2011

A Comparison Of Eighth Grade Reading Proficiency On State Assessments With The National Assessment Of Educational Progress

Kathryn B. Dyer
University of Central Florida

A COMPARISON OF EIGHTH GRADE READING PROFICIENCY ON STATE ASSESSMENTS WITH THE NATIONAL ASSESSMENT OF EDUCATIONAL PROGRESS

by

KATHRYN B. DYER
B.A. Stetson University, 1999
M.Ed. Stetson University, 2004

A dissertation submitted in partial fulfillment of the requirements
for the degree of Doctor of Education
in the Department of Educational Research, Technology, and Leadership
in the College of Education
at the University of Central Florida
Orlando, Florida

Spring Term
2011

Major Professors: Rosemarye Taylor and Debbie Hahs-Vaughn

ABSTRACT

The National Assessment of Educational Progress is a nationwide assessment administered every other year to eighth grade students in the United States in reading and mathematics. The purpose of this study was to compare the results of 2009 eighth grade state reading assessment proficiency percentages to NAEP proficiency percentages.

Primarily, this study examined whether a predictive relationship existed between state and NAEP proficiency percentages. Subsequent research questions analyzed the extent to which a relationship existed for subgroups (race/ethnicity, English Language Learners, low socioeconomic status, and students with disabilities) and while controlling for census regions.

It was found that a predictive relationship does exist between state and NAEP proficiency percentages for eighth grade students who took these reading assessments in 2009. The correlations between the variables were consistently high; however, the relationships were not significant for all subgroups nor for all census regions.

It was determined that NAEP and state assessment proficiency percentages are not well suited to direct comparisons. Recommendations for practice included the development of nationwide common assessments, standards, and proficiency scales.

This dissertation is dedicated to the village

ACKNOWLEDGMENTS

I am pleased to thank Dr. Rosemarye Taylor, co-chair of my dissertation committee, who was forever a kind and patient guide throughout the process of writing the dissertation. Her calm words kept me on track and moving forward. I would also like to thank Dr. Debbie Hahs-Vaughn, co-chair of my dissertation committee, who was kind enough to teach me the statistics I needed to know to fill in my learning gaps. Her painstaking editing has improved this study beyond measure. I would like to take this opportunity to express my respect for the work of Dr. William Gordon II whose work inspired this study. He was kind enough to meet with me during the initial stages to help me to form my ideas and to share the wisdom he learned while completing his own dissertation. I would also like to thank Dr. George Pawlas and Dr. Jeffrey Kaplan for giving me good ideas to further my research, for their insight and guidance.

I would be remiss not to thank my cohort group from the University of Central Florida at Daytona Beach: John Carr, Kelly Carter, Julian Jones, Julie Roseboom, John Shelby, and especially Teresa Marcks. The support system that we provided to each other has value beyond measure. Thanks also to Jen Williams, a fellow assistant principal and good friend at Pine Ridge High, who has always been willing to listen and support me through it all. I most appreciate having a principal, Tom Russell, who always expressed interest in my work even when I talked about studentized residuals and multicollinearity.

Lastly, I would like to thank the most critical members of the village who got me through my doctoral work: my family. My father, John Bentley, took the time to read through several edits of my paper—even the parts he disagreed with! My mother, Deanie

Bentley, provided tireless support and encouragement. My children, Luke and Isaac, also kept me going and kept pushing me to get my work done as quickly as possible so that I could spend more time with them! And finally, my husband, Mike, was the biggest help of all. His considerate and consistent support made it possible for me to further my education. I love and appreciate all he does for our family.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER ONE: INTRODUCTION

In an effort to improve student achievement and increase school accountability, the No Child Left Behind Act of 2001 (NCLB) mandated that each state develop its own assessment system to measure student progress toward proficiency (U.S. Department of Education, 2002). Each state set its own proficiency standards and then developed assessments to determine the percent of students achieving those standards. Accordingly, the achievement that students demonstrate on those assessments served to determine whether schools had enough proficient students for the state to conclude that the school made Adequate Yearly Progress (AYP). Because the percentage of students meeting proficiency standards varies widely, debate has occurred about the differences in rigor among state assessments and the impact these differences may have on perceptions about the percentage of schools in each state making AYP. Those who do not understand AYP tend to interpret failure to make AYP as an indicator that students are not learning at the schools. Some educators and legislators assert that a national assessment would give a more uniform look at the progress of each state's schools toward proficiency standards (FDOE, 2010; Taylor & Gordon, in press). In part, the NAEP was created for this purpose (National Center for Education Statistics [NCES], 2009d).

The National Assessment of Educational Progress (NAEP) was given to sampled students in each state as another way of measuring each state's progress toward proficiency. One U.S. Department of Education guide to NAEP (2003) suggests that

large gaps between student achievement on NAEP and state assessments should prompt investigations concerning the rigor of state standards and/or the validity of the assessments being used. Research has been conducted using the NAEP as a way to determine the rigor of state assessments (Bandeira de Mello et al., 2009; Carnoy & Loeb, 2002; Ercikan, 1997; Gordon, 2009; Kolen & Brennan, 2004; Linn & Kiplinger, 1995; Prowker & Camilli, 2007; Taylor & Gordon, in press; Waltman, 1997). Further research to determine the relationship of student subgroup performance on state assessments and NAEP was warranted. In addition, further research must be done to evaluate the use of a national assessment such as NAEP for the purpose of evaluating school achievement (Gordon).

Theoretical Framework

In 2001, NCLB ushered in an absolute shift in the way schools are managed and measured. However, accountability measures began long before NCLB. In fact, Hansen (1993) traces accountability practices back to 1st century Greek historian Plutarch. Accountability in education was an aspect of President Martin Van Buren's presidency as well. Subsequently, the federal Department of Education was established in 1867 during Andrew Jackson's tenure with the following mission: "to collect such statistics and acts as shall show the condition and progress of education in the several states and territories" (Hansen, p. 12). Hansen asserts that the federal Department of Education was unable to fulfill its mission for most of its existence due to lack of funding; however, in recent years, more federal funds have been to directed toward the accountability cause (Hansen).

Accountability also enjoyed a resurgence during the Sputnik era when the Department of Education sponsored Project TALENT, a project created to use a large sample of schools to analyze student performance "on uniform objective and traditional tests, against such variables as levels of expenditure, size of classes, qualifications of teachers, and student socioeconomic background" (Hansen, p. 12). Project TALENT findings were later used by U.S. Commissioner of Education Francis Keppel in 1965 to justify the need for the Elementary and Secondary Education Act and Title I, which would provide special funding for economically disadvantaged students. It was Project TALENT data that enabled Keppel to show that students from lower socioeconomic levels were performing below students of like age and grade (Hansen).

Manno (2004) cites the mid-1980s as the time when "states and districts in the United States…[began to create] new approaches to school-, district-, and state-level accountability systems that included consequences—rewards and sanctions for performance" (p. 27). With the publication of *A Nation At Risk* (1983), the United States began to see a greater focus on school accountability measures at all levels (U.S. National Commission on Excellence in Education).

The accountability process was designed not merely as a means to measure schools, but primarily as a response to philosophies about the improvements accountability itself could bring to education. Hansen (1993) spells out what he has determined are the four basic assumptions about accountability. First, he says, is the belief that "stricter accountability requirements lead to improvements in education"

(Hansen, p. 13). Second, is the assumption that "meaningful educational improvements can be effected through legislatively mandated accountability" (Hansen, p. 13). Third, he stipulates the assumption that "the most appropriate focal point for accountability-driven reform is the individual school" (Hansen, p. 14). Last, Hansen asserts his fourth assumption: "broad involvement of the school and district community is essential for successful accountability-driven school reform" (p. 15).

Of these four assumptions, Hansen (1993) believes the first assumption to be a necessary condition for all the others to be true. Rustique-Forrester (2005), a scholar writing about school policies in England, echoes Hansen's belief about the first assumption of accountability: "The central assumption of many contemporary accountability schemes is that by holding schools, districts, teachers, and students responsible for results on a range of achievement and performance measures, teaching will improve and expectations for students will rise" (p. 2). About the validity of this assumption there was much debate. Research supports conclusions for both proponents and opponents of the idea that accountability brings achievement increases (Decker & Bolt, 2008).

Accountability measures increased again in 1994 when the Improving America's Schools Act (IASA) was passed. The IASA required states to set rigorous standards and create assessments to measure the achievement of those standards (1994). This new act held both schools and their districts responsible for ensuring that all students achieved, but the act had no major consequences for schools whose data did not measure up to its

challenge. Consequently, Goertz (2005) reports that state responses to IASA were "uneven."

Partially as an antidote to the variety of responses to the IASA, NCLB was enacted at the direction of the U.S. Department of Education (2002). Along with the creation of NCLB came an expansion of the federal role in state educational processes and new requirements for states—testing more, setting higher and more standard goals, and levying sanctions when schools failed to meet the goals (Goertz, 2005). Despite this intent toward uniformity, striking differences remain in the way states have interpreted and responded to NCLB. Differences notwithstanding, NCLB creators clearly increased education standards for states and included provisions for sanctions in an effort to bring about student achievement increases (Goertz, 2005; USDOE, 2002).

NCLB sets forth a number of requirements all with the ultimate goal of improving schools for all students: "The purpose of this title is to ensure that all children have a fair, equal, and significant opportunity to obtain a high-quality education and reach, at a minimum, proficiency on challenging state academic achievement standards and state academic assessments" (U.S. Department of Education, 2001, Section 1001). Specifically, NCLB requires that states create their own accountability systems, including defined standards, regular testing of students, distinct proficiency baselines, and a determination about what scores will be considered "proficient" on these assessments. In addition, schools must make "adequate yearly progress" (AYP) to meet their proficiency

goals. States may set their own AYP goals each year, as long as by 2014 they have set the goal at 100 percent proficiency (U.S. Department of Education, 2002).

NCLB allows states to establish their curriculum standards and standardized tests, thus providing that states establish the difficulty level of their curriculum and tests. States are also able to set their own proficiency levels for each test; therefore, each state determines which students will pass and which will not. In addition, states are given the latitude to set their baseline for proficiency, and their own annual measurable objectives (AMOs) for each year as they decide how their state will make progress toward 100 percent AYP in 2014 (NCLB, 2002, Sec.1111 (2) (F)). In addition, states may decide how many students must be in a subgroup before that group's scores are included in AYP decisions. Each of these decisions play a large role in determining how many of a state's schools will not make AYP each year (Porter, Linn, & Trimble, 2005).

Peterson and Hess (2005) recorded the wide differences between state accountability systems and commented on "the perverse reality" that the AYP results of states with more rigorous accountability systems look worse than states with more lenient accountability systems (p. 53). States with especially rigorous assessments who set a higher cut score for proficiency would understandably classify many more of their schools as in need of improvement; conversely, states with less rigorous assessments and a lower cut score for proficiency would classify fewer schools in need of improvement (Peterson & Hess, p. 53).

Hess (2005) found that "by 2005, some states had virtually no schools identified as needing improvement while other states identified close to 80% of theirs" (pp. 54-55). Casserly (2004) remarked:

> these disparities do not reflect genuine differences in student learning; ...schoolchildren in Boston and San Diego perform similarly on the NAEP, yet 31 percent of Boston's schools are in the improvement process, compared with just 18 percent of San Diego's. Instead, the disparities are the result of Congress's decision to let the states define their own standards of performance. (p. 32)

The 2008 report from the Center on Education Policy suggested "that states with lower standards and easier tests will find it easier to meet the goal of 100% proficiency" (p. 7).

Since 1969, the National Assessment of Educational Progress (NAEP) has been administered to sampled students in every state to serve as a "common yardstick" that measures the academic progress of America's students over time (NCES, 2009c, p. 1). Overseen by the U.S. Commissioner of Education Statistics, students in grades four, eight, and twelve are tested in the areas of "mathematics, reading, science, writing, the arts, civics, economics, geography, and U.S. history" (NCES, 2009c, p. 1). NAEP results are reported nationally, by region, and by state; however, they are not reported for schools or individual students (NCES, 2008, p. 1).

In addition, federal law requires that NAEP instruments be regularly externally evaluated to ensure their reliability (NCES, 2010a). Consequently, the results of NAEP assessments are potentially useful in comparing differences in student achievement by

state and by region, as well as in comparing differences in student achievement over time (NCES, 2010a). Participation in NAEP testing was voluntary for schools unless they receive Title I funds, in which case their participation was required by federal law (NCES, 2010a). Title I, which was concerned with "Improving the Academic Achievement of the Disadvantaged," was established in 1965 with the Elementary and Secondary Schools Act. Title I was enacted "to ensure that all children have a fair, equal, and significant opportunity to obtain a high-quality education and reach, at a minimum, proficiency on challenging State academic achievement standards and state academic assessments" (Title I, Section 101). One way the federal government sought to achieve this purpose was by "distributing and targeting resources sufficiently to make a difference to local educational agencies and schools where needs are greatest;" (Title I, Section 1001 (5)). According to NAEP legislation, schools that receive Title I funds must participate in NAEP testing as a condition of receiving federal funds.

Because NAEP was administered in all 50 states, researchers have an additional set of data with which to compare state accountability systems. Although NAEP and state assessment data are often discussed in conjunction with one another, few researchers have studied the relationships between 8th grade reading proficiency percentages. Most NAEP to state assessment comparisons have been conducted using mathematics data. In one comparative study, Gordon (2009) analyzed the relationship between 2007 NAEP eighth grade reading proficiency percentages and 2007 state reading assessment proficiency percentages and found that state assessment proficiency percentages could be

used as predictors for NAEP proficiency percentages, as well as to predict the performance of some subgroups. Gordon found that the proficiency percentages of low socioeconomic students could only be predicted for one of the four census regions tested by NAEP; however, a strong correlation was found between NAEP proficiency percentages and state assessment proficiency percentages for nonwhite students in each of the four census regions of the country. Despite finding a predictive relationship between the two measures, Gordon determined there were significant disparities between the percent of students performing at proficiency on NAEP and state assessments. In all cases, the percent of students proficient on state assessments exceeded the percent proficient on NAEP by at least 10 percentage points, with most exceeding by at least 25 percent. In the most extreme case, Tennessee students scored at 92% proficient on their state exam, but just 26% scored at proficient on NAEP.

Statement of the Problem

Although NCLB has the goal of ensuring the quality of all schools in the United States, disparities between state assessments and state AYP calculation formulas may misrepresent student performance. NCLB guidelines require that states pattern their tests after NAEP. In fact, one function of NAEP staff was assisting states as they created their own assessments in part by helping states compare their assessments to NAEP as a way of determining validity (Vinovskis, 1998). Furthermore, NCLB requires that each state set challenging standards and create an accountability system to track its schools progress

toward AYP (2002). Furthermore, states are statutorily required to ensure that their definitions of AYP are based on statistically reliable and valid scores, that their assessments are high quality, and that their assessments are consistent with nationally recognized standards (NCLB, 2002).

Each state was given the latitude to design its own test, determine its own proficiency starting point, and decide on the rate of progress that must be made to reach 100% proficiency by 2014. With that in mind, one can see the number of schools meeting AYP (as determined by the number of students performing at proficiency) has much to do with the rigor of these calculations. To be sure, differences in the number of schools meeting AYP from state to state may also show a difference in student achievement. The problem is—without a level playing field, a common assessment, or a way to measure the rigor of the proficiency-determining instrument—one cannot be sure whether proficiency percentage differences from state to state are due to student achievement differences in reading or testing design. A comparison of the percentage of students demonstrating proficiency according to NAEP and state assessment data may show the relationship between the two metrics.

The NAEP offers the potential to provide data to states about the rigor of their assessments; however, more study on the relationship between the NAEP and state assessment reading proficiency percentages must be conducted to gather information. Most of the existing comparisons between NAEP and state assessments compare mathematics data (Carnoy & Loeb, 2002; Ercikan, 1997; Linn & Kiplinger, 1995;

Prowker & Camilli, 2007; Waltman, 1997). Once more study is conducted, legislators and educators will have more information to use in determining whether changes should be made to standards and/or assessments.

Purpose of the Study

This study builds on Gordon's (2009) study, which examined the relationship between 2007 NAEP eighth grade reading proficiency percentages and 2007 state reading assessment proficiency percentages. The purpose of this study was to determine if there was a predictive relationship between 2009 NAEP eighth grade reading assessment proficiency percentages and 2009 eighth grade reading state assessment proficiency percentages. Additionally, data were disaggregated into the four census regions of NAEP to make comparisons between the total populations of each region, as well as certain AYP subgroups. In particular, this study extended Gordon's study to also control for percentage of subgroups meeting proficiency on state assessments. In addition, this study was designed differently to include aggregate census region data, as well as data specific to each census region.

Research Questions

The following research questions were used to guide this study:

1. To what extent can the percentage of eighth grade students demonstrating proficiency on the 2009 NAEP reading assessment be predicted by the percentage of eighth grade students demonstrating proficiency on 2009 state reading assessments?
2. To what extent can the percentage of eighth grade students demonstrating proficiency on the 2009 NAEP reading assessment be predicted by the percentage of eighth grade students demonstrating proficiency on 2009 state reading assessments, controlling for census regions defined by NAEP?
3. To what extent can the percentage of eighth grade students demonstrating proficiency on the 2009 NAEP reading assessment be predicted by the percentage of eighth grade students demonstrating proficiency on 2009 state reading assessments, controlling for the percent proficient on the state exam for each of the five major racial/ethnic groups identified as subgroups by NCLB: American Indian, Asian, Hispanic, Black, White; and controlling for census regions defined by NAEP?
4. To what extent can the percentage of eighth grade ELL students demonstrating proficiency on the 2009 NAEP reading assessment be predicted by the percentage of eighth grade ELL students demonstrating proficiency on 2009 state reading assessments, controlling for census regions defined by NAEP?
5. To what extent can the percentage of eighth grade students qualifying for free and reduced lunch who demonstrate proficiency on the 2009 NAEP reading

assessment be predicted by the percentage of eighth grade students qualifying for free and reduced lunch who demonstrate proficiency on 2009 state reading assessments, controlling for census regions defined by NAEP?

6. To what extent can the percentage of eighth grade students with disabilities demonstrating proficiency on the 2009 NAEP reading assessment be predicted by the percentage of eighth grade students with disabilities demonstrating proficiency on 2009 state reading assessments, controlling for census regions defined by NAEP?
7. On average does the difference between the percentage of eighth grade students demonstrating proficiency on state and NAEP reading assessments change from 2007 to 2009?

Definition of Terms

Adequate Yearly Progress (AYP): "Under the Elementary and Secondary Education Act (ESEA), reauthorized as No Child Left Behind in 2002, each state has developed and implemented measurements for determining whether its schools and local educational agencies (LEAs) are making adequate yearly progress (AYP). AYP was an individual state's measure of progress toward the goal of 100 percent of students achieving to state academic standards in at least reading/language arts and math. It sets the minimum level of proficiency that the state, its school districts, and schools must achieve each year on annual tests and related academic indicators. Parents whose

children are attending Title I (low-income) schools that do not make AYP over a period of years are given options to transfer their child to another school or obtain free tutoring (supplemental educational services)" (U.S. Department of Education, "Adequate Yearly Progress," 2009, p. 1).

AYP Subgroups: "Each school district and school must report their AYP on student bodies as a whole, but also by four different subgroups: Economically disadvantaged; Special education; Limited English Proficient students (also known as ELL---English Language Learners); and Students from major racial/ethnic groups" (Public Education Network and National Coalition for Parent Involvement in Education, 2010).

National Assessment of Educational Progress (NAEP): "NAEP, or the National Assessment of Educational Progress, is often called the 'Nation's Report Card.' It is the only measure of student achievement in the United States where you can compare the performance of students in your state with the performance of students across the nation or in other states. NAEP, sponsored by the U.S. Department of Education, has been conducted for over 30 years. The results are widely reported by the national and local media" (National Center for Education Statistics, "Nation's Report Card: FAQ," 2010, p. 1).

No Child Left Behind Act (NCLB): "The Elementary and Secondary Education Act (ESEA), reauthorized as the No Child Left Behind Act of 2002, is the main federal law affecting education from kindergarten through high school. ESEA is built on four

principles: accountability for results, more choices for parents, greater local control and flexibility, and an emphasis on doing what works based on scientific research" (U.S. Department of Education, "No Child Left Behind," 2009, p. 1).

