HIGH SCHOOL EDITION
Journal of Student Research

Variation in Student Achievement: Effects of NCLB on Standards-Based Accountability in Arizona

Avinash Thakur[1] and Nicole Yates[#]

[1]Avon Grove High School, West Grove, PA, USA
[#]Advisor

ABSTRACT

Over the years, the United States government has assumed a larger role in educational policy. The No Child Left Behind Act (NCLB) is one of the most notable examples of an increase in federal involvement in education. Through the creation of accountability systems, NCLB required states to test in reading and mathematics. Prior research has maintained that states with pre-existing test-based accountability remained relatively unchanged by the passage of the Act. States without test-based accountability measures pre-NCLB may have been more likely to be affected by its implementation (e.g., Arizona). In order to test this claim, I examined student achievement in Arizona pre-, during, and post-NCLB under AZ LEARNS and the A-F Accountability System over certain years (1996, 2000, 2005, 2009, and 2015). Through the use of data from the Nation's Report Card and the Arizona Department of Education, I sought to answer two key questions: (1) How do Arizona's achievement scores pre-, during, and post-NCLB compare to other accountability types? and (2) In what way do Arizona school ratings vary during and after NCLB? By conducting a series of independent sample *t*-tests, I compared Arizona's state scale scores to states with moderate accountability, and I compared school ratings under the accountability systems. These analyses reject the null hypothesis that there is no significant difference associated with school ratings under the two accountability systems, therefore supporting the prior stated claim. The results offer insight into Arizona's accountability and are valuable for future evidence-based policymaking.

Introduction

The No Child Left Behind Act (NCLB), signed into law by President Bush in 2002, largely expanded the federal role in public school education. The Act placed measures on states to institute their own standards and administer testing in math and reading from third through eighth grade. At the time of its passing, there was a growing worry that the United States was falling behind in academic performance and needed to better compete within the global economy (Wilgus, 2019). NCLB aimed to reach full proficiency across all students by the year 2014. According to the U.S. Department of Education, a central part of the law was its focus on reducing long-existing achievement gaps and measuring improvement by disaggregating student performance data (*USDOE*, 2002).

Although well-intended in its purpose, this bipartisan piece of legislation has faced increased criticism over its negative effects, including the encouragement of states to "lower their standards" (Ryan, 2004, p. 11). Schools were expected to show progress each year in achievement, but if a district failed to meet performance goals, they faced the possibility of losing federal funding among other sanctions. In a report prepared for the U.S. Department of Education, the results of making Adequate Yearly Progress were summarized: "Stable national rates of making AYP from 2003–04 to 2005–06 mask the fact that some states' rates of making AYP rose substantially while other states' rates fell substantially" (Taylor et al., 2010, p. 49).

The report also drew a connection between state accountability measures and AYP outcomes, suggesting that results are heavily dependent upon a state's accountability system. From 2003-04 to 2005-06, Arizona had a 16% drop in the number of schools making AYP (Taylor et al., 2010). Adequate Yearly Progress (AYP) refers to the improvement that an individual school or district is expected to reach each year, and these varied results indicate a connection between state and school accountability. The state is held accountable for the performance of its students, and this idea has been reflected in the law. For example, NCLB required "highly qualified" status for teachers, which included having a bachelor's degree and meeting criteria set by the state (Birman et al., 2007).

In Arizona, Adequate Yearly Progress was calculated through three main measures: proficiency, number of students assessed, and additional factors. Firstly, a school must meet annual measurable objectives (AMOs) set by the state, which differ by grade and subject. AMOs report the growth in the annual amount of students passing AIMS (Arizona's Instrument to Measure Standards), Arizona's standardized test. The school had to test 95 percent of enrolled students and in addition to this, there were graduation requirements for high schools and attendance requirements for elementary schools. In Arizona, all subgroups were treated the same for AYP calculations and, as a result, schools with a larger population of minority students or students from lower socioeconomic backgrounds may struggle to reach AYP targets.

In this paper, I examine Arizona's student achievement pre-, during, and post-NCLB under AZ LEARNS and the A-F Accountability System. AZ LEARNS was created to satisfy Proposition 301, a state statute passed by Arizona voters in November 2001. Under this accountability system, each K-8 school was given an Arizona LEARNS Achievement Profile that was based on factors such as AIMS scores. Schools were rated and placed into one of the following categories: failure to meet standards, underperforming, performing, performing plus, highly performing, and excelling. In 2010, The A-F Accountability System was put into effect through the passage of the A-F Letter Grade System by the state legislature. The main difference between the A-F Letter Grade System and AZ LEARNS is that A-F places a larger emphasis on the academic improvement of students each year. The A-F Accountability System has two components: a growth score (based on student improvement) and a composite score (based on the percent of students passing AIMS and other factors), which both constitute 50% of the calculation. Schools are given a rating from A through F, with "C" marking an average performance level and "D" marking a below-average performance (Arizona Department of Education, 2012).

Context

Prior research has found that states with test-based accountability measures already in place before NCLB remained relatively unchanged by the passage of the Act (Dee & Jacob, 2011). I hypothesized that states without test-based accountability measures pre-NCLB were more likely to have been affected by its implementation. Thus, I wanted to examine NCLB's effect on one such state and investigate whether this claim holds true. Dee and Jacob (2011) contains a table that lists all the states that had consequential accountability before NCLB, but Arizona is not one of them. Furthermore, according to Husband and Hunt (2015), NCLB had a focus on reducing the Black-white achievement gap. However, in Arizona, the Black-white achievement gap is one of the worst in the country (Roberts, n.d.). In fact, African Americans scored lower on all the sections of the 2008 AIMS test when compared to White students, with only 56% passing the math section (Morel-Seytoux, 2009). Arizona also ranks close to the bottom of the nation in terms of spending per pupil for K-12 education (Jimenez-Castellanos & Martinez, 2014). All of this is important to note since Arizona's achievement and educational outcomes may be linked to its educational system as a whole. For these reasons, I chose to focus the forthcoming analysis on Arizona.

Study and Research Questions

In this paper, I situated Arizona in the context of NCLB, seeking to examine the effects of standards-based accountability within the state. This study looks at the relationship between student achievement (in eighth-grade mathematics) and accountability over the years 1996, 2000, 2005, 2009, and 2015. For the purpose of this study, 1996 and 2000 were considered pre-NCLB, 2005 and 2009 during NCLB, and 2015 post-NCLB. I compared the ratings of schools in Maricopa County under AZ LEARNS and the A-F Accountability System, as well as Arizona test scores, in order to identify any differences over time. I predicted that noticeable changes would be seen in Arizona across this time period. Ultimately, I aimed to answer the following research questions: (1) How do Arizona's achievement scores pre-, during, and post-NCLB compare to other accountability types? and (2) In what way do school ratings vary during and after NCLB?

Controversy Regarding NCLB Policy

The purpose of this paper is not to advocate for or against NCLB; research has already largely addressed this issue. Instead, I hope to identify trends associated with Arizona over an accountability timeline stretching from before to after NCLB's institution. The No Child Left Behind Act (NCLB), as it was first conceived, seemed reasonable for the betterment of educational quality across the country. Those in favor of the law point out that it expands the involvement of parents in their children's education (Peterson & West, 2003), while critics assert that the high-stakes testing created through NCLB is detrimental to school culture (Tingey, 2009).

Literature Review

In this section, I look at the existing literature regarding NCLB as it pertains to the relevance of this paper. I review the prior literature in several key ways before drawing the significance of my research.

NCLB and Standards-Based Accountability

Prior research considering standards-based accountability (SBA) is abundant due to this being a major part of the law. Many sources detailed experiences of individual stakeholders as a result of accountability (Hamilton et al., 2007, Wronowski & Urick, 2019, Olivant, 2015, Winstead, 2011). This included looking at the impact on educators to teach a subject or the impact on teachers in being able to encourage a certain classroom environment (Winstead, 2011, Olivant, 2015). Hamilton et al. (2007) defined an SBA system as "standards, assessments, and consequences" (p. xvii). This simplistic but wholesome definition encapsulates the main components of an accountability system.

This is evident by looking at the provisions under NCLB. First off, states were required to test in reading and math at least once in high school and from grades 3-8. They also had to set annual measurable objectives (AMOs) and establish AYP standards in line to reach 100% proficiency by 2014. In an effort to drive improvement, schools that continually failed to make AYP faced consequences. In the situation where a school failed to meet AYP for two consecutive years, students were given the opportunity to transfer to a different school at the district's expense. Schools that did not make AYP goals for three years in a row had to provide Supplemental Educational Services (SES), such as free tutoring or remedial programs. For four years of failure, schools faced corrective actions such as having to replace the staff or update the curriculum. Finally, a school that missed AYP targets for a fifth consecutive year had to change its control, which included undergoing a state takeover or turning into a charter school (Le Floch et al., 2007). Clearly, NCLB accountability was influenced by a few key principles, which can further manifest in state accountability systems.

Hamilton et al. (2007) concluded that the accountability systems created as a response to NCLB were different across a couple of states, attributing a lot of the differences to how involved the state was in SBA measures pre-NCLB.

A recurring theme in the literature is this acknowledgment that accountability looks different between states and that accountability systems were impacted by the state's involvement in such measures before NCLB. Hanushek and Raymond (2002) pointed out, "The basic premise of virtually all proposed school accountability systems is that student performance should be the key element" (p. 1). This is also, arguably, the reason why accountability is so important: to foster improvements in student achievement. However, the results of student performance due to accountability measures remain vastly different across states. Mawhinney (2013) analyzed the state takeover of 11 Baltimore City schools in Maryland and how that has been affected through the involvement of both state and local efforts. It reported, "The influences on a state's consequential accountability regime prior to NCLB were salient to the trajectories they subsequently took in implementing the legislation's mandates" (p. 2). This once again emphasizes how state accountability systems under NCLB were inextricably linked to pre-NCLB accountability. Furthermore, Hanushek and Raymond (2005) recognized, "States began experimenting with school accountability systems during the 1980s, but the decade of 1990s began the age of accountability" (p. 306). This sets the pre-NCLB years as starting particularly from 1990, suggesting that accountability existed prior to NCLB but was solidified through NCLB requirements.

