeley, CA: University of California, Berkeley.

Camper, N. K. (1978). Testing, guidance and curriculum: The impact of progressive education in Waltham, Massachusetts, 1918-1968. *Educational Studies*, *9*(2), 159-171.

Cargile, E. (2012, May 04). Tests' price tag $90 million this year. KXAN News.com. Retrieved from http://www.kxan.com/dpp/news/investigations/staars-price-tag-90-million-this-year

Casbarro, J. A. (2005). *Test anxiety and what you can do about it: A practical guide for teachers, parents, and kids*. New York: NPR Inc.

Cawelti, G. (2006). The side effects of NCLB. *Educational Leadership*, *64*(3), 64.

Center on Education Policy. (2008, February). Instructional time in elementary education. *Child Today*, *32*(2), 50-53.

Chang, M. L. (2009). An appraisal perspective of teacher burnout: Examining the emotional work of teachers. *Educational Psychology Review*, *21*(3), 193-218

Chingos, M. (2012). *Strength in numbers, state spending on K-12 assessment systems*. The Brookings review. Washington, DC: Brookings Institution.

Clarke, M. M., Madaus, G. F., Horn, C. L., & Ramos, M. A. (March 01, 2000). Retrospective on educational testing and assessment in the 20th century. *Journal of Curriculum Studies*, *32*(2), 159-81.

Cochran, W. G. (1977). *Sampling techniques*. New York: Wiley.

Congress, U. S. (1965). Elementary and Secondary Education Act (ESEA). Public Law, 89-10.

Cowie, M., & Croxford, L. (2007). *Intelligent accountability: Sound-bite or sea-change*? Centre for Educational Sociology, University of Edinburgh.

Darling-Hammond, L. (2005). *Policy and change: Getting beyond bureaucracy*. In Extending educational change (pp. 362-387). Springer Netherlands.

Darling-Hammond, L. (2007a). Evaluating No Child Left Behind. *The Nation* (May 21, 2007).

Darling-Hammond, L. (2009). President Obama and education: The possibility for dramatic improvements in teaching and learning. *Harvard Educational Review*, *79*(2), 210-223.

Darling-Hammond, L. (2010). *The flat world and education: How America's commitment to equity will determine our future*. New York: Teachers College Press.

Dee, T. S., Jacob, B., & Schwartz, N. L. (2013). The effects of NCLB on school resources and practices. *Educational Evaluation and Policy Analysis*, *35*(2), 252-279.

Dewey, J. (1900). *The school and society*. Chicago, IL: University of Chicago.

Dewey, J. (1916). *Democracy and education*. New York: Macmillan.

Dewey, J. (1933). *How we think: A restatement of the relation of reflective thinking to the educative process*. Lexington, MA: Heath.

Dewey, J. (1938). *Experience and education*. New York: Macmillan.

Dobbelaer, S. G. (2010, April). Do high-stakes assessments improve learning? [Electronic version]. *Education and Urban Society*, *42*(4), 450-472.

Education Policy Analysis Archives, *13*(24). Retrieved July 18, 2013 from http://epaa.asu.edu/epaa/v13.n4/

Educational products. In S. A. Courtis (Ed.), (YEAR). *The Measurement of Educational Products* (17th Yearbook of the National Society for the Study of Education, Pt. 2. pp. 16–24). Bloomington, IL: Public School.

Education Commission of the States & United States. (2004). *ECS report to the nation: State implementation of the No Child Left Behind Act: Respecting diversity among states*. Denver, CO: Education Commission of the States.

Educational Testing Service. (2006). *Keeping our edge: Hispanic Americans speak on education and competitiveness*. Princeton, NJ: Educational Testing Service.

Edwards, D., & Mercer, N. (1987). *Common knowledge: The development of understanding in the classroom*. London: Methuen/Routledge.

Ellison, S. (2012). Intelligent Accountability: Re-thinking the concept of accountability in the popular discourse of education policy. *Journal of Thought*, *47*(2), 19-41.

Ercikan, K. (2006). Developments in assessment of student learning. In P. A. Alexander & P. H. Winne (Eds.), *Handbook of educational psychology* (2nd ed.), 929–952. Mahwah, NJ: Lawrence Erlbaum Associates.