Proficient (defined by NAEP): **"**One of the three NAEP achievement levels, representing solid academic performance for each grade assessed. Students reaching this level have demonstrated competency over challenging subject matter, including subject-matter knowledge, application of such knowledge to real-world situations, and analytical skills appropriate to the subject matter" (National Center for Education Statistics, "Glossary," 2009, p. 1.

Proficient (defined by NCLB): NCLB A.1.1111(b)(1)(D)(ii)(I-III) specifies that states must develop "challenging student academic achievement standards that are aligned with the State's academic content standards; describe two levels of high achievement (proficient and advanced) that determine how well children are mastering the material in the State academic content standards; and describe a third level of achievement (basic) to provide complete information about the progress of the lower-achieving children toward mastering the proficient and advanced levels of achievement" (U.S. Department of Education, No Child Left Behind Act of 2001, 2002, p. 1).

Methodology

The following details represent a concise summary of this study's methodology; additional details are provided in Chapter 3.

Population and Sample

The population tested includes the eighth grade students sampled by 2009 NAEP tests and the eighth grade students chosen by a state's AYP formula who completed that state's qualifying exams in the same year. All data analyzed were aggregated to the state level, rather than the individual student level because the data provided by the National Center for Education Statistics and the U.S. Department of Education was publicly available in that fashion.

Data Collection

Data from NAEP and state assessments were available through the U.S. Department of Education. All states were required to report their assessment results to the U.S. Department of Education, where the results were then housed on the DOE website in Consolidated State Performance Reports (U.S. Department of Education, 2010). NAEP results were available through the National Center for Education Statistics, which was financed by the DOE (2010b). Once collected, the data were analyzed using SPSS 16.0 software.

Data Analysis

Using 2009 eighth grade state NAEP reading proficiency percentages as the dependent variable and 2009 eighth grade state assessment reading proficiency percentages as the independent variable, Research Question One will be analyzed using simple regression. In all questions controlling for NAEP census region, the Northeast

region will be used as the reference region because this region was typically the highest performing. Research Questions Two through Six will be computed using multiple regression. Research Question Two will control for NAEP census region in the model. Research Question Three will follow that of the second question, however will control for census region, as well as the proficiency percentages of each of the five major racial/ethnic groups identified as subgroups by NCLB. Research Question Four will examine proficiency of ELL students; Research Question Five will examine proficiency of FRL students; and Research Question Six will examine proficiency of SWD. Finally, Research Question Seven will descriptively compare the differences among states between the percentage of eighth grade students demonstrating proficiency on state and NAEP reading assessments from 2007 to 2009.

Limitations

This study includes data aggregated to the state level for eighth grade students enrolled in public schools in all 50 United States. Since the data were aggregated, interpretations can be made at only the state level, not the individual student or school level. In addition, not all school types were included in this study. Also, each state had different rules for which students were included in its test. The sample of students used for NAEP testing was often small and may not have represented a school district (or state) accurately. Moreover, differences in the difficulty of each state's test may affect the percentage of students demonstrating proficiency. Another limitation was that the state

tests were not calibrated to each other, they were not required to assess the same content standards, so one could not assume they have measured the same content. Such differences in standards covered from state to state have not been factored into this analysis.

Significance of the Study

The results of this study will provide useful information for legislators, policymakers, and members of the education community as they consider how to develop guidelines for assessing school achievement. NCLB guidelines require that states pattern their tests after NAEP (2002). Study results may show a relationship between the nationally recognized NAEP test and each state's test results. In addition, this study may show a relationship in the percentage of students considered proficient on state and NAEP assessments and may also show differences in state and national proficiency standards. A relationship between NAEP and state assessments would demonstrate a high degree of shared variance, or concurrent validity. In turn, a high degree of concurrent validity could mean that these assessments are measuring similar constructs. In addition, a high degree of shared variance may suggest that NAEP and state assessments are measuring similar standards. Given that each state's assessment results are required to be patterned after NAEP, significant disparities between percent proficient on each assessment would be of interest to those who seek to hold states accountable for producing the high-quality, rigorous assessments NCLB requires.

The percentage of a state's schools that make AYP will depend partly on student performance, test rigor, and state AYP calculations. Comparing differences in performance on NAEP and state assessments may show that differences in the standards and rigor of each state's assessments are playing a role in the number of schools making AYP. This study may help those legislators, policymakers, and educators who are considering the possibility and the ramifications of a national (rather than state) assessment for AYP calculations or contribute to the discussion about national standards.

CHAPTER TWO: REVIEW OF LITERATURE

Introduction

This chapter presents key reasons for investigating a possible relationship between state assessments and NAEP. Herein is a discussion of the central issues surrounding accountability and research on the use of accountability practices to improve student achievement. Following is a historical overview of school accountability practices, NLCB requirements for assessments, interpretations of AYP, the history and role of NAEP, the methodology and results of studies linking NAEP and state assessment results, as well as the influence of NAEP on state assessments. The process followed in conducting this literature review included study of relevant history, Department of Education and NAEP guidelines, legal statutes and acts pertaining to education reform, and current research on assessments and studies relating assessments. Primary search engines used included Web of Science, World Cat, Dissertation Theses and Abstracts, EBSCO Host, and the ERIC database. In addition, University of Central Florida library resources were examined.

Historical Overview of School Accountability Practices

The No Child Left Behind Act of 2001 substantially changed the way schools are managed and measured. However, accountability measures began long before NCLB. In

fact, some sources (Hansen, 1993) trace accountability practices back to 1st century Greek historian Plutarch, who said:

> Fathers themselves ought every few days to test their children and not rest their hopes on the disposition of a hired teacher: for even those persons will devote more attention to the children if they know they must from time to time render an account. (p. 11)

Early America

Centuries later, American education pioneer Henry Barnard was able to convince President Martin Van Buren to ask questions about national literacy levels on the census of 1840 (Hansen, 1993). Then in 1845, Horace Mann promoted the idea of giving common assessments to Boston schoolchildren to make judgments about Boston schools (Hansen). Subsequently, the U.S. Department of Education was established in 1867 during Andrew Jackson's tenure with the mission of collecting statistics that would show educational progress in the states (Hansen). Hansen asserts that the Department of Education was unable to fulfill its mission for most of its existence due to lack of funding; however, in recent years, more federal funds have been directed toward accountability.

Sputnik Era and Project TALENT

Accountability also enjoyed a resurgence during the Sputnik era as the United States sought educational reform to keep pace with Russian technological progress

(Hansen, 1993). During that same time period, the Department of Education sponsored Project TALENT. Before this project, American school evaluation had been based primarily on input variables (such as number of students taught), rather than output variables (such as grades or assessment scores) (Hansen). Although Project TALENT was not created as a way to measure schools, its focus was on outputs. The project used a large sample of schools to analyze student performance on tests as compared to class size, cost per student, teacher qualifications, and student socioeconomic status (Hansen). Project TALENT findings were later used by U.S. Commissioner of Education Francis Keppel in 1965 to justify the need for the Elementary and Secondary Education Act (which would later evolve into NCLB) and Title I, which would provide special funding for economically disadvantaged students. It was Project TALENT data that enabled Keppel to show that students from lower socioeconomic levels were performing below students of like age and grade (Hansen).

Effects of *A Nation At Risk* on School Accountability

Just after *A Nation At Risk* was published in 1983, the United States began to see a greater focus on school accountability measures at all levels (U.S. National Commission on Excellence in Education). Manno (2004) cites the mid-1980s as the time when policymakers introduced the idea of accountability systems with rewards and sanctions. Some states instituted charter laws as a way to bring competition among schools that would put pressure on schools by giving parents choices about which school their child would attend (Manno, 2004). Public response to *A Nation at Risk* substantially

increased pressure on the federal government to become more involved in educational achievement.

Improving America's Schools Act

Accountability measures increased further in 1994 when the Improving America's Schools Act (IASA) was passed. The IASA required states to set rigorous standards and create assessments to measure the achievement of those standards (1994). This new act held both schools and districts responsible for ensuring that all students achieved, but the act had no major consequences for schools when their data did not measure up to the challenge. Consequently, Goertz (2005) reports that state responses to IASA were "uneven":

> Although all states developed assessments, standards, performance reporting, and, in most cases, consequences for performance, states found different ways to define what it meant for schools to succeed, what indicators to include in their definition of success, and what the consequences would be. (p. 73)

No Child Left Behind Act of 2001

Partially as an antidote to the variety of responses to IASA, the No Child Left Behind Act of 2001 (NCLB) was created out of the Elementary and Secondary Education Act of 1965. Along with the creation of NCLB came an expansion of the federal role in state educational processes and new requirements for states—testing more, setting higher and more standard goals, and levying sanctions when schools failed to meet the goals

(Goertz, 2005). Despite this intent toward uniformity, striking differences remain in the way states have interpreted and responded to NCLB.

Requirements of the No Child Left Behind Act of 2001

The No Child Left Behind Act of 2001 has changed the way educational leaders and politicians look at school reform. NCLB sets forth a number of requirements all with the ultimate goal of improving schools for all students: "The purpose of this title is to ensure that all children have a fair, equal, and significant opportunity to obtain a high-quality education and reach, at a minimum, proficiency on challenging state academic achievement standards and state academic assessments" (U.S. Department of Education, 2001, Section 1001). Furthermore, NCLB has transformed the way people talk about schools, measure educational achievement, and approach challenges (Hess, 2005).

Because another major goal of NCLB was to close achievement gaps between high and low socioeconomic students and between white and nonwhite students, many leaders of large urban districts supported the 2001 act. As a group, Casserly (2004) noted, most urban educators saw NCLB laws as the key piece of legislation in educational history. NCLB's aims to increase achievement, close gaps, and toughen accountability mirrored the goals that urban educators had themselves been working toward (Casserly). At the same time, the sanctions included within this law also made this same group of leaders very nervous (Casserly).

To determine which schools meet the mark and which need assistance, NCLB has

brought about significantly more testing. NCLB states that students in grades 3 through 8, as well as students in at least one grade level in high school, be given state assessments yearly (USDOE, 2002).

Adequate Yearly Progress (AYP)

Although NCLB defined subject areas and grade levels in which accountability assessments are required, the Act gave states freedom to create their own accountability systems, including defined standards, regular testing of students, distinct proficiency baselines, and a determination about what scores would be considered "proficient" on these assessments. In addition, the law required schools to make "adequate yearly progress" (AYP) to meet their proficiency goals; however, states had the liberty to set their own AYP goals each year, as long as by 2014 they set the goal at 100 percent proficiency. NCLB further specified that states have a three-stage plan for their schools that are not meeting state AYP proficiency targets. These stages must include improvement, corrective action, and restructuring.

This Act also provided that policymakers in each state establish curriculum standards and standardized tests, but allowed states to establish the difficulty level of their curriculum and tests. States were also able to set the proficiency level for each test; therefore, each state determined which students would pass and which would not. In addition, states were given the latitude to set the baseline for proficiency and the annual measurable objectives (AMOs) to determine state guidelines for progress toward 100

percent AYP in 2014. States also decided the number that constitutes a subgroup before that subgroup's proficiency percentages were included in AYP decisions. Each of these decisions played a large role in determining how many of a state's schools would not make AYP each year.

Hess (2005) noted that Peterson and Hess (2005) "documented the immense disparity in the rigor of state accountability systems, and the perverse reality that NCLB's AYP requirements make school performance look worse in states with more demanding accountability systems" (p. 53). States with especially rigorous assessments who set a high cut score for proficiency would understandably classify many more of their schools as in need of improvement; conversely, states with less rigorous assessments with a low cut score for proficiency would classify fewer schools in need of improvement (Hess).

Hess (2005) found that "by 2005, some states had virtually no schools identified as needing improvement while other states identified close to 80% of theirs" (pp. 54-55). Casserly (2004) remarked:

> "these disparities do not reflect genuine differences in student learning; ...schoolchildren in Boston and San Diego perform similarly on the National Assessment of Educational Progress, yet 31 percent of Boston's schools are in the improvement process, compared with just 18 percent of San Diego's. Instead, the disparities are the result of Congress's decision to let the states define their own standards of performance" (p. 32).

In addition, decisions about minimum size in a subgroup population and whether to use single scores or confidence intervals played a large role in determining how many schools were subject to sanctions. When Porter et al. (2005) studied Kentucky data to determine the impact of having various AMO baselines, AYP trajectories, subgroup population numbers, and various confidence intervals, they found enormous differences in the number of schools identified in need of improvement based on the method chosen. Using the most lenient plan, 31% of their 2003 schools were identified, versus 90% using the strictest plan (Porter et al., p. 37).

While some states have responded to challenges presented by NCLB by shifting their AMO goals or setting low proficiency targets, other states and advocacy groups have taken more vocal measures to express their displeasure with the Act. A look at what happens when schools failed to meet AYP goals demonstrates why some states have been motivated to avoid such interventions.

Failing to Make AYP

NCLB specified that schools failing to meet AYP goals for two consecutive years must enter the "improvement" stage. During this stage, recommended interventions are writing comprehensive school improvement plans, implementing research-based programs, and contracting with consultants for additional staff development and/or tutoring services. Districts must provide assistance to their schools as they seek to accomplish their goals. Once a school enters the improvement stage, parents are given the opportunity to send their children to another school. If a school does not make AYP after

its first year in the improvement stage, NCLB guidelines allow parents to take advantage of district-provided tutoring services at no cost to the family. If a school did not make AYP after its second year in the improvement stage, the school was entered into the corrective action stage.

In the corrective action stage, the district was to become even more involved in helping to ensure the success of the school. Features of this stage could include the removal of staff, mandated curricula, the revocation of school administrative authority, and the extension of teaching time. Schools that did not make AYP after the first year in the corrective action stage then became subject to the "restructuring" stage (Mintrop & Trujillo, 2005, p. 2).

Restructuring could take the form of "reconstitution, state takeover, conversion into a charter, transfer to a private management company, and similarly radical measures. Thus, a school that fails to improve for five consecutive years ceases to exist in its original form according to NCLB" (Mintrop & Trujillo, 2005, p. 2). Such measures have drawn criticism from some groups. NCLB met some opposition from the urban leaders who had supported it due to the emphasis of the Act on sanctions; in their eyes, they "see less energy being devoted at the federal and state levels to raising achievement than to implementing the law's sanctions" (Casserly, 2004, p. 36). Implementation of NCLB varies widely by state.

In fact, by 2006, 27 percent of school districts in the U.S. had failed to make AYP for two or more consecutive years; in Florida, that same number was a staggering 72

percent of school districts (McLester, p. 20). By 2007, the U.S. Department of Education estimated that more than 1,200 schools had not met their states' requirements for AYP for five straight years; in addition, 800 more schools had not met their states' AYP requirements for four years in a row (Hoff, 2007). Such a difference among states and as compared to the national average prompts one to ask whether current assessments allow a fair comparison of schools making AYP.

National Assessment of Educational Progress (NAEP) Requirements

Since 1969, the National Assessment of Educational Progress (NAEP) has been administered to sampled students in the United States to serve as a "common yardstick" that measures the academic progress of America's students over time (National Center for Education Statistics, "NAEP Overview," 2009, p. 1). Although the test has evolved to become a respected metric after which many states pattern their accountability assessments, a brief history of NAEP demonstrates this was not always the case (Jones & Olkin, 2004).

The Origin of a National Assessment

When the idea of such an assessment was codified in the Elementary and Secondary Education Act of 1965, it was highly criticized by the American Association of Superintendents and teacher groups such as the National Council for Teachers of English (Vinovskis, 1998). Many educators feared that U.S. Commissioner of Education

Francis Keppel was exaggerating concerns over the lack of available data on American academics to strengthen federal power over education (Vinovskis). U.S. founding fathers who established the groundwork for education gave full responsibility to individual states and regarded any federal intervention as intrusive (Bourque, 2009). Because of this suspicion, Keppel and other proponents of a national assessment endorsed the idea of reporting test results in four geographic regions, rather than at the state level. NAEP reports test results in four census regions: Northeast, South, Midwest, and West.

In 1964, Ralph W. Tyler, the chairman of the Exploratory Committee on Assessing the Progress of Education (the initial committee appointed to study the idea of a national assessment), assured that reporting results regionally would be a way to ensure that states were not compared to one another (Vinovskis, 1998). At the time, it was important to legislators and educators that NAEP be designed so that states' rights were protected (Bourque, 2009). Although reporting by geographic region rather than by state resolved this issue, many argued that regional data also made NAEP results ambiguous and less useful for their original intent—improving education.

From Regional to State Accountability

In the 1980s, NAEP reform efforts centered on the issue of developing an assessment with results reported by state rather than region. When Secretary of Education William J. Bennett brought 22 governmental and educational leaders together in 1986 to form a NAEP study group headed by Tennessee Governor Lamar Alexander and Spencer Foundation President H. Thomas James, the Alexander-James leaders recommended

changes for NAEP. According to Vinovskis (1998), the group believed that the national and regional information on student achievement that NAEP provided was not sufficient for the purpose of holding states accountable for the responsibility of educating that had been entrusted to them. They sought a test that would make possible comparisons between states. In the mid- to late-1980s, many governors and state legislators began to welcome the idea of a state-level NAEP because they needed data behind the public fervor to improve schools that followed the popularization of *A Nation at Risk* (Vinovskis). Accordingly, the Alexander-James group proposed expanding NAEP to include fourth, eighth, and twelfth grades in a range of subjects.

However, when the Department of Education asked the National Academy of Education to form a special committee to review the Alexander-James report, Chairman Robert Glaser opposed this expansion:

> We are concerned about… state-by-state comparisons of average test scores. Many factors influence the relative rankings of states, districts, and schools. Simple comparisons are ripe for abuse and are unlikely to inform meaningful school improvement efforts…. The ability of a state or locality to examine its progress over time is much more informative than the comparison with other states or localities at any one point in time.… The simple ranking of geographic units by achievement levels is rarely informative. Not surprisingly, schools with greater resources and fewer problem students routinely fill the upper ranks. So what have we learned? (Vinovskis, 1998, p. 15)

Despite some remaining opposition to a national assessment with state-level data from the National Parent-Teachers Association and the American Association of School Administrators, Congress enacted legislation to bring about state-level NAEP with the Augustus F. Hawkins-Robert T. Stafford Elementary and Secondary School Improvement Amendments of 1988 (Public Law 100-297). With this new legislation, NAEP was statutorily required to collect and report data by state on reading and mathematics at least every two years, on writing and science at least every four years, and on history, geography, and other subject areas at least every six years (Public Law 100-297, Sec. 3404. (a) (i) (2) (A) (i-iv)). This amendment also expanded NAEP from being administered solely to 9-, 13- and 17-year-olds to being administered to fourth-, eighth-, and twelfth-grade students. As of 2010, NAEP continues to test students in these three grade levels in the areas of "mathematics, reading, science, writing, the arts, civics, economics, geography, and U.S. history" (NCES "NAEP Overview," p. 1). NAEP results are reported nationally, by region, and by state; however, they are not [publicly] reported for schools or individual students (NCES, "NAEP Technical Documentation," 2008, p. 1).

Governance of NAEP

Because power over a national assessment has always been a contested issue, NAEP's founders had an interest in maintaining a sense of distance between Congress, the U.S. Department of Education, and NAEP. Initially, NAEP was monitored by an educational research program, the National Center for Educational Research and

Development. Over the years, monitoring power over NAEP has shifted. In the early 1970s, the National Center for Educational Statistics (NCES) became the new monitoring body for NAEP. Although NCES continues to play an important role with NAEP, Congress wished to create an additional layer between U.S. agencies and NAEP.

In 1978, that wish manifested itself in the creation of a 17-member Assessment Policy Committee. Although the assessment was overseen by the U.S. Commissioner of Education Statistics and various aspects of assessment administration and analysis are contracted out, the Assessment Policy Committee was given the role of being a non-governmental body to oversee NAEP (Vinovskis, 1998). This committee was composed of "two representatives of business and industry, three from the general public, four classroom teachers, two state legislators, two school district superintendents, one state governor, one chair of a state board of education, one chair of a local school board, and one chief state school officer" (Vinovskis, p. 9). Public Law 95-561 created this committee to be responsible for NAEP's design and to ensure its "validity, effectiveness, and utilization" (Vinovskis, p.9).