Hamilton et al. (2012) detailed a few challenges regarding NCLB requirements. One concern the authors outlined is that high-stakes testing may lead to a decreased focus on the underlying standards since the tests may not accurately reflect "all of the knowledge and skills expressed in the standards" (p. 162). As a result, the authors noted that "strong sanctions" (p. 162) usually lead to undesirable outcomes by changing the practices that educators can use. This supports the idea that oftentimes sanctions place too much pressure on schools. Shin (2022) expressed that "practices such as teaching to the test and curriculum narrowing, along with the increased rates of test anxiety, resulted under NCLB" (p. 56) due to test scores being of paramount importance to making AYP. Thus, tests may be limited in the extent to which they can assess an individual student's capabilities. Additionally, Desimone (2013) suggested that before NCLB, standards-based reform was more beneficial and "less punitive" (p. 59). Evidently, past studies have identified shortcomings of standards-based accountability, but on the other hand, Spurrier et al. (2020) identified some positives. The authors summed up the goal behind this type of accountability as, "…To ensure that public schools are helping all students meet high academic standards, regardless of their backgrounds" (p.1). The article mentioned the benefits of higher expectations for all students and better transparency with school data being publicly available. Throughout the literature, there is consistent evidence that under NCLB, there was an increased emphasis on testing, with standardized testing being one of the primary indicators for student performance.

Student Achievement under NCLB

Dee and Jacob (2011) utilized interrupted time series analysis (ITS) to analyze the effects of NCLB on student achievement. One of the limitations of this design, however, is that any other event around the same time period could impact the validity of the results. Specifically, around the same time that NCLB was implemented, there was a "sex abuse scandal in Catholic schools" (p. 43). The study by Dee and Jacob concluded that there was no noticeable improvement in English for grades 4 or 8, but improvement was observed in eighth-grade mathematics, especially among lower-performing groups. In contrast, Husband and Hunt (2015) found that data from the National Assessment of Educational Progress (NAEP) detailed overall improvements for minority and grade 4 students, but there were "mixed results for middle school and high school students" (p. 223). Moreover, according to Stullich et al. (2007), although achievement gaps were getting better, there were not many recent changes. The long-term NAEP trend was that there were "significant declines in black-white and Hispanic-white achievement gaps, but recent changes in achievement gaps often were not statistically significant" (p. xxiv). The conflicting results between different studies signify how difficult it is to come to a standard conclusion. Thus, my examination of eighth-grade achievement in mathematics serves to add to this existing body of literature.

NCLB in Arizona

Research on NCLB in Arizona has tended to focus on one specific aspect or subgroup. For instance, there is a lot of research on English Language Learners (ELL), in part since Arizona has a large population of ELL students. However, a greater reason for the inclusion of ELL students in the literature is because Arizona's policies regarding language have especially impacted this demographic. Proposition 203, passed in 2000, eliminated bilingual education programs- requiring English instruction in Arizona. Students who did not know English were placed in an English immersion program to learn the language within a year's time. Wright and Choi (2006) conducted a survey of third-grade ELL teachers. The teachers were asked questions about NCLB, Arizona LEARNS, and Proposition 203. The researchers concluded that Proposition 203 did not really improve ELL students' education. Further, according to Wright and Choi (2005), the policies caused confusion across schools within the state. It was also found that in over half of the 40 schools selected for the study, according to the teachers, ELL students were not given the "testing accommodations" (p. iv) that NCLB provides. A separate study looked at the connection between performance on the Arizona English Language Learner Assessment (AZELLA) and Arizona Instrument to Measure Standards (AIMS) among a group of ELL students. The findings revealed that AZELLA is not accurate for measuring a student's ability to do well in the classroom because at "higher grade levels," it "becomes less predictive of academic achievement" (Garcia et al., 2010, p. 13). AZELLA is used to determine if ELL students are ready to leave their immersion program, so the fact that the test overpredicts performance is alarming.

Another area researched is Native American contexts. Similar to the studies of English Language Learners (ELL), much of the research on this topic has to do with how NCLB and Arizona's language policies have impacted Native American culture. For example, Combs and Nicholas (2012) detailed how although Proposition 203 concerned ELL students, Native Americans were affected by it since they wouldn't be able to as easily enroll in "indigenous language revitalization programs" (p. 106) that they may have been a part of before the law was passed. A separate article discussed heritage language programs in California, Arizona, and Texas. With regards to Arizona, it was mentioned that Proposition 203 applied to public schools and not schools run by any tribe, but over time, state education officials became increasingly strict about reservations following the requirements (Wright, 2007). This is important because it shows a change over time in American Indian education due to Arizona policy. Lastly, one research paper has examined American Indian achievement in Arizona before and after NCLB from 2000-2006 (Garcia, 2008). Using Ordinary Least Squares (OLS) regression, Garcia calculated the percentage point change each year in students scoring at or above proficient on Arizona's standardized test. The results indicated an overall improvement in American Indian achievement. However, if a "2005 test score spike" (p. 149) is excluded, then the achievement is drastically lower. This explains a limitation of simply relying on state achievement scores since if outliers aren't accounted for, the achievement levels may purport to be much higher than they actually are or even perfectly in line with AYP goals.

Research has also been done on accountability in Arizona. McKinney (2008) performed a series of tests to analyze the impact Arizona LEARNS and NCLB have had on middle school principals. The idea behind the study was that school leadership is equally important to standards-based accountability since varying behavior can contribute to different school outcomes. The results show that accountability measures have made it harder for principals to best allocate the limited time in a school day. Furthermore, Standerfer (2004) discussed changes in "policy and practice" (p. III) that have been observed in Arizona schools as a result of

Arizona LEARNS and NCLB. It found that changes occurred in areas such as “Assessment Practices” and “Scheduling/Grouping Practices” (p. III). The fact that these studies are together analyzing Arizona LEARNS and NCLB highlights the connection between state accountability and NCLB mandates. Finally, in a study investigating how size of enrollment and type of school (district or charter) affect accountability, it was found that Arizona charter schools were more likely “to both close and improve at higher rates relative to district schools” (Milliman, 2016, p. 85). The authors were cautious to point out that the strength of these results is undermined when specifically looking at instances with different “student demographics” (p. 73). The non-inclusion of schools with more equal racial/ethnic composition is a drawback to their study. Throughout the literature, there is evidence supporting the idea that under NCLB and Arizona LEARNS, school administrators have had to change their practices in one or more ways.

Significance

My research adds to the existing literature in a few key ways. First, it looks at longer-term trends by examining a time period that focuses on before, during, and after NCLB was passed. Most studies have either observed longer-term trends in the nation’s NAEP scores or shorter-term trends in state standardized test scores, but I am observing longer-term trends in state-level NAEP scores. The problem with relying on state standardized test data is that the scores are subject to be influenced by any policy changes, but a national sample of data is more reliable and can still be used to compare achievement across states if state NAEP scores are incorporated. It is also much more accurate to use state-level NAEP scores when making comparisons to other states, as opposed to using each state’s respective standardized test scores. This is because the standardized test scores of a given state may be affected by its accountability system. Second, I am examining school ratings during and after NCLB by comparing the two accountability systems of Arizona at different times. Due to the general nature of differences between accountability systems, research is limited on such comparisons. However, the A-F Accountability System was modeled off of AZ LEARNS, so by examining this, I am helping to fill a gap within the existing literature. In fact, an empirical literature review on NCLB was conducted and it concluded that “subsequent research might examine how states are implementing alternative accountability plans and interventions after receiving a waiver” (Husband & Hunt, 2015). Third, since there have been unclear results in the literature regarding middle school NCLB achievement, I have specifically chosen to examine achievement in eighth-grade mathematics.

Methods

To better gain insight into student achievement in Arizona, significance tests were applied to examine whether differences remain over a certain period of time.

Data Collection

For this study, I incorporated data from two key sources: the National Assessment of Educational Progress (NAEP) and the Arizona Department of Education. The National Assessment of Educational Progress (NAEP), also known as the Nation’s Report Card, publishes meaningful results on educational achievement each year. In fact, it has been used to “verify the results of statewide assessments" (*USDOE*, 2002). Its addition in several other papers further supports its reliability (Taylor et al., 2010, Dee and Jacob, 2011, Le Floch et al., 2007, Hanushek and Raymond, 2002, Hanushek and Raymond, 2005, Spees et al., 2016, Blank, 2011, Nelson et al., 2004). Since the information from this source is nationally representative, I included it to detail changes in Arizona’s scores over the years 1996, 2000, 2005, 2009, and 2015. Specifically, secondary data obtained from this source is used to answer the first research question. For research question 2, I selected two pieces of data from the Arizona Department of Education. The first contained ratings of Arizona schools by AZ LEARNS classification and the second was the A-F Letter Grades of Arizona schools during the 2018-2019 school year.

Data Preparation

For research question 1, Arizona's scores were compared to states with moderate accountability. In order to make such a comparison, state-level scale scores (in eighth-grade math) from the National Assessment of Educational Progress (NAEP) were utilized. For each of the years (1996, 2000, 2005, 2009, & 2015), I recorded Arizona's scale scores as well as the scale scores of Wisconsin, Nevada, Georgia, and Oregon. The reason for the selection of these states in addition to Arizona is because these are states that had moderate accountability before NCLB (Dee & Jacob, 2011). Finally, I also recorded the national average scale scores for the same years.

Table 1. State-level NAEP Scores for Arizona and states of Moderate Accountability Type.