Flattau, P. E., Bracken, J., Institute for Defense Analyses, Science and Technology Policy Institute (Rand Corporation), & United States. (2007). *The National Defense Education Act of 1958: Selected outcomes*. Washington, DC: Institute for Defense Analyses, Science & Technology Policy Institute.

Flumerfelt, S. F. (2008, April). Is lean appropriate for schools? [Electronic version]. In S. Flumerfelt (Ed.), White papers. The Pawley Lean Institute.

Forte, E. (2010). Examining the assumptions underlying the NCLB federal accountability policy on school improvement. *Educational Psychologist*, *45*(2), 76-88.

Fuller, B., Wright, J., Gesicki, K., & Kang, E. (2007). Gauging growth: How to judge No Child Left Behind?. *Educational Researcher*, *36*(5), 268-278.

Gardner, H. (1993). *Multiple intelligences: The theory in practice:[a reader]*. Basic books.

Gay, L. R. (1987). *Educational research: Competencies for analysis and application,* (3rd ed.). Columbus, OH: Merrill.

Goddard, H. H. (1913). Standard method for giving the Binet test. *Training Schools, 10*, 9-11.

Gotbaum, B. (2002, November 21). *Pushing out at-risk students: An analysis of high school discharge figures*. The Public Advocate for the City of New York. Retrieved from http://publicadvocategotbaum.com/

Gravetter, F. J., & Wallnau, L. B. (2004). *Statistics for the Behavioral Sciences*. Belmont, CA: Wordsworth.

Greaney, V., & Kellaghan, T. (2012). *Implementing a national assessment of educational achievement*. Washington, DC: The World Bank.

Guisbond, L., & National Center for Fair & Open Testing (FairTest). (2012). NCLB's Lost Decade for Educational Progress: What Can We Learn from this Policy Failure? National Center for Fair & Open Testing (FairTest). P.O. Box 300204, Jamaica Plain, MA 02130. Tel: 617-477-9792; Web site: http://www.fairtest.org.

Guskey, T. R. (2005). Mapping the Road to Proficiency. *Educational leadership*, *63*(3), 32-38.

Hamilton, L. S., Stecher, B. M., & Yuan, K. (2012). Standards-based accountability in the United States. *Education*, *3*(2), 149-170.

Hamilton, L., Dunbar, S., & Linn, R. (2003). 2 Assessment as a policy tool. Review of Handbook. Retrieved from http://www.occ.gov/publications/publications-by-type/comptrollers- handbook/sampmeth.pdf

Haney, W. (1984) Testing reasoning and reasoning about testing. *Review of Educational Research*, *54*(4), 597-654.

Haney, W. M., Madaus, G. F., & Lyons, R. (1993). *The fractured marketplace for standardized testing*. Boston, MA: Kluwer-Academic Publishers.

Haney, W., & National Institute of Education (U.S.). (1981). *Thinking about test development*. Washington, DC: National Institute of Education.

Hargrove, K. (January 01, 2012). From the classroom: Advocating acceleration. *Gifted Child Today*, *35*, 1, 72-73.

Hamilton, L. S., Stecher, B. M., & Klein, S. P. (2002). *Making sense of test-based accountability in education*. Santa Monica, CA: Rand Corporation.

Harris, D. N. (2007). High-flying schools, student disadvantage, and the logic of NCLB. *American Journal of Education*, *113*(3), 367-394.

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

Hemphill, F. C., & Vanneman, A. (2011). Achievement gaps: How Hispanic and White students in public schools perform in mathematics and reading on the National Assessment of Educational Progress. NCES 2011-459. National Center for Education Statistics.

Herman, J. L., Haertel, E., & National Society for the Study of Education. (2005). *Uses and misuses of data for educational accountability and improvement*. Chicago, IL: NSSE.

Higgins, B., Miller, M., & Wegmann S. (2006). Teaching to the test…not! Balancing best practice and testing requirements in writing. *The Reading Teacher*, *60*(4), 310-319. Retrieved from http://www.reading.org/publications/journals/

Holcombe, R., Jennings, J., & Koretz, D. (2013). The roots of score inflation: An examination of opportunities in two states' tests. In G. Sunderman (Ed.), *Charting reform, achieving equity in a diverse nation*, 163-189. Greenwich, CT: Information Age Publishing.

Hollingworth, L. (2007). Five ways to prepare for standardized tests without sacrificing best practice. *Reading Teacher*, *61*(4), 339-342. doi:10.1598/RT.61.4.7

Hornof, M. (2008). Reading tests as a genre study. *Reading Teacher*, *62*(1), 69-73. doi:10.1598?RT.62.1.8

Hout, M., & Elliott, S. W. (Eds.). (2011). *Incentives and test-based accountability in education*. National Academies Press.

Hout, M., Elliott, S., & Frueh, S. (2012). Do High-Stakes Tests Improve Learning? Test-based incentives, which reward or sanction schools, teachers, and students based on students' test scores, have dominated US education policy for decades. But a recent study suggests that they should be used with caution and carefully evaluated. *Issues in Science and Technology*, *29*(1), 33.

H.R. 3989--112th Congress: *Student success act*. (2012). In www.GovTrack.us Retrieved September 23, 2013, from http://www.govtrack.us/congress/bills/112/hr3989.

House Bill 5, 83(R), 83R 6400 PAM-D. (2013).

House Research Organization. (2013). *HB 5 bill analysis 3/26/2013.* State of Texas, U.S. Retrieved from http://www.hro.house.state.tx.us/pdf/ba83r/hb0005.pdf.

Humes, K., Jones, N. A., & Ramirez, R. R. (2011). *Overview of Race and Hispanic Origin, 2010*. US Department of Commerce, Economics and Statistics Administration, US Census Bureau.

Jacobsen, R., & Young, T. V. (2013). The New Politics of Accountability Research in Retrospect and Prospect. *Educational Policy*, *27*(2), 155-169.

Jencks, C. (1979). Who Gets Ahead? The Determinants of Economic Success in America.

Jensen, A. (1969). How Much Can We Boost IQ and Scholastic Achievement? *Harvard Educational Review*, February 1969, No. 1, pp. 1

Jensen, A. (1992). Understanding in Terms of Information Processing, *Educational Psychology Review*, *4*(3), pp. 271-308, particularly pp. 299-300.

Johnson, W., McGue, M., & Iacono, W.G. (2007). Socioeconomic status and school grades: Placing their association in broader context in a sample of biological and adoptive families. *Intelligence*, *35*, 526-541.

Jolly, J. L. (2009). The National Defense Education Act, current STEM initiative, and the gifted. *Gifted Child Today 32*(2), 50-53.

Jolly, J., & Kettler, T. (2008). Gifted Education Research 1994-2003: A Disconnect Between Priorities and Practice. *Journal of Education for the Gifted*, *31*(4), 427-446.

Joncas, M., & Foy, P. (2011). *Sample Design in TIMSS and PIRLS*. Retrieved from http://timssandpirls.bc.edu/methods/pdf/TP_Sampling_Design.pdf

Jones, M. G., Jones, B. D., & Hargroves, T. Y. (2003). *The unintended consequences of high-stakes testing*. Lanham, MD: Rowman & Littlefield.

Kantor, H., & Lowe, R. (2006). From New Deal to no deal: No Child Left Behind and the devolution of responsibility for equal opportunity. *Harvard Educational Review*, *76*(4), 474-502.

Kim, J. S., & Sunderman, G. L. (2005). Measuring academic proficiency under the No Child Left Behind Act: Implications for educational equity. *Educational Researcher*, *34*(8), 3-13.

Kobrin, J. L., Patterson, B. F., Shaw, E. J., Mattern, K. D., & Barbuti, S. M. (2008). *Validity of the SAT® for predicting first-year college* grade *point average* (College Board Research Report No. 2008-5). New York: College Board.

Koch, M. J., & DeLuca, C. (2012). Rethinking validation in complex high-stakes assessment contexts. *Assessment in Education: Principles, Policy & Practice*, *19*(1), 99-116.

Kolb, D. A. (1984). *Experiential learning*. Englewood Cliffs, NJ: Prentice Hall.

Koretz, D. (2008). A measured approach. *American Educator*, *32*(2), 18-39.

Koretz, D. (2010). *Some implications of current policy for educational measurement.* Retrieved from http://www.k12center.org/publications.html

Koretz, D. (2013). Commentary on E. Haertel, How Is Testing Supposed to Improve Schooling? *Measurement: Interdisciplinary Research and Perspectives*, *11*(1-2), 40-43.

Koretz, D., & Hamilton, L.S. (2006). Testing for accountability in K-12. In R.L. Brennan (Ed.), *Educational Measurement* (4th ed.), 531-578. Westport, CT: American Council on Education/Praeger.

Krejcie, R. V., & Morgan, D. W. (1970). Determining sample size for research activities. *Educational and Psychological Measurement*, *30,* 607-610.

Kress, S., Zechmann, S., & Schmitten, J. M. (2011). When performance matters: The past, present, and future of consequential accountability in public education. *Harv. J. on Legis.*, *48*, 185.

Lai, E. R., & Waltman, K. (2008). Test preparation: Examining teacher perceptions and practices. *Educational Measurement: Issues & Practice*, *27*(2), 28–45.

Lay, J. C., & Stokes-Brown, A. K. (2009). Put to the test understanding differences in support for jhigh-stakes testing. *American Politics Research*, *37*(3), 429-448.

Leary, M. R. (1995). *Behavioral research methods*. Pacific Grove, CA: Brooks/Cole Publishing Company.

Lee, J., & Lee, Y. S. U. (2012). Effects of Testing. *International Guide to Student Achievement*, 416, New York, Routledge.

Lehtonen, R. & Pahkinen E. (2004). *Practical methods for design and analysis of complex surveys.* New York: Wiley and Sons.

Levin, R. A. (1991). The debate over schooling: Influences of Dewey and Thorndike. *Childhood Education*, *68*(2), 71-75.

Loyens, S., Rikers, R., & Schmidt, H. (2007). Students' conceptions of distinct constructivist assumptions. *European Journal of Psychology of Education*, *22*(2), 179-199.

Mackintosh, N. J. (1998). *IQ and human intelligence*. Oxford: University Press.

Martin, M. O., Mullis, I., & Chrostowski, S. (2003). *TIMSS 2007 international science report*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Mayrowetz, D. (2009). Instructional practice in the context of converging policies teaching mathematics in inclusive elementary classrooms in the standards reform era. *Educational Policy*, *23*(4), 554-588.

McCallister, B., & Plourde, L. (2008). Enrichment curriculum: Essential for mathematically gifted students. *Education*, *129*(1), 40-49.

McKinsey and Company. (2009). *The economic impact of the achievement gap in America's schools: Summary of findings*. New York: McKinsey & Company, Special Sector Office.

McNeil, L. M. (2005). Faking equity: High-stakes testing and the education of Latino youth. *Leaving children behind: How "Texas-style" accountability fails Latino youth*, 57-111.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher*, *18*(2), 5-11.

Messick, S., & Educational Testing Service. (1990). *Validity of test interpretation and use*. Princeton, NJ: Educational Testing Service.

Moe, T. M. (2002). *Politics, control, and the future of school accountability*. Program on Educational Policy and Governance, Harvard University.

Murphy, M., & Schulz, W. (2006). *Sampling for national surveys in education*. Retrieved from http://www.mceetya.edu.au/verve/_resources/Sampling_for_National_Surveys_in Education.pdf

Musoleno, R. R., Malvern, P. A., & White, G. P. (2010). Influences of high-stakes testing on middle school mission and practice. *RMLE Online*, *34*(3).

National Commission on Excellence in Education. (1983). *A Nation at risk*. Washington, DC: Congressional Research Service.

National Center for Education Statistics. (2009). *The Nation's Report Card: Mathematics 2009* (NCES 2010-451). Institute of Education Sciences, U.S. Department of Education, Washington, DC.

National Center for Education Statistics. (2009). *The nations report card: Reading 2009* (NCES 2010-458). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.

National Center for Education Statistics. (2013). *Nations report card glossary.* Retrieved from http://nces.ed.gov/nationsreportcard/glossary.aspx

National, C. F. E. S. (2012). *Digest of education statistics 2011*. Sl: Claitors Publishing.

National Education Statistics Center. (2007). *NAEP 2007 sample design*. Retrieved from http://nces.ed.gov/nationsreportcard/tdw/sample_design/2007/sampdsgn_2007.aspNational

National Education Statistics Center. (2009). *Trends in international math and science study* (TIMSS). Retrieved from http://nces.ed.gov/timss/

National Education Statistics Center. (2011). *Selecting schools for participation in state-level NAEP*. Retrieved from http://www.ode.state.or.us/initiatives/naep/naep_sampling_factsheet.pdf

Nelson, M., & Eddy, R. M. (2008). Evaluative thinking and action in the classroom. *New Directions for Evaluation*, *117*, 37-46. Retrieved from http://www.interscience.wiley.com.ezp.waldenulibrary.org/jpages/0271-0560/

New Zealand., & Maths Technology. (2010). National standards: School sample monitoring & evaluation project: Survey of principals & analysis of reports formats: *Report to the Ministry of Education*. Wellington, NZ: Ministry of Education.

Nichols, S. L., & Berliner, D. C. (2008). Testing the Joy out of Learning. *Educational Leadership*, *65*(6), 14-18.

Nichols, S. L., & Valenzuela, A. (2013). Education policy and youth: Effects of policy on practice. *Theory into Practice*, *52*(3), 152-159.

Nichols, S., Glass, G., & Berliner, D (2012). High-stakes testing and student achievement: Updated analyses with NAEP data. *Education Policy Analysis Archives*, *20*.

No Child Left Behind (NCLB, 2001) Congress, U. S. (2001). Public Law 107-110: In 107th Congress. Retrieved June (Vol. 12), 2006.

O'Connor, C., Hill, L., & Robinson, S. R. (2009, March). Who's at risk in school and what's race got to do with it? *Review of Research in Education*, *33*(1), 1-34. doi:10.3102/0091732X08327991

Obiakor, F. E., & Beachum, F. D. (2005). *Urban education for the 21st century: Research, issues, and perspective*. Springfield: Charles C Thomas Publisher, LTD.

OECD (2004). *Learning for tomorrow's world*. First results from PISA 2003. Paris: OECD.

Organisation for Economic Co-operation and Development. (2010). Education at a glance 2010: OECD indicators. Paris, France: OECD Pub.

Palmer, D., & Rangel, V. (2011). High stakes accountability and policy implementation: Teacher decision making in bilingual classrooms in Texas. *Educational Policy*, *25*(4), 614-647, doi: 10.1177/0895904810374848

Pandina S. T., Callahan, C. M., & Urquhart, J. (2008). Paint-by-number teachers and cookie-cutter students: The unintended effects of high-stakes testing on the education of gifted students. *Roeper Review*, *31*(1), 40-52.

Patterson, B. F., & Mattern, K. D. (2011). Validity *of the SAT® for predicting first-year grades: 2008 SAT® validity sample*, (College Board Statistical Report 2011-5). *New York: The College Board.*

Patterson, B. F., & Mattern, K. D. (2012). *Validity of the SAT® for predicting first-year Grades: 2009 SAT® validity sample*, New York: The College Board.

Patterson, B. F., Mattern, K. D., & Kobrin, J. L. (2009). Validity of the SAT® for predicting FYGPA: 2007 SAT® validity sample. College Board Statistical Report.

Peine, Marie & Coleman, Laurence. (2010). The phenomenon of waiting in class. *Journal of Gifted Education*, *34*(2), 22-244.

Popham, W. J. (1987). The merits of measurement-driven instruction. *The Phi Delta Kappan*, *68*(9), 679-682.

Popham, W. J. (2001). *The truth about testing: An educator's call to action*. Alexandria, Va: Association for Supervision and Curriculum Development.

Queensland Studies Authority. (2008). NAPLAN. *Outcomes: All Queensland schools.* Brisbane, Qld: Queensland Studies Authority.

Rao, P. S. R.S. (2000). Sampling Methodologies: With applications. Chapman & Hall/CRC.

Ravitch, D. (2010, June 14). Why I changed my mind: Choice and accountability sounded good on paper, but in reality, they have failed. *The Nation*, 20-24.

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

Ravitch, D. (2011). *The death and life of the great American school system: How testing and choice are undermining education*. New York: Basic Books.

Ravitch, D. (2012). *No student left untested*. New York: Review Blog.

Reardon, S. F. (2011). The widening academic achievement gap between the rich and the poor: New evidence and possible explanations. Whither opportunity, 91-116. *Research in Education, 27*, 25.

Rising Above the Gathering Storm Committee (U.S.), National Academy of Sciences (U.S.), National Academy of Engineering, & Institute of Medicine (U.S.). (2010). *Rising above the gathering storm, revisited: Rapidly approaching category 5*. Washington, DC: National Academies Press.

Robson, C. (2002). *Real world research: A resource for social scientists and practitioner-researchers*. Oxford, UK: Blackwell Publishers.

Rothstein, R., Jacobsen, R., & Wilder, T. (2008). *Grading education: Getting accountability right*. Washington, DC: Teachers College Press.

Sacks, P. (1999). *Standardized minds: The high price of America's testing culture and what we can do to change it*. Cambridge, Mass: Perseus Books.

Saenz, W. (2010). A correlational study between the amount of property wealth behind each student attending Florida district schools and the academic proficiency among 149150 5th grade White, Black, and Hispanic students in reading within the 67counties of Florida (Doctoral dissertation, University of Central Florida, 2010). Retrieved from http://proquest.umi.com/

Sahlberg, P. (2006). Raising the bar: How Finland responds to the twin challenge of secondary education. *Profesorado, 10*(1), 1-26.

Sahlberg, P. (2007). Education policies for raising student learning: The Finnish approach. *Journal of Education Policy*, *22*(2), 147–171.

Sahlberg, P. (February 01, 2010). Rethinking accountability in a knowledge society. *Journal of Educational Change*, *11*(1), 45-61.

Sahlberg, P., & Hargreaves, A. (2011). *Finnish lessons: What can the world learn from educational change in Finland*? New York: Teachers College Press.

Savola, L. (2012). Assessment in Finnish Schools. *Journal of Mathematics Education at Teachers College*, *3*, 40–44.

Schraw, G. (2010). No school left behind. *Educational Psychologist*, *45*(2), 71-75.

Scotland. (2012). *Scottish survey of literacy and numeracy 2011 (Numeracy): Highlights from Scotland's results*. Edinburgh: Scottish Government.

Sloane, F. C., & Kelly, A. E. (2003). Issues in high-stakes testing programs. *Theory Into Practice*, *42*(1), 12-17.

Snyderman, M., & Rothman, S. (1988). *The IQ controversy, the media and public policy*. New Brunswick, NJ: Transaction Books.

Sokal, M. M. (1987). Psychological testing and American society 1890-1930. In This volume had its origins in a symposium of the same title, organized and chaired by Michael M. Sokal, at the 150th National Meeting of the American Association for the Advancement of Science, held in New York on May 29, 1984. Rutgers University Press.

Spring, J. H. (2008). *The American school: From the puritans to no child left behind*. Boston, MA: McGraw-Hill.

Texas Association of School Administrators. (2008). Public Education Visioning Institute. *Creating a new vision for public education in Texas*. Retrieved from http://www.tasb.org/legislative/documents/vpevi.pdf..

Texas Association of School Administrators. (2013). *Resolution concerning high stakes, standardized testing of Texas public school students.* Retrieved from http://www.tasanet.org/cms/lib07/TX01923126/Centricity/Domain/111/sampleresolution.pdf

Texas Education Agency (2008). Technical Digest 2007-2008. Retrieved November 1 2012 from http://www.tea.state.tx.us/student.assessment/techdigest/yr0708/

Texas Education Agency (2012). Snapshot 2012: Summary tables state totals. Retrieved August 13 2013 from http://ritter.tea.state.tx.us/perfreport/snapshot/2012/state.html

Texas Education Agency. (2013), Retrieved from http://www.tea.state.tx.us/ayp/

Texas Education Agency, (n.d.a). STAAR Resources. Retrieved from http://www.tea.state.tx.us/student.assessment/staar/

Texas High Performance Schools Consortium Act, Tex Stat. 1 C, 7, Education Code 7.0561 (2011).

College Board, The SAT® Report (2012). *The SAT® report on college & career readiness: 2012*. Retrieved from http://media.collegeboard.com/homeOrg/content/pdf/sat-report-college-career-readiness-2012.pdf.

Thomas, J. Y., & Brady, K. P. (2005). The elementary and secondary education act at 40: Equity, accountability, and the evolving federal role in public education. *Review of Research in Education*, *29*, 51-67

Thomas, R. M. (2013). High-stakes testing: Coping with collateral damage. New York: Routledge.

Thompson, S. K. (2012). *Sampling*. Hoboken, NJ: John Wiley and Sons.

Thorndike, E. L. (1918). The nature, purposes, and general methods of measurement of

Thorndike, Robert L. 1942. Two screening tests of verbal intelligence. *Journal of Applied Psychology, 26,* 128–35.

Timar, T. B. & Maxwell-Jolly, J. (Eds) (2012). *Narrowing the achievement gap: Perspective and strategies for challenging times*. Cambridge, MA: Harvard Education Press.

Time Out from Testing. (2012). *National resolution.* Retrieved from http://timeoutfromtesting.org/nationalresolution/2012

Topol, B., Olson, J., & Roeber, E. (2010*). The cost of new higher quality assessments: A comprehensive analysis of the potential costs for future state assessments*. Stanford, CA: Stanford University, Stanford Center for Opportunity Policy in Education.

United States Department of Education (2004). A guide to education and No Child Left Behind. Retrieved March 1, 2012 from http://www2.ed.gov/nclb/overview/intro/guide/guide.pdf

U.S. Department of Education. (2008). A nation accountable: Twenty-five years after A *Nation at Risk*. Retrieved from http://www.ed.gov/rschstat/research/pubs/accountable*I*

U.S. Department of Education. (2009). *Race to the top program: Executive summary*. Washington, DC: U.S. Department of Education. United States Department of the Treasury (1998). Sampling Methodologies: Comptroller's

U.S. Government Accountability Office (2003). *Characteristics of tests will influence expenses: Information sharing may help states realize efficiencies*. Washington DC: U.S. Government Accountability Office.

United States. (1983). A nation at risk: *The imperative for educational reform: A report to the nation and the secretary of education.* United States Department of Education. Washington, D.C: National Commission on Excellence in Education.

U.S. Census (n.d.). *Developing sampling techniques*. Retrieved from

United States. (USDO, 2010). College and career ready standards and assessments. Washington, DC: U.S. Dept. of Education.

Vu, P. (2008, January 17). Do state tests make the grade? *Stateline: The daily news service of the pew charitable trusts.* Retrieved from http://www.pewstates.org/projects/stateline/headlines/do-state-tests-make-the-grade-85899387452.

Vygotsky, L. S. (1978). *Mind in society*. Cambridge, MA: Harvard University Press.

Wang, L., Beckett, G. H., & Brown, L. (2006). Controversies of standardized assessment in school accountability reform: A critical synthesis of multidisciplinary research evidence. *Applied Measurement in Education*, *19*(4), 305-328.

Wardrop, J. L. (1976). *Standarized testing in the schools: Uses and roles*. Monterey, Calif: Brooks/Cole Pub. Co.

Wei, X. (March 01, 2012). Are more stringent NCLB state accountability systems associated with better student outcomes? An analysis of NAEP results across states. *Educational Policy, 26*, 2, 268-308.

White, K. R. (1982). The relation betwen socioeconomic status and academic achievement. *Psychological Bulletin*, *91*, 461-481

Wixson, K. K., Dutro, E., & Athan, R. G. (2003). The challenge of developing content standards. *Review of Research in Education*, *27*, 69-107.

Wong, K. K. (2013). Politics and governance evolving systems of school accountability. *Educational Policy*, *27*(2), 410-421.

World Bank (2007). International Program for Development Evaluation Training. Retrieved from http://www.worldbank.org/oed/ipdet/modules/M_09-na.pdf

Wright, T. and Farmer, J. (2000). A bibliography of selected statistical methods and development related to census 2000. Retrieved from http://www.census.gov/history/pdf/Wright.pdf

Yang, C. (2008). The effects of TIMSS sampling weights on inference accuracy when utilizing structural equations models. Retrieved November 15, 2012 from http://www.iea.nl/fileadmin/user_upload/IRC/IRC_2008/Papers/IRC2008_Yang1.pdf

Yeung, W., & Conley, D. (2008). Black-White achievement gap and family wealth. *Child Development*, *79*(2), 303-324. DOI: 10.1111/j.1467-8624.2007.01127

Zimmerman, B. J., & Schunk, D. H. (2003). *Educational psychology: A century of contributions*. Mahwah, NJ: L. Erlbaum Associates.