In 1988, the Elementary and Secondary School Improvement Amendments replaced the Assessment Policy Committee with the 20-member National Assessment Governing Board. This Board was much like the prior committee except that the new body contains an additional governor of a different political party, an additional chief state school officer, one school district superintendent rather than two, three classroom teachers rather than four, and one representative of business rather than two. Newly

specified members include two curriculum specialists, two testing experts, one nonpublic school administrator, and two principals (Public Law 100-297, Sec. 3403, (2) (C) (5) (B) (i-xiii)).

With the shifting in overseeing governing bodies, as well as the involvement of multiple government agencies it is not surprising that role confusion has been an issue for these groups. Vinovskis (1998) asserted that there has been continued tension between NAGB and NCES as to which group has authority over which issues.

Role of NAEP

In addition to the role confusion that has existed between the various groups charged with aspects of control over NAEP, there has also been continued debate about the role NAEP should play in education. Bourque & Hambleton (1993) highlighted this confusion and called NAEP's assessment frameworks "a delicate balance between what is and what should be or will be" (p. 42). Overall, the results of NAEP assessments have been useful in comparing differences in student achievement by state and by region, as well as in comparing differences in student achievement over time (NCES, "The Nation's Report Card: FAQ," 2010). Participation in NAEP testing was voluntary for schools unless they received Title I funds, in which case their participation was required by federal law (NCES, "Nation's Report Card: FAQ", 2010). In addition, federal law requires that NAEP instruments be regularly externally evaluated to ensure their reliability (NCES, "Nation's Report Card: FAQ", 2010).

No Child Left Behind guidelines specifically limit the role of NAEP. Data from NAEP are not to be used "to rank, compare, or otherwise evaluate individual students or teachers or to provide rewards or sanctions for individual students, teachers, schools or local educational agencies" (Public Law 107-110, Sec. 411 (B) (4) (A)). Nor are NAEP assessment results to be used "to establish, require, or influence the standards, assessments, curriculum, including lesson plans, textbooks, or classroom materials, or instructional practices of States or local educational agencies" (Sec. 411 (B) (4) (B)). Although this definition limits the role of NAEP to an extent, the federal dollars appropriated to finance the NAEP assessment system underscore its importance.

Financing NAEP

As NAEP has become of more interest to legislators, Congressional coffers have allocated increasingly large amounts to the task of compiling "The Nation's Report Card." Before 1968, the funding to establish a national assessment came from the Carnegie Foundation (Vinovskis, 1998). The year 1968, however, marked a turning point for federal dollars to begin pouring into national assessments; in that year, $372,358 was appropriated by the federal government. In 1969, the federal appropriation was $1 million. In 1970, Congress put forward $2.4 million. In 1972, the figure rose to $4.5 million (Vinovskis). In 1973, $6 million was given; however, 1979 saw a decrease in funding to $4.3 million and 1980 marked a further decline to $3.9 million, which remained the appropriation until 1984.

When funds were allocated for 2003, more than $112 million was earmarked to fund NAEP testing and its National Assessment Governing Board (Public Law 107-279, Sec. 305 (A) (1) (A-B)).

Proficiency Definitions for NAEP Assessments

Although the concept of proficiency levels sounds innocuous enough, the function of these levels has been a great source of debate. When NAEP was reauthorized in 1988 with Public Law 100-297, the NAGB was given the duty of identifying appropriate achievement goals, which led to the creation of an achievement level system (Bourque & Hambleton, 1993). Subsequently, when the NAGB came up with the original definition of "proficient" in 1990, the Board was criticized for reaching beyond what could be measured by one sitting of NAEP. The initial definition was as follows:

> **Proficient.** This central level represents solid academic performance for each grade tested—4, 8, and 12. It will reflect a consensus that students reaching this level have demonstrated competency over challenging subject matter and are well prepared for the next level of schooling. For grade 12, the proficient level will encompass a body of subject-matter knowledge and analytical skills, cultural literacy, and insight that all high school graduates should have for democratic citizenship, responsible adulthood, and productive work. (Bourque, 2009, p. 8)

As cited in Bourque (2009), Linn et al. (1991) and Stufflebeam et al. (1991) felt that the above definition of proficiency included "predictive statements… that could not be

validated using NAEP data" (Bourque, 2009, p. 10). As a result, the proficiency definition, as well as the definitions for each achievement level, were rewritten:

> **Proficient.** This level represents solid academic performance for each grade assessed. Students reaching this level have demonstrated competency over challenging subject matter, including subject-matter knowledge, application of such knowledge to real-world situations, and analytical skills appropriate to the subject matter. (Bourque, p. 11)

This definition of proficiency for NAEP was approved and remains in use.

While it may seem that the differences in the two definitions were a purely semantic matter, they were hotly debated because their principal functions were to "serve policy decision-making efforts at the local, state, and federal levels…and [to] serve as a model for state assessments under NCLB" (Bourque, 2009, p. 26). When achievement levels were addressed in NCLB, it was required that states also set standards for their state assessments using NAEP achievement levels (Basic, Proficient, and Advanced (NCLB, 2002). Examining relationships between proficiency levels on NAEP and state assessments was one way that researchers have sought to compare assessment results.

Studies on the Relating of Scores from NAEP and State Assessments

There are a variety of methodologies that can be used when seeking to compare the results of distinct tests. Although different researchers vary in the way they have defined their terms, assessment-relating studies have been conducted for a number of years. In

general, these researchers have suggested caution about the use of results from relating distinct assessments, simply because it is difficult to compare the performance of different students taking different tests with different standards. However, assessment-relating studies continue to be conducted because stakeholders want to be able to compare states' performance on assessments.

Mislevy (1992) asked the important question: "Can we measure progress toward national goals if different students take different tests?" (p. 72). He concluded that the answer to this question is not simple. However, Mislevy offered some thoughts about what is possible when conducting studies to relate assessments. He found that it is possible to make projections about the performance of a subgroup on an assessment using the results from another assessment. Still, he cautioned that the results could be unstable and might shift over time. Thus, Mislevy recommended that assessment-relating analyses be conducted regularly to strengthen findings (Mislevy).

The Concept and Brief History of Relating Distinct Assessments

One of the key questions for those seeking to relate assessment results centered around whether the assessments have enough in common to be related. Kolen and Brennan (2004) developed four features to analyze when determining whether and/or how tests results can be related:

Inferences: To what extent are scores for the two tests used to draw similar inferences?

Constructs: To what extent do the two tests measure the same constructs?

> Populations: To what extent are the two tests designed to be used with the same population?
>
> Measurement conditions: To what extent do the tests share common measurement conditions, including, for example, test length, test format, administration conditions, etc. (p. 224)

Using these criteria, Kolen and Brennan concluded that NAEP and state assessments were dissimilar in all areas.

When Kolen (2004) researched the concept and history of relating assessments, he studied the contributions of Mislevy (1992) and Linn (1993) to the body of knowledge about relating assessments. Kolen referenced the Mislevy/Linn framework, which provided a conceptual model for relating assessments based on four methods: equating, calibration, statistical moderation, and projection. Mislevy and Linn discussed using regression methodology as a way to project, or predict scores from one test by using scores from another assessment (Mislevy, 1992; Linn, 1993). Specifically, Linn said that the degree to which one assessment is comparable to another depends on how similar are the tests' questions, conditions, and cognitive complexity (Linn). Both Mislevy and Linn emphasized that the results from projection studies were situation, time, and group dependent. For that reason, this study investigated the relationship between assessment results for each subgroup, rather than relying on the results of the total population. Dorans (2004) of Educational Testing Service also suggested that regression methodology be used when comparing results among tests that measure different

constructs.

Dorans (2004) offered another way to approach relating assessments. He discussed three ways that scores from different assessments can be compared: "equating, concordance, and prediction of expected scores" (Dorans, p. 227). Dorans explained the framework he developed to determine whether tests could be related: Equating, he says, has the goal of producing interchangeable scores. To equate scores, two assessments must measure the same construct and must be expressed using the same metric (Dorans). Concordant scores, however, are placed on a common metric and must be symmetric but need not be interchangeable. According to Dorans, the least restrictive, least demanding method to relate assessments is to use the expected or predicted score. When predicting scores, Dorans reported that the goal is to minimize any error in the prediction. Dorans described three processes that one must go through when evaluating comparisons between assessment scores. First, the similarity of the construct being measured by the two assessments must be determined. Second, the researcher should conduct a correlation to determine the strength of the relationship between the two assessments. Third, the researcher should determine the degree to which the relationship varies across population subgroups. Dorans stipulated that the construct must be the same, the correlation high, and the subgroup results invariant. This study included each of Dorans's three processes as it sought to evaluate the relationship between NAEP and state reading assessments for eighth grade students.

Though Dorans (2004) detailed three different methods for relating scores from

different assessments, other researchers have not been as confident about the validity of results obtained from assessment-relating studies. In his text *Uncommon measures: Equivalence and linkage among educational tests,* Feuer (1999) was generally cautious about the ability to relate scores from different assessments. This work set the tone for those that followed. He brought up a variety of factors that might affect the validity of assessment-relating studies: assessment content, assessment format, measurement error, and assessment result use, to name a few. Feuer said the analysis required to achieve valid comparative results would not be economically feasible for most states.

Pommerich, Hanson, Harris, and Sconing (2004) saw a cause for concern when assessment-relating studies were conducted between tests whose scores could not be equated. They suggested that there would be a greater likelihood that related scores could be misused or misinterpreted; they cautioned that it would be harder to use the results in a clear-cut way. Despite these reservations, Pommerich et al. do see merit in relating test results for distinct tests—provided that caution is used and their guidelines followed. Pommerich et al. proposed four steps to follow when conducting the relating process:

1. Choosing an appropriate type and methodology for relating assessments
2. Relating scores and computing summary measures
3. Evaluating the quality of the relationship and determining what to report
4. Making recommendations for the interpretation and use of the assssement-relating results (p. 248)

In addition, this group of researchers recommended being careful to report only those

results about which there can be the least misinterpretation and also making specific recommendations about how assessment-relating results can be used. They expressed concern that those who obtain relating results are often unskilled in assessment methodology and may not realize when they are misinterpreting or misusing results.

Studies Involving the Relating of NAEP and State Assessments

Other analysts, such as Linn and Kiplinger (1995), see potential in relating scores to compare assessment results. Linn and Kiplinger studied eighth grade mathematics scores on NAEP and state assessments. Specifically, they studied whether relationships between tests varied for males and female subgroups. Although Linn and Kiplinger expressed reservations about the usefulness of comparing these assessments for very high or very low scores; they found their assessment-relating results to be useful for making estimates about average state performance.

In one relating study, Ercikan (1997) discussed the accuracy of comparing NAEP and state assessments. It was suggested that caution be used when comparing proficiency percentages on these metrics because they all test different standards. Ercikan's study used eighth grade mathematics scores as a basis for comparison. Waltman (1997) also conducted a relating study to determine the relationship between fourth grade mathematics scores from NAEP and the Iowa Test of Basic Skills (ITBS). Waltman concluded that her results were more stable than those reported in her colleagues' (Ercikan; Linn & Kiplinger, 1995) studies, perhaps because fourth graders demonstrate less variability in motivation from test to test (Waltman, 1997).

In another study, Carnoy and Loeb (2002) investigated the relationship between NAEP mathematics scores and the strength of state accountability systems. When they compared the results of 1996-2000 NAEP fourth and eighth grade mathematics results for all states, Carnoy and Loeb found that students in high-accountability states significantly outperformed students in states with less rigorous accountability standards (2002). To determine this relationship, Carnoy and Loeb created a zero- to five-point index to quantify the relative strength of state accountability systems and compared that number to NAEP scores. Although Carnoy and Loeb's study did not take specific state assessment scores into account, their study does have important implications. This study's results suggest that one will find a positive relationship between NAEP proficiency percentages and assessment proficiency percentages in states that have higher accountability standards and correspondingly more difficult tests. Conversely, their study suggests that lower accountability standards and less rigorous assessments would lead to lower NAEP proficiency percentages.

Other researchers, Prowker and Camilli (2007), built on the work of Feuer (1999). They cautioned that there was no way to account for the variety of policy factors that affect performance on state and NAEP assessments and stated that it was not possible to obtain an accurate picture of states' achievement results simply by comparing NAEP proficiency percentages among states. Prowker and Camilli used a method that involved looking at individual NAEP items, rather than composite scores. By using differential item functioning in their study of states' performance NAEP mathematics items, they

were able to make conclusions about states' performance on NAEP as compared to their level of proficiency on NAEP. For instance, Prowker and Camilli concluded that Texas students perform better on lower-difficulty items, while California students showed better performance on higher-difficulty items. Interestingly, California's overall proficiency score was lower than Texas's, but this method offered a way to deconstruct performance with greater detail. By looking at the items on which student scored best, Prowker and Camilli then made inferences about which content standards each state had emphasized. These two researchers believed that differential item functioning provides a more complete method for understanding states' performance on NAEP (Prowker & Camilli, 2007).

Gordon (2009) compared NAEP and state definitions of proficiency in his study of 2007 eighth-grade reading proficiency percentages. Specifically, he analyzed the relationship between 2007 NAEP eighth grade reading proficiency percentages and 2007 state reading assessment proficiency percentages and found that state assessment proficiency percentages could be used as predictors for NAEP proficiency percentages, as well as to predict the performance of some subgroups. Gordon found that the proficiency percentages of low socioeconomic students could be predicted for only one of the four census regions tested by NAEP; however, a strong correlation was found between NAEP and state assessment proficiency percentages for nonwhite students in all regions of the country (Gordon).

Despite finding a predictive relationship between the two measures, Gordon determined there were significant disparities between the percent of students performing at proficiency on NAEP and state assessments. In all cases, the percent of students proficient on state assessments exceeded the percent proficient on NAEP by at least 10 percentage points, with most exceeding by at least 25 percent. In the most extreme case, Tennessee students scored at 92% proficient on their state exam, but just 26% scored at proficient on NAEP (Gordon, 2009). Although Tennessee has adopted proficiency levels in accordance with NCLB legislation that are the same as NAEP's proficiency levels, there was clearly a difference in rigor between the two assessments' definition of what skills a "proficient" student can perform. Accordingly, "because of the variation in assessments and where proficiency is set, state to state comparisons are not meaningful" (Taylor & Gordon, in press, p. 3).

In their 2009 effort to create more meaningful comparisons, Bandeira de Mello, Blankenship, and McLaughlin (2009) published a report for NCES called, "Mapping State Proficiency Standards Onto NAEP Scales: 2005-2007." This report details the work that NCES has done toward "mapping each state's standard for proficient performance onto a common scale" (Bandeira de Mello et al., p. v). In short, each state's proficient score was given a numeric value (235 or 263, for example) according to the level of rigor required to reach that level of achievement on the NAEP assessment. In establishing this common scale, "the level of achievement required for proficient performance in one state can then be compared with the level of achievement required in another state" and

proficiency standards among states can be compared (Bandeira de Mello et al., p. v). When Bandeira de Mello et al. compared each state's assessment results to NAEP results in eighth grade reading, the correlation was approximately .7 or more with most states. Moreover, this correlation statistic held true for at least half the states in reading at both fourth and eighth grade levels (Bandeira de Mello et al., p. 21).

The Influence of NAEP on State Assessments

NAEP has had a profound influence on state assessments not only because NCLB required that NAEP levels be used for states to pattern their own after, but also because the assessment gave states more than 30 years of experience with testing issues from which to learn. In fact, one function of NAEP staff was assisting states as they created their own assessments (Vinovskis, 1998). In 1984, the National Assessment Policy Committee (the committee charged with overseeing NAEP before NAGB) agreed to help states compare their state level assessments to NAEP as a way of determining validity (Vinovskis).

NCLB requires that each state set challenging standards and create an "accountability system that will be effective in ensuring that all local educational agencies, public elementary schools, and public secondary schools make adequate yearly progress" (Public Law 107-110, Sec. 1111 (2) (A)). Furthermore, states are statutorily required to ensure that " 'adequate yearly progress'… [was] defined… in a manner that… is statistically valid and reliable" (Sec. 1111 (2) (C) (ii)). In addition, states must demonstrate that they have implemented "high-quality, yearly student academic

assessments" (Sec. 1111 (3) (A)) that will be used as the primary means of determining adequate yearly progress. Such assessments must "be used for purposes for which such assessments are valid and reliable, and be consistent with relevant, nationally recognized professional and technical standards" (Sec. 1111 (3) (A) (iii)). Moreover, these state assessments may be used only if the state provides "evidence from the test publisher or other relevant sources that the assessments used are of adequate technical quality for each purpose required under this Act… and [if] such evidence is made public… upon request" (Sec. 1111 (3) (A) (iv)).

NCLB legislation provides that states develop performance standards by using NAEP's performance standards as a guide. Specifically, the law requires that tests be designed to yield scores on at least three achievement levels (below proficient, at proficient, above proficient) to show a spectrum of student performance results. Bourque's (2009) review of performance standards among states found that "12 states use a 5-level system, 29 use a 4-level system, 10 use a 3-level system, and 1 uses a 6-level system" (p. 23). One issue of discrepancy arises in where states position the "proficient" level among their levels. Bourque determined that "states with three or four levels…positioned… 'Proficient' at the second highest level" whereas 9 of the 13 states using 5 or 6 levels "positioned… 'Proficient' at the third highest level" (p. 23). Bourque notes that the difference in levels among assessments and the corresponding placement of the proficiency level has "the likely effect of depressing the definition of Proficient…. [and] the definition of Proficient can vary from state to state" (p. 23). It was Bourque's

opinion that having a consistent definition and positioning of proficiency "would go a long way to resolving the disparity between NAEP results for the states and the states' performance on their approved NCLB assessments" (p. 23).

Recent methods developed by NCES (Bandeira de Mello et al., 2009) have made it possible for NAEP to serve as a metric against which state assessments can be measured and compared. In this area of comparing states, however, NCES must tread lightly due to prohibitions in NCLB legislation that limit the use of NAEP for comparing states (Public Law 107-110, Sec. 411 (B) (4) (A)).

Notwithstanding, mapping the assessments in the way described by Bandeira de Mello et al. "offers an approximate way to assess the relative rigor of the states' adequate yearly progress (AYP) standards" and "the NAEP scale equivalent score representing the state's proficiency standards can be compared to indicate the relative rigor of those standards" (p. v). Given the limitations NCLB has placed on using NAEP for comparing state assessments, Bandeira de Mello et al. offer the following qualification: "The term rigor as used here does not imply a judgment about state standards. Rather, it is intended to be descriptive of state-to-state variation in the location of the state standards on a common metric" (p. v). Neither Gordon's study nor this study uses the kind of mapping method developed by Badeira de Mello et al.; however, the Bandeira de Mello study was similar to this study in that both use NAEP to allow for comparison of state assessments. All three studies use NAEP as a measure to compare state assessment proficiency results.

One parent's guide to NAEP discusses the usefulness of the assessment in the

following manner: “NAEP data will highlight the rigor of standards and tests for individual states: If there is a large discrepancy between children’s proficiency on a state’s tests and their performance on NAEP, that would suggest that the state needs to take a closer look at its standards and assessments and consider making improvements” (U.S. Department of Education, 2003, p. 14).

These words from the parent guide seem to oversimplify the relationship between NAEP and state assessments, which are after all totally different assessments each developed to measure often disparate standards. Nevertheless, this message from a parent guide exemplifies the public perception about performance on NAEP and state assessments: That is, the public expects that what is proficient on a state’s test will also be proficient on NAEP.

The work of Hombo (2003) explored the difficulty of comparing results of state assessments and offered NAEP as a practical way to quantify comparisons. Hombo noted: “Some states use… commercial tests while others have developed their own state assessments. Because the assessments are not comparable, comparisons across states cannot be validly made” (p. 59). In the face of this difficulty, Hombo suggested that “NAEP provides the missing common measure of student achievement so that state-to-state comparisons can be made” (p. 59). In other words, “NAEP has a new role: to act as a serious discussion tool in evaluating results of state assessments, and in providing a common base for comparisons between states” (p. 59). Here, Hombo’s approach demonstrates a more practical, informed look at comparing assessment scores while still

underscoring the importance of NAEP.