	Arizona	Wisconsin	Nevada	Georgia	Oregon	National Average
1996	267.87	282.85		262.47	276.34	270.51
2000	268.58		264.94	265.36	280.06	271.83
2005	274.31	288.54	269.91	272.19	282.24	277.52
2009	277.33	288.14	274.15	277.56	285.04	281.67
2015	282.92	289.08	275.17	278.71	282.81	281.28

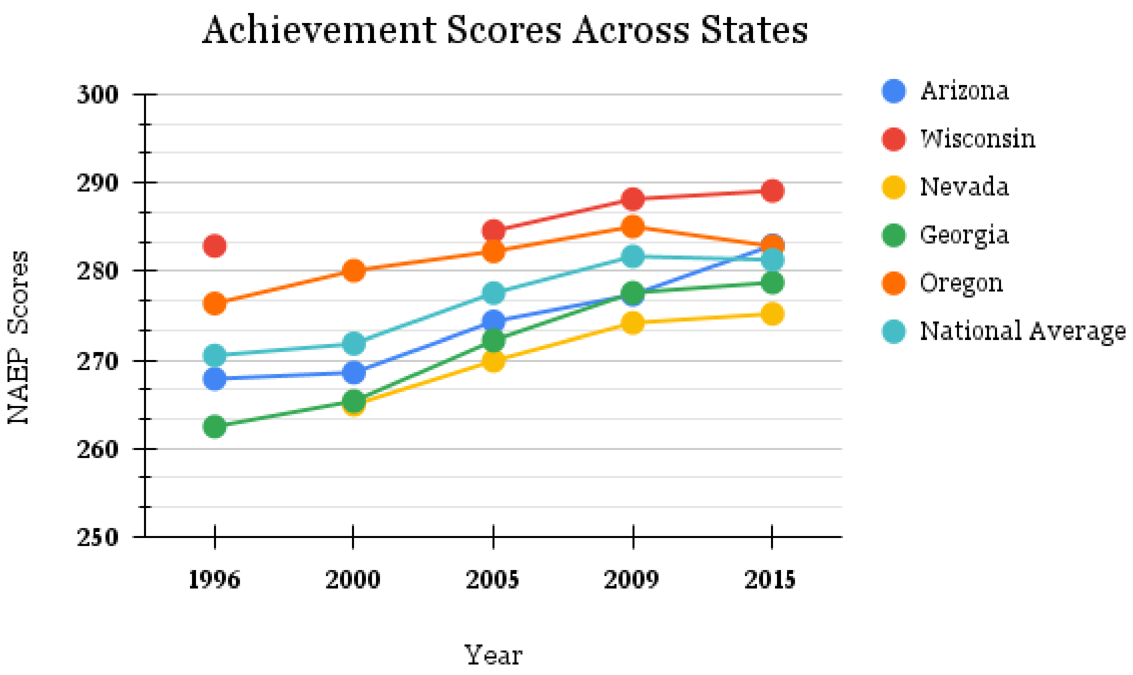


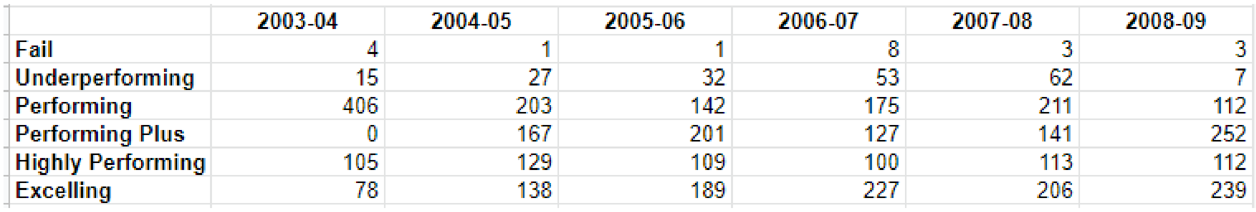
Figure 1. State-level NAEP Scores for Arizona and States of Moderate Accountability.

As seen above, there seems to be an overall improvement in state scale scores over the years. There are also a few missing values that were not available due to reporting standards not being met. The last preparation of data involved information from the Arizona Department of Education. One piece of data contained ratings of Arizona schools by AZ LEARNS classification from 2003-2009, and the other had the A-F Letter Grades of Arizona schools during the 2018-2019 school year. The first piece of data was limited to just schools in Maricopa County, whereas the second piece of data consisted of schools all across Arizona. Hence, to set up for a valid comparison, I filtered down the second piece of data by county so that I was left with the schools just in Maricopa County. I effectively constructed a new smaller dataset to work with.

Table 2. Total Number of Schools by A-F Rating in Maricopa County (2018-2019 Year).

	Traditional K-8 Schools	Traditional 9-12 Schools	Traditional Hybrid Schools	Alternative Schools
A	248	45	28	6
B	258	46	17	36
C	193	16	7	16
D	49	3	2	4
F	12	0	1	2

I should also point out that although this piece of data offers information based on school category (Traditional K-8 Schools, Traditional 9-12 Schools, Traditional Hybrid Schools, Alternative Schools), the most relevant column is Traditional K-8 Schools. This is because the other piece of data regarding AZ LEARNS classification consists of K-8 schools. Hence, it follows that both pieces of data must be consistent in this way as well in order to establish a basis for comparison. Just for reference, here is the information provided by the other piece of data:



	2003-04	2004-05	2005-06	2006-07	2007-08	2008-09
Fail	4	1	1	8	3	3
Underperforming	15	27	32	53	62	7
Performing	406	203	142	175	211	112
Performing Plus	0	167	201	127	141	252
Highly Performing	105	129	109	100	113	112
Excelling	78	138	189	227	206	239

Figure 2. Total Number of Schools by AZ LEARNS Classifications in Maricopa County.

Analysis

In this part of the methods section, the following information pertains to the analytical strategy utilized for each individual research question.

Research Question 1

To answer the first research question, I compared Arizona's state-level NAEP scores (in eighth-grade math) to Georgia, a state that had moderate accountability measures prior to NCLB. As seen in Figure 1, there were a couple of other states that fell under this moderate accountability provision, but the reason I chose Georgia is because it is the only state out of the four that did not have school repercussions (Dee & Jacob, 2011). This is an interesting basis for further comparison since with "school repercussions," states have more authority in making sure that students are held accountable through the enactment of consequences. To make a comparison between Arizona and Georgia's scores, I conducted an independent samples *t*-test to determine if the two samples were statistically significant. The null hypothesis being tested was that the difference between the means of the two samples was zero. The two samples of data were the relevant two columns from Table 1. Using the SPSS software, I ran a *t*-test between these two columns of data.

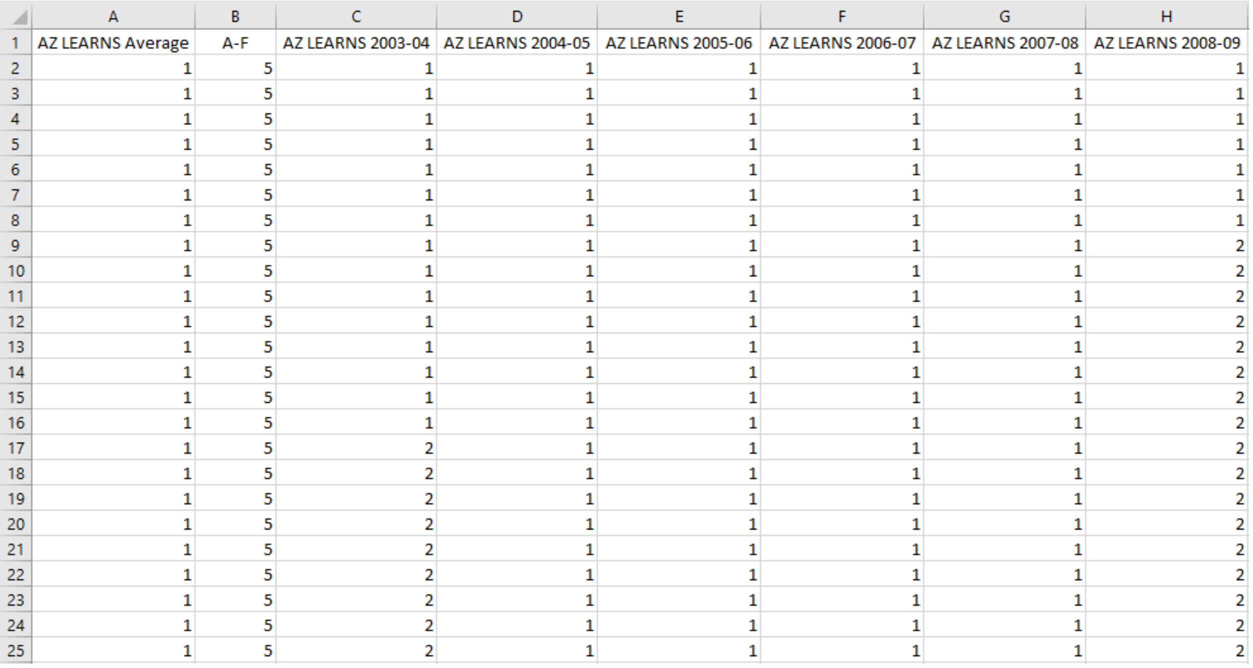
Research Question 2

Through a series of independent sample *t*-tests, I evaluated school ratings both during and after NCLB under AZ LEARNS and the A-F Accountability System. I was careful about making a comparison between these two systems since they each have their own rating categories. Under AZ LEARNS, schools were rated in one of the following categories: Fail, Underperforming, Performing, Performing Plus, Highly Performing, and Excelling. On the other hand, under the A-F Accountability System, schools are given a letter grade from A through F. To account for these differences, I created a common scale ranging from 1 to 5. Each rating category of both systems was given a whole number value between 1 and 5, with 5 representing the best performance rating and 1 representing the worst. For example, for the A-F Accountability System, a 5 was assigned to A, a 4 was assigned to B, 3 was assigned to C, 2 was assigned to D, and 1 was assigned to F. Similarly, for AZ LEARNS, values were assigned as the following- *Underperforming*:1, *Performing*:2, *Performing Plus*:3, *Highly Performing:*4, *Excelling*:5. It should be noted that as part of this scale system, the Fail category was not included. Schools rated Fail are simply schools that have been Underperforming for three consecutive years, so this is the reason for the exclusion of this category. Next, for the A-F data (Table 2), focusing on the Traditional K-8 Schools column, I assigned the relevant number from 1-5 for each of the schools. For example, I assigned a 5 for each of the 248 schools rated A, creating a new column of data with 5 appearing 248 times (in 248 cells). I performed the same steps with the rest of the A-F data, and the result was a new column of data consisting of 248 fives, 258 fours, 193 threes, 49 twos, and 12 ones. The importance of this new column is that it represents all of the A-F data.

Likewise, 6 new columns were constructed for the AZ LEARNS data (Figure 2). Since the AZ LEARNS data is across multiple years, whereas the A-F data represents one year, I constructed a unique column for each school year (2003-2009). For example, for the year 2003-04, I created a new column of data consisting of 15 ones, 406 twos, 0 threes, 105 fours, and 78 fives. I performed the same procedure for the rest of the years (2004-05, 2005-06, 2006-07, 2007-08, 2008-09), creating a new column of data for each of them.

Lastly, I constructed one final column of data that represented a 2003-2009 AZ LEARNS Average. First, I counted the total number of schools in each category across all the years (as such): *Underperforming*:196, *Performing*: 1249, *Performing Plus:*888, *Highly Performing*:668, *Excelling*:1077. I then divided each of these values by six to calculate the average number of schools in each category between 2003 and 2009 (rounded to the nearest whole number). This resulted in the following: *Underperforming:*33, *Performing*:208, *Performing Plus*:148, *Highly Performing:*111, *Excelling*:180. I used these values to construct a column consisting of 33 ones, 208 twos, 148 threes, 111 fours, and 180 fives.

With all the data well organized into appropriate columns, in order to make a comparison, I conducted a series of independent sample *t*-tests between two samples of data using the SPSS software. Since sample sizes and variances were unequal between the groups, a set of Welch's *t*-tests was performed. The null hypothesis was that the difference between the means of the two samples was zero. The first *t*-test compared the A-F column with the AZ LEARNS Average column. The rest of the performed *t*-tests compared the A-F column with each individual AZ LEARNS column. In other words, the remaining *t*-tests were conducted between the following samples: 1) A-F column and AZ LEARNS 2003-04 column 2) A-F column and AZ LEARNS 2004-05 column 3) A-F column and AZ LEARNS 2005-06 column 4) A-F column and AZ LEARNS 2006-07 column 5) A-F column and AZ LEARNS 2007-08 column 6) A-F column and AZ LEARNS 2008-09 column. It should be noted that this research question could technically be answered just by comparing the A-F column with the AZ LEARNS Average column, even without these last remaining *t*-tests. However, the reason for the inclusion of these *t*-tests was to ensure reliability and account for outliers that may have affected the AZ LEARNS Average calculation. Comparing each individual year would expose differences between the years and makes the most sense because it is consistent with the A-F data, which only represents one year (2018-19). For clarity, the constructed columns can be seen in the figure below along with the first couple of values.



	A	B	C	D	E	F	G	H
1	AZ LEARNS Average	A-F	AZ LEARNS 2003-04	AZ LEARNS 2004-05	AZ LEARNS 2005-06	AZ LEARNS 2006-07	AZ LEARNS 2007-08	AZ LEARNS 2008-09
2	1	5	1	1	1	1	1	1
3	1	5	1	1	1	1	1	1
4	1	5	1	1	1	1	1	1
5	1	5	1	1	1	1	1	1
6	1	5	1	1	1	1	1	1
7	1	5	1	1	1	1	1	1
8	1	5	1	1	1	1	1	1
9	1	5	1	1	1	1	1	2
10	1	5	1	1	1	1	1	2
11	1	5	1	1	1	1	1	2
12	1	5	1	1	1	1	1	2
13	1	5	1	1	1	1	1	2
14	1	5	1	1	1	1	1	2
15	1	5	1	1	1	1	1	2
16	1	5	1	1	1	1	1	2
17	1	5	2	1	1	1	1	2
18	1	5	2	1	1	1	1	2
19	1	5	2	1	1	1	1	2
20	1	5	2	1	1	1	1	2
21	1	5	2	1	1	1	1	2
22	1	5	2	1	1	1	1	2
23	1	5	2	1	1	1	1	2
24	1	5	2	1	1	1	1	2
25	1	5	2	1	1	1	1	2

Figure 3. Columns for Research Question 2.

Strengths

Especially for research question 3, the samples collected and approaches used are novel. Although they have not been widely tested in previous research, the research methods have been, namely the use of significance tests like the student's *t*-test. The novelty of this research also adds to its importance and encourages further research to be done on this topic in the future. Another strength in this research design is the large amount of data that has been incorporated, which would help support more reliable results. Finally, since Maricopa County is Arizona's most populous county, the results are likely to be more representative of the state since the data used (for research question 2) is of Maricopa County.

Results

Research Question 1

The mean score for Arizona (*N* = 5) was 274.20 (*SD* = 6.27). Georgia (*N* = 5) was associated with numerically smaller scores *M* = 271.26 (*SD* = 7.21) comparatively. To test the hypothesis that the mean scores of Arizona and Georgia were not statistically different from each other, an independent samples *t*-test was performed. From Table 3, it can be seen that the distributions of the two groups were sufficiently normal in order to conduct a *t*-test (i.e., skew < |2.0| and kurtosis < |9.0|; Schmider, Ziegler, Danay, Beyer, & Bühner, 2010). Furthermore, the homogeneity of variances assumption was both tested and satisfied through Levene's F test, *F*(8) = .30, *p* = .597. A statistically significant effect was not associated with the independent samples *t*-test, *t*(8) = .69, *p* = .511. The mean scores of the groups were not significantly different, meaning the null hypothesis cannot be rejected. 0.435 was the estimated Cohen's *d* value, which, according to Cohen's (1992) guidelines, is a small effect. In Figure 4, a graphical representation of the 95% confidence level and respective means is displayed.

Table 3. Descriptive Statistics Associated with State-level NAEP Scores.

	N	*M*	*SD*	Skew	Kurtosis
Arizona	5	274.20	6.27	0.44	-1.16
Georgia	5	271.26	7.21	-0.24	-2.54

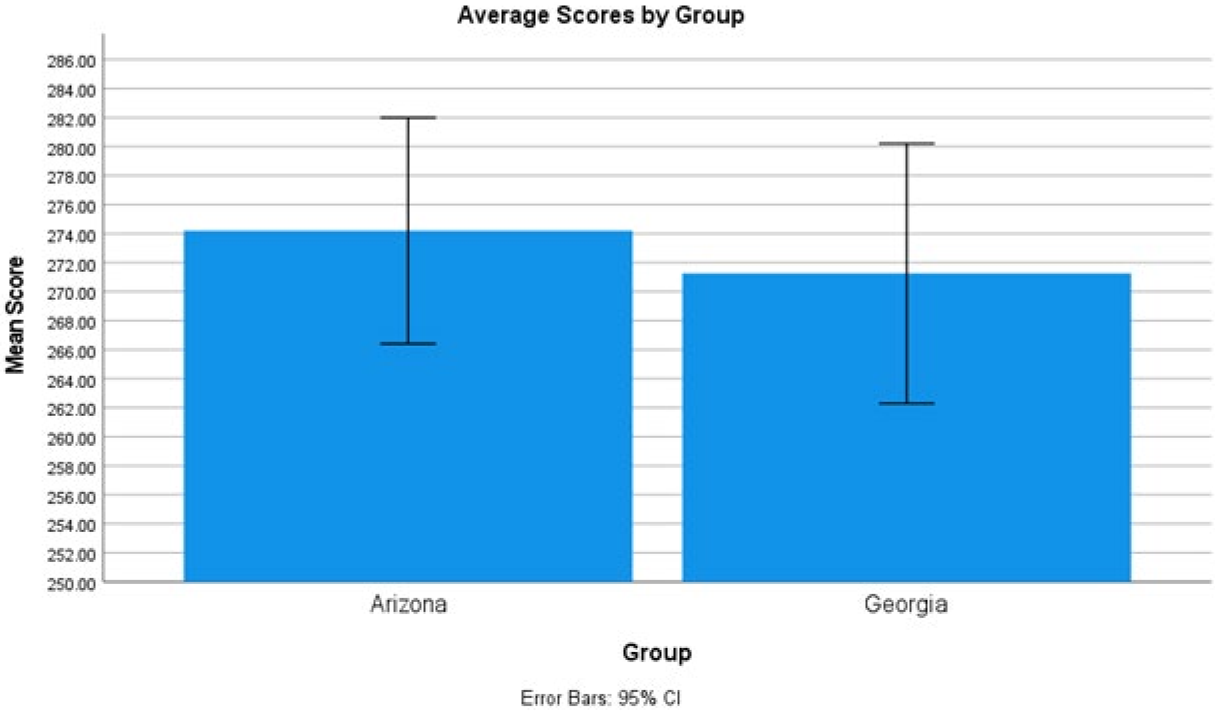


Figure 4. Average Scores by Group (95% Confidence Interval).

Research Question 2

A-F and AZ LEARNS Average

The A-F group (*N* = 760) had a mean rating of 3.90 (*SD* = 0.99). The AZ LEARNS Average group (*N* = 680) was associated with numerically lower ratings *M* = 3.29 (*SD* = 1.28) comparatively. A Welch's *t*-test was used to test the hypothesis that the mean ratings of the A-F and AZ LEARNS Average groups were not statistically different from each other. Since variances and sample sizes were unequal between the groups, I relied on a Welch's *t*-test instead of a student's *t*-test. From Table 4, it can be seen that the distributions of the two groups were sufficiently normal in order to conduct a *t*-test (i.e., skew < |2.0| and kurtosis < |9.0|; Schmider, Ziegler, Danay, Beyer, & Bühner, 2010). The Welch's *t*-test demonstrated a statistically significant effect, *t*(1269.866) = 9.982, *p* < .001. The mean ratings of the groups were significantly different, meaning the null hypothesis should be rejected. 0.534 was the estimated Cohen's *d* value, which, according to Cohen's (1992) guidelines, is a medium effect. In Figure 5, a graphical representation of the means and 95% confidence level is seen.

Table 4. Descriptive Statistics Associated with School Ratings.

	N	*M*	*SD*	Skew	Kurtosis
A-F	760	3.90	0.99	0.59	0.24
AZ LEARNS AVG	680	3.29	1.28	0.07	1.33

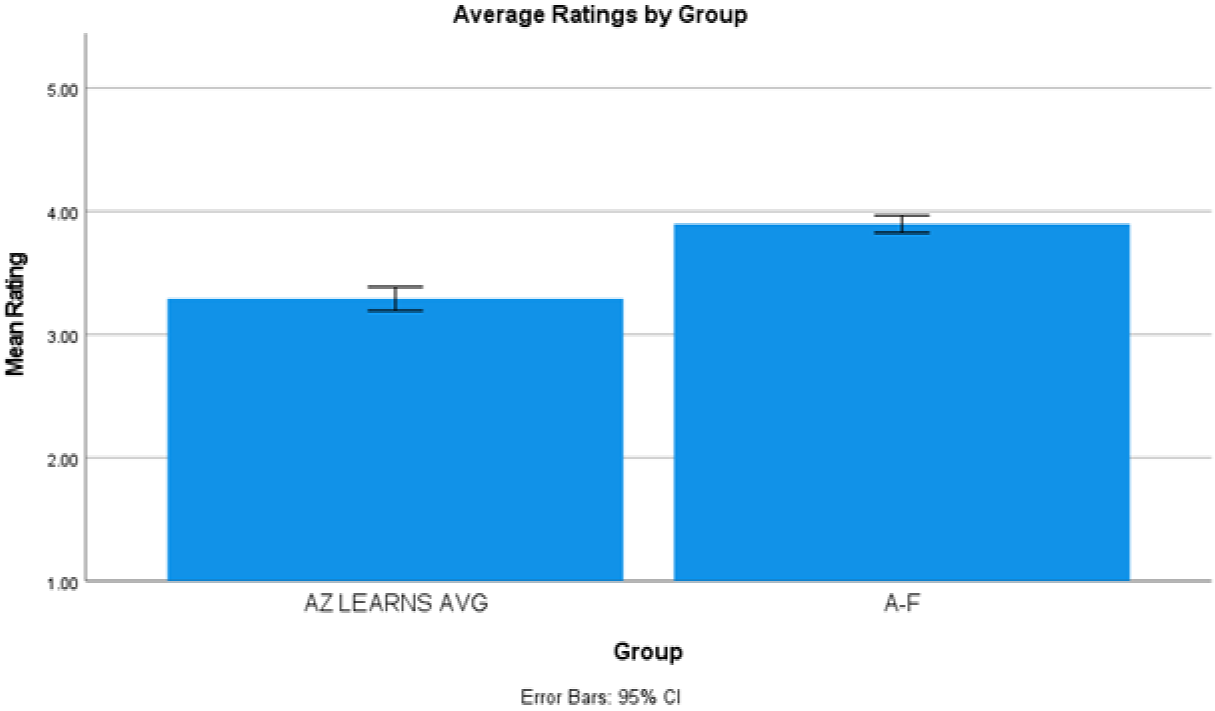


Figure 5. Average Ratings by Group (95% Confidence Interval).

A-F and AZ LEARNS 2003-04

The A-F group (*N* = 760) had a mean rating of 3.90 (*SD* = 0.99). The AZ LEARNS 2003-04 group (*N* = 604) was associated with numerically lower ratings *M* = 2.71 (*SD* = 1.17) comparatively. A Welch's *t*-test was used to test the hypothesis that the A-F and AZ LEARNS 2003-04 groups were not statistically different from each other. Since variances and sample sizes were unequal between the groups, I relied on a Welch's *t*-test instead of a student's *t*-test. From Table 5, it can be seen that the distributions of the two groups were sufficiently normal in order to conduct a *t*-test (i.e., skew < |2.0| and kurtosis < |9.0|; Schmider, Ziegler, Danay, Beyer, & Bühner, 2010). The Welch's *t*-test showed a statistically significant effect, *t*(1174.061) = 19.867, *p* < .001. The mean ratings of the groups were significantly different, meaning the null hypothesis should be rejected. 1.105 was the estimated Cohen's *d* value, which, according to Cohen's (1992) guidelines, is a large effect. In Figure 6, a graphical representation of the means and 95% confidence level is displayed.

Table 5. Descriptive Statistics Associated with School Ratings.

	N	*M*	*SD*	Skew	Kurtosis
A-F	760	3.90	0.99	0.59	0.24
AZ LEARNS 2003-04	604	2.71	1.17	0.97	0.67

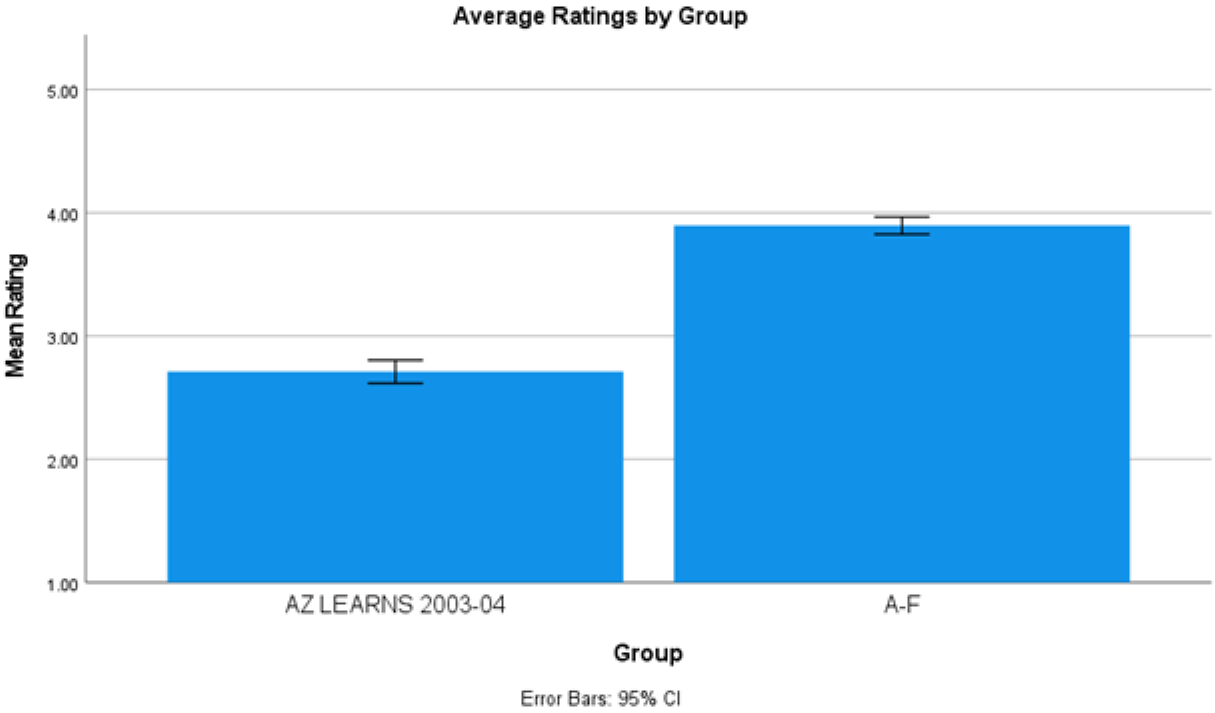


Figure 6. Average Ratings by Group (95% Confidence Interval).

A-F and AZ LEARNS 2004-05

The A-F group (*N* = 760) had a mean rating of 3.90 (*SD* = 0.99). The AZ LEARNS 2004-05 group (*N* = 664) was associated with numerically lower ratings *M* = 3.22 (*SD* = 1.20) comparatively. A Welch's *t*-test was used to test the hypothesis that the A-F and AZ LEARNS 2004-05 groups were not statistically different from each other. Since variances and sample sizes were unequal between the groups, I relied on a Welch's *t*-test instead of a student's *t*-test. From Table 6, it can be seen that the distributions of the two groups were sufficiently normal in order to conduct a *t*-test (i.e., skew < |2.0| and kurtosis < |9.0|; Schmider, Ziegler, Danay, Beyer, & Bühner, 2010). The Welch's *t*-test showed a statistically significant effect, *t*(1283.092) = 11.451, *p* < .001. The mean ratings of the groups were significantly different, meaning the null hypothesis should be rejected. 0.616 was the estimated Cohen's *d* value, which, according to Cohen's (1992) guidelines, is a medium effect. In Figure 7, a graphical representation of the means and 95% confidence level is displayed.

Table 6. Descriptive Statistics Associated with School Ratings.

	N	*M*	*SD*	Skew	Kurtosis
A-F	760	3.90	0.99	0.59	0.24
AZ LEARNS 2004-05	664	3.22	1.20	0.14	1.17

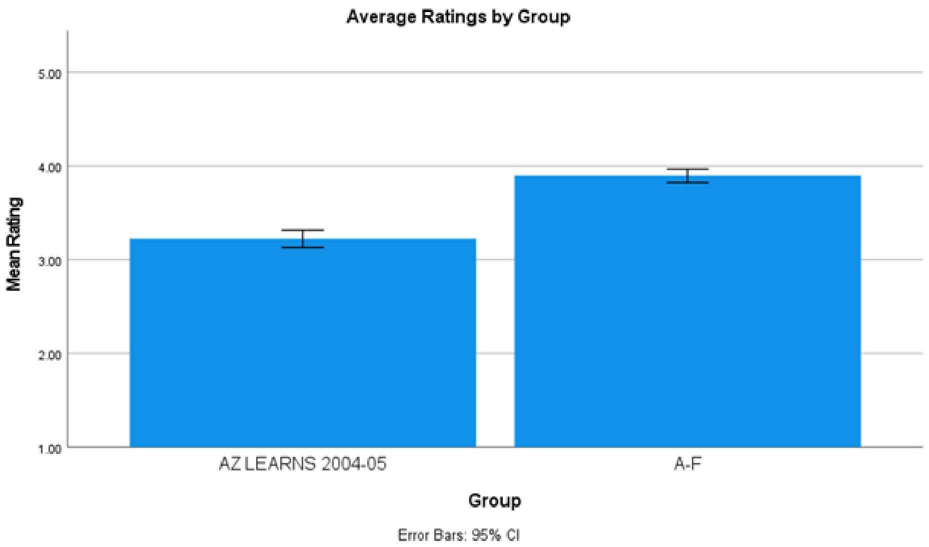


Figure 7. Average Ratings by Group (95% Confidence Interval).

A-F and AZ LEARNS 2005-06

The A-F group (*N* = 760) had a mean rating of 3.90 (*SD* = 0.99). The AZ LEARNS 2005-06 group (*N* = 673) was associated with numerically lower ratings *M* = 3.42 (*SD* = 1.23) comparatively. A Welch's *t*-test was used to test the hypothesis that the A-F and AZ LEARNS 2005-06 groups were not statistically different from each other. Since variances and sample sizes were unequal between the groups, I relied on a Welch's *t*-test instead of a student's *t*-test. From Table 7, it can be seen that the distributions of the two groups were sufficiently normal in order to conduct a *t*-test (i.e., skew < |2.0| and kurtosis < |9.0|; Schmider, Ziegler, Danay, Beyer, & Bühner, 2010). The Welch's *t*-test showed a statistically significant effect, *t*(1284.875) = 8.057, *p* < .001. The mean ratings of the groups were significantly different, meaning the null hypothesis should be rejected. 0.432 was the estimated Cohen's *d* value, which, according to Cohen's (1992) guidelines, is a small effect. In Figure 8, a graphical representation of the means and 95% confidence level is displayed.

Table 7. Descriptive Statistics Associated with School Ratings.

	N	*M*	*SD*	Skew	Kurtosis
A-F	760	3.90	0.99	0.59	0.24
AZ LEARNS 2005-06	673	3.42	1.23	0.08	1.13

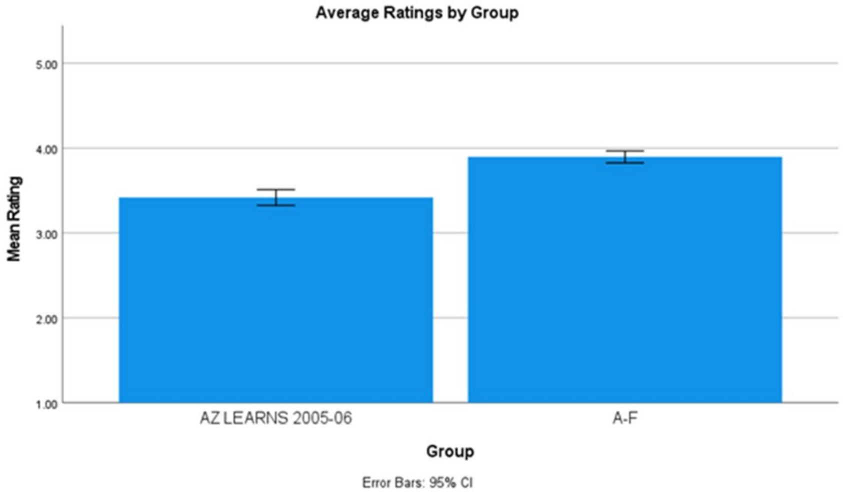


Figure 8. Average Ratings by Group (95% Confidence Interval).

A-F and AZ LEARNS 2006-07

The A-F group (N = 760) had a mean rating of 3.90 (SD = 0.99). The AZ LEARNS 2006-07 group (N = 682) was associated with numerically lower ratings M = 3.40 (SD = 1.37) comparatively. A Welch's t-test was used to test the hypothesis that the A-F and AZ LEARNS 2006-07 groups were not statistically different from each other. Since variances and sample sizes were unequal between the groups, I relied on a Welch's t-test instead of a student's t-test. From Table 8, it can be seen that the distributions of the two groups were sufficiently normal in order to conduct a t-test (i.e., skew < |2.0| and kurtosis < |9.0|; Schmider, Ziegler, Danay, Beyer, & Bühner, 2010). The Welch's t-test showed a statistically significant effect, *t*(1221.146) = 7.794, p < .001. The mean ratings of the groups were significantly different, meaning the null hypothesis should be rejected. 0.418 was the estimated Cohen's *d* value, which, according to Cohen's (1992) guidelines, is a small effect. In Figure 9, a graphical representation of the means and 95% confidence level is displayed.

Table 8. Descriptive Statistics Associated with School Ratings.

	N	*M*	*SD*	Skew	Kurtosis
A-F	760	3.90	0.99	0.59	0.24
AZ LEARNS 2006-07	682	3.40	1.37	0.15	1.38

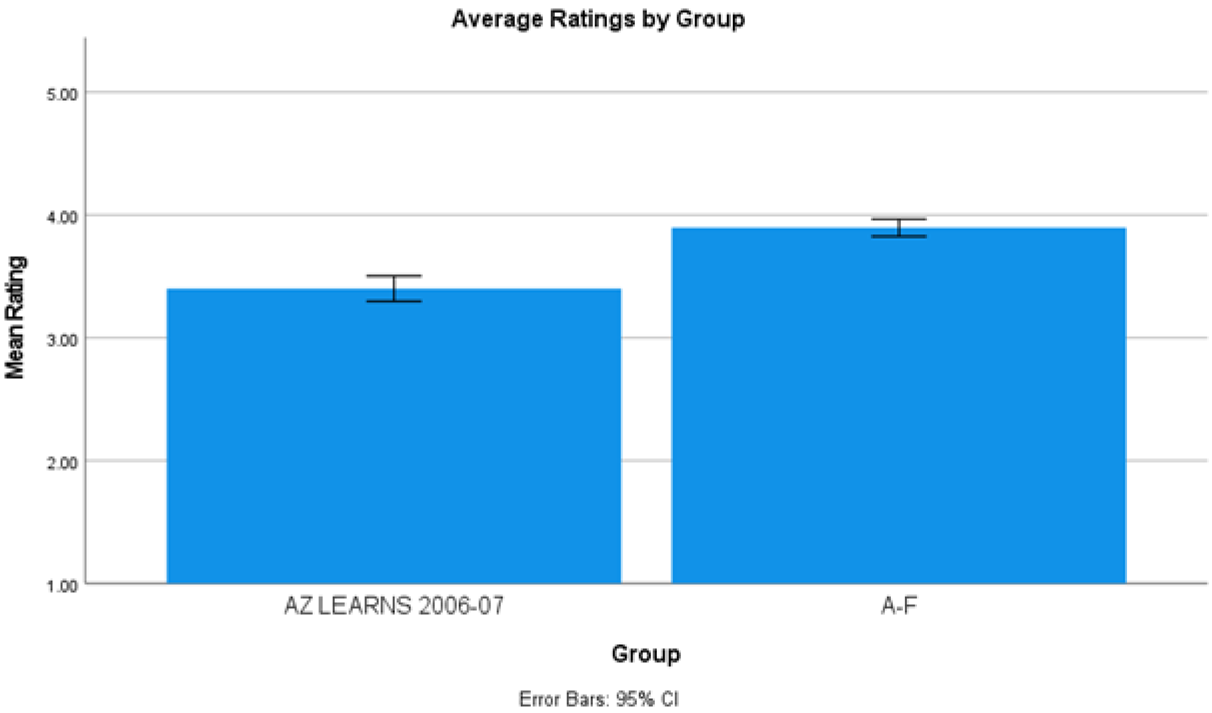


Figure 9. Average Ratings by Group (95% Confidence Interval).

A-F and AZ LEARNS 2007-08

The A-F group (*N* = 760) had a mean rating of 3.90 (*SD* = 0.99). The AZ LEARNS 2007-08 group (*N* = 733) was associated with numerically lower ratings *M* = 3.26 (*SD* = 1.36) comparatively. A Welch's *t*-test was used to test the hypothesis that the A-F and AZ LEARNS 2007-08 groups were not statistically different from each other. Since variances and sample sizes were unequal between the groups, I relied on a Welch's *t*-test instead of a student's *t*-test. From Table 9, it can be seen that the distributions of the two groups were sufficiently normal in order to conduct a *t*-test (i.e., skew < |2.0| and kurtosis < |9.0|; Schmider, Ziegler, Danay, Beyer, & Bühner, 2010). The Welch's *t*-test showed a statistically significant effect, *t*(1333.516) = 10.348, *p* < .001. The mean ratings of the groups were significantly different, meaning the null hypothesis should be rejected. 0.539 was the estimated Cohen's *d* value, which, according to Cohen's (1992) guidelines, is a medium effect. In Figure 10, a graphical representation of the means and 95% confidence level is displayed.

Table 9. Descriptive Statistics Associated with School Ratings.

	N	*M*	*SD*	Skew	Kurtosis
A-F	760	3.90	0.99	0.59	0.24
AZ LEARNS 2007-08	733	3.26	1.36	0.003	1.36

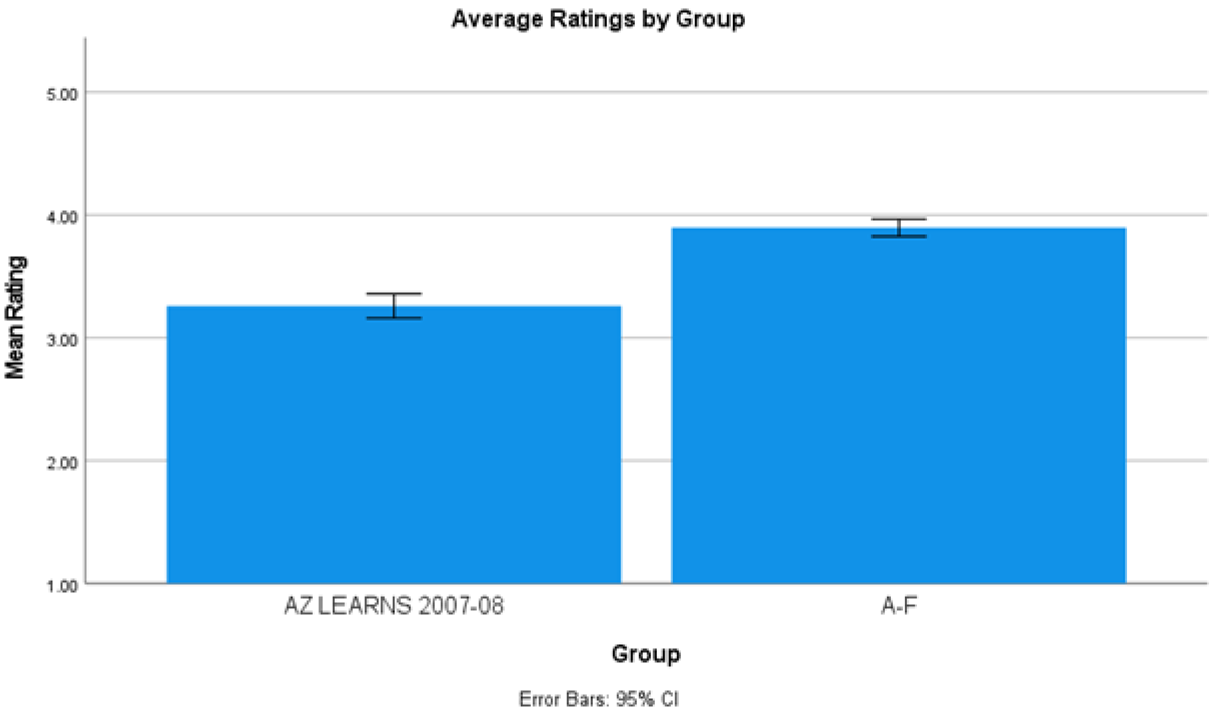


Figure 10. Average Ratings by Group (95% Confidence Interval).

A-F and AZ LEARNS 2008-09

The A-F group (*N* = 760) had a mean rating of 3.90 (*SD* = 0.99). The AZ LEARNS 2008-09 group (*N* = 722) was associated with numerically lower ratings *M* = 3.64 (*SD* = 1.12) comparatively. A Welch's *t*-test was used to test the hypothesis that the A-F and AZ LEARNS 2008-09 groups were not statistically different from each other. Since variances and sample sizes were unequal between the groups, I relied on a Welch's *t*-test instead of a student's *t*-test. From Table 10, it can be seen that the distributions of the two groups were sufficiently normal in order to conduct a *t*-test (i.e., skew < |2.0| and kurtosis < |9.0|; Schmider, Ziegler, Danay, Beyer, & Bühner, 2010). The Welch's *t*-test showed a statistically significant effect, *t*(1432.923) = 4.607, *p* < .001. The mean ratings of the groups were significantly different, meaning the null hypothesis should be rejected. 0.240 was the estimated Cohen's *d* value, which, according to Cohen's (1992) guidelines, is a small effect. In Figure 11, a graphical representation of the means and 95% confidence level is displayed.

Table 10. Descriptive Statistics Associated with School Ratings.

	N	*M*	*SD*	Skew	Kurtosis
A-F	760	3.90	0.99	0.59	0.24
AZ LEARNS 2008-09	722	3.64	1.12	0.09	1.25

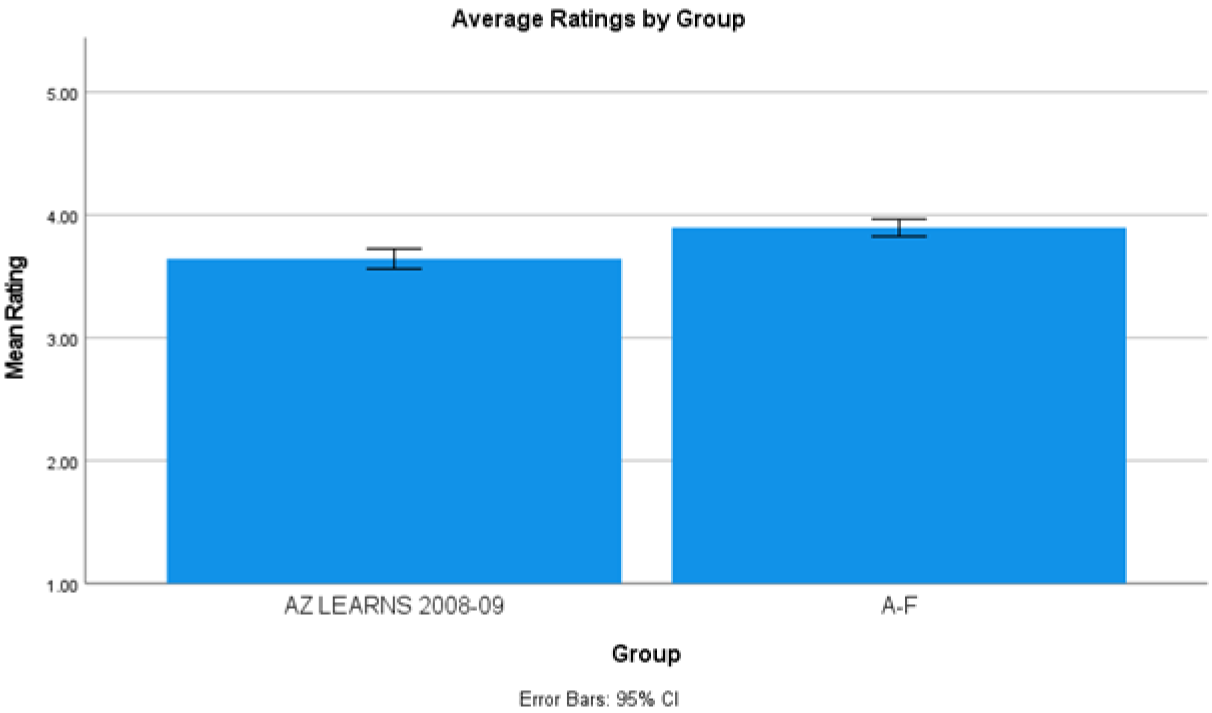


Figure 11. Average Ratings by Group (95% Confidence Interval).

Discussion

The results consistently indicate a statistically significant difference associated with school ratings under AZ LEARNS and the A-F Accountability System. However, they do not support a statistically significant difference associated with state-level NAEP scores of Arizona and Georgia (different accountability types). These findings are important because they suggest that No Child Left Behind may have affected the implementation of standards-based accountability in Arizona. This is in line with Dee and Jacob (2011) and supports my original claim that states without test-based accountability measures pre-NCLB are more likely to have been affected by its implementation. I speculate that as a result of the provisions maintained under NCLB, Arizona might have changed elements of its accountability to reflect certain changes. However, I am careful not to attribute the findings to the impact of NCLB, since that goes beyond what can be concluded through this study and involves other types of analyses (e.g., interrupted time series). For research question 2 (as can be seen in Table 5), the largest difference between the mean ratings exists in the comparison between the A-F group (M = 3.90) and the AZ LEARNS 2003-04 group (M = 2.71). The difference between the mean ratings is 1.19, which is a larger mean difference than any of the other comparisons done for this research question. Interestingly enough, Garcia et al. (2010) found a 2004-2005 test score spike in Arizona scores that is responsible for irregularities and a skewing of achievement rates. Although not confirmed, this spike may, in part, explain the large mean difference between the accountability groups in 2003-04. In the scenario that elements of the accountability systems were updated during that year to reflect future changes (i.e., test score spikes), it would make sense for there to, as a result, be a larger mean difference.

It is also noticeably important that for the same comparison (between the AZ LEARNS 2003-04 group and the A-F group), a large effect size was estimated. This is expected since there was a large mean difference, but it also attests to the strength of this specific comparison. A large effect size, as in this case, supports the research finding as having a practical significance.

Additionally, an examination of the mean differences of the groups, in the comparisons conducted under research question 2, (A-F and AZ LEARNS 2003-04: *1.19*, A-F and AZ LEARNS 2004-05: *0.68*, A-F and AZ LEARNS 2005-06: *0.48*, A-F and AZ LEARNS 2006-07: *0.50*, A-F and AZ LEARNS 2007-08: *0.64*, A-F and AZ LEARNS 2008-09: *0.26*) reveal somewhat of a decreasing pattern over the years, although unpredictable. Because of this, a connection cannot be drawn regarding whether the mean ratings under AZ LEARNS have gradually increased over time. The difference between the mean ratings in the comparison between the A-F group (*M* = 3.90) and the AZ LEARNS Average group (*M* = 3.290) is 0.61. Compared to the other mean differences, this value is somewhere in between. However, it is strikingly similar to the average of those mean differences
((1.19 + 0.68 + 0.48 + 0.50 + 0.64 + 0.26) / 6 = 0.625). This similarity suggests that the results are fairly accurate and also shows how the ratings of the AZ LEARNS Average group may be skewed by the results of any particular year. This is important to note because it supports the idea that accountability systems are volatile and subject to various factors.

To touch on an earlier point, for research question 1, there was no statistically significant difference associated with the state-level NAEP scores of Arizona and Georgia. The relationship between the scores of the accountability groups hints that actual achievement scores may largely vary from state to state. Even within the same accountability range (weak, moderate, strong), states can have different policies that shape elements like student achievement and classroom environment.

Overall, my results fit well within the existing literature and provide insight into Arizona's achievement during the NCLB era. There are implications for future evidence-based policymaking in terms of best fostering student achievement. From these findings, it can be understood that the A-F Accountability System likely has higher mean ratings than AZ LEARNS, which seems promising for the state since A-F is the current system. However, much is to be done to build off of this. The state could benefit from incorporating more indicators of student success, including projects and hands-on learning opportunities. In addition, it should be taken into account that the data used to compare the accountability systems consisted of all Maricopa County schools. Although strong in terms of size, I am cautious not to make any claims or generalizations for the entire state since there are other counties as well that may be different from Maricopa. As a result, further research may incorporate data from various counties to conduct an even larger analysis. It could also consider differences in achievement between states of the same accountability type and identify the factors responsible for these differences.

Conclusion

The No Child Left Behind Act (NCLB) altered the educational landscape by setting a new sequence of accountability. NCLB focused on standardized testing as a primary indicator of student success and pushed states to set their own academic standards. My research supports a statistically significant difference associated with school ratings under Arizona's two accountability systems. Both AZ LEARNS and the A-F Letter Grade System incorporated elements of NCLB as part of evaluating overall performance, but A-F looked more at student improvement each year as opposed to simply the percentage of students scoring at a certain level.

Moving forward, our educational system could benefit from involving students in the educational process and valuing student improvement as much as test scores. This is a rather unique concept since "student achievement," as outlined by NCLB, was essentially scores on state standardized tests. However, in today's world, redefining "student achievement" to include an individual student's improvement, both in terms of higher scores and other academic indicators, would provide more reliable information on student success. If "student achievement" involved student improvement instead of this just being another additional factor to consider, the chances of reducing achievement gaps could also possibly be higher. Overall, as evident by the observed statistically significant difference, it is likely that Arizona was strongly affected by NCLB, guiding its standards-based accountability. Furthermore, since the mean rating of the A-F group (M = 3.90) was larger than the mean ratings of any of the AZ LEARNS groups, I speculate that students are performing better under the new accountability system. This could potentially be in part because A-F places a heavier weight on student improvement as a part of their calculations.

Finally, not much can be concluded regarding Arizona's achievement scores when compared to other accountability types. My research does not support a statistically significant difference associated with the state-level NAEP scores of Arizona and Georgia, and although these states are of different accountability types, not much can be said since this is only one comparison and the sample size is small. However, from this analysis, I suppose that Arizona and Georgia may share some elements of their accountability systems. Further research may conduct multiple comparisons of the scores of states of different accountability types or might compare the factors involved (in deciding the accountability type of a given state) in an attempt to evaluate its significance.

Limitations

The main potential limitation in the research design pertains to research question 1. The sample size of the data was relatively small, so it may be rather difficult to establish strong conclusions from the results. Something else that should be closely examined is the approach used to compare Arizona LEARNS with the A-F Accountability System. These two systems are characteristically different, and this study does not attempt to force similarity between the two or draw a connection where not possible. Instead, through the creation of a scale system, the purpose is to compare the means of the samples. This scale system was uniquely designed to establish a basis of comparison in the first place since not only do both accountability systems consist of different ratings, but they also have an unequal amount. The A-F Accountability System has five possible ratings while AZ LEARNS has six.

Acknowledgments

I would like to thank my mentor for helping me with this project. I would also like to give credit to how2stats for the arrangement of my Results section.

References

Arizona Department of Education (n.d.). *AZ LEARNS Classifications For Maricopa County Schools*. Retrieved from https://www.arizonaindicators.org/az-learns/

Arizona Department of Education (2012). AZ Learns and A-F Letter Grades. *National Center on Assessment and Accountability for Special Education*. Retrieved from https://www.ncaase.com/docs/AZ_Learns_and_A-F_Letter_Grades040512.pdf

Arizona Secretary of State. (2000). *Proposition 301*. Retrieved from https://apps.azsos.gov/election/2000/Info/pubpamphlet/english/prop301.htm

Arizona State Board of Education (2019). *2018-2019 A-F Letter Grades*. Retrieved from https://azsbe.az.gov/f-school-letter-grades

Birman, B., Le Floch, K. C., & Klekotka, A. (2007). Evaluating teacher quality under no child left behind. https://doi.org/10.7249/RB9287

Blank, R. K. (2011). Closing the Achievement Gap for Economically Disadvantaged Students? Analyzing Change since No Child Left Behind Using State Assessments and the National Assessment of

Educational Progress. *Council of Chief State School Officers*. Retrieved from https://files.eric.ed.gov/fulltext/ED518986.pdf

Cohen, J. (1992). A power primer, *Psychological Bulletin*, *112*, 155-159. https://doi.org/10.1037//0033-2909.112.1.155

Combs, M. C., & Nicholas, S. E. (2012). The effect of Arizona language policies on Arizona Indigenous students. *Language Policy*, *11*(1), 101-118. https://doi.org/10.1007/s10993-011-9230-7

Dee, T. S., & Jacob, B. (2011). The impact of No Child Left Behind on student achievement. *Journal of Policy Analysis and management*, *30*(3), 418-446. https://doi.org/10.3386/w15531

Desimone, L. M. (2013). Reform before NCLB. *Phi Delta Kappan*, *94*(8), 59-61. https://doi.org/10.1177/003172171309400811

Garcia, D. R. (2008). Mixed Messages: American Indian Achievement Before and Since the Implementation of No Child Left Behind. *Journal of American Indian Education*, *47*(1), 136–154. http://www.jstor.org/stable/24398510

Garcia, E., Lawton, K., & Diniz de Figueiredo, E. H. (2010). Assessment of Young English Language Learners in Arizona: Questioning the Validity of the State Measure of English Proficiency. *UCLA: The Civil Rights Project / Proyecto Derechos Civiles*. Retrieved from https://escholarship.org/uc/item/6xx644b5

Hamilton, L. S., Stecher, B. M., Marsh, J. A., McCombs, J. S., Robyn, A., Russell, J. L., Naftel, S., & Barney, H. (2007). *Standards-Based Accountability Under No Child Left Behind: Experiences of Teachers and Administrators in Three States* (1st ed.). RAND Corporation. http://www.jstor.org/stable/10.7249/mg589nsf

Hamilton, L. S., Stecher, B. M., & Yuan, K. (2012). Standards-based accountability in the United States: Lessons learned and future directions. *Education inquiry*, *3*(2), 149-170. https://doi.org/10.3402/edui.v3i2.22025

Hanushek, E. A., & Raymond, M. E. (2002). Lessons About the Design of State Accountability Systems. Retrieved from https://files.eric.ed.gov/fulltext/ED477342.pdf

Hanushek, E. A., & Raymond, M. E. (2005). Does school accountability lead to improved student performance?. *Journal of Policy Analysis and Management: The Journal of the Association for Public Policy Analysis and Management*, *24*(2), 297-327. https://doi.org/10.1002/pam.20091

how2stats. (2014, March 14). *Independent Samples t-test - Writing Up Results* [Video]. YouTube. https://www.youtube.com/watch?v=WA7Ysxd-91E

Husband, T., & Hunt, C. (2015). A review of the empirical literature on No Child Left Behind from 2001 to 2010. *Planning and Changing*, *46*(1/2), 212.

Jimenez-Castellanos, O., & Martinez, D. (2014). Arizona. *Journal of Education Finance*, *39*(3), 247-249. https://www.muse.jhu.edu/article/539787.

Le Floch, K. C., Martinez, F., O'Day, J., Stecher, B., Taylor, J., & Cook, A. (2007). State and Local Implementation of the" No Child Left Behind Act." Volume III--Accountability under" NCLB" Interim Report. *US Department of Education*. Retrieved from https://files.eric.ed.gov/fulltext/ED499023.pdf

Legal Information Institute. (2001). No Child Left Behind Act of 2001. *Cornell Law School*. Retrieved from https://www.law.cornell.edu/wex/no_child_left_behind_act_of_2001

Mawhinney, H. B. (2013). Reactive Sequences in the Evolution of Maryland’s Consequential Accountability Regime. *Educational Policy*, *27*(2), 279–306. https://doi.org/10.1177/0895904812472723

McKinney, S. (2008). *An analysis of the influence of No Child Left Behind and Arizona Learns on middle-school principal leadership behaviors and responsibilities*. The University of Arizona. Retrieved from http://hdl.handle.net/10150/194025

Milliman, S. (2016). Charter schools, enrollment size, and educational accountability: A preliminary Arizona analysis. *Journal of School Choice*, *10*(1), 73-95. https://doi.org/10.1080/15582159.2015.1134238

Morel-Seytoux, S. (2009). Minority Student Progress Report 2009: A Snapshot of Arizona's Educational Achievement. *Arizona Commission for Postsecondary Education*. Retrieved from https://files.eric.ed.gov/fulltext/ED517262.pdf

Nelson, F. H., Rosenberg, B., & Van Meter, N. (2004). Charter School Achievement On The 2003 National Assessment Of Educational Progress. *Education Policy Studies Laboratory*, *Arizona State University College of Education*. Retrieved from https://files.eric.ed.gov/fulltext/ED483349.pdf

Olivant, K. F. (2015). "I am not a format": Teachers' experiences with fostering creativity in the era of accountability. *Journal of Research in Childhood Education*, *29*(1), 115-129. https://doi.org/10.1080/02568543.2014.978920

Peterson, P. E., & West, M. R. (Eds.). (2003). *No Child Left Behind?: The Politics and Practice of School Accountability*. Brookings Institution Press. http://www.jstor.org/stable/10.7864/j.ctvb6v789

Roberts, N. (n.d.). Black Arizonans Battling To Close The Education Achievement Gap. *The Block*. Retrieved from https://theblockcharlotte.com/27358/black-arizonans-battling-to-close-the-education-achievement-gap/

Ryan, J. E. (2004). The perverse incentives of the no child left behind act. *NYUL Rev.*, *79*, 932. https://doi.org/10.2139/ssrn.476463

Schmider, E., Ziegler, M., Danay, E., Beyer, L., & Bühner, M. (2010). Is it really robust? Reinvestigating the robustness of ANOVA against violations of the normal distribution assumption. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *6*, 147-151. https://doi.org/10.1027/1614-2241/a000016

Shin, E. (2022). *No Child Left Behind Act: The Impact of Standards-based Accountability* (Doctoral dissertation). Retrieved from https://hdl.handle.net/2104/11875

Spees, L. P., Potochnick, S., & Perreira, K. M. (2016). The academic achievement of Limited English Proficient (LEP) youth in new and established immigrant states: Lessons from the National Assessment of Educational Progress (NAEP). *Education Policy Analysis Archives*, *24*. 99. https://doi.org/10.14507/epaa.24.2130

Spurrier, A., Alderman, C., O'Neal Schiess, J., & Rotherham, A. J. (2020). The impact of standards-based accountability. *Washington, DC: Bellwether Education Partners.* Retrieved from https://files.eric.ed.gov/fulltext/ED606418.pdf

Standerfer, L. A. (2004). *Modifications of policy and practice in Arizona school districts in attempting to comply with No Child Left Behind and AZ LEARNS*. Arizona State University.

Stullich, S., Eisner, E., & McCrary, J. (2007). National assessment of title I, final report: Volume I: Implementation. *US Department of Education (National Center for Education Statistics Institute of Education Sciences, 2007)*, *1*. Retrieved from https://ies.ed.gov/ncee/pdf/20084012_rev.pdf

Taylor, J., Stecher, B., O'Day, J., Naftel, S., & Le Floch, K. C. (2010). State and Local Implementation of the" No Child Left Behind Act". Volume IX--Accountability under" NCLB". *US Department of Education*. Retrieved from https://files.eric.ed.gov/fulltext/ED508912.pdf

The Nation's Report Card. (2015). 2015 Mathematics Grades 4 and 8 Assessment Report Cards: Summary Data Tables for National and State Average Scores and Achievement Level Results. *National Center for Education Statistics*. Retrieved from https://www.nationsreportcard.gov/reading_math_2015/files/2015_Results_Appendix_Math.pdf

The Nation's Report Card. (2019). NAEP State Profiles. *National Assessment of Educational Progress.* Retrieved from https://www.nationsreportcard.gov/profiles/stateprofile?chort=2&amp;sub=MAT&amp;sj=&amp;sfj=NP&amp;st=MN&amp;year=2019R3

Tingey, R. A. (2009)."High-Stakes Testing Under The No Child Left Behind Act: How Has It Impacted School Culture?". *Theses and Dissertations*. 1864. Retrieved from https://scholarsarchive.byu.edu/etd/1864

US Department of Education. (1996). NAEP 1996 Mathematics: Report Card for the Nation and the States. *National Center for Education Statistics*. Retrieved from https://nces.ed.gov/nationsreportcard/pdf/main1996/97488.pdf

US Department of Education. (2000). The Nation's Report Card: Mathematics 2000. *National Center for Education Statistics*. Retrieved from https://nces.ed.gov/nationsreportcard/pdf/main2000/2001517.pdf

US Department of Education. (2002). Fact sheet on the major provisions of the conference report to HR 1, the No Child Left Behind Act. Retrieved from https://www2.ed.gov/nclb/overview/intro/factsheet.html

US Department of Education. (2005). The Nation's Report Card: Mathematics 2005. *National Center for Education Statistics*. Retrieved from https://nces.ed.gov/nationsreportcard/pdf/main2005/2006453.pdf

US Department of Education. (2009). The Nation's Report Card: Mathematics 2009. *National Center for Education Statistics*. Retrieved from https://nces.ed.gov/nationsreportcard/pdf/main2009/2010451.pdf

Wilgus, G. (2019). From a nation at risk to no child left behind to race to the top: The US response to global competition. In *Investment in early childhood education in a globalized world* (pp. 107-158). Palgrave Macmillan, New York. https://doi.org/10.1057/978-1-137-60041-7_4

Winstead, L. (2011). The impact of NCLB and accountability on social studies: Teacher experiences and perceptions about teaching social studies. *The Social Studies*, *102*(5), 221-227. https://doi.org/10.1080/00377996.2011.571567

Wright, W. E. (2007). Heritage language programs in the era of English-only and No Child Left Behind. *Heritage Language Journal*, *5*(1), 1-26. https://doi.org/10.46538/hlj.5.1.1

Wright, W. E., & Choi, D. (2005). Voices from the Classroom: A Statewide Survey of Experienced Third-Grade English Language Learner Teachers on the Impact of Language and High-Stakes Testing Policies in Arizona. *Language Policy Research Unit*. Retrieved from https://files.eric.ed.gov/fulltext/ED508521.pdf

Wright, W. E., & Choi, D. (2006). The impact of language and high-stakes testing policies on elementary school English language learners in Arizona. *Education Policy Analysis Archives*, *14*(13), 1-75. https://doi.org/10.14507/epaa.v14n13.2006

Wronowski, M. L., & Urick, A. (2018). Examining the relationship of teacher perception of accountability and assessment policies on teacher turnover during NCLB. *Education Policy Analysis Archives*, *27*(86). https://doi.org/10.14507/epaa.27.3858