A Statistical Analysis of the Effectiveness of Mathematics Standards Reform in Texas

By

Jacob Martin

A Thesis Submitted in Partial Fulfillment

of the Requirements for the Degree

Master of Science

in Mathematics

West Texas A&M University

Canyon, TX

August 2020

Abstract

The Texas Education Agency adopted a revised set of mathematics Texas Essential Knowledge and Skills in April 2012 in order to be more aligned with the Common Core State Standards. These standards were fully implemented for the 2014-2015 school year in Kindergarten through eighth-grade classrooms. This project aimed to determine if these new standards have made an impact on Texas students' achievement in sixth and seventh grade mathematics. Using scores from the State of Texas Assessment of Academic Readiness (STAAR) from the year prior to implementation (2014) and the most recent spring administration (2019) a linear regression analysis was conducted to determine how the scores have been impacted since the new standards have gone into effect. In this study, sixth and seventh grade STAAR scores served as the data set for a linear regression model. The data analysis shows that the state-wide average score has had a statistically significant increase after the new standards became the focus of mathematics education; additionally, there has been significant change in average scores for most of the examined subpopulations.

Approved:

Chairman, Thesis Committee	Date
Member, Thesis Committee	Date
Member, Thesis Committee	Date
Department Head/Direct Supervisor	Date
Dean, Academic College	Date
Dean, Graduate School	Date

Table of Contents

I. Introduction

In April 1983, President Ronald Reagan delivered a report, entitled *A Nation at Risk: The Imperative for Education Reform,* spurring a new approach to education in the United States. Since this report was delivered, there has been a culture of educational reform that has included programs such as *No Child Left Behind* and other attempts at standards-based revision. For nearly two decades there have been multiple shifts in educational ideologies that worked to improve the quality of education in American schools (Ravich, 2020). The most recent major reform that has taken place in the United States education system is the *Common Core State Standards* (CCSS).

The CCSS was a reform effort that was contibuted to by a group of educators funded by the National Governors Association (NGA) and the Council of Chief State School Officers (CCSSO) (Bidwell, 2014). The program, proposed in 2007, is a set of national standards that was intended to allow all 50 states to educate students with the same set of educational goals, however a standardized curriculum was not included (Common Core State Standards Initiative [CCSSI], n.d.). This program was adopted by many states throughout the country; however, it was not proposed without criticism.

Many states did not adopt the CCSS, or reversed the decision to use the standards, and opted to reconfigure their own standards (EdGate Correlation Services, 2018). Texas was a notable state, due to its large population, that did not adopt the standards, choosing to revise the Texas Essential Knowledge and Skills (TEKS) instead. In April 2012, the

Texas Education Agency (TEA) adopted a new set of mathematics standards that was to be fully implemented as early as the 2014-2015 school year. This allowed the state to maintain full state level control on educational standards, while bringing them in line with standards used in a large portion of the rest of the country.

This study aimed to determine if the realignment of state standards has impacted the statewide average scores of the State of Texas Assessments of Academic Readiness (STAAR), the current state-level standardized test utilized in Texas public schools. Therefore, this study will add to the literature focusing on the effects of CCSS and similar efforts on standardized testing in the United States. The data used to aid in this determination is from the 2013-2014 and 2018-2019 school years and was obtained from the TEA. By doing so, this will allow insight into the effectiveness of the standards realignment by examining test scores from before and after the standards implementation. Furthermore, the research will examine the magnitude of the realignment effects by examining two of the most heavily impacted grade levels, focusing on test scores from sixth and seventh grade mathematics (Coppell Independent School District [CISD], 2013).

II. Background

Dating back to the early 1980s, there has been several attempts at standards-based reform in United States public education (Boyd, 2014). This was sparked by the release of the report *A Nation at Risk* by the National Committee of Excellence in Education

(Boyd, 2014). This report indicated a concern regarding the quality of education in America and the ability of the country to keep pace in educational metrics with the rest of the world (NCEE, 1983).

After the release of *A Nation at Risk,* there have been numerous attempts at educational reform in the United States. In 1994, Congress passed the Improving America's Schools Act (IASA) that required states to have a set of educational standards in place (United States Department of Education [DOE], 1995). The important consideration is the fact that there is a difference between educational standards and curriculum. (Faiella, 2018), where curriculum provides the "how" in delivering knowledge, whereas standards provide the "what" to teach. (Faiella, 2018)

Several years later, President George W. Bush's administration spearheaded another massive update in American education reform with No Child Left Behind (NCLB). This effort updated a key element of the 1994 IASA accountability measures. The accountability measures included in the IASA allowed for the monitoring of how successful American schools are at delivering the knowledge required by the standards to their students (DOE, 1995). The new measures tied school and district funding to benchmarks regarding "adequate yearly progress," a measure to monitor how students were performing on state and national level standardized testing (DOE, 2006). Ultimately, these updated accountability measures led to the implementation of a "high-stakes" testing culture.

This new culture of testing exposed another problem in American education. Under IASA and NCLB states were free to determine individual sets of standards from state to state (Redfield & Sheinker, 2004). This caused conflict in the frame that

educational state standards had a vast amount of variation from state-to-state for supposed identical subject areas (Hunt Institute, 2008). As such, the reliability of NCLB accountability measures became less certain (Hunt Institute, 2008). With this the National Governors Association (NGA) and Council of Chief State School Officers (CCSSO) set out to close the state to state gaps in educational standards. This led to the creation of the Common Core State Standards (CCSS) a new system that would allow states to share a common set of educational standards. While not required, the Department of Education (DOE) did encourage the adoption of the CCSS, with initiatives like the Race to the Top fund, but did not require adoption of the standards (DOE, 2016). This meant that states were welcome to use their own standards or to adopt the new CCSS.

While Texas did not officially adopt the measures created by the NGA and CCSSO, shortly after the introduction of the program, the state developed and adopted its own set of new standards. The Texas Essential Knowledge and Skills (TEKS) for mathematics were revised in 2012 and set to be fully implemented by the beginning of the 2014-2015 school year (Texas Education Agency [TEA], 2012). This study will focus on the new set of standards and seek to determine if these new standards have had a significant impact on student success on state's standardized testing, the State of Texas Assessment of Academic Readiness (STAAR).

The implementation of the revised mathematics standards in Texas was not without issue. As in the past, one of the goals of this reform was to ensure that the standards of the mathematics education in Texas public schools reached the necessary levels of rigor to ensure effective education (TEA, n.d.). This would lead one to believe that, by increasing rigor, students would show improved performance on the STAAR.

The goal of this research is to statistically determine if the goal of reform was accomplished by comparing scores from 2014, the last year under the previous standards, and 2019, the most recent administration of the yearly exams. These years were chosen in order to use the most recent scores earned by students who learned under the old standard set and compare them to scores that would be earned by students who have largely only taken standardized tests under the new standards set. This will be accomplished by examining the scores of two of the most heavily impacted grade levels by the standards realignment, sixth and seventh grade (CISD, 2013). These grade levels were so heavily impacted because the sixth-grade mathematics standards were the most changed by realignment with 55 percent of the standards being new to that grade level (CISD, 2013). Seventh grade also had a significant amount of change to the grade level standards with the adoption of the new standards (TEA, 2012). Ultimately, this study wanted to examine two consecutive grade levels that were heavily impacted.

III. Literature Review

In consulting the literature for this study, there was a limited amount of research using statistical analysis methods regarding the updated TEKS. This was especially true for literature concerning statistical analysis of the standards. That is not to say that there is no literature of this variety, just a limited amount and what was found was not related to the same subject area as this study (citation, Mathis for example). This, however, does not mean that there is not ample literature to support a study into this subject area. Fortunately, the TEKS, while not being exactly the Common Core State Standards, are

closely aligned with the CCSS (Smith, n.d.). As previously stated, there is a very close relationship between the TEKS and CCSS. Given this relationship and the existence of ample literature studying CCSS, there is enough supporting literature to aid in the statistical analysis of the effectiveness of the previous standards versus the update TEKS. Additionally, there have been similar studies, which will be discussed in this section, that investigated the effectiveness of related Texas educational policies in the past that can help account for previous related research regarding this state's practices.

CCSS based non-empirical research

While there has been a fair amount of research regarding the effects of CCSS, it is important to note that some of these investigations are non-empirical in nature. However, the studies that focus on the contents rather than the effectiveness of the CCSS are relevant despite not being focused on the TEKS because there is a close association between the two sets of standards. While this type of research may not seem like the best sources to examine, this does allow a comparison of the effectiveness of the implementation of the sets of standards between the CCSS and the TEKS, both empirically and otherwise.

Non-empirical studies, such as the qualitative study conducted by Cristol & Ramsey, 2014, are incredibly important to the research because they provide a look into the effectiveness of the standards implementation. These authors investigated the implementation of CCSS at four different districts throughout the country of which a summary of their findings will be provided. Starting with the Kenton County, Kansas school district; the research found that the district fully embraced CCSS and fully developed a district wide curriculum around the CCSS. This allowed for a seemingly

rather painless transition in all aspects of the implementation, from the classroom to district level standardized testing, all the way to professional development and educator evaluations. The same success did not seem to be afforded by the Metro Nashville Public Schools in Tennessee, however. Metro Nashville was an early adopter of CCSS, ahead of the state's textbook alignment timeline. This left the district with a struggle to provide CCSS aligned lessons due to the extensive alterations that needed to be made to the district's existing curriculum (Cristol & Ramsey, 2014). Additionally, the state's educator evaluations were not CCSS aligned leaving only classroom observations at the district level to account for CCSS alignment. The investigation continues to examine two more districts, one in New York and one in Nevada, with similar mixed results to the previously mentioned districts (Cristol & Ramsey, 2014).

Additionally, a study conducted by Polikoff et al. (2011), examined how well state level standardized testing was aligned to the respective state's standards prior to the possible adoption of CCSS. These authors determined, that during the push for standards-based testing under No Child Left Behind, there was generally a massive disconnect between the standards as written and the standards as tested. In the study, it was determined that, on average, 27 percent of mathematics standardized tests being perfectly aligned with their respective state standards (Polikoff et al., 2011). The authors suggest that "coherence is the core principle underlying standards-based educational reforms. Assessments aligned with content standards are designed to guide instruction and raise achievement" (p. 965). Meanwhile, their research indicates that, prior to CCSS being heavily implemented throughout the country, there was a clear lack of coherence in pre-CCSS educational environments.

Examining sources such as these allows for a more personal look at how the CCSS are being implemented and whether this is being accomplished effectively. Additionally, it allows insight into issues that pre-CCSS standards-based assessment had under NCLB. These sources may seem to have questionable relevance, but as mentioned previously aligning standards and testing is important in standards-based reform. (Polikoff et al., 2011) Therefore, examining the effectiveness of implementation and the alignment of standardized tests is important when judging the effectiveness of the standards altogether. Looking at the standards that the TEKS are closely aligned with, and earlier standardized testing efforts, gives an important frame of reference for evaluating the TEKS themselves.

CCSS based empirical research

When focusing on empirical studies, it is important to note there have been several studies comparing standardized test results prior to the adoption of the CCSS and after. These studies provide data-based insight as to the effectiveness of the implementation of the standards. This information is very important because, as research conducted for this study has found, there has not been a similar study comparing the results for Texas specifically. There are studies that examine the differences to determine if they are statistically significant for multiple individual states and several that examine the entire United States. One primary issue with studies of this nature is that they indicate Texas as a non-adopting state, which while technically true is a bit misleading because of the similarities between the CCSS and the TEKS (Smith, n.d.).

In the examined studies focusing on the CCSS mathematics standards, researchers tended to focus on NAEP mathematics results or ACT results. For example, a study

conducted by Hamilton (2015), focused on the impact of the new standards on ACT math scores for eleventh grade high school students in a West Tennessee school district. In this study, Hamilton (2015) determined that a statistically significant difference did exist between pre-CCSS implementation and post implementation ACT scores in the examined school district. However, the author continued to point out that the difference may not be completely due to CCSS implementation. This is, in part, due to the comparisons being done using the year immediately prior to and the year immediately after CCSS implementation in the investigated district. Hamilton (2015) indicated that one possible source of the difference could be, at least in part, because of additional resources and efforts provided to and by educators in the initial year of CCSS use in the district. Regardless of if the CCSS was owed total attribution of the significant difference in scores, the author indicates that the one point gain on average ACT scores between the two ACT administrations may not be seen as a sufficient enough increase in scores to justify the cost of the implantation of CCSS (Hamilton, 2015).

Another study, that used ACT scores as a model to determine the effectiveness of CCSS, was conducted by Xu and Cepa (2015). This study examined three separate cohorts of students in Kentucky, following each cohort from eighth grade through their sitting for the ACT towards the end of their eleventh-grade year. The cohorts differed in the fact that they had differing exposure to the Kentucky Core Academic Standards (KCAS), which is the CCSS-aligned standards set used in the state. The first cohort, students who sat for the ACT in 2011, had no exposure to KCAS prior to taking the exam. The second and third cohorts had one- and two-years exposure, respectively, prior to taking the ACT. This allowed the authors to use other exams taken prior to the ACT, in

the form of state level standardized exams and the now discontinued PLAN test (administered by ACT, Inc.), to be used as a baseline for student performance on the ACT. The authors determined that each cohort made significant progress towards college and career readiness from the respective previous cohort. However, the progress did not appear to be significant enough to be fully attributable to the KCAS transition. The authors indicate that this could be due to other changes made during the students' respective education from one cohort to the next (Xu & Cepa, 2015).

Finally, Faiella (2018) examined the results of the NAEP and used the results to determine the effects of CCSS implementation. In this study, the author examined results of the NAEP from state to state, largely looking to compare if there is a significant difference in change from states that were CCSS adopter states and non-adopter states. The author only accounted for official CCSS adoption, not situations like Texas where the state created its own standards that are closely aligned with the CCSS standards. The author used data from the 2011, 2013, and 2015 administrations of the fourth-grade mathematics NAEP exam. Faiella (2018) utilized the fourth-grade exam because it is the earliest grade level in which students sit for this test. Using the mentioned administrations allowed for an adequate comparison, due to the stages in which CCSS was adopted throughout the country. In 2011, only Kentucky had implemented the CCSS and was in its first year of using the new standards when the NAEP was administered in November (CCSSI, n.d.). For 2013, twenty-three states had previously implemented CCSS, while seven states were in their first year when the test was administered in December (CCSSI, n.d.). Ultimately this study showed a negative correlation between CCSS implementation and NAEP scores for adopting states. However, the author indicates that due to

limitations of the study, he would caution against considering the study to be conclusive regarding the impact of the CCSS, but rather, encourages further investigation into the matter.

Overall, there are not many empirical studies regarding the impact on standardized test scores of the CCSS standards. Even more so, there are fewer studies that focus primarily on state-level standardized testing. In the research, only one such study was found which considered this kind of testing which was conducted by Mathis (2016). Mathis' research does examine state-level standardized testing, investigating the effects of the English CCSS on English II exams in North Carolina. Therefore, Mathis' research establishes a precedent for this type of study. This allows this research to add to existing literature by examining state-level mathematics standardized testing and investigating a non-adopter state that is still closely aligned with the CCSS. Additionally, of the empirical studies discussed in this section, all three indicated the results to be inconclusive regarding the effectiveness of the CCSS. This study aims to add to the literature by examining a broader time range between the final year before standards realignment and examining post implementation test scores, indicating that the results may be somewhat more conclusive than previous studies.

Research not based on CCSS

Due to the lack of research based strictly on the TEKS, there is not much relevant information that is based on Texas students. To help counteract this, this study will discuss the research that investigated other efforts in Texas. These efforts can include many different takes on educational reform in the state such as the implementation of the popular curriculum management system CSCOPE, the transition in standardized testing

from the Texas Assessment of Knowledge and Skills (TAKS) to the STAAR, and other changes to educational guidelines in the state.

Davis and Willson (2015) conducted a study regarding test-centric literacy instruction. Their research, while not based on mathematics instruction, is relevant to this study because it considered the effects of the instructional practice known as "teaching to the test." The researchers focused on the effects of the transition from the TAKS to the STAAR, especially how the transition made the practice of test-centric instruction more prevalent than before. Davis and Willson (2015) determined that test-centric instructional methods were long used before the introduction of the STAAR, however the transition highlighted their existence in a way that had never been accomplished previously. Ultimately, participating educators who used these practices realized that these were not the most efficient methods of literacy instruction but used them because they allow students to flourish on standardized testing. Meanwhile, even though these practices were used before the testing transition and were continued through the process, the transition did cause disruption and uncertainty regarding these methods and other parts of the education process (Davis & Willson, 2015). These elementary and middle school educators were uncertain of what to expect the test to look like, there was confusion regarding the rigor level of in-class instruction, and finally the transition left teachers wondering how it would affect their "day-to-day instructional decision making" (Davis & Willson, 2015, p. 358). This research highlights the confusion and disruption of teaching practices that occured from educational reform. As mentioned by teachers interviewed by the authors, when the standardized testing model is changed, a degree of uncertainty impacts their instructional decisions (Davis & Willson, 2015). Ultimately it effects how

the teachers plan their lessons and leads to questions of if the right content is being taught (Davis & Willson, 2015).

One of the many educational policy reforms that has happened in Texas over the past fifteen years is CSCOPE (2006). CSCOPE was a popular, albeit controversial (Merritt, 2011), curriculum management system used by educators, making it a similar idea to CCSS even though CCSS was only a set of standards. This system provided educators who used the system with lesson plans, a scope and sequence document related to the TEKS, as well as many other resources that allowed teachers to not only provide education to their students, but to share a common structure with other teachers throughout the state (Gulick, 2010). In a similar study to the one being conducted, Merritt (2011) focused on Texas state-level mathematics standardized testing. In his study, Merritt sought to determine if districts that took advantage of the offerings of CSCOPE showed to have any statistical difference on mathematics TAKS scores from those that did not use the system. In doing so, Merritt examined archival TAKS data from the 2007-2008, 2008-2009, and 2009-2010 school years for grades three through eight. He then proceeded to compare mean passing percentages of districts using the system versus those that did not use it. To aid in focusing on the study, Merritt used data specifically from mid-sized (at the time, AAA or 430-989 students) districts throughout the state breaking down overall results, as well as results of selected subpopulations of interest. The research conducted by Merritt goes on to determine that there was a positive significant difference overall for schools that implemented CSCOPE. However, when the author further investigated the analysis, it began to show that was not true for every grade level. When comparing results for third grade students, schools that did not use the

system performed significantly higher. Inversely, seventh and eight grade students performed significantly higher in schools that did implement CSCOPE. Meanwhile, the data indicated that there was no significant difference for students in fourth through sixth grade. In conclusion, this research indicates a positive difference in favor of the use of CSCOPE while admitting there are limitations to the results.

The goal of the research for the present study is to add to the literature regarding educational reform in the state of Texas. While there is a good amount of research regarding educational strategy and testing policy in Texas, there seems to be a lack of statistically backed research regarding the effectiveness of the 2012 revision of the TEKS. Therefore, this study will look to contribute to existing literature by providing, if not the first, one of very few statistically driven examinations into this revision and whether it has been effective.

IV. Research Questions

The purpose of this research is to identify and understand the changes, if any, to student achievement on standardized testing in Texas due to the TEKS realighment that went into effect for the 2014-2015 school year. Therefore, this research seeks to answer the following questions:

Research Question 1

Is there a statistically significant difference in mean scale score between the STAAR administrations before and after TEKS realignment went into effect across all studied demographics?

Research Question 2

Is there a statistically significant difference in mean scale score between the STAAR administrations before and after TEKS realignment went into effect based on student gender?

Research Question 3

Is there a statistically significant difference in mean scale score between the STAAR administrations before and after TEKS realignment went into effect between students of each ethnicity?

Research Question 4

Is there a statistically significant difference in mean scale score between the STAAR administrations before and after TEKS realignment went into effect between students based on Economically Disadvantaged status?

V. Methodology

This study was conducted utilizing quantitative methodology, examining archival data obtained from the Texas Education Agency. This data encompasses the 2013-2014 and 2018-2019 mathematics STAAR administrations for sixth and seventh grade students. This allows for a statistical analysis of standardized test scores, which is a

common measure of student achievement that is easily made publicly available by the TEA. These administrations were chosen for several reasons, including that it allows comparison of the final STAAR administration before the TEKS realignment and the most recent administration at the time of writing. (Note: the 2019-2020 administration was cancelled because of the COVID-19 outbreak). Additionally, these grade levels were chosen because the sixth-grade mathematics standards were the most heavily impacted by realignment with 55 percent of the standards being new to that grade level (CISD, 2013). This allowed the examination of consecutive grade levels that would be some of the most heavily affected, if not the most, by the standards being changed due to the amount of change in the sixth grade standards.

There were many possible choices, in this study, for the dependent variable considered for each grade level and administration. According to the TEA, the raw score for each test is the number of correct answers; while the scale score considers the agency's perceived difficulty of the test. The reasoning behind this, as explained by the TEA, is because a 7 out of 10 score on a calculus test is vastly different from the same score on a basic multiplication test (TEA, 2018). Given the point that test difficulty should be taken into consideration, it seems most appropriate for the research to focus on the scale score. The TEA provides a technical digest that describes the process of determining how the STAAR test is created each year. As part of this publication, the process in which the scale scores are determined from year to year is outlined.

The TEA technical digest shows that the Rasch Partial Credit Model (RPCM) is the model used to develop the standard scale score utilized as the standard for the STAAR each year. The RPCM is uni-dimensional in nature and allows for responses

recorded in two or more ordered categories to be analyzed. There have been successful applications of this model to a wide variety of measurement problems in statewide testing. The RPCM allows person and item parameters to be separated, as well as sufficient statistics which leads to conjoint additivity. This feature enables objective comparisons of persons and items; therefore, each set of parameters are conditioned out of the estimation procedure for the other. The RPCM is the simplest of all item response models for ordered categories containing only two sets of parameters according to Masters and Wright (1997): one for persons and one for items. All parameters in the RPCM are locations on a variable, distinguishing it from other models that include item discrimination or dispersion parameters. The RPCM is specifically useful when measuring item or assessment criterion performances in two or more ordered categories with the intention to combine results across items or criteria to obtain measures on some underlying variable, such as STAAR scale scores over multiple years and administrations. Successful applications of the RPCM have been reported in a wide variety of areas of interest including measures of critical thinking, the diagnosis of mathematics errors, and statewide testing programs, as well as many other uses.

Field testing questions to use on future versions of the STAAR is the first step in determining the scale scores (TEA, 2014; 2018). After a question has been field tested, it is evaluated using the statistical model known as the Rasch Partial-Credit Model (RPCM) to scale and equate the difficulty of the item (TEA, 2014; 2018). This is the first step in a multi-faceted equating process used to create the scale scores used to determine Texas students' content mastery.

When used in this process, the RPCM follows the equation or a similarly expressed equation with the same end goal (TEA, 2014;2018). The TEA states that the RPCM maintains a one-to-one relationship (TEA, 2014; 2018). This means that each scale score is uniquely associated with a raw score (TEA, 2014; 2018). This model was chosen by the TEA, because it is flexible in the way it accommodates multiple-choice data and multiple response category data (TEA, 2014; 2018). Once the RPCM is applied to each test item they are placed into an item bank to possibly be used on future assessments.

Once a test form is developed it undergoes pre-equating, the next step in the equating process, in order to place the form on the Rasch scale so that a table establishing a link between raw scores and scale scores may be produced (TEA, 2014;2018). The difficulty of a test form can be estimated because the difficulty of each item was established in advance using the RPCM (TEA, 2014;2018). The final step in completing the equating process is known as post-equating. After administration, a test form will have the post-equating constant applied ($t_{a,b} = \frac{\sum_{i=1}^{k}\left(d_{i,a} - d_{i,b}\right)}{k}$), which is done in order to transform the Rasch difficulty to reflect the current test item (TEA, 2014;2018). This is necessary because the difficulty of the item may not be the same as newer instruction practices are put in place or the presentation of a question is altered, such as formatting, wording, positioning, et cetera.

Ultimately, the scale score is the most appropriate comparison because once the performance standards are established, they are maintained on all subsequent test forms (TEA, 2014;2018). The STAAR 3-8 standards were first applied to the spring 2012

administrations (TEA, 2015) and the overall process for determining the scale score remains the same. Therefore, the scores from the 2014 and 2019 administrations have comparable scales, despite the difference in test length and the apparent difference upon visual inspection.

The independent variable for the study is the administration year. Using the administration year as the independent variable is key because doing so essentially allows the research to determine the effectiveness of standards realignment. For further exploration into the matter, multiple demographic control variables were established for this study. Student gender, ethnicity, and economically disadvantaged status were all control variables that allow the research to explore the effectiveness of standards realignment further than a base level understanding of state-wide impact.

When originally examining the data, each administration's data set contained more than 350,000 results. Therefore, this experiment was run using a population sample calculated using STATA statistical software. STATA has a built-in power and sample size command that allows users to determine a necessary sample size in order to obtain the desired experiment power and confidence level. In this experiment it was determined that a 95 percent confidence level, which leads to a 5 percent significance level, regarding the difference in mean STAAR scores from 2014 and 2019 would be desirable. It is important to note, because of multiple statistical tests (ten tests in all) taking place with the same data set, that applying Bonferroni's correction would be necessary. Therefore, each statistical test would have to return a p-value of less than .005 to be considered significant. This means that the minimum sample size for each administration is 2271 samples for sixth grade scores and 30827 for seventh grade scores, in order to obtain the

desired experimental confidence. It would normally be rather interesting that there would be such a disparagement in the required the minimum sample sizes, however it is easily explained in this case. For the overall population, being both administrations for each respective grade level, the standard deviation is much lower and a much larger change in average test scores for the sixth-grade exams than the corresponding seventh-grade exams.

After determining the necessary sample size, the data will be examined using STATA statistical software. The data will be entered using a coded variable strategy, using zero through four to represent the possible answers for ethnicity (White, Asian, Black, Hispanic, and "others" to include Native American, Pacific Islander, and Two or more races), zero or one to represent gender (male or female), zero or one economically disadvantaged status (no or yes), and finally zero or one to represent the standards set for the corresponding administration (old or new).

After the appropriate data has been input to STATA, the software will be used to conduct a linear regression analysis to examine the difference in scores regarding different sets of standards. This will allow the research to determine if the independent variable, which is the subject of the primary research question in this study, should be considered significant. After this test is run, to examine the effects on test scores that the standards realignment was potentially responsible for each control variable's respective subpopulations, a 99.5 percent confidence interval will be examined. This is a form of statistical Post-Hoc testing, that allows one to examine how a specific subpopulation of a control or factor variable is affected by the independent variable in a data set.

The data analysis in this research took place in several stages. Ultimately to answer the main research question, the data was examined using a coded variable strategy and the linear regression model:

$$Scale\,Score = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon .$$

Where each β_0 is the mean score constant and each following β_i is the corresponding regression coefficient for each variable. Additionally x_1 represents the standards set, x_2 indicates the variable used for student gender, x_3 is the variable used for student socioeconomic status, and x_4 is the consideration for a student's ethnicity. As mentioned, the analysis took place in several stages, due to the desire to make sure that adding in each control variable did not seem to make an improper impact on the independent variable. Therefore, at first the analysis was run only including the independent variable. The analysis continued by checking each control variable with just the independent variable, then two control variables, then finally by analyzing the full model.

Each demographical research question was answered by examining the aforementioned 99.5 percent confidence intervals. This is easiest to determine reliably by hand using the following equation:

$$99.5\%\ \text{Confidence Interval} = \bar{X} \pm 2.807 \frac{\sigma}{\sqrt{n}},$$

where $\bar{X}$ is the difference in mean scale score for each administration for the respective subpopulation, σ is the subpopulation's standard deviation for the combined scores of both administrations, $\sqrt{n}$ is the square root of the number of samples based on which confidence interval is currently being determined, and 2.807 is the critical Z-value for a

normal distribution where $\alpha/2 = 0.0025$. This allows the research to analyze how the scores were affected by the changing of the TEKS for each major subpopulation, but also to determine how different portions of other subpopulations typically perform on the tests versus the study's control group (e.g. Hispanic female students who are not economically disadvantaged versus White male students who are not economically disadvantaged).

The use of the regression model will allow for the primary goal to be accomplished by running one statistical test for each grade level, because three of the research questions listed above are related to control variables, rather than differing dependent variables. Additionally, the regression model allows for a more in depth look at some of the more nuanced questions of interest, including those that may not be addressed in this particular research.

VI. Results

As mentioned previously, the main research question was addressed by first examining the response variable and the independent variable. Doing this yielded a statistically significant difference for both sixth and seventh grade exams. The mean scale score shows a positive regression coefficient for both grade levels at this stage, with sixth grade scores showing an average improvement of approximately 15 points while seventh grade scores showed just under a seven-point average increase. The next stage involved checking the model with each control variable added separately (year and gender, year and ethnicity, and year and economically disadvantaged status). This did not change the

results, as far as the difference in scores from one set of standards to the next being indicated as statistically significant. The only difference in this stage is a change to regression coefficient regarding the standards. Next, the model was examined with the independent variable, gender, and each other control variable separately. This again did not show any change in statistical significance from the model that only considered the independent variable and any one of the control variables. Then finally the full regression model was analyzed. Again, each grade level showed significance indicating a positive change in mean scale score from one set of standards to the next. In the end, when examining the full model, the mean scale score for sixth grade students showed an increase of 19.716 points, while the average increase for seventh grade scores was 7.42 points. Both increases were indicated to be significant due to both results having a p-value of 0.000, while significance was indicated by a p-value of less than .005 due to examining the data with $\alpha = 0.05$ and Bonferroni's correction that was necessary because of ten overall models being examined. Therefore, the data indicates the rejection of the null hypothesis for the primary research question. This would seem to indicate that the TEKS realignment has had the desired effect for sixth and seventh grade students.

There could be an argument against the validity of using the scale score as the response variable in this study. Simply put, the scales from both years do not match point for point and the length of the tests are not the same. For example, the 2014 sixth grade administration had 52 questions and a scale score range (that follows a normal probability curve) of 949 to 2138. Meanwhile, the 2019 sixth grade administration only had 38 questions with a normally distributed scale score range of 1038 to 2186. However, this potential issue can be cleared up thanks to the previously discussed use of the RPCM and

information regarding the differing performing categories that students can be placed in. For the 2014 administration of the STAAR, there were three different sets of performance standards published each with three categories. The sets of standards were meant to allow a phase-in period while students acclimated to the differences between STAAR and its predecessor, the TAKS, before the final recommended standards were put into place. Each set of standards contained three simple categories: *unsatisfactory performance*, *satisfactory performance*, and *advanced performance*. The 2019 performance categories consisted of *did not meet grade-level expectations*, *approaches grade-level expectations*, meets grade-level expectations, and *masters grade-level expectations*. These reporting categories become important because the scale score corresponding to, *meets grade-level* or *satisfactory* are at similar places on the normal curve used for the scale score, while the score required for *advanced* or *masters* is also similarly or identically placed. (Note: The *approaches* category does not match any of the performance categories for any of the sets of performance standards from 2014). From there, it can easily be determined if the trends in performance categories of the entire population, not just the tested sample, match the indications of the analysis of the sample data. When examining the trends in the performance categories for sixth grade, it was determined that sixth grade students saw a reduction in the number of students who had unsatisfactory performances as 56.661 percent did not meet expectations in 2019, down from 61.927 percent in 2014. However, there was an opposite trend in both the percentage of students who met or mastered expectations in 2019. In 2014, 21.041 percent of sixth grade students met expectations, while 17.033 percent achieved masters expectations. Meanwhile, 24.306 percent of sixth grade students met expectations and 19.038 mastered expectations in 2019. Additionally,

there were similar trends for seventh grade as 2014 saw 71.6 percent of students not meeting expectations with 18.254 percent and 10.147 percent achieving satisfactory or advanced performances, respectively. Meanwhile, in 2019 the percentages for each performance category was 65.286 percent, 21.524 percent, and 13.197 percent respectively. Between the trends in the score categories being positive matching the increases in mean scale scores and the application of the RPCM to establish the scales, the validity of using the scale score as the response variable should be much less questionable.

Now that the primary research question has been answered, with justification for using the chosen response variable provided, the focus will shift to the subsequent questions. These questions will be answered in presentation order starting with gender, then focusing on ethnicity, and finishing up with economically disadvantaged status. In order to examine these, a confidence interval was figured for each major subpopulation such as sixth-grade Hispanic students. For the given example, the confidence interval can be calculated using the following equation:

$$99.5\%\,\text{Confidence Interval} = 17.4062 \pm 2.807\frac{151.6445}{\sqrt{2341}},$$

which will allow the research to determine how sixth-grade Hispanic students did in 2019 compared to their 2014 counterparts. A similar model will be used to allow a comparison of test scores for each subpopulation of each demographic group, continuing to examine sixth grade and seventh grade results separately.

When examining the full model, girls outscored boys by an average of 8.865 points in sixth grade and 2.283 points for seventh grade combined across both test administrations. These, however, were not considered to be significant differences as the

p-values for these particular statistics were 0.072 and 0.221 respectively, much higher than the p-value considered for significance of .0125. This leads to a question of whether either gender showed to have a significant difference in their scores between the 2014 and 2019 administrations. To examine this, the necessary confidence interval was calculated, starting with sixth grade female students using:

$$99.5\%\,\text{Confidence Interval} = 18.2964 \pm 2.807\frac{168.6664}{\sqrt{2228}}.$$

For sixth grade female students, the data showed an average increase of 18.2964 points in the mean scale score for the 2019 administration. This leads to obtaining a 99.5 percent confidence interval of approximately 8.2661 to 28.3267-point average increase, when rounding to the fourth decimal place. Because zero is not included in this confidence interval, this increase should be deemed statistically significant. Meanwhile, seventh-grade female students showed a mean scale score improvement of 8.8217 points, also being deemed significant with a confidence interval of 5.0652 to 12.5782-point average increase. Male students also were shown to have significance in the average increase of mean scale scores. Sixth grade males increased by 12.7475 points on average, resulting in a confidence interval ranging from 2.619 to 22.876. Meanwhile, seventh grade males had an average increase of 4.8448 points, with a confidence interval of 1.186 to 8.5038. With this information the second research question would lead to the null hypothesis being rejected because all four subpopulations were deemed significant.

When examining the full model, across both administrations, there was a good amount of variation between how different ethnicities performed on the STAAR. For both grade levels, white students were used as the control group, therefore the average

score differences presented will be for each ethnicity as compared to the aforementioned white students.

For sixth grade students, taking both administrations into account, Asian and Other ethnicity (Native American, Pacific Islander, and Two or more races) score higher on average, but not in a statistically significant manner with p-values of 0.021 and 0.176, respectively. Meanwhile, when examining the full model, Black and Hispanic students were indicated to have statistically significant differences with Black students scoring 57.284 points lower and Hispanic students scoring 35.024 points lower on average than white students. The p-value for those groups was 0.000 for both.

However, to answer the third research question, it is more prudent to consider each subpopulation's respective confidence interval when considering that ethnicity's average difference in score between administrations. Examining these confidence intervals will allow the research to determine if each ethnicity subpopulation showed a significant difference in mean scale score between test administrations. Each grade level will be examined separately, in the same order, starting with White students, then examining Asians, followed by Black students, then Hispanics, and finally all other ethnicities.

Beginning with sixth grade white students, the data shows an average increase of 17.7938 points from 2014 to 2019, this leads to obtaining a confidence interval of 3.9761 to 31.6115 indicating a statistically significant growth. For Asian students in sixth grade, there was an average decrease of 21.8658 points, however this leads to a confidence interval of -84.2739 to 40.5423. Because the preceding confidence interval does not include zero, the decrease in average scale score for sixth-grade Asian students should

not be considered statistically significant. Sixth grade black students had an average increase of 30.1385 points between the 2014 and 2019 administrations. The confidence interval for this subpopulation spanned an average increase of 13.9818 points to 46.2952 points and therefor should be considered significant.

Next, examining sixth grade Hispanic students shows an average increase of 17.4062 points, which is deemed significant for this subpopulation with a confidence interval of 8.6085 to 26.2039. Finally, the last sixth grade ethnicity group to be examined is the group composed of each other possibly represented ethnicity. This subpopulation showed an average increase of 64.2979 points, which indicates statistical significance with a confidence interval ranging from 7.7489 to 120.8469 points. Having examined each ethnic subpopulation for sixth grade, four of the five were determined to be statistically significant. Therefore, for sixth grade students, the data indicates to reject the null hypothesis for the research question regarding ethnicity.

Moving to examining the seventh-grade results by ethnic group, each subpopulation showed to perform worse than white students across both administrations. Asian students scored 17.198 points worse on average, while Black students performed worse by an average difference of 30.636 points. Hispanic students scored 13.779 points lower on average than their white counterparts, meanwhile the average difference for each other student was a 46.752 decrease. Each of these should be deemed statistically significant with a p-value of 0.000 for all four groups. For white students, the average difference between scores for the 2014 and 2019 administrations shows a decrease of 7.9446 points. This difference should be considered statistically significant as the associated confidence interval spans -13.4723 to -2.4169 points. Seventh grade Asian

students showed an average increase of 2.2312 points in the 2019 administration. However, unlike the difference for White students, the confidence interval of -18.983 to 23.4454 for this statistic indicates that this is not statistically significant. The 2019 administration of the STAAR showed an average increase of 23.7514 points for Black students. The associated confidence interval, which indicates a statistically significant increase, calculates as 17.7954 to 29.7074. Hispanic students had a confidence interval of 9.8202 to 16.09 to show that their increase of 12.9551 points on average should be considered statistically significant. Finally, each other ethnicity showed an average decrease of 2.2087 points. The related confidence interval is -24.6814 to 20.264, therefore this decrease should not be considered statistically significant. Having examined each ethnic subpopulation for seventh grade, three of the five were determined to be statistically significant. The statistically significant difference for White, Black, and Hispanic seventh grade students indicates to reject the null hypothesis.

The last research question that needs to be examined is to determine if the change in TEKS has had a statistically significant impact on STAAR scores based on economically disadvantaged status. Regarding the full model, which again includes both administrations, economically disadvantaged students scored lower in both grade levels, with these students scoring 55.475 points lower on the sixth-grade test on average, while the seventh-grade average difference was an average of 11.474 points lower. Both of these differences are to be considered statistically significant with the p-value being 0.000 for both groups. In order to determine if the change in the TEKS had an effect on the average score for both students who are or are not economically disadvantaged, the following reduced model where economically disadvantaged status was isolated will be

examined: Starting with sixth grade students who are not considered economically disadvantaged, these students showed an increased score of 17.2719 points on average. The associated confidence interval for this increase was 4.1988 to 30.345 and therefore the difference should be considered statistically significant. Meanwhile, the seventh-grade students who are not considered economically disadvantaged had an average decrease of 9.7477 points. This difference should be considered significant, with a confidence interval of -14.6876 to -4.8078. When considering students who are listed as economically disadvantaged, sixth grade students showed a 19.3295-point average increase. The associated confidence interval in this case is 11.6734 to 26.9856 and therefore the difference should be considered statistically significant. The difference for students who are considered economically disadvantaged in seventh grade also showed an increase for the 2019 STAAR administration, on average scoring 18.9195 points higher than their 2014 counterparts. This difference also showed to significantly different with the associated confidence interval of 16.1413 to 21.6977. With the fact that all four subpopulations showed to be statistically significant in their score difference, it is clear that the data indicated to reject the null hypothesis for the fourth and final research question.

The analysis of the data provided by the TEA shows that there is clearly some benefit to altering the TEKS to the standards that were adopted in 2012 and put into effect for the 2014-2015 school year. Per the analysis there is a statistically significant difference in at least one subpopulation of each control group. Because of this, it is clear because the data indicated it was necessary to reject each of the four null hypotheses. Most importantly the primary research question's null hypothesis was rejected, so

therefore it can be determined that statewide the average score for both sixth and seventh grade went up in a statistically significant manner from 2014 to 2019. This supports the idea that the new TEKS are making a difference for Texas students, which is further shown to be true by the fact that the percentage of students meeting grade level expectations or exceeding them is also on the rise.

VII. Discussion and Conclusion

Leading into the implementation of the updated mathematics standards, Texas had been considered to have a mediocre education system for several years, receiving a grade of C plus or lower on the American Legislative Exchange Council (ALEC) student performance and education policy grades index in four consecutive years (ALEC, 2015). Additionally, the state had seen a steady decline on NAEP rankings, falling to 18th in 2013 from 8th 2009 (ALEC, 2015). However, with the implementation of the new TEKS on a similar schedule to the similarly designed CCSS, it would be possible that Texas students would be able to close those gaps, assuming the new TEKS are effective. As shown in the previous section, it appears that the new set of standards should be considered effective but there are some details that should be further discussed, especially when considering the differences in performances based on demographic control variables.

An interesting performance difference worth noting is the one associated with gender, whereas female students showed a stronger significant growth between the two

STAAR administrations than their male counterparts. Thus, there is a possible limitation to the validity of this finding; because research has shown that boys, on average, perform better on standardized math tests, reverse from the English Language Arts exams (Reardon et al., 2018). This could possibly be because girls tend to perform better on tests with more open-ended questions while boys do better on more multiple-choice reliant assessments (Stanford Graduate School of Education, 2019). The reason a possible limitation to this study exists is because both the sixth and seventh grade STAAR tests were fourteen questions shorter in 2019 than they were in 2014, meanwhile examining both released STAAR tests reveals that the number of open-ended questions remained the same on both tests between the administrations. While this could be an indicator of why females showed a larger average increase, the fact that both male and female students showed a significant difference, leads to question whether this could be a concern regarding experimental validity.

Racial bias is considered a long running issue in many, if not all, forms of standardized testing, especially for Black and Latinx students (Rosales, 2018). The data examined for this research does not show a difference in the case of these particular students performing worse than their White counterparts. Interestingly, the data indicates that Asians and all other races perform better on the sixth-grade exams than white students, meanwhile they performed worse on the seventh-grade test. Because of the existence of these racial biases, there is a concern that the research regarding the growth based on ethnicity from one set of standards to the next could be inadequate. However, this concern is fairly easy to negate. This is due to the fact that for both grade levels, the null hypothesis was rejected; because for both grade levels, Hispanic and Black students

showed a statistically significant increase between the 2014 and 2019 administrations. In summary, in both grade levels, both of the ethnic groups most marginalized by standardized testing (Rosales, 2018) showed significant growth after the realigned standards were put into effect, therefore that should largely clear up concerns of racial bias being a limiting factor on this research.

The last of the major concerns regarding possible limitations of this research would be regarding how historically low socio-economic status (SES) students typically perform worse on standardized testing. As with minority ethnic groups, there is a history of low SES students performing worse on standardized testing than their counterparts (Dixon-Roman et al., 2013). However, in both grade levels, students who are deemed Economically Disadvantaged, or more appropriately low SES, demonstrated a statistically significant growth between the two STAAR administrations. Meanwhile, students not deemed low SES only showed a significant improvement in sixth grade, yet had a statistically significant decrease in seventh grade. So, given that economically disadvantaged students had a better track record for growth between the 2014 and 2019 STAAR administrations, this possible limitation should be much less concerning than it may normally would be.

All of these findings lead to many different questions of interest in educational research. For example, with the history of males performing better on mathematics exams then female students, why is the growth stronger for females? Is it because of the fact that girls perform better on exams with more open-ended question, as presented by Stanford University, that they performed better on the 2019 exam which had a higher percentage of the questions being open-ended in nature (Stanford Graduate School of Education,

2019)? Is something about the standards realignment aiding female students to excel at a quicker rate than male students? This is a point of interest where further research could be incredibly beneficial for educators. The same idea exists for why minority ethnic groups saw the growth presented in this research. With the history of racial bias in standardized testing, could that issue be starting to decline? Could the new standards make mathematics education seem less daunting for minorities? Research examining why these ethnicity based performance gaps are closing could lead to another major improvement for educators, especially if they were aided by the standards realignment.

Additionally, it would be interesting to explore further into why the gaps closed as strongly as they did between students of different Socio-Economic Status groups. This was possibly the most important difference when examining subpopulations, as low SES students saw a significant increase in both grade levels while seventh-grade students not considered Economically Disadvantaged results indicated a significant decrease. There could be many reasons for this, however if this change is because (or at least partially due to) the standards realignment it could be a point of interest for educational researchers. While seventh-grade students who are not considered Economically Disadvantaged having a significant decrease in test scores is a cause for concern, the fact that there was closure in the performance gap could be considered a success. Therefore, it would be interesting to see if this change could be linked to the standards realignment or perhaps to some other influence. If a connection could be determined, that could allow researchers to determine how to continue closing the gap without negatively impacting low SES students.

Overall, there is no guarantee that the adoption of the realigned math TEKS is the sole reason behind the STAAR score growth shown in this research or even attribute to this growth. The growth could be because of differing teaching practices in the classroom, shorter test forms in 2019, or any number of other changes to classroom procedure in Texas mathematics classrooms. However, there is also no guarantee that the new standards are not a contributing or the primary factor in these changes. The data, without question, indicates a significant growth in the average mean scale score from 2014 to 2019. The data also shows that there has been an increase in students meeting or exceeding the prescribed grade level standards as set by the TEA, along with a corresponding decrease in students not meeting these goals. Therefore, the results indicate that standards realignment has been effective should be considered reliable.

The results of this research are somewhat contradictory to previous research regarding the related CCSS (Hamilton, 2015; Faiella, 2018). Where the research conducted by Hamilton (2015) and Faiella (2018) ultimately showed inconclusive results, the results of this research indicate a positive correlation between Texas standards realignment and student achievement. This not only has implications for Texas educators, but possibly educators throughout the country. If Texas realigned their standards to be close to that of the CCSS and has shown a positive growth, while research regarding CCSS has proven inconclusive, perhaps it would be beneficial for educators, and students alike, in Texas if the state continued to use these standards. Additionally, if Texas standards are closely aligned with CCSS and are proving more effective, it would seem to imply that maybe other states could stand to improve performance by taking the TEKS into consideration when examinging their own standards. However, since there is a

record of some student groups showing decreased performance in Texas, it may be considered beneficial both in the state, and throughout the country, if the TEA were to continue working with and improving upon the current standards to prevent slightly mixed results in the future. Ultimately, if student performance can continue to show significant increases with further use of the current or similar standards, perhaps the United States as a whole can eventually begin to close the gap in mathematics performance with the rest of the world.

REFERENCES

Chingos, M., Blagg, K., Recht, H., Baird, C., Tilsley, A., & Forney, E. (2016, June 22). America's Gradebook. https://apps.urban.org/features/naep/.

Coppell Independent School District *Mathematics / New 2014-15 TEKS*. / New 2014-15 TEKS. (2013). https://www.coppellisd.com/page/9235.

Cristol, K., & Ramsey, B. S. (2014, February). Common Core in the Districts: An Early Look at Early Implementers. Thomas B. Fordham Institute.

Davis, D. S., & Willson, A. (2015). Practices and Commitments of Test-centric Literacy Instruction: Lessons From a Testing Transition. *Reading Research Quarterly*, *50*(3), 357–379. https://doi.org/10.1002/rrq.103

Dixon-Roman, E. J., Everson, H., & Mcardle, J. J. (2013, May). *(PDF) Race, Poverty and SAT Scores: Modeling the Influences of Family Income on Black and White High School Students' SAT Performance*. ResearchGate. https://www.researchgate.net/publication/280232788_Race_Poverty_and_SAT_Scores_Modeling_the_Influences_of_Family_Income_on_Black_and_White_High_School_Students'_SAT_Performance.

Faiella, C. F. A. (2018). *The Relationship Between the Common Core State Standards and Math Achievement at the State Level* (thesis). ProQuest.

Gulick, J. (2010, August 15). New curriculum system CSCOPE to bring big changes to schools in Lubbock, across state. *Lubbock Avalanche-Journal*.

Hamilton, K. D. (2015). *The impact of Common Core State Standards Initiative on math Act scores of west Tennessee high school students* (dissertation).

Hunt, J. B., & Rizzo, J. A. (2008, October). *The Hunt Institute's Blueprint for Education Leadership*. http://www.hunt-institute.org/wp-content/uploads/2015/04/Blueprint_Number_2.pdf.

Ladner, D. M. (2015, November 10). *Report Card on American Education: State Education Rankings*. American Legislative Exchange Council. https://www.alec.org/publication/report-card-on-american-education-20thedition/.

Linden, W. J. van der., Hambleton, R. K., Masters, G. N., & Wright, B. D. (1997). The Partial Credit Model. In *Handbook of modern item response theory* (pp. 101–121). essay, Springer.

Mathis, C. G. (2016). *English Common Core State Standards and Student Academic Achievement: A Time-Series Quasi-Experimental Study* (dissertation).

Merritt, B. R. (2011). *CSCOPE's effect on Texas' state mandated standardized test scores in mathematics* (dissertation).

The National Commission on Excellence in Education. (1983). *A nation at risk: the imperative for educational reform*.

Ravitch, D. (2020, February 1). *The Education Reform Movement Has Failed America*. Time. https://time.com/5775795/education-reform-failed-america/.

Polikoff, M. S., Porter, A. C., & Smithson, J. (2011). How Well Aligned Are State Assessments of Student Achievement With State Content Standards? *American Educational Research Journal*, *48*(4), 965–995. https://doi.org/10.3102/0002831211410684

Reardon, S. F., Kalogrides, D., Fahle, E. M., Podolsky, A., & Zárate, R. C. (2018). The Relationship Between Test Item Format and Gender Achievement Gaps on Math and ELA Tests in Fourth and Eighth Grades. *Educational Researcher*, *47*(5), 284–294. https://doi.org/10.3102/0013189x18762105

Redfield, D., & Sheinker, J. (2004, September). Issues Paper. The Council of Chief State School Officers.

Riley, R. W. (1994). *The Improving America's Schools Act of 1994*. Archived: The Improving America's Schools Act of 1994. https://www2.ed.gov/offices/OESE/archives/legislation/ESEA/brochure/iasa-bro.html.

Rosales, J. (2018). *The Racist Beginnings of Standardized Testing*. NEA. http://www.nea.org/home/73288.htm.

Smith, J. (n.d.). TEKS Alignment to Common Core. Smith Curriculum and Consulting.

Staff, T. W. (2014, May 31). *The Common Core backlash*. The Week - All you need to know about everything that matters. https://theweek.com/articles/446535/common-core-backlash.

Stanford Graduate School of Education. (2019, July 18). *Question format may impact how boys and girls score on standardized tests, Stanford study finds*. Stanford Graduate School of Education. https://ed.stanford.edu/news/question-format-may-impact-how-boys-and-girls-score-standardized-tests-stanford-study-finds.

Texas Education Agency. (n.d.). *TEKS Review and Revision*. Texas Education Agency. https://tea.texas.gov/academics/curriculum-standards/teks-review/teks-review-and-revision.

Texas Education Agency. (2012). *Mathematics Texas Essential Knowledge and Skills*. Texas Education Agency. https://tea.texas.gov/academics/curriculum-standards/teks-review/mathematics-texas-essential-knowledge-and-skills.

Texas Education Agency. (2015). *Technical Digest 2014-2015*. Texas Education Agency. https://tea.texas.gov/student-assessment/testing/student-assessment-overview/technical-digest-2014-2015.

Texas Education Agency. (2015, January). State of Texas Assessments of Academic Readiness (STAAR®) Standard Setting Questions and Answers.

Texas Education Agency. (2018, April). Texas Assessment Program Frequently Asked Questions.

Texas Education Agency. (2019). *Technical Digest 2018-2019*. Texas Education Agency. https://tea.texas.gov/student-assessment/testing/student-assessment-overview/technical-digest-2018-2019.

US Department of Education (ED). (2006, May 3). *The Facts About...Making Gains Every Year*. https://www2.ed.gov/nclb/accountability/ayp/yearly.html.

US Department of Education (ED). (2016, July 19). Race to the Top Fund. https://www2.ed.gov/programs/racetothetop/index.html.

US Department of Education (ED). (2017, May 25). *Federal Role in Education*. Home. https://www2.ed.gov/about/overview/fed/role.html.

Xu, Z., & Cepa, K. (2015, March). Getting College and Career Ready during State Transition toward the Common Core State Standards. National Center for Analysis of Longitudinal Data in Education Research.