Summary

The central question of accountability was whether those who entrust educators to teach students are getting the achievement results they desire. In financial terms, those who financially support education are interested in evaluating whether they are getting their money's worth, and whether their money was being directed appropriately. Accountability assessments have been used as a way to allow money holders to distribute funds as a reward for good performance or as a way to help low performers.

In education, accountability has long been used as a way to ensure that schooling was taking place as planned. Literacy levels were a national question as early as 1840, and assessments were used to make judgments about schools by 1845 (Hansen, 1993). School accountability has been an interest of the U.S. federal government at least as early as 1867 when the U.S. Department of Education was established under Andrew Jackson with the mission to collect facts and figures to show the state of schools. International pressures brought accountability to a more serious level during the Sputnik era with Project TALENT. This project was one of the first to disaggregate results to show the performance of economically disadvantaged students and the need for more funding in underperforming areas (Hansen). With the publication of *A Nation at Risk* in 1983 came the popularization of the idea of using rewards and sanctions as part of the accountability process (Manno, 2004). Since then, U.S. accountability efforts have focused on steadily

increasing pressure on schools (Goertz, 2005). In 1994, the IASA required states to be accountable to standards, but allowed each state to set its own standards; as such, resulting standards were characterized by unevenness (Goertz, 2005). In 2001, the inception of NCLB led to an increase in sanctions against underperforming schools, but some of the unevenness between state results present with the IASA remains in NCLB (Goertz, 2005).

NCLB's main goal was high quality education for all and closing achievement gaps, a goal many urban educators were initially excited about (Casserly, 2004). NCLB defined subject areas and grade levels in which accountability assessments are required but allows states to create their own accountability requirements and to determine what AYP means in their states. States must also create their own sanctions for states not performing up to par; however, NCLB does suggest some recommended interventions. This decision to increase the federal role in education but to allow states the freedom to create their own accountability assessments has resulted in widely varying interpretations about what constitutes AYP from state to state (Hess, 2005).

Differences in middle school reading proficiency percentages highlight AYP divergence issues (Gordon). Differences in the percentage of schools identified as being in need of state intervention prompt one to question whether current assessment system was allowing a fair comparison of schools making AYP (Casserly, 2004; Hess, 2005; Hoff, 2007; Porter et al, 2005).

Key components of change that works include teacher commitment and trust among all stakeholders (Burke, 1996; Mintrop & Trujillo, 2005); however, these each take more time than accountability timelines allow. All the resources spent on intervention demonstrate the importance of ensuring that accountability assessments are accurately identifying struggling schools from state to state, rather than allowing an over- or under-rigorous assessment to squander state and federal dollars. Some believe one key part of this includes standardizing assessments (Jones & Olkin, 2004).

All but three states have opposed NCLB laws in some fashion, either by applying for waivers or by refusing to comply with the law. State and federal legislators alike have concerns that there are not enough government funds to implement NCLB mandates. The high stakes of failing to make AYP and the highly charged public reactions have prompted some to look for national assessments such as NAEP that might standardize the accountability process (Jones & Olkin, 2004).

Although some look toward NAEP as a possible solution to the variation among state assessment systems, NAEP must walk a fine line in discussions about state-to-state comparisons. While NAEP was crafted with the capability of comparing state achievement, specific legislative language prohibits the use of NAEP for this purpose. When NAEP was first created, educational leaders feared that the assessment would one day be used as a tool for federal intervention and would limit states' rights (Bourque, 2009). Initially, NAEP proficiency percentages were reported by the census regions only; not until 1988 was the authority granted to report scores at the state level (Vinovskis,

1998). Because of the power this national assessment contains for policymakers, NAEP governance has been an issue since the test's beginning (Vinovskis, 1998). In addition, what should be the proper role of this national assessment has continued to be debated throughout its existence (Vinovskis, 1998; Bourque & Hambleton, 1993). Nevertheless, hundreds of millions of dollars are spent administering this exam (Public Law 107-279, Sec. 305 (A) (1) (A-B)).

Details over what constitutes "proficiency" and which scale should be used to measure it have been a recurring theme in NAEP history (Bourque & Hambleton, 1993; Bourque, 2009). Once a definition was settled, NCLB creators required that states pattern their own proficiency scales after the one in use by NAEP (NCLB, 2002). Despite this requirement, states vary widely in their use of proficiency results, and in their reporting of proficiency percentages as well (Gordon, 2009). As one way to make sense of the range of proficiency percentages reported by state assessments, researchers have sought to tie state assessment proficiency percentages to NAEP proficiency percentages.

The result was a growing body of research and recommendations on relating assessments, specifically on relating state and NAEP tests (Bandeira de Mello et al., 2009; Carnoy & Loeb, 2002; Dorans, 2004; Ercikan, 1997; Feuer, 1999; Gordon, 2009; Hombo, 2003; Kolen & Brennan, 2004; Linn, 1993; Linn & Kiplinger, 1995; Mislevy, 1992; Pommerich et al., 2004; Prowker & Camilli, 2007; Taylor & Gordon, in press; Waltman, 1997). Despite numerous recommendations that NAEP and state assessments

are a poor fit for relating (Kolen & Brennan, 2004), researchers have persisted in making these comparisons.

Assessment-relating studies and their results aside, NAEP has had a profound impact on state assessments. Thirty years of assessment experience offered much from which state assessment designers could learn (Vinovskis, 1998). Therefore, the NCLB requirement that states pattern their assessments after NAEP does not come as a surprise.

CHAPTER THREE: METHODOLOGY

Introduction

The primary goal of this study was to determine to what extent eighth grade state reading assessment proficiency percentages could be used to predict eighth grade NAEP reading assessment proficiency percentages. Seven research questions that relate to this issue were developed to determine whether predictions could be made across subgroups and across census regions. The methodology used to test these research questions is detailed within.

First, the statement of the problem that gave rise to this study is restated. Next, the research questions are presented, as well as the population and sample used for this study. An in-depth exploration of the instrumentation for NAEP and state assessments is given. An analysis of representative assessments for each of the four census regions is included as follows: New York, Northeast; Texas, South; Illinois, Midwest; and California, West. This analysis is followed by data source information and a detailed discussion of data analysis procedures to be used with each question. Last, a full list of variables is stated.

Statement of the Problem

Although NCLB has the goal of ensuring the quality of all schools in the United States, disparities between state assessments and state AYP calculation formulas may misrepresent student performance. NCLB guidelines required that states pattern their

tests after NAEP. In fact, one function of NAEP staff was assisting states as they created their own assessments in part by helping states compare their assessments to NAEP as a way of determining validity (Vinovskis, 1998). Furthermore, NCLB required that each state set challenging standards and create an accountability system to track its schools' progress toward AYP (USDOE, 2002). Furthermore, states were statutorily required to ensure that their definitions of AYP were based on statistically reliable and valid scores, that their assessments were high quality, and that their assessments were consistent with nationally recognized standards (USDOE).

Each state was given the latitude to design its own test, determine its own proficiency starting point, and decide on the rate of progress needed to reach 100% proficiency by 2014. With that in mind, one can see that the number of schools meeting AYP (as determined by the number of students performing at proficiency) had much to do with the rigor of these calculations. To be sure, differences in the number of schools meeting AYP from state to state may also show a difference in student achievement. The problem is—without a level playing field, a common assessment, or a way to measure the rigor of the proficiency-determining instrument—one cannot be sure whether proficiency percentage differences from state to state are due to student achievement differences in reading or testing design. A comparison of the percentage of students demonstrating proficiency according to NAEP and state assessment data may show the relationship between the two metrics.

The NAEP offers the potential to provide data to states about the rigor of their assessments; however, more study on the relationship between the NAEP and state assessment reading proficiency percentages must be conducted to gather information. Once more study is conducted, legislators and education will be equipped with the information they need to determine what kinds of changes are necessary to standards and/or assessments, if any.

Research Questions

The following research questions guided this study:

1. To what extent can the percentage of eighth grade students demonstrating proficiency on the 2009 NAEP reading assessment be predicted by the percentage of eighth grade students demonstrating proficiency on the 2009 state reading assessment?
2. To what extent can the percentage of eighth grade students demonstrating proficiency on the 2009 NAEP reading assessment be predicted by the percentage of eighth grade students demonstrating proficiency on the 2009 state reading assessment, controlling for census regions defined by NAEP?
3. To what extent can the percentage of eighth grade students demonstrating proficiency on the 2009 NAEP reading assessment be predicted by the percentage of eighth grade students demonstrating proficiency on the 2009 state reading assessment, controlling for each of the five major racial/ethnic groups identified

as subgroups by NCLB: American Indian, Asian, Hispanic, Black, White; and controlling for census regions defined by NAEP?

4. To what extent can the percentage of eighth grade ELL students demonstrating proficiency on the 2009 NAEP reading assessment be predicted by the percentage of eighth grade ELL students demonstrating proficiency on the 2009 state reading assessment, controlling for census regions defined by NAEP?
5. To what extent can the percentage of eighth grade students qualifying for free and reduced lunch who demonstrate proficiency on the 2009 NAEP reading assessment be predicted by the percentage of eighth grade students qualifying for free and reduced lunch who demonstrate proficiency on 2009 state reading assessments, controlling for census regions defined by NAEP?
6. To what extent can the percentage of eighth grade students with disabilities demonstrating proficiency on the 2009 NAEP reading assessment be predicted by the percentage of eighth grade students with disabilities demonstrating proficiency on the 2009 state reading assessment, controlling for census regions defined by NAEP?
7. On average does the difference between the percentage of eighth grade students demonstrating proficiency on state and NAEP reading assessments change from 2007 to 2009?

Population and Sample

The population tested included the eighth grade students sampled by 2009 NAEP tests and the eighth grade students chosen by a state's AYP formula who completed that state's qualifying exams in the same year. All data analyzed were aggregated to the state level, rather than the individual student level.

Instrumentation

Data from NAEP and state assessments were available through the U.S. Department of Education. All states were required to report their assessment results to the U.S. Department of Education, where the results were then housed on the USDOE website in Consolidated State Performance Reports (U.S. Department of Education [USDOE], 2010). NAEP results were available through the National Center for Education Statistics, which was financed by the DOE (2010b). Once collected, the data were analyzed using SPSS 16.0 software.

NAEP

The eighth grade NAEP reading test is an assessment instrument designed to measure patterns in educational accomplishment over time. Those who oversee its administration, the National Assessment Governing Board (NAGB), aim to balance the need for a stable, yet current metric. The NAGB tries to keep changes in the assessment design to a minimum, while also ensuring that new questions are added to reflect changes

in curriculum standards (National Center for Education Statistics [NCES], 2001). The 2009 reading test questions were intended to measure reading comprehension and vocabulary: Specifically, they were intended to measure students' ability to locate and recall, integrate and interpret, and critique and evaluate. The assessment contained multiple-choice questions, as well as short and extended constructed-response items. Students read approximately two passages and responded to questions based on their reading (NCES, 2009c).

There were 39 questions on the 2009 eighth grade NAEP reading assessment. There were two subscales in this assessment: literary (containing 19 questions) and informational (containing 20 questions). Students received a subscale score on each of these measures, and these scores were then weighted to provide the composite score, which was measured on a scale of 0-500 (NCES, 2009c). Students who scored 243 and below were designated as performing at the "Basic" achievement level; those who scored 281 and below were designated as "Proficient"; those who scored 323 and above were designated "Advanced" (NCES, 2010d). See Table 1 for a description of performance indicators for each of the above achievement levels (NCES, 2010d).

Table 1. Descriptions of NAEP Achievement Levels and Score Ranges

	Description of Achievement Levels
Basic (243-280)	"Eighth-grade students performing at the Basic level should be able to locate information; identify statements of main idea, theme, or author's purpose; and make simple inferences from texts. They should be able to interpret the meaning of a word as it is used in the text. Students performing at this level should also be able to state judgments and give some support about content and presentation of content."
Proficient (281-322)	"Eighth-grade students performing at the Proficient level should be able to provide relevant information and summarize main ideas and themes. They should be able to make and support inferences about a text, connect parts of a text, and analyze text features. Students performing at this level should also be able to fully substantiate judgments about content and presentation of content."
Advanced (above 322)	"Eighth-grade students performing at the Advanced level should be able to make connections within and across texts and to explain causal relations. They should be able to evaluate and justify the strength of supporting evidence and the quality of an author's presentation. Students performing at the Advanced level also should be able to manage the processing demands of analysis and evaluation by stating, explaining, and justifying."

NCES, 2010d

The multiple-choice items were designed with 4 answer choices. Short constructed-response items gave 4-7 lines for student answers; responses to these short items were scored to determine whether students have full comprehension, partial comprehension, or little or no comprehension. Extended constructed-response items gave approximately 10 lines for student answers; responses were judged to be extensive, essential, partial or unsatisfactory. One multiple-choice item read as follows:

What does the poem mainly describe?
A. A personal experience
B. An unusual dream
C. The danger of alligators
D. Traveling in Florida (NCES, 2010c)

One short constructed-response item asks students to consider the following:

The following lines are from the poem:

I drank up until the moment it came
crashing toward me,
its tail flailing like a bundle of swords,
slashing the grass,
and the inside of its cradle-shaped mouth gaping,
and rimmed with teeth—

Choose an image from these lines and explain what it shows about the speaker's experience with the alligator. (NCES, 2010c)

An extended response-item for the same poem asked students to provide an explanation: "Describe what happens to the speaker of the poem and explain what this experience makes the speaker realize" (NCES, 2010c).

Scores on the 2009 eighth grade NAEP reading assessment were calculated based on item-response theory. Using this theoretical model, answers from a sample set of questions were used to determine how students would have been likely to respond to similar questions on the same standard. Student responses on a given set of items were used to determine the likelihood that students will score correctly on similar items assessing the same standard (NCES, 2010e).

In 2009, the NAEP population sample consisted of 161,000 eighth grade students.

The test was given to students in the winter of 2009 in all 50 states of the United States, as well as its territories.

The most recent NAEP Technical Documentation (NCES, 2008) found the weighted alpha reliability for the eighth grade reading assessment to be between .66 and .76, depending on the sample. The weighted alpha reliability score represents Cronbach's coefficient alpha when the calculation equally weights all student responses (NCES, 2008).

NCLB provisions specifed that an appointed 26-member National Assessment Governing Board oversaw the creation and administration of NAEP, as well as ensured that NAEP scores were reliable and valid. In addition, Congress called for continual reevaluations of NAEP via the convening of expert external panels, such as the National Academy of Sciences. In 2009, an external evaluation of NAEP was conducted by The Buros Center for Testing with the University of Massachusetts's Center for Educational Assessment and the University of Georgia (NCES, 2010).

NCLB Assessment Guidelines for States

NCLB required that each state set challenging standards and create an "accountability system that will be effective in ensuring that all local educational agencies, public elementary schools, and public secondary schools make adequate yearly progress" (Public Law 107-110, Sec. 1111 (2) (A)). Furthermore, states were statutorily required to ensure that " 'adequate yearly progress'… [is] defined… in a manner that… is statistically valid and reliable" (Sec. 1111 (2) (C) (ii)). In addition, states must have

demonstrated that they have implemented "high-quality, yearly student academic assessments" (Sec. 1111 (3) (A)) as the primary means of determining adequate yearly progress. Such assessments must "be used for purposes for which such assessments are valid and reliable, and be consistent with relevant, nationally recognized professional and technical standards" (Sec. 1111 (3) (A) (iii)). Moreover, these state assessments were to be used only if the state provides "evidence from the test publisher or other relevant sources that the assessments used are of adequate technical quality for each purpose required under this Act… and [if] such evidence is made public… upon request" (Sec. 1111 (3) (A) (iv)).

State Assessments

In order to compare state assessment instruments, the state with the largest population in each census region was analyzed. New York represented the Northeast census region, Texas the South, Illinois the Midwest, and California the West. Although every state's test differed somewhat, those highlighted offer examples of the types of assessments and were likely fairly representative of assessments of other states in their region.

New York (Northeast Census Region)

The reading skills of eighth-grade students in New York were tested using the New York State Testing Program's (NYSTP) eighth Grade English Language Arts Operational Test. This assessment was designed to measure the skills and standards

taught in New York's schools and to determine which schools were making adequate yearly progress (New York State Department of Education [NYDOE], 2009a). Teachers also used this test as a diagnostic assessment to determine students' strengths and weaknesses, as well as concomitant interventions. Specifically, the English Language Arts test measured student proficiency in reading, writing, and listening. Proficiency was scored on a four-level scale: Level I, Not Meeting Learning Standards; Level II, Partially Meeting Learning Standards; Level III, Meeting Learning Standards; and Level IV, Meeting Learning Standards with Distinction (NYDOE, 2009a). Scores were also given a Standard Performance Index number to measure students' knowledge of certain skills and standards. The standards being assessed were information and understanding, literary response and expression, and critical analysis and evaluation (NYDOE, 2009a).

New York's eighth Grade English Language Arts Operational Test was a criterion-referenced test containing 26 multiple-choice questions and eight constructed-response items. The multiple-choice items were designed with four answer choices. Two constructed-response items asked students to respond in a table diagram; four constructed-response items asked students to respond in seven lines for student answers; and two long constructed-response questions asked students to respond in an essay. One multiple-choice item from the NYSTP: English Language Arts Test Book One Grade 8 (NYDOE, 2009b) read as follows:

What is the most likely reason the author writes that the pine branches "snapped from burning trees with a cannon-like sound"?
A. to give objects human-like qualities
B. to provide hints about a future event
C. to make a comparison for dramatic effect
D. to explain how the characters are feeling (p. 5)

In a listening section, students were asked to listen to an article and then answer questions such as the following question: "What does Bernie Krause mean when he states that every living thing has a "sound signature"? Give two examples of living things that have sound signatures. Use details from the article to support your answer" (NYDOE, 2009c, p.4). Last, in the essay writing section of the NYSTP: English Language Arts Test Book Two Grade 8 (NYDOE, 2009c), students were given the following writing prompt:

> Write an essay in which you describe three challenges Bernie Krause has faced when recording nature sounds. Explain Krause's responses to the challenges. Then explain what his responses reveal about him. Use details from the article to support your answer. (p. 5)

New York's eighth grade reading/language arts test was given to students in New York classrooms over a two-day period in January 2009. To ensure the validity of the test, the NYDOE asked educators from diverse backgrounds to review the test materials. The state made an effort to represent teachers from various experience levels, geographic regions, genders, and ethnicities. The test had a stratified alpha reliability weighting of .86 for the eighth grade reading assessment (NYDOE, 2009a).

Texas (South Census Region)

The reading skills of eighth-grade students in Texas were tested using the Texas Assessment of Knowledge and Skills (TAKS) Reading Grade 8 Exam. This assessment wass designed to measure the degree to which students were learning the state-mandated curriculum, the Texas Essential Knowledge and Skills (Texas Education Agency: Student Assessment Division [TEA], 2004). Scores were given a scale score (ranging from 1162 to 2734) to measure student knowledge of certain skills and standards (TEA).

Proficiency was also scored on a three-level scale: Did Not Meet the Standard (below 2100 scale score), "unsatisfactory performance, below state passing standard, insufficient understanding of the TEKS reading curriculum"; Met the Standard (2100 scale score and higher), "satisfactory performance, at or above state passing standard, a sufficient understanding of the TEKS reading curriculum"; or Commended Performance (2400 scale score and higher), "high academic achievement, considerably above state passing standard, a thorough understanding of the TEKS reading curriculum" (TEA).

Reading selections for the TAKS were approximately 700 to 1,000 words for eighth grade students. Specifically, the TAKS Reading Exam assessed student proficiency in four objectives: basic understanding of culturally diverse material, knowledge of literary elements, analysis using reading strategies, and analysis using critical-thinking skills for textual analysis (TEA, 2004). TAKS Grade 8 Reading exam was a criterion-referenced test containing 48 multiple-choice questions: 12 items test basic understanding, 10 test literary elements, 10 test reading strategies, and 16 test

critical-thinking skills. The multiple-choice items each contained four answer choices. Two multiple-choice items from the *Grade 8 Reading Texas Assessment of Knowledge and Skills Information Booklet* (TEA, 2004) read as follows:

> **The fact that Lekeni's father trusts him to care for the family's cattle makes Lekeni proud because –**
> A. the work often involves great danger
> B. the cattle represent his family's wealth
> C. only warriors are allowed to herd cattle
> D. the work is part of his training to become a warrior (p. 23)
>
> **The author's choice of words in paragraph 1 of this story creates a mood of**
> A. anticipation
> B. uncertainty
> C. concern
> D. triumph (p. 26)

The TAKS reading test was given to eighth grade students in Texas classrooms during March 2009. During its development, input was sought from teachers, administrators, business people, parents, college faculty, professional organizations, and content area experts (TEA). The test had a stratified alpha reliability weighting of .887 for the eighth grade reading assessment (TEA, 2009).

Illinois (Midwest Census Region)

The reading skills of eighth-grade students in Illinois were tested using the Illinois Standards Achievement Test (ISAT) Reading Grade 8 Exam. This assessment was designed to measure the degree to which students were learning the Illinois Learning Standards (Illinois State Board of Education Division of Assessment [ISBE], 2009). The

test measured student knowledge achievement of two goals: the ability to "read with understanding and fluency" and "read and understand literature representative of various societies, eras, and ideas" (ISBE).

Students were scored on a four-level scale: Academic Warning (120-179), Below Standards (180-230), Meets Standards (231-277), and Exceeds Standards (278+). Specifically, the test assessed four strands of reading knowledge: vocabulary development, reading strategies, reading comprehension, and literature (ISBE, 2009).

The ISAT Grade 8 Reading exam contained 70 multiple-choice questions and two extended-response questions; 20 of the multiple-choice items were pilot items and were not used in calculating the student's score. Of those 70 items, 40 were criterion-referenced items, and 30 were norm-referenced items. The test was given in three segments of 45-minutes each, although students who were engaged in answering test items may take up to an additional 10 minutes. The extended response items were scored using a four-point holistic rubric. The multiple-choice items each contained four answer choices. Two multiple-choice items from *ISAT sample book 2009: Grade 8* (ISBE, 2009) read as follows:

> **In line 6, when the speaker says, "I see things others don't," she most likely means—**
> A. people often overlook what's around them
> B. people don't pay attention when their picture is taken
> C. cameras are the most accurate form of record keeping
> D. the camera lens is like a microscope (p. 13)

What is the meaning of tension in paragraph 5?
A. suspense
B. stretching
C. emotional strain
D. a measuring device (p. 18)

Last, in the essay writing section of *ISAT sample book 2009: Grade 8* (ISBE, 2009), students were given the following writing prompt:

> In the story, the author describes the behavior of adults at a little league game. Explain why adults behave as they do in this story. Use information from the story and your own observations and conclusions to support your answer. (p. 24)

The ISAT reading test was given to eighth grade students in Illinois classrooms during the spring of 2009. The test was developed by Illinois educators, Illinois Department of Education leaders, and curriculum experts. Items are screened for bias during item writing, item review, and data review. This eighth grade reading assessment was found to have a Cronbach's alpha reliability weighting of .92 (ISBE, 2009).

California (West Census Region)

California used the California Standards Test (CST) in English-Language Arts to determine the reading skills of its eighth-grade students (California Department of Education [CDOE], 2009b). This assessment was developed to measure students' knowledge of the California content standards and to determine which of California's schools were making adequate yearly progress (CDOE, 2009b). Specifically, the English Language Arts test measured student proficiency in literary response and analysis;

reading comprehension; word analysis, fluency, and systematic vocabulary development; writing strategies; and written and oral English language conventions (CDOE, 2009a). Scores were given a Performance Level Scale Score Range to quantify and compare student proficiency on a scale from 150 to 600. Proficiency was scored on a five-level scale: advanced (150-234), proficient (235-299), basic (300-349), below basic (350-406), and far below basic (407-600) (CDOE, 2009a).

The CST for eighth grade English-Language Arts was a criterion-referenced test containing 75 selected-response items (CDOE, 2009a). Selected-response items may have included true/false, matching, and multiple-choice items (CDOE, 2004). The selected-response items each gave students 4 possible answer choices. Two examples of representative CST selected-response items from the STAR (CDOE, 2009c) read as follows:

> **Which summary of paragraph 3 of Document A is the most accurate?**
> A. Students can do a better job of grading than teachers can.
> B. Teachers should be paid higher salaries for grading.
> C. Teachers can devote more time to teaching duties if students do the grading.
> D. Students learn more from one another than from teachers.
>
> **Read these lines from "My Fingers."**
> *Frail of an eggshell, Pull of a string,*
>
> **These lines suggest that the speaker**
> A. is a very small child.
> B. cannot see very well.
> C. appreciates life's little details.
> D. is a painter or photographer. (Search Engine results)

The CST in eighth grade English-Language Arts was given to approximately 465,000

students in California classrooms in the spring of 2009 (CDOE, 2010a). To ensure a valid testing program, the California State Board of Education developed a policy requiring that test items be reviewed through cooperation between K-12 and postsecondary educators (California State Board of Education, 2001). In addition, the test also uses Cronbach's alpha reliability testing to determine reliability. In 2009, California's reading test for eighth grade students had a reliability weighting of .94 (CDOE, 2010b).

Data Source

Data from NAEP and state assessments were available through the U.S. Department of Education. All states were required to report their assessment results to the U.S. Department of Education, where the results were then housed on the DOE website in Consolidated State Performance Reports (USDOE, 2010). NAEP results were available through the National Center for Education Statistics, which was financed by the DOE (2010b).

In July 2010, the NAEP reading performance data for eighth grade students were accessed through the National Center for Education Statistics database (NCES, 2010b). In addition, at the same time state assessment data were obtained through the U.S. Department of Education's Consolidated State Performance Reports (USDOE, 2010). Once collected, the data were analyzed using SPSS 16.0 software.

Data Analysis

Using 2009 eighth grade State NAEP reading proficiency percentages as the dependent variable and 2009 eighth grade state assessment reading proficiency percentages as the independent variable, the first research question was analyzed using simple regression. In all questions controlling for NAEP census region, the Northeast region was used as the reference region. Research Questions Two through Six were computed using multiple regression. Research Question Two controlled for NAEP census region in the model. Research Question Three followed that of the second question; however, it controlled for census region, as well as the proficiency percentages of each of the five major racial/ethnic groups identified as subgroups by NCLB. Research Question Four examined proficiency of ELL students; Research Question Five examined proficiency of students who qualify for free and reduced lunch (FRL); Research Question Six examined proficiency of students with disabilities. Finally, Research Question Seven compared the differences among states between the percentage of eighth grade students demonstrating proficiency on state and NAEP reading assessments from 2007 to 2009.

Once the data were obtained from each of the websites above, the data were downloaded, entered into a table for analysis, and subsequently analyzed using SPSS16.0.

Research Question One

Research question one asked: "To what extent can the percentage of eighth grade students demonstrating proficiency on the 2009 NAEP reading assessment be predicted

by the percentage of eighth grade students demonstrating proficiency on the 2009 state reading assessment?" A simple linear regression analysis was conducted to examine the relationship between the percentage of eighth grade students demonstrating proficiency on the 2009 state reading assessment and the percentage of eighth grade students demonstrating proficiency on the 2009 NAEP reading assessment and to determine if the NAEP proficiency percentage could be predicted by the state assessment proficiency percentage. A scatterplot for the two variables was analyzed to determine if a relationship was present. Unstandardized residuals were reviewed for normality, as well as histograms and Q-Q plots. Skewness and kurtosis results were analyzed to test for normality, as well as Shapiro-Wilk test results. A regression equation was sought to determine the relationship between eighth grade NAEP reading proficiency percentage and eighth grade state reading assessment proficiency percentage. The independent variable was the state assessment proficiency percentage, while the NAEP proficiency percentage was the dependent variable. An alpha level of .05 was used to determine the significance of the regression. A Pearson correlation coefficient was used to determine the relative strength of the predictive relationship.

Research Question Two

Research Question Two asked: "To what extent can the percentage of eighth grade students demonstrating proficiency on the 2009 NAEP reading assessment be predicted by the percentage of eighth grade students demonstrating proficiency on the 2009 state reading assessment, controlling for census regions defined by NAEP?" A

multiple linear regression analysis was conducted to examine the relationship between the percentage of eighth grade students scoring proficient and above on NAEP and state reading assessments, controlling for NAEP census region. The independent variables were: 1) the state assessment proficiency percentage; and 2) the census region. The Northeast census region was used as the reference category because this region was typically the highest performing on NAEP. A scatterplot was analyzed to determine if a relationship was present. Unstandardized residuals were reviewed for normality, as well as histograms, Q-Q plots, and box plots. Skewness and kurtosis results were analyzed to test for normality, as well as Shapiro-Wilk test results. A regression equation was sought to determine the relationship between eighth grade NAEP reading proficiency percentage and eighth grade state reading assessment proficiency percentage, as well as to determine if the relationship was similar in each NAEP census region. An alpha level of .05 was used to determine the significance of the regression. A Pearson correlation coefficient was used to determine the relative strength of the predictive relationship.

Research Questions Three, Four, Five, and Six

Research Questions Three, Four, Five and Six explored the same question as Research Question Two; however, each question explored the question as it relates to a different AYP subgroup (race/ethnicity, ELL, free and reduced lunch status (FRL), and students with disabilities). Question Three asked: "To what extent can the percentage of eighth grade students demonstrating proficiency on the 2009 NAEP reading assessment be predicted by the percentage of eighth grade students demonstrating proficiency on the

2009 state reading assessment, controlling for the percentage of students proficient in each of the five major racial/ethnic groups identified as subgroups by NCLB: American Indian, Asian, Hispanic, Black, White; and controlling for census regions defined by NAEP?" Research Question Four asked: "To what extent can the percentage of eighth grade ELL students demonstrating proficiency on the 2009 NAEP reading assessment be predicted by the percentage of eighth grade ELL students demonstrating proficiency on the 2009 state reading assessment, controlling for census regions defined by NAEP?" Research Question Five asked: "To what extent can the percentage of eighth grade students qualifying for free and reduced lunch who demonstrate proficiency on the 2009 NAEP reading assessment be predicted by the percentage of eighth grade students qualifying for free and reduced lunch who demonstrate proficiency on 2009 state reading assessments, controlling for census regions defined by NAEP?" Research Question Six asked: "To what extent can the percentage of eighth grade students with disabilities demonstrating proficiency on the 2009 NAEP reading assessment be predicted by the percentage of eighth grade students with disabilities demonstrating proficiency on the 2009 state reading assessment, controlling for census regions defined by NAEP?"

The same statistical analyses conducted for Research Question Two were also conducted for Research Questions Three, Four, Five, and Six. Thus, the data analysis procedures are not reiterated here, with the exception that a different AYP subgroup was analyzed in each question.

Research Question Seven

Research Question Seven asked: "On average does the difference between the percentage of eighth grade students demonstrating proficiency on state and NAEP reading assessments change from 2007 to 2009?" For this question, the differences between the NAEP and state average percent proficient scores for eighth grade reading in 2007, as well as the difference in 2009, were calculated for each state. The states with the greatest differences between the state and national proficiency scores in 2007 were compared with the states with the greatest differences in 2009. In addition, the states with the lowest differences in 2007 and 2009 were examined.

Variables

Data from 2007 and 2009 were collected using SPSS as follows:

Table 2. Variables collected for this study

Percentage of eighth grade student group meeting or exceeding proficiency on NAEP	*Percentage of eighth grade student group meeting or exceeding proficiency on State Tests*	*Other*
Total in 2007	Total in 2007	
Total in 2009	Total in 2009	Name of state
Male in 2009	Male in 2009	NAEP census region
Female in 2009	Female in 2009	
American Indian in 2009	American Indian in 2009	
Asian in 2009	Asian in 2009	
Black in 2009	Black in 2009	
Hispanic in 2009	Hispanic in 2009	
White in 2009	White in 2009	
Low SES in 2009	Low SES in 2009	
ELL in 2009	ELL in 2009	
SWD in 2009	SWD in 2009	

Summary

This chapter restated the purpose of this research and presented each of the seven research questions to be analyzed. The population and sample (eighth grade students who sat for the NAEP reading assessment, as well as eighth grade students who sat for the reading assessment given by their states) was stated. Assessment instrumentation details for NAEP and representative states were examined. The states with the highest student population in its respective census regions were chosen for study: New York, Northeast; Texas, South; Illinois, Midwest; and California, West.

Data sources used to obtain NAEP and state assessments proficiency percentages were also discussed in this chapter. Lastly, data analysis procedures and relevant variables for each of the research questions were presented. Simple and multiple regression studies, in addition to a study of differences, were performed to study the research questions. Results of the data analysis undertaken are presented in the following chapter.

CHAPTER FOUR: ANALYSIS

Introduction

This study intended to investigate to what extent eighth grade NAEP reading assessment proficiency percentages could be predicted using eighth grade state reading assessment proficiency percentages for the four census regions defined by NAEP, as well as for each AYP subgroup—race and ethnicity, free and reduced lunch status (FRL), and students with disabilities (SWD). This study also aimed to analyze the difference between NAEP and state assessment proficiency percentages from 2007 to 2009. The purpose of the study was achieved by using simple and multiple regression analyses to investigate whether a predictive relationship existed between NAEP and state assessment proficiency percentages. In addition, difference testing was used to examine differences between proficiency percentages over time. This chapter details the data analysis results for the seven stated research questions.

Purpose of the Study

This study builds on Gordon's (2009) study, which examined the relationship between 2007 NAEP and state eighth grade reading assessment proficiency percentages. The purpose of this study was to determine if there was a predictive relationship between 2009 NAEP and state eighth grade reading assessment proficiency percentages. Additionally, data were disaggregated into the four census regions of NAEP to make

comparisons between the total populations of each region, as well as certain AYP subgroups. In particular, this study extended Gordon's study to control for percentage of subgroups meeting proficiency on state assessments.

Research Questions

The following research questions guided this study:

1. To what extent can the percentage of eighth grade students demonstrating proficiency on the 2009 NAEP reading assessment be predicted by the percentage of eighth grade students demonstrating proficiency on the 2009 state reading assessment?
2. To what extent can the percentage of eighth grade students demonstrating proficiency on the 2009 NAEP reading assessment be predicted by the percentage of eighth grade students demonstrating proficiency on the 2009 state reading assessment, controlling for census regions defined by NAEP?
3. To what extent can the percentage of eighth grade students demonstrating proficiency on the 2009 NAEP reading assessment be predicted by the percentage of eighth grade students demonstrating proficiency on the 2009 state reading assessment, controlling for each of the five major racial/ethnic groups identified as subgroups by NCLB: American Indian, Asian, Hispanic, Black, White; and controlling for census regions defined by NAEP?
4. To what extent can the percentage of eighth grade ELL students demonstrating proficiency on the 2009 NAEP reading assessment be predicted by the percentage

of eighth grade ELL students demonstrating proficiency on the 2009 state reading assessment, controlling for census regions defined by NAEP?

5. To what extent can the percentage of eighth grade students qualifying for free and reduced lunch who demonstrate proficiency on the 2009 NAEP reading assessment be predicted by the percentage of eighth grade students qualifying for free and reduced lunch who demonstrate proficiency on 2009 state reading assessments, controlling for census regions defined by NAEP?
6. To what extent can the percentage of eighth grade students with disabilities demonstrating proficiency on the 2009 NAEP reading assessment be predicted by the percentage of eighth grade students with disabilities demonstrating proficiency on the 2009 state reading assessment, controlling for census regions defined by NAEP?
7. On average does the difference between the percentage of eighth grade students demonstrating proficiency on state and NAEP reading assessments change from 2007 to 2009?

Population and Sample

The population tested included the eighth grade students sampled by 2009 NAEP tests and the eighth grade students chosen by a state's AYP formula who completed that state's qualifying exams in the same year. All data analyzed were aggregated to the state level, rather than the individual student level.

Analysis of Data

Using 2009 eighth grade State NAEP reading proficiency percentages as the dependent variable and 2009 eighth grade state assessment reading proficiency percentages as the independent variable, the first research question will be analyzed using simple regression. In all questions controlling for NAEP census region, the Northeast region will be used as the reference region. The Northeast census region was used as the reference category because this region was typically the highest performing on NAEP. Research Questions Two through Six will be computed using multiple regression. The second research question will control for NAEP census region in the model. The third research question will follow that of the second question, however will control for census region, as well as the proficiency percentages of each of the five major racial/ethnic groups identified as subgroups by NCLB. The fourth question will examine proficiency of ELL students; the fifth will examine proficiency of students who qualify for free and reduced lunch; and the sixth will examine proficiency of students with disabilities. Finally, the last research question will compare the differences among states between the percentage of eighth grade students demonstrating proficiency on state and NAEP reading assessments from 2007 to 2009.

Once the data were obtained from each of the websites above, the data were downloaded, entered into a table for analysis, and subsequently analyzed using SPSS16.0.

Research Question One

To what extent can the percentage of eighth grade students demonstrating proficiency on the 2009 NAEP reading assessment be predicted by the percentage of eighth grade students demonstrating proficiency on the 2009 state reading assessment?

A simple linear regression analysis was conducted to examine the relationship between the percentage of eighth grade students demonstrating proficiency on the 2009 state reading assessment and the percentage of eighth grade students demonstrating proficiency on the 2009 NAEP reading assessment to determine if the NAEP proficiency percentage could be predicted by the state assessment proficiency percentage. The null hypothesis was that the regression coefficient was equal to zero.

Research Question One: Testing Assumptions

An initial review of scatterplots and casewise diagnostics indicated there were no cases to be removed. Simple linear regression assumptions were tested and met. The scatterplot for the two variables indicated a relatively linear relationship because the data points fell into a mostly straight line with a positive slope. Figure 1 (Appendix F) illustrates that as state assessment proficiency percentages increased, NAEP proficiency percentages increased as well. A scatterplot of studentized residuals to predicted values (Appendix F, Figure 2) indicated assumptions of linearity were also met, since the data points fell primarily within the range of +/- 2.

Unstandardized residuals were reviewed and found to be normally distributed. The histogram (Appendix F, Figure 3) and Q-Q plots (Appendix F, Figure 4) indicated the distribution was what would be expected when normally distributed. In addition,

skewness (.050) and kurtosis (-1.003) statistics indicated normality (since they fell within an absolute value of 2), as did non-significant Shapiro-Wilk tests ($W = .965$, $df = 51$, $p = .138$).

A scatterplot of studentized residuals to case number indicated assumption of independence was met, since the points fell randomly with no apparent pattern to the points (Appendix F, Figure 5).The scatterplot of unstandardized residuals to unstandardized predicted *Y* also indicated independence because the data points fell randomly (Appendix F, Figure 6). A scatterplot of studentized residuals to unstandardized predicted values suggested that homogeneity of variance was a reasonable assumption, because there was no pattern to the data points and they were randomly scattered around zero (Appendix F, See Figure 7).

Research Question One: Regression Results

The percentage of students meeting proficiency on eighth grade state reading assessments was a good predictor of the percentage of students meeting proficiency on eighth grade NAEP reading assessments, F (1, 49) = 8.915, $p < .004$. The regression equation for predicting eighth grade NAEP reading proficiency percentage as it relates to eighth grade state reading assessment proficiency percentage was:

Eighth grade NAEP reading proficiency percentage =

15.664 + (.213)(Eighth grade state reading assessment proficiency percentage)

The average percentage of students meeting proficiency on eighth grade NAEP reading assessments was 15.665%, controlling for the percentage proficient on the state

reading assessment. Every one unit increase in eighth grade state reading assessment proficiency percentage resulted in an average increase in eighth grade NAEP reading proficiency percentage of .213.

Using Cohen's (1988) guidelines to measure correlation strength, a correlation of .1 or less was considered small, .3 or less moderate, and .5 or more large. Accordingly, accuracy in predicting eighth grade reading NAEP proficiency percentage was moderately strong, with a correlation between NAEP and state assessment percentages of .392. Table 3 illustrates that approximately 15% (R^2 = .154) of the variation in the percentage of students meeting proficiency on eighth grade NAEP reading assessments was accounted for by its linear relationship with state assessment proficiency percentages.

Table 3. Simple Regression

	Model	*B*	*Standard Error*	*t*	*p*
1	(Intercept)	15.665	5.239	2.990	.004
	State_All_2009	.213	.071	2.986	.004
	R	.392			
	R^2	.154			
	ΔF	8.915			
	ΔR^2	.154			
	$p\Delta F$	.004			

There was a significant relationship between percent proficient on eighth grade state reading assessments and percent proficient on eighth grade NAEP reading assessments. The correlation between the percent proficient on these two assessments was moderate (Cohen, 1988). For every one percent increase on the state assessment, the NAEP increases by approximately .21%.

Research Question Two

To what extent can the percentage of eighth grade students demonstrating proficiency on the 2009 NAEP reading assessment be predicted by the percentage of eighth grade students demonstrating proficiency on the 2009 state reading assessment, controlling for census regions defined by NAEP?

A multiple linear regression analysis was conducted to examine the relationship between the percentage of eighth grade students scoring proficient and above on NAEP and state reading assessments, controlling for NAEP census region. The Northeast census region was chosen as the reference category, as this region is typically the highest performing on NAEP and represents a standard closer to that which NCLB statutes require. The null hypothesis was that the regression coefficients were equal to zero.

Research Question Two: Testing Assumptions

Initial review of Cook's distance (between .000 and .123), centered leverage values (.039 and .154), and plots suggested there were no outliers (See Table 4).

Table 4. Residuals Statistics

	Minimum	*Maximum*	*Mean*	*Std. Deviation*	*N*
Cook's Distance	.000	.123	.022	.026	51
Centered Leverage Value	.039	.154	.078	.027	51

Multiple linear regression assumptions were tested and met. The partial regression plot for the dependent to independent variables (Appendix F, Figure 8) indicated the variables are linearly related—as state reading proficiency percentage increases, NAEP reading proficiency percentage increases, while controlling for census region. Scatterplots of studentized residuals to unstandardized predicted values (Appendix F, Figure 9) and to the independent variable (Appendix F, Figure 10) indicated assumption of linearity was also met, as all values were located within a band of +/- 2.

Unstandardized residuals were reviewed for normality. Skewness (.195) and kurtosis (-.895) statistics for these unstandardized residuals indicated normality (because they were less than the absolute value of 2), as did non significant Shapiro-Wilk tests ($W = .932$, $df = 20$, $p = .171$). The histogram and Q-Q plots for unstandardized residuals indicated normality as well (Appendix F, Figures 11 and 12). The boxplot of unstandardized residuals indicated no outliers (Appendix F, Figure 13). A scatterplot of studentized residuals to the independent variable (Appendix F, Figure 14) indicated the assumption of independence was met, since the points fell randomly with no apparent pattern to the points.

In addition, scatterplots of studentized residuals to unstandardized predicted *Y* (Appendix F, Figure 15) and studentized residuals to case number (Appendix F, Figure 16) suggested that the assumption of independence was appropriate because the data points fell randomly. Moreover, the scatterplot of studentized residuals to unstandardized predicted values (Appendix F, Figure 15) suggested that homogeneity of variance was not violated, as the predicted values did not increase with increased residual values. Tolerance was greater than .10 (.975); variance inflation factor was less than 10 (1.025) there were not multiple eigenvalues close to zero (2.860, 1.000, 1.000, 0.127, and .013); and only one of the condition indices was greater than 15 (1.000, 1.691, 1.691, 4.737, and 15.085). Thus, there does not seem to be a problem with multicollinearity.

Research Question Two: Regression Results

Census region and state reading assessment proficiency percentages were good predictors of NAEP reading proficiency percentages , F (4, 46) = 14.624, $p < .001$. The regression equation for predicting NAEP reading proficiency percentage as a result of state reading assessment proficiency percentage and census region is shown in Table 5 and expressed as follows:

Eighth grade NAEP reading proficiency percentage = 24.813 + .184(state assessment proficiency percentage) – 11.484 (South) – 4.551 (Midwest) – 8.456 (West)

The model shown in Table 5 suggested that when controlling for census region, the average NAEP proficiency percentage was about 25 percent proficient. Every one percent change in eighth grade state reading assessment proficiency percentage resulted

in an average increase in eighth grade NAEP reading proficiency percentage of .184. Relative to the Northeast region, other regions of the country have lower percentages of students proficient on the eighth grade NAEP . Specifically, states in the South have about 11.5% fewer, states in the Midwest have about 4.5% fewer, and states in the West have about 8.5% fewer eighth grade students demonstrating proficiency on the 2009 NAEP reading assessment. Accuracy in predicting NAEP reading proficiency percentage by census region was strong with a multiple correlation coefficient of .748. About 56% (R^2 = .560) of the variance in NAEP reading proficiency percentage was accounted for by the regression model in Table 5.

Table 5. Multiple Regression

	Model	*B*	*Standard Error*	*t*	*P*
1	(Intercept)	24.813	4.200	5.980	.000
	State_All_2009	.184	.054	3.423	.001
	South	-11.484	1.875	-6.126	.000
	Midwest	-4.551	2.009	-2.265	.028
	West	-8.456	1.975	-4.281	.000
	R	.748			
	R^2	.560			
	F	14.624			

When controlling for census regions, there was a significant relationship between percent proficient on eighth grade state reading assessments and percent proficient on eighth grade NAEP reading assessments. The correlation between the percent proficient on these two assessments was strong (Cohen, 1988). For every one percent increase on the state assessment, the NAEP increases by less than .20%. The relationship between percent proficient was significant in each of the four census regions. The Northeast demonstrated the highest proficiency percentages, followed by the Midwest, West, and South, in that order.

Research Question Three

To what extent can the percentage of eighth grade students demonstrating proficiency on the 2009 NAEP reading assessment be predicted by the percentage of eighth grade students demonstrating proficiency on the 2009 state reading assessment, controlling for each of the five major racial/ethnic groups identified as subgroups by NCLB: American Indian, Asian, Hispanic, Black, White; and controlling for census regions defined by NAEP?

A multiple linear regression analysis was conducted to examine the relationship between the proficiency percentages of eighth grade students on NAEP and state reading assessments, controlling for each of the five major racial/ethnic groups identified as subgroups by NCLB and controlling for NAEP census region. The Northeast census region was chosen as the reference category, as this region is typically the highest performing on NAEP and represents a standard closer to that which NCLB statutes require. The null hypothesis was that the regression coefficients were equal to zero.

Research Question Three: Testing Assumptions

An initial review of Cook's distance (between .000 and .144) suggested no outliers. However, while one of the centered leverage values was .055, the other value at .548 was high enough to suggest the possibility of outliers (See Table 6). Furthermore, several scatterplots suggested there were outliers.

Table 6. Residuals Statistics

	Minimum	*Maximum*	*Mean*	*Std. Deviation*	*N*
Cook's Distance	.000	.144	.024	.033	49
Centered Leverage Value	.055	.547	.184	.091	49

Multiple linear regression assumptions were tested. The partial regression plot for the dependent to independent variables (Appendix F, Figure 17) indicated a somewhat linear relationship because the data points fell into a mostly straight line with a positive slope. As state reading proficiency percentage increases, NAEP reading proficiency percentage increases, while controlling for the proficiency percentages of race/ethnicity subgroups and census regions. Conversely, the scatterplots of studentized residuals to unstandardized predicted values (Appendix F, Figure 19) and to the primary independent variable (Appendix F, Figure 20) showed little apparent linearity. In addition, some of the studentized residuals indicated outliers, as there were a few plot points with values outside the band of +/- 2. Each scatterplot of studentized residuals to the independent variables representing race/ethnicity (Appendix F, Figures 20-24) also showed little

apparent linearity. In addition, each scatterplot showed some of the studentized residuals indicating outliers; in each scatterplot, there were a few plot points with values outside the band of +/- 2.

Unstandardized residuals were reviewed for normality. Skewness (.374) and kurtosis (.147) statistics for these unstandardized residuals indicated normality (because they were less than the absolute value of 2), as did non significant Shapiro-Wilk tests (W = .965, df = 49, p = .147) (Table 9). The histogram and Q-Q plots for unstandardized residuals indicated normality as well (Appendix F, Figures 25 and 26). The boxplot of unstandardized residuals indicated no outliers (Appendix F, Figure 27). A scatterplot of studentized residuals to the primary independent variable (Appendix F, Figure 28) indicated the assumption of independence was met, since the points fell randomly with no apparent pattern to the points. In addition, scatterplots of studentized residuals to unstandardized predicted Y (Appendix F, Figure 29) and studentized residuals to case number (Appendix F, Figure 30) suggested that the assumption of independence was appropriate, because the data points fell randomly.

Furthermore, the scatterplot of studentized residuals to unstandardized predicted values (Appendix F, Figure 29) suggested that homogeneity of variance was not violated, as the predicted values did not increase with increased residual values. Furthermore, there was no pattern to the data points and they were randomly scattered around zero. An examination of the tolerance values and variance inflation factors revealed a potential problem with multicollinearity. While six of the nine values for tolerance were acceptable

with values greater than 1.0, three of the nine values for tolerance were less than .10 (State All = .052, State Hispanic = .076, and State Black = .049). Moreover, the same three variables with tolerance values suggesting multicollinearity also had variance inflation factors greater than 10 (State All = 19.238, State Hispanic = 13.102, and State Black = 20.538). Eigenvalues and condition indices also indicated a potential problem with multicollinearity. There were six eigenvalues close to zero (.042, .010, .004, .003, .002, and .001), and three of the condition indices were greater than 30 (51.199, 68.984, 98.566). Thus, there seemed to be a problem with multicollinearity. In order address the issue of multicollinearity, the model was first run with all variables included. Subsequently, the Hispanic variable was removed in an attempt to correct the muticollinearity of the Hispanic and Black variables. The model, as is, is presented first and is then followed by analysis that addresses the multicollinearity.

Research Question Three: Regression Results

Census region, race/ethnicity, and state reading assessment proficiency percentages were statistically significant predictors of NAEP reading proficiency percentages, F (9, 39) = 8.576, $p < .001$. The regression equations for predicting NAEP reading proficiency percentage as a result of state reading assessment proficiency percentage, race/ethnicity, and census region are shown in Table 7 and expressed as follows:

Eighth grade NAEP reading proficiency percentage =

14.759 + .952 (state assessment proficiency percentage) – .041 (American Indian)

+ .091 (Asian) – .185 (Hispanic) – .263 (Black) – .339 (White) – 6.197 (South) – 3.942 (Midwest) – 3.892 (West)

The model shown in Table 7 suggested that when controlling for census region and race/ethnicity, most results were not statistically significant. However, the model did suggest that every one percent change in eighth grade state reading assessment proficiency percentage resulted in an average increase in eighth grade NAEP reading proficiency percentage of .952. Relative to the Northeast region, one other region of the country had a lower percentage of students proficient in the eighth grade NAEP reading proficiency percentage that was statistically significant. Specifically, states in the South have about 6.197% fewer students demonstrating proficiency on the 2009 NAEP reading assessment.

Table 7. Multiple Regression

Model	*B*	*Standard Error*	*T*	*P*
(Intercept)	14.759	8.907	1.657	.106
State_All_2009	.952	.210	4.433	.000
State_Amer_Ind	-.041	.089	-.459	.649
State_Asian	.091	.115	.792	.433
State_Hispanic	-.185	.137	1.351	.184
State_Black	-.263	.163	1.619	.113
State_White	-.339	.169	2.004	.052
South	-6.197	2.127	2.913	.006
Midwest	-3.942	2.005	1.966	.056
West	-3.892	2.286	1.703	.097
R	.815			
R^2	.664			
F	8.576			

Accuracy in predicting NAEP reading proficiency percentage by state proficiency percentage while controlling for census region and race/ethnicity appeared strong, with a multiple correlation coefficient of .815. About 66% (R^2 = .664) of the variance of NAEP reading proficiency percentage was accounted for by the regression model summarized in Table 7.

When controlling for race, ethnicity, and census regions, there was a significant relationship between percent proficient on eighth grade state reading assessments and percent proficient on eighth grade NAEP reading assessments. Although there was a predictive relationship between these two assessments, it was not an equal relationship. The correlation between the percent proficient on these two assessments was strong (Cohen, 1988). The relationship between percent proficient was significant in only the South census region, where percent proficient was lower than all other regions.

Research Question Three: Testing Assumptions with Outliers Removed

Once the Hispanic and White variables were removed to remedy multicollinearity, a review of Cook's distance (between .000 and .164) was conducted and suggested no outliers. However, while one of the centered leverage values was .046, the other value at .402 was high enough to suggest the possibility of outliers (See Table 8). Furthermore, several scatterplots suggested there were outliers.

Table 8. Residuals Statistics

	Minimum	*Maximum*	*Mean*	*Std. Deviation*	*N*
Cook's Distance	.000	.164	.024	.037	49
Centered Leverage Value	.046	.402	.143	.067	49

Multiple linear regression assumptions were tested. The partial regression plot for the dependent to independent variables (Appendix F, Figure 31) indicated a somewhat linear relationship because the data points fell into a mostly straight line with a positive slope. As state reading proficiency percentage increases, NAEP reading proficiency percentage increases, while controlling for the proficiency percentages of race/ethnicity subgroups and census regions. Conversely, the scatterplots of studentized residuals to unstandardized predicted values (Appendix F, Figure 32) and to the primary independent variable (Appendix F, Figure 33) showed little apparent linearity. In addition, some of the studentized residuals indicated outliers, as there were a few plot points with values outside the band of +/- 2. Each scatterplot of studentized residuals to the independent variables representing race/ethnicity (Appendix F, Figures 34-36) also showed little apparent linearity. In addition, each scatterplot showed some of the studentized residuals indicating outliers; in each scatterplot, there were a few plot points with values outside the band of +/- 2.

Unstandardized residuals were reviewed for normality. Skewness (.429) and kurtosis (-.358) statistics for these unstandardized residuals indicated normality (because they were less than the absolute value of 2), as did non significant Shapiro-Wilk tests (*W*

= .963, *df* = 49, *p* = .131). The histogram and Q-Q plots for unstandardized residuals indicated normality as well (Appendix F, Figures 37 and 38). The boxplot of unstandardized residuals indicated no outliers (Appendix F, Figure 39). A scatterplot of studentized residuals to the primary independent variable (Appendix F, Figure 40) indicated the assumption of independence was met, since the points fell randomly with no apparent pattern to the points. In addition, scatterplots of studentized residuals to unstandardized predicted *Y* (Appendix F, Figure 41) and studentized residuals to case number (Appendix F, Figure 42) suggested that the assumption of independence was appropriate, because the data points fell randomly.

Furthermore, the scatterplot of studentized residuals to unstandardized predicted values (Appendix F, Figure 41) suggested that homogeneity of variance was not violated, as the predicted values did not increase with increased residual values. Furthermore, there was no pattern to the data points and they were randomly scattered around zero. An examination of the tolerance values and variance inflation factors revealed a potential problem with multicollinearity. While five of the seven values for tolerance were acceptable with values greater than 1.0, two of the seven values for tolerance were less than .10 (State All = .080 and State Black = .068). Moreover, the same two variables with tolerance values suggesting multicollinearity also had variance inflation factors greater than 10 (State All = 12.467 and State Black = 14.698). Eigenvalues and condition indices also indicated a potential problem with multicollinearity. There were four eigenvalues close to zero (.034, .009, .003, and .001) and two of the condition indices

were greater than 30 (42.183 and 71.695). Thus, there continued to be a problem with multicollinearity.

Research Question Three: Regression Results With Variable Removed

In an effort to correct multicollinearity, the Hispanic and White sets of proficiency percentages for state reading assessments (State_Hispanic and State_White) were removed. These variables were chosen because values showed that the Hispanic and White data sets were too closely related to the Black data set (r = .91 and .885, respectively). With the Hispanic and White variables removed, census region, race/ethnicity, and state reading assessment proficiency percentages remained statistically significant predictors of NAEP reading proficiency percentages, F (7, 41) = 8.914, p < .000. The regression equations for predicting NAEP reading proficiency percentage as a result of state reading assessment proficiency percentage, race/ethnicity (with the Hispanic and White variables removed), and census regions are shown in Table 9 and expressed as follows:

Eighth grade NAEP reading proficiency percentage =

9.080 + .679 (state assessment proficiency percentage) – .099 (American Indian)

+ .072 (Asian) – .374 (Black) – 7.355 (South)

– 4.583 (Midwest) – 4.666 (West)

The model shown in Table 9 suggests that when controlling for census region and race/ethnicity, most results were statistically significant. Removing the Hispanic and White variables did cause an increase in the number of subgroups whose results were

statistically significant. The model suggested that every one percent change in eighth grade state reading assessment proficiency percentage results in an average increase in eighth grade NAEP reading proficiency percentage of .679. Relative to the Northeast region, two other regions of the country had a lower percentage of students proficient in the eighth grade NAEP reading proficiency percentage that were statistically significant. Specifically, states in the South have about 7.355% fewer students and states in the Midwest have about 4.583% fewer students demonstrating proficiency on the 2009 NAEP reading assessment than states in the Northeast. In addition, it was found to be statistically significant that both Black students had about 3.74% fewer students demonstrating proficiency.

Table 9. Multiple Regression

Model	*B*	*Standard Error*	*T*	*P*
(Intercept)	9.080	9.013	1.007	.320
State_All_2009	.679	.179	3.788	.000
State_Amer_Ind	-.099	.085	-1.168	.249
State_Asian	.072	.119	.604	.549
State_Black	-.374	.146	-2.562	.014
South	-7.355	2.119	-3.471	.001
Midwest	-4.583	1.964	-2.333	.025
West	-4.666	2.403	-1.942	.059
R	.777			
R^2	.603			
F	8.914			

With the Hispanic and White variables removed, accuracy in predicting NAEP reading proficiency percentage by state proficiency percentage (while controlling for census region and race/ethnicity) appeared strong (although it decreased slightly), with a multiple correlation coefficient of .777. About 60% (R^2 = .603 of the variance in NAEP reading proficiency percentage was accounted for by the regression model summarized in Table 8.

Once the Hispanic and White variables had been removed, the number of significant relationships increased. When controlling for race, ethnicity, and census regions, there was a significant relationship between percent proficient on eighth grade state reading assessments and percent proficient on eighth grade NAEP reading assessments. The correlation between the percent proficient on these two assessments was strong (Cohen, 1988). For every one percent increase on the state assessment, the NAEP increases by about 1%. The relationship between percent proficient was significant in the South and Midwest census regions, as well as the Black subgroup, where percent proficient was lower than all other regions.

Research Question Four

To what extent can the percentage of eighth grade ELL students demonstrating proficiency on the 2009 NAEP reading assessment be predicted by the percentage of eighth grade ELL students demonstrating proficiency on the 2009 state reading assessment, controlling for census regions defined by NAEP?

A multiple linear regression analysis was conducted to examine the relationship between the percentage of eighth grade ELL students scoring proficient on NAEP and

state reading assessments, controlling for NAEP census region. The Northeast census region was chosen as the reference category, as this region is typically the highest performing on NAEP and represents a standard closer to that which NCLB statutes require. The null hypothesis was that the regression coefficients were equal to zero.

Research Question Four: Testing Assumptions

An initial review of Cook's distance (between .000 and .658) suggested no outliers. However, while one of the centered leverage values was .054, the other value at .268 was high enough to suggest the possibility of outliers (See Table 10). Furthermore, several scatterplots suggested the presence of outliers.

Table 10. Residuals Statistics

	Minimum	*Maximum*	*Mean*	*Std. Deviation*	*N*
Cook's Distance	.000	.658	.052	.146	27
Centered Leverage Value	.054	.268	.148	.069	27

Multiple linear regression assumptions were tested but not met. Partial regression plots for the dependent to independent variables (Appendix F, Figures 43-46) suggested linearity, with a relatively straight horizontal line. As state reading proficiency percentage increases for ELL students, NAEP reading proficiency percentage may increase slightly for ELL students as well, while controlling for census region. Conversely, the scatterplots of studentized residuals to unstandardized predicted values (Appendix F, Figure 47) and

to the primary independent variable (Appendix F, Figure 48) showed little apparent linearity. In addition, some of the studentized residuals indicated outliers, as there were a few plot points with values outside the band of +/- 2.

Unstandardized residuals were reviewed and found to be normally distributed. Skewness (1.038) indicated normality, since it fell within an absolute value of 2; however, the kurtosis (8.058) statistic for these unstandardized residuals indicated abnormality, as it was greater than the absolute value of 2. Furthermore, statistically significant Shapiro-Wilk tests added another indicator of abnormality ($W = .780$, $df = 27$, $p = .000$). The histogram and Q-Q plots for unstandardized residuals also indicated abnormality, with extreme outliers apparent (Appendix F, Figures 49 and 50). The boxplot of unstandardized residuals shown in Figure 51 indicated two outliers (case numbers 41 and 44). A scatterplot of studentized residuals to the primary independent variable (Appendix F, Figure 52) indicated the assumption of independence was met, since the data points fell randomly. In addition, scatterplots of studentized residuals to unstandardized predicted *Y* (Appendix F, Figure 47) and studentized residuals to case number (Appendix F, Figure 53) suggested that the assumption of independence was appropriate, because the data points fell randomly. Moreover, the scatterplot of studentized residuals to unstandardized predicted values (Appendix F, Figure 47) suggested that homogeneity of variance was not violated, as the predicted values did not increase with increased residual values, there was no pattern to the data points, and they were randomly scattered around zero.

Tolerance values were greater than .10 (.703, .398, .448, and .441), and variance inflation factors were less than 10 (1.422, 2.511, 2.230, and 2.267). Most eigenvalues were not close to zero (2.773, 1.030, and 1.000); however, two values were close to zero (Midwest ELL = .120 and West ELL = .077). None of the condition indices was greater than 15 (1.000, 1.641, 1.665, 4.798, and 6.013). There does not seem to be a problem with multicollinearity.

However, the issue of outliers was important to address. In order address the issue, the model was first run with all outliers included. Subsequently, the outliers were removed in an attempt to correct the abnormality caused by the outliers. The model, as is, is presented first and is then followed by analysis that addresses the outliers.

Research Question Four: Regression Results

Census region and state reading assessment proficiency percentages for ELL students were statistically significant predictors of NAEP reading proficiency percentages for ELL students, F (4, 22) = 4.792, $p < .006$. The regression equation for predicting NAEP reading proficiency percentage for ELL students as a result of state reading assessment proficiency percentage for ELL students and census region is shown in Table 11 and expressed as follows:

Eighth grade NAEP reading proficiency percentage for ELL students =

1.502 + .064(state assessment proficiency percentage) +2.708 (South)

+ .910 (Midwest) - 1.213 (West)

Table 11. Multiple Regression

	Model	*B*	*Standard Error*	*t*	*P*
1	(Intercept)	1.502	1.583	.948	.353
	State_ELL	.064	.033	1.928	.067
	South	2.011	2.011	1.346	.192
	Midwest	.910	2.138	.426	.675
	West	-1.213	1.704	-.712	.484
	R	.682			
	R^2	.466			
	F	4.792			

The model shown in Table 11 illustrates that there were no statistically significant predictors. This finding suggests that the results for the outcome were similar regardless of proficiency by census region or proficiency for ELL. Although no individual results were significant, accuracy in predicting NAEP reading proficiency percentage by census region was strong with a multiple correlation coefficient of .682. About 47% (R^2 = .466) of the variance of NAEP reading proficiency percentage was accounted for by the regression model summarized in Table 11.

Research Question Four: Testing Assumptions with Outliers Removed

Once outliers were removed (case numbers 41 and 44), an review of Cook's distance (between .000 and .422) suggested no outliers. However, while one of the centered leverage values was .051, the other value at .364 was high enough to suggest the possibility of outliers (See Table 12). On the other hand, scatterplots no longer suggested the presence of severe outliers.

Table 12. Residuals Statistics

	Minimum	*Maximum*	*Mean*	*Std. Deviation*	*N*
Cook's Distance	.000	.422	.066	.101	25
Centered Leverage Value	.051	.364	.160	.085	25

Once outliers were removed, multiple linear regression assumptions improved; upon retesting, more assumptions were met. Partial regression plots for the dependent to independent variables (Appendix F, Figures 55-58) suggested linearity. As state reading proficiency percentage increases for ELL students, NAEP reading proficiency percentage may increase slightly for ELL students as well, while controlling for census region. The scatterplots of studentized residuals to unstandardized predicted values (Appendix F, Figure 59) and to the primary independent variable (Appendix F, Figure 60) showed apparent linearity. In addition, some of the studentized residuals may have indicated outliers, as there were a few plot points with values that appeared slightly outside the band of +/- 2.

Unstandardized residuals were reviewed and found to be normally distributed. Skewness (.483) and kurtosis (-.312) statistics indicated normality (since they fell within an absolute value of 2). Once the outliers were removed, Shapiro-Wilk results were no longer statistically significant, and no longer indicated abnormality ($W = .959$, $df = 25$, $p = .398$. Likewise, the histogram and Q-Q plots for unstandardized residuals were corrected to indicate abnormality, with outliers no longer apparent (Appendix F, Figures 61 and 62). The boxplot of unstandardized residuals shown in Appendix F, Figure 63

indicated no outliers (case numbers 41 and 44). A scatterplot of studentized residuals to the primary independent variable (Appendix F, Figure 64) indicated the assumption of independence was met, since the data points fell randomly. In addition, scatterplots of studentized residuals to unstandardized predicted *Y* (Appendix F, Figure 59) and studentized residuals to case number (Appendix F, Figure 65) suggested that the assumption of independence was appropriate, because the data points fell randomly. Moreover, the scatterplot of studentized residuals to unstandardized predicted values (Appendix F, Figure 66) suggested that homogeneity of variance was not violated; the predicted values did not increase with increased residual values, there was no pattern to the data points, and they were randomly scattered around zero.

Tolerance values were greater than .10 (.747, .499, .442, and .465), and variance inflation factors were less than 10 (1.338, 2.006, 2.261, and 2.150). Most eigenvalues were not close to zero (2.762, 1.026, and 1.000); however, two values were close to zero (Midwest ELL = .131 and West ELL = .082). None of the condition indices was greater than 15 (1.000, 1.641, 1.662, 4.600, and 5.800). Hence, there does not seem to be a problem with multicollinearity.

Research Question Four: Regression Results With Outliers Removed

Once the two outliers (case numbers 41 and 44) were removed, results changed slightly. Census region and state reading assessment proficiency percentages for ELL students remained statistically significant predictors of NAEP reading proficiency percentages for ELL students, F (4, 20) = 14.918, $p < .000$. The regression equation for

predicting NAEP reading proficiency percentage for ELL students as a result of state reading assessment proficiency percentage for ELL students and census region is shown in Table 13 and expressed as follows:

Eighth grade NAEP reading proficiency percentage for ELL students = 1.642 + .057(state assessment proficiency percentage) +2.419 (South) + 1.106 (Midwest) - 1.161 (West)

Table 13. Multiple Regression

	Model	*B*	*Standard Error*	*t*	*P*
1	(Intercept)	1.642	.726	2.262	.035
	State_ELL	.057	.016	3.483	.002
	South	2.419	.924	2.618	.016
	Midwest	1.106	.981	1.127	.273
	West	-1.161	.771	-1.506	.148
	R	.865			
	R^2	.749			
	F	14.918			

The model shown in Table 13 illustrates that there were two statistically significant predictors: State ELL score and South census region. The model suggests that when controlling for census region and for state ELL, the average NAEP ELL proficiency percentage was about 2% proficient. Every one percent change in eighth

grade state reading assessment proficiency percentage resulted in an average increase in eighth grade NAEP reading proficiency percentage of .057. Relative to the Northeast region, the South census region was the only region of the country that had significantly higher percentages of students proficient on the eighth grade NAEP reading assessment. Specifically, states in the South have about 2.4% more eighth grade students demonstrating proficiency on the 2009 NAEP reading assessment. Accuracy in predicting NAEP reading proficiency percentage by census region was strong, with a multiple correlation coefficient of .865. About 75% (R^2 = .749) of the variance in NAEP reading proficiency percentage was accounted for by the regression model summarized in Table 11.

When controlling for census regions, there was a significant relationship between the percent of ELL eighth grade students proficient on state reading assessments and the percent of ELL eighth grade students proficient on NAEP reading assessments. The correlation between the percent of ELL students proficient on these two assessments was strong (Cohen, 1988). For every one percent increase on the state assessment, the NAEP increases by approximately .06%. The relationship between percent of ELL students proficient was not significant in any of the census regions, unless the two outliers are removed. Once the outliers were removed, it was determined that percent proficient for eighth grade ELL students in the South census region was significantly higher than in the Northeast, as well as higher than the other regions.

Research Question Five

To what extent can the percentage of eighth grade students qualifying for free and reduced lunch who demonstrate proficiency on the 2009 NAEP reading assessment be predicted by the percentage of eighth grade students qualifying for free and reduced lunch who demonstrate proficiency on 2009 state reading assessments, controlling for census regions defined by NAEP?

A multiple linear regression analysis was conducted to examine the relationship between the proficiency percentages of eighth grade FRL students on NAEP and state reading assessments, controlling for NAEP census region. The Northeast census region was chosen as the reference category, as this region is typically the highest performing on NAEP and represents a standard closer to that which NCLB statutes require. The null hypothesis was that the regression coefficients were equal to zero.

Research Question Five: Testing Assumptions

An initial review of Cook's distance (between .000 and .147) suggested no outliers. Neither of the centered leverage values (.039 and .145) was high enough to suggest the possibility of outliers (See Table 14). Furthermore, several scatterplots suggested the presence of outliers.

Table 14. Residuals Statistics

	Minimum	*Maximum*	*Mean*	*Std. Deviation*	*N*
Cook's Distance	.000	.147	.022	.030	51
Centered Leverage Value	.039	.145	.078	.028	51

Multiple linear regression assumptions were tested but not met. Partial regression plots for the dependent to independent variables (Appendix F, Figures 67-70) did not seem to suggest linearity, since the points fell randomly. As state reading proficiency percentage increases for FRL students, NAEP reading proficiency percentage may increase slightly as well, while controlling for census region. Moreover, the scatterplots of studentized residuals to unstandardized predicted values (Appendix F, Figure 71) and to the primary independent variable (Appendix F, Figure 72) showed little apparent linearity, since the data points fell randomly. In addition, some of the studentized residuals indicated outliers, as there were a few plot points with values outside the band of +/- 2.

Unstandardized residuals were reviewed and found to be normally distributed. Skewness (.560) and kurtosis (.067) statistics indicated normality (since they fell within an absolute value of 2). Nonsignificant Shapiro-Wilk test results indicated normality (W = .968, df = 51, p = .175). The histogram and Q-Q plots for unstandardized residuals also indicated possible abnormality, with one outlier apparent (Appendix F, Figures 73 and 74). The boxplot of unstandardized residuals shown in Figure 50 indicated one outlier (case number 27). A scatterplot of studentized residuals to the primary independent variable (Appendix F, Figure 76) indicated the assumption of independence was met, since the points fell randomly with no apparent pattern to the points. In addition, scatterplots of studentized residuals to unstandardized predicted Y (Appendix F, Figure 77) and studentized residuals to case number (Appendix F, Figure 78) suggested that the assumption of independence was appropriate, because the data points fell randomly. In

addition, the scatterplot of studentized residuals to unstandardized predicted values (Appendix F, Figure 77) suggested that homogeneity of variance was not violated; the predicted values did not increase with increased residual values, there was no pattern to the data points, and they were randomly scattered around zero.

Tolerance values were greater than .10 (.957, .506, .548, and .549), and variance inflation factors were less than 10 (1.044, 1.974, 1.826, and 1.823). Most eigenvalues were not close to zero (2.848, 1.001, and 1.000); however, two values were close to zero (Midwest FRL = .123 and West Low FRL = .029). None of the condition indices was greater than 15 (1.000, 1.687, 1.688, 4.821, and 9.952). There does not seem to be a problem with multicollinearity.

However, the issue of the outlier was important to address. In order address the issue, the model was first run with the outlier included. Subsequently, the outlier was removed in an attempt to correct the abnormality caused by the outlier. The model, as is, is presented first and is then followed by analysis that addresses the outlier.

Research Question Five: Regression Results

Census region and state reading assessment proficiency percentages for FRL students were statistically significant predictors of NAEP reading proficiency percentages for FRL students, F (4, 46) = 4.547, $p < .004$. The regression equation for predicting NAEP reading proficiency percentage for low SES students as a result of state reading assessment proficiency percentage for low SES students and census region is shown in Table 15 and expressed as follows:

Eighth grade NAEP reading proficiency percentage for low SES students = 18.291 + .029(state assessment proficiency percentage) - 5.087 (South) - .925 (Midwest) - 2.925 (West)

Table 15. Multiple Regression

	Model	*B*	*Standard Error*	*T*	*P*
1	(Intercept)	18.291	2.084	8.778	.000
	State_FRL_Eligible	.029	.032	.907	.369
	South	-5.087	1.392	-3.655	.001
	Midwest	-.925	1.487	-.622	.537
	West	-2.925	1.446	-2.022	.049
	R	.532			
	R^2	.283			
	F	4.547			

The model shown in Table 15 suggests that when controlling for census region and when controlling for state's percent FRL, the average NAEP proficiency percentage for FRL students was about 18% proficient. Relative to the Northeast region, two other regions of the country have lower percentages of FRL students proficient in the eighth grade NAEP reading proficiency percentage. Specifically, states in the South have about 5.082% fewer, and states in the West have about 2.925% fewer, eighth grade students demonstrating proficiency on the 2009 NAEP reading assessment.

Accuracy in predicting NAEP reading proficiency percentage by census region was moderately strong with a multiple correlation coefficient of .532. About 28% (R^2 = .283) of the variance in NAEP reading proficiency percentage was accounted for by the regression model in Table 15.

Research Question Five: Testing Assumptions with Outlier Removed

Once the outlier was removed (case number 27), a review of Cook's distance (between .000 and .109) suggested no outliers. Neither of the centered leverage values (.039 and .152) was high enough to suggest the possibility of outliers (See Table 16). Furthermore, several scatterplots suggested the presence of outliers.

Table 16. Residuals Statistics

	Minimum	*Maximum*	*Mean*	*Std. Deviation*	*N*
Cook's Distance	.000	.109	.023	.030	50
Centered Leverage Value	.039	.152	.080	.029	50

Even with the outlier removed, multiple linear regression assumptions were tested but not met. Partial regression plots for the dependent to independent variables (Appendix F, Figures 79-82) did not suggest linearity, as the points fell randomly. As state reading proficiency percentage increases for FRL students, NAEP reading proficiency percentage may increase slightly for as well, while controlling for census region. Moreover, the scatterplots of studentized residuals to unstandardized predicted values (Appendix F, Figure 83) and to the primary independent variable (Appendix F, Figure 84) showed little apparent linearity, since the data points fell randomly. In addition, some of the studentized residuals indicated outliers, as there were a few plot points with values just slightly outside the band of +/- 2.

Unstandardized residuals shown in were reviewed and found to be normally distributed. Skewness (.344) and kurtosis (-.362) statistics indicated normality (since they fell within an absolute value of 2). Once the outlier was removed, nonsignificant Shapiro-Wilk test results continued to indicate normality ($W = .968$, $df = 50$, $p = .186$). On the other hand, the histogram and Q-Q plots for unstandardized residuals no longer indicated abnormality (Appendix F, Figures 85 and 86). The boxplot of unstandardized residuals shown in Figure 87 (Appendix F) did not indicate outliers. A scatterplot of studentized residuals to the primary independent variable (Appendix F, Figure 88) indicated the assumption of independence was met, since the points fell randomly with no apparent pattern to the points. In addition, scatterplots of studentized residuals to unstandardized predicted *Y* (Appendix F, Figure 89) and studentized residuals to case number (Appendix F, Figure 90) suggested that the assumption of independence was appropriate, because the data points fell randomly. Moreover, the scatterplot of studentized residuals to unstandardized predicted values (Appendix F, Figure 89) suggested that homogeneity of variance was not violated; the predicted values did not increase with increased residual values, there was no pattern to the points, which were randomly scattered around zero.

Tolerance values were greater than .10 (.951, .511, .551, and .564), and variance inflation factors were less than 10 (1.052, 1.955, 1.815, and 1.773). Most eigenvalues were not close to zero (2.845, 1.001, and 1.000); however, two values were close to zero (Midwest FRL = .125 and West Low FRL = .029). None of the condition indices was

greater than 15 (1.000, 1.686, 1.687, 4.765, and 9.905). Overall, there did not seem to be a problem with multicollinearity.

Research Question Five: Regression Results with Outlier Removed

Once the outlier (case number 27) was removed, census region and state reading assessment proficiency percentages for FRL students remained statistically significant predictors of NAEP reading proficiency percentages for FRL students, F (4, 45) = 5.457, $p < .001$. The regression equation for predicting NAEP reading proficiency percentage for low SES students as a result of state reading assessment proficiency percentage for low SES students and census region is shown in Table 17 and expressed as follows:

Eighth grade NAEP reading proficiency percentage for low SES students = 18.878 + .018(state assessment proficiency percentage) - 5.014 (South) - .851 (Midwest) – 3.643 (West)

Table 17. Multiple Regression

	Model	*B*	*Standard Error*	*T*	*P*
1	(Intercept)	18.878	1.940	9.729	.000
	State_FRL_Eligible	.018	.030	.615	.542
	South	-5.014	1.289	-3.889	.000
	Midwest	-.851	1.378	-.617	.540
	West	-3.643	1.362	-2.675	.010
	R	.572			
	R^2	.327			
	F	5.457			

The model shown in Table 17 suggests that when controlling for census region and when controlling for state's percent FRL, the average NAEP proficiency percentage for FRL

students was about 19% proficient. Relative to the Northeast region, other regions of the country have lower percentages of FRL students proficient in the eighth grade NAEP reading assessment. Specifically, states in the South have about 5.014% fewer, and states in the West have about 3.643% fewer, eighth grade students demonstrating proficiency on the 2009 NAEP reading assessment. Accuracy in predicting NAEP reading proficiency percentage by census region was moderately strong, with a multiple correlation coefficient of .572. About 33% (R^2 = .327) of the variance in NAEP reading proficiency percentage was accounted for by the regression model in Table 17.

When controlling for census regions, there was a significant relationship between percent of FRL eighth grade students proficient on state reading assessments and percent of FRL eighth grade students proficient on NAEP reading assessments. The correlation between the percent of FRL students proficient on these two assessments was strong (Cohen, 1988). For every one percent increase on the state assessment, the NAEP increases by approximately .02%. The relationship between percent of FRL students proficient was significant in only the South and West census regions. Percent proficient for eighth grade FRL students in the South census region was significantly lower than in the Northeast. In addition, percent proficient for eighth grade students in the West census region were significantly lower than the Northeast. Removing outliers did not bring about a change in significance levels.

Research Question Six

To what extent can the percentage of eighth grade students with disabilities (SWD) demonstrating proficiency on the 2009 NAEP reading assessment be predicted by the percentage of eighth grade SWD demonstrating proficiency on the 2009 state reading assessment, controlling for census regions defined by NAEP?

A multiple linear regression analysis was conducted to examine the relationship between the percentage of eighth grade SWD students scoring proficient and above on NAEP and state reading assessments, controlling for NAEP census region. The Northeast census region was chosen as the reference category, as this region is typically the highest performing on NAEP and represents a standard closer to that which NCLB statutes require. The null hypothesis was that the regression coefficients were equal to zero.

Research Question Six: Testing Assumptions

An initial review of Cook's distance (between .000 and .205) suggested no outliers. Neither of the centered leverage values (.040 and .168) was high enough to suggest the possibility of outliers (See Table 18). Additionally, several scatterplots suggested the presence of outliers.

Table 18. Residuals Statistics

	Minimum	*Maximum*	*Mean*	*Std. Deviation*	*N*
Cook's Distance	.000	.205	.021	.036	51
Centered Leverage Value	.040	.168	.078	.027	51

Multiple linear regression assumptions were tested but not met. Partial regression plots for the dependent to independent variables (Appendix F, Figures 91-94) did not seem to suggest linearity, as the data points fell randomly. As state reading proficiency percentage increases for SWD students, NAEP reading proficiency percentage may increase slightly for SWD students as well, while controlling for census region. Furthermore, the scatterplots of studentized residuals to unstandardized predicted values (Appendix F, Figure 95) and to the primary independent variable (Appendix F, Figure 96) showed little apparent linearity. In addition, some of the studentized residuals indicated outliers, as there were a few plot points with values outside the band of +/- 2.

Unstandardized residuals were reviewed and found to be normally distributed. Skewness (1.246) and kurtosis (1.385) statistics indicated normality (since they fell within an absolute value of 2), as they were within an absolute value of two. Statistically significant Shapiro-Wilk test results indicated abnormality ($W = .892$, $df = 51$, $p = .000$). The histogram and Q-Q plots for unstandardized residuals also indicated possible abnormality, with several outliers apparent (Appendix F, Figures 97 and 98). The boxplot of unstandardized residuals shown in Appendix F, Figure 99 indicated two outliers (case numbers 21 and 31). A scatterplot of studentized residuals to the primary independent variable (Appendix F, Figure 100) indicated the assumption of independence was met, since the data points fell randomly with no apparent pattern to the points. In addition, scatterplots of studentized residuals to unstandardized predicted *Y* (Appendix F, Figure 101) and studentized residuals to case number (Appendix F, Figure 102) suggested that

the assumption of independence was appropriate, because the points fell randomly. In addition, the scatterplot of studentized residuals to unstandardized predicted values (Appendix F, Figure 101) suggested homogeneity of variance was not violated, as the predicted values did not increase with increased residual values, there was no pattern to the data points, and they were randomly scattered around zero .

Tolerance values were greater than .10 (.924, .505, .538, and .549), and variance inflation factors were less than 10 (1.082, 1.979, 1.860, and 1.821). Most eigenvalues were not close to zero (2.777, 1.006, and 1.000); however, two values were close to zero (Midwest SWD = .147 and West SWD = .071). None of the condition indices was greater than 15 (1.000, 1.662, 1.666, 4.353, and 6.269). There does not seem to be a problem with multicollinearity.

However, the issue of outliers was important to address. In order address the issue, the model was first run with all outliers included. Subsequently, the outliers were removed in an attempt to correct the abnormality caused by the outliers. The model, as is, is presented first and is then followed by analysis that addresses the outliers.

Research Question Six: Regression Results

Census region and state reading assessment proficiency percentages for SWD students were statistically significant predictors of NAEP reading proficiency percentages for SWD students, F (4, 46) = 5.278, $p < .001$. The regression equation for predicting NAEP reading proficiency percentage for SWD students as a result of state reading

assessment proficiency percentage for SWD students and census region is shown in Table 19 and expressed as follows:

Eighth grade NAEP reading proficiency percentage for SWD students = 8.947 + .019 (state assessment proficiency percentage) - 4.769 (South) - 3.344 (Midwest) - 5.243 (West)

Table 19. Multiple Regression

	Model	*B*	*Standard Error*	*t*	*P*
1	(Intercept)	8.947	2.084	7.202	.000
	State_SD	.019	.032	.779	.440
	South	-4.769	1.392	-3.941	.000
	Midwest	-3.344	1.487	-2.565	.014
	West	-5.243	1.446	-4.175	.000
	R	.561			
	R^2	.315			
	F	14.624			

The model shown in Table 19 suggested that when controlling for census region and when controlling for state's percent SWD, the average NAEP proficiency percentage for SWD students was about 9% proficient. Relative to the Northeast region, other regions of the country have lower percentages of SWD students proficient in the eighth grade NAEP reading proficiency percentage. Specifically, states in the South have about 4.769% fewer, states in the Midwest have about 3.344% fewer, and states in the West have about 5.243% fewer eighth grade students demonstrating proficiency on the 2009 NAEP reading assessment.

Accuracy in predicting SWD students' NAEP reading proficiency percentage by census region was moderately strong, with a multiple correlation coefficient of .561. About 32% (R^2 = .315) of the variance of NAEP reading proficiency percentage was accounted for by the regression model in Table 19.

Research Question Six: Testing Assumptions with Outliers Removed

Once the outliers were removed (case numbers 21 and 31), a review of Cook's distance (between .000 and .181) suggested no outliers. Neither of the centered leverage values (.042 and .170) was high enough to suggest the possibility of outliers (See Table 20). However, several scatterplots suggested the presence of outliers.

Table 20. Residuals Statistics

	Minimum	*Maximum*	*Mean*	*Std. Deviation*	*N*
Cook's Distance	.000	.181	.022	.031	49
Centered Leverage Value	.042	.170	.082	.030	49

Multiple linear regression assumptions were tested but not met. Partial regression plots for the dependent to independent variables (Appendix F, Figures 103-106) did not seem to suggest linearity, as the data points fell randomly. As state reading proficiency percentage increases for SWD students, NAEP reading proficiency percentage may increase slightly for SWD students as well, while controlling for census region. Furthermore, the scatterplots of studentized residuals to unstandardized predicted values (Appendix F, Figure 107) and to the primary independent variable (Appendix F, Figure 108) showed little apparent linearity. In addition, some of the studentized residuals

indicated outliers, as there were a few plot points with values well outside the band of +/- 2.

With the outliers removed, unstandardized residuals were reviewed and found to be normally distributed. Skewness (.866) and kurtosis (.167) statistics indicated normality, as they were within an absolute value of two. Significant Shapiro-Wilk test results continued to indicate abnormality ($W = .919$, $df = 49$, $p = .002$). The histogram and Q-Q plots for unstandardized residuals also indicated possible abnormality, with several outliers apparent (Appendix F, Figures 109 and 110). The boxplot of unstandardized residuals shown in Appendix F, Figure 111 did not indicate outliers (case numbers 21 and 31). A scatterplot of studentized residuals to the primary independent variable (Appendix F, Figure 112) indicated the assumption of independence was met, since the data points fell randomly with no apparent pattern to the points. In addition, scatterplots of studentized residuals to unstandardized predicted *Y* (Appendix F, Figure 113) and studentized residuals to case number (Appendix F, Figure 114) suggested that the assumption of independence was appropriate, because the points fell randomly. In addition, the scatterplot of studentized residuals to unstandardized predicted values (Appendix F, Figure 113) suggested homogeneity of variance was not violated, as the predicted values did not increase with increased residual values, there was no pattern to the data points, and they were randomly scattered around zero .

Tolerance values were greater than .10 (.917, .478, .501, and .518), and variance inflation factors were less than 10 (1.090, 2.090, 1.995, and 1.930). Most eigenvalues

were not close to zero (2.784, 1.006, and 1.000); however, two values were close to zero (Midwest SWD = .139 and West SWD = .071). None of the condition indices was greater than 15 (1.000, 1.664, 1.669, 4.476, and 6.269). There does not seem to be a problem with multicollinearity.

Research Question Six: Regression Results With Outliers Removed

After removing the two outliers (case numbers 21 and 31), census region and state reading assessment proficiency percentages for SWD students remained statistically significant predictors of NAEP reading proficiency percentages for SWD students, F (4, 44) = 5.788, $p < .001$. The regression equation for predicting NAEP reading proficiency percentage for SWD students as a result of state reading assessment proficiency percentage for SWD students and census region is shown in Table 21 and expressed as follows:

Eighth grade NAEP reading proficiency percentage for SWD students = 8.541 + .003 (state assessment proficiency percentage) - 4.276 (South) – 2.244 (Midwest) – 4.322 (West)

Table 21. Multiple Regression

	Model	*B*	*Standard Error*	*T*	*P*
1	(Intercept)	8.541	1.024	8.342	.000
	State_SD	.003	.020	.139	.890
	South	-4.276	1.031	-4.157	.000
	Midwest	-2.244	1.098	-2.042	.047
	West	-4.322	1.053	-4.106	.000

R	.587
R^2	.345
F	5.788

The model shown in Table 21 suggests that when controlling for census region and when controlling for state's percent SWD, the average NAEP proficiency percentage for SWD students was about 9% proficient. Relative to the Northeast region, other regions of the country have lower percentages of SWD students proficient in the eighth grade NAEP reading proficiency percentage. Specifically, states in the South have about 4.276% fewer, states in the Midwest have about 2.244% fewer, and states in the West have about 4.322% fewer eighth grade students demonstrating proficiency on the 2009 NAEP reading assessment.

Accuracy in predicting SWD students' NAEP reading proficiency percentage by census region was moderately strong, with a multiple correlation coefficient of .587. About 35% (R^2 = .345) of the variance of NAEP reading proficiency percentage was accounted for by the regression model in Table 21.

When controlling for census regions, there was a significant relationship between percent of SWD eighth grade students proficient on state reading assessments and percent of SWD eighth grade students proficient on NAEP reading assessments. The correlation between the percent of SWD students proficient on these two assessments was strong (Cohen, 1988). For every one percent increase on the state assessment, the NAEP increases by approximately .01%. The relationship between percent of SWD students

proficient was significant in all census regions. Proficiency percentages for eighth grade SWD students in the South census region were significantly lower than in the Northeast. In addition, percent proficient for eighth grade students in the West census region was significantly lower than the Northeast, but higher than the South. Eighth grade SWD students in the Northeast demonstrated the highest proficiency percentages, followed by the Midwest, South, and West, in that order. Removing outliers did not bring about a change in significance levels.

Research Question Seven

On average does the difference between the percentage of eighth grade students demonstrating proficiency on state and NAEP reading assessments change from 2007 to 2009?

A comparison of differences was conducted to examine the discrepancies between the percentage of eighth grade students demonstrating proficiency on state and NAEP reading assessments in 2007 and 2009. First, aggregate proficiency percentages for all states on both NAEP and state assessments in 2007 (Appendix B) and 2009 (Appendix C) were reviewed.

In 2007, the mean proficiency percentage on NAEP was 29.98%, and the median was 31. Census region averages were as follows: Northeast, 36% proficient and above; Midwest, 32% proficient and above; South, 27% proficient and above; and West, 25% proficient and above. In 2009, the mean proficiency percentage on NAEP was 31.10%, and the median was 32. Census region averages were as follows: Northeast, 38%

proficient and above; Midwest, 34% proficient and above; South, 28% proficient and above; and West, 26% proficient and above.

In 2007 for state assessments, the mean proficiency percentage was 69.46%, and the median was 72%. In 2009 for state assessments, the mean proficiency percentage was 72.57%, and the median was 71%.

Research Question Seven: Top Scoring States

In 2007, the state with the highest eighth grade NAEP reading proficiency percentage was Massachusetts, with 43% of its students demonstrating proficiency. In 2009, the states with the highest eighth grade NAEP reading proficiency percentages were Connecticut and Massachusetts, with 43% of their students demonstrating proficiency.

In 2007, the state with the highest eighth grade state reading assessment proficiency percentages was Tennessee, with 92% of its students demonstrating proficiency. In 2009, the state with the highest eighth grade state reading assessment proficiency percentage was Nebraska, with 92% of its students demonstrating proficiency.

As shown in Table 22, there was no commonality among top scorers on the NAEP and top scorers on state reading assessments in 2007. As shown in Table 23, there was no commonality among top scorers on the NAEP and top scorers on state reading assessments in 2009.

Table 22. A Comparison of the Top 10 Scoring States for Eighth Grade NAEP and State Reading Assessment Proficiency percentages in 2007

State	*State Assessment Percent Proficient*	*State*	*NAEP Assessment Percent Proficient*
Tennessee	92.1	Massachusetts	43
Nebraska	90.7	Vermont	42
Georgia	88.9	Montana	39
North Carolina	87.9	New Jersey	39
Texas	87.5	Connecticut	37
Colorado	86.6	Maine	37
Idaho	85.8	Minnesota	37
Wisconsin	84.1	New Hampshire	37
Illinois	80.9	South Dakota	37
Utah	80.8	Ohio	36

Table 23. A Comparison of the Top 10 Scoring States for Eighth Grade NAEP and State Reading Assessment Proficiency percentages in 2009

State	*State Assessment Percent Proficient*	*State*	*NAEP Assessment Percent Proficient*
Nebraska	95.2	Massachusetts	43
Texas	94.3	Connecticut	43
Georgia	93.8	New Jersey	42
Tennessee	92.6	Vermont	41
Idaho	91.5	Pennsylvania	40
Colorado	88.5	New Hampshire	39
Virginia	87.4	Montana	38
Kansas	85.4	Minnesota	38
Wisconsin	85.2	South Dakota	37
Illinois	83.4	Ohio	37

Research Question Seven: Bottom Scoring States

In 2007, the states with the lowest eighth grade NAEP reading proficiency percentages were New Mexico and Mississippi, with 17% demonstrating proficiency. In 2009, the state with the lowest eighth grade NAEP reading proficiency percentage was Mississippi, with 19% demonstrating proficiency.

In 2007, the state with the lowest eighth grade state reading assessment proficiency percentage was South Carolina, with 35% demonstrating proficiency. In 2009, the state with the lowest eighth grade state reading assessment proficiency percentage was California, with 48% demonstrating proficiency.

A comparison of 2007 bottom scorers in Table 24 reveals that half of the poorest performing states on NAEP also showed low state proficiency percentages. However, the following five states reported poor state proficiency percentages but did not perform in the bottom 10 on NAEP: Florida, Missouri, New York, Rhode Island, and South Carolina. Of these five states, all demonstrated proficiency percentages in the middle third on NAEP, except South Carolina. The precise NAEP ranking of each state was as follows: Florida (34), Missouri (27), New York (24), Rhode Island (36), and South Carolina (40).

A comparison of 2009 bottom scorers in Table 25 reveals that seven of the poorest performing states on NAEP also showed low state proficiency percentages. However, the following three states reported poor state proficiency percentages but did not perform in the bottom 10 on NAEP: Florida (30), Missouri (18), and Rhode Island

(35). Of these three states, all demonstrated proficiency percentages in the middle third on NAEP, except Missouri (which was in the top third). The precise NAEP ranking of each state was as follows: Florida (30), Missouri (18), and Rhode Island (35).

Table 24. A Comparison of the Bottom 10 Scoring States for Eighth Grade NAEP and State Reading Assessment Scores in 2007

State	*State Assessment Percent Proficient*	*State*	*NAEP Assessment Percent Proficient*
Washington, D.C.	17	Washington, D.C.	12
South Carolina	35	Mississippi	17
California	42	New Mexico	17
Missouri	43	Louisiana	19
Florida	49	Hawaii	20
Mississippi	52	California	21
New Mexico	56	Alabama	21
Nevada	57	Nevada	22
New York	57	West Virginia	23
Rhode Island	58	Arizona	24

Table 25. A Comparison of the Bottom 10 Scoring States for Eighth Grade NAEP and State Reading Assessment Scores in 2009

State	*State Assessment Percent Proficient*	*State*	*NAEP Assessment Percent Proficient*
Washington, D.C.	46	Washington, D.C.	14
California	48	Mississippi	19
Mississippi	48	Louisiana	20
Missouri	50	West Virginia	22
Florida	54	New Mexico	22
West Virginia	61	Nevada	22
Nevada	61	Hawaii	22
Louisiana	61	California	22
Rhode Island	62	South Carolina	24
New Mexico	63	Alabama	24

Research Question Seven: A Comparison of Differences in Proficiency Percentages

In Appendix D, the difference between the 2007 eighth grade NAEP and state reading assessment percent proficient are presented. On average, the mean difference between the two assessments in 2007 was 39 points. As shown in Table 26, the state with the greatest difference in 2007 was Tennessee, with a 66% difference between the percent of students demonstrating proficiency NAEP (26%) and the percent demonstrating proficiency on its own state exam (92%). The territory with the least difference in 2007 was Washington, D.C., with a 5% difference between the percent of students demonstrating proficiency on NAEP (12%) and the percent demonstrating proficiency on its own state exam (17%). The state with the least difference in 2007 was South Carolina, with a 10% difference between the percent of students demonstrating proficiency on NAEP (25%) and the percent demonstrating proficiency on its own state exam (35%). Twenty-one states had differences greater than 41 points. Six states had differences less than 25 points.

In Appendix E, the difference between the 2009 eighth grade NAEP and state reading assessment percent proficient are presented. On average, the mean difference between the two assessments in 2009 was 41 points. As shown in Table 26, the state with the greatest mean difference in 2009 was Texas, with a 67% difference between the percent of students demonstrating proficiency on NAEP (27%) and the percent demonstrating proficiency on its own state exam (94%). The state with the least

difference in 2007 was Missouri, with a 16% difference between the percent of students demonstrating proficiency on NAEP (34%) and the percent demonstrating proficiency on its own state exam (50%). Twenty-three states had differences greater than 41 points. Two states had differences less than 25 points. As shown in Table 26, seven of the 10 states with the greatest differences in percent proficient and above in 2007 remained in the top ten for 2009.

Table 26. A Comparison of Percentage Differences Between NAEP and State Eighth Grade Reading Assessments in 2007 and 2009

State	*Difference Between State and NAEP Percent Proficient in 2009*	*State*	*Difference Between State and NAEP Percent Proficient in 2007*
Texas	67.3	Tennessee	66.1
Georgia	66.8	Georgia	62.9
Tennessee	64.6	North Carolina	59.9
Nebraska	60.2	Texas	59.5
Idaho	58.5	West Virginia	57.2
Colorado	56.5	Nebraska	55.7
Virginia	55.4	Idaho	53.8
Alaska	54.7	Alaska	52.2
South Carolina	53.5	Oklahoma	51.7
Kansas	52.4	Colorado	51.6

Mean Difference	41.5	Mean Difference	39.5

Summary

In this chapter, an introduction was given regarding the analysis and statistical tests that were to be discussed. This was followed by a restatement of the purpose of study. In addition, the seven research questions were presented as a whole. The demographics pertaining to the study, as well as data analysis methods, were stated. Next, the results of each research question were detailed.

CHAPTER FIVE: CONCLUSIONS AND RECOMMENDATIONS

Introduction

In the previous chapter, data findings were presented, reported, and analyzed. Chapter Five is comprised of a summary of the study, discussion of the findings, limitations, implications for practice, recommendations for further research, and conclusions. The latter sections of this chapter are intended to explore the concepts of this study so as to permit the use of findings by legislators and policymakers, school leaders, and practitioners. Limitations of this study are provided to caution readers about the best use of these results to prevent misuse or misunderstanding. Implications of this study are examined to give policymakers informed recommendations about the use of NAEP and state assessments for accountability purposes. Suggestions for further research are given to offer ideas for those who wish to conduct future studies on the relationship of state and national assessments. Finally, concluding remarks are made to synthesize the contribution of this study to current knowledge on the relating of NAEP and state assessments and the use of these assessments for NCLB accountability purposes.

Summary of the Study

This chapter begins with a summary of the purpose and significance of the study and is followed by major findings related to the predictive relationships found between eighth grade NAEP and state reading assessment proficiency percentages. Conclusions

from the study's findings are discussed in relation to criteria to determine if assessment-relating studies should be conducted. In addition, conclusions are discussed in light of the implications of these findings for NCLB and AYP policy.

The purpose of this study was to determine if there was a predictive relationship between 2009 NAEP eighth grade reading assessment proficiency percentages and 2009 eighth grade reading state assessment proficiency percentages. Additionally, data were disaggregated into the four census regions of NAEP to make comparisons between the total populations of each region and specified AYP subgroups. In particular, this study extended Gordon's (2009) study to also control for percentage of subgroups meeting proficiency on state assessments.

Simple and multiple regression analyses, in addition to descriptive analyses of percent proficient from 2007 to 2009, were performed to study the seven research questions. Results from the first six research questions indicated that it was possible to use eighth grade state reading assessment proficiency percentages to construct a predictive model for eighth grade NAEP reading assessment proficiency percentages. Moderate to strong positive correlations were found for the total population, census regions, ELL students, low SES students, and students with disabilities.

Analysis of the regression equations for each census region showed patterns of performance. Difference in percent proficient data from Research Question Seven demonstrated wide disparities between NAEP and state assessment results for both 2007 and 2009. Data results from this study were consistent with results found by Gordon

(2009) and other researchers (Hess, 2005; Taylor & Gordon, in press) who have found wide disparities between state and NAEP proficiency percentages. A complete discussion of the results of each question is presented below.

When correlation strengths among variables were compared (*R* values ranging from .392 to .865 as seen in Table 23), it was evident that a moderate to strong relationship exists between NAEP and state reading assessment proficiency percentages among eighth grade students. Generally, the models suggest that as percent proficient on the state increases by 1%, percent proficient on the NAEP increases from .003% to .964%, holding all else in the model constant. Proficiency percentages were consistently lower on NAEP reading assessments than on state reading assessments.

Table 27. A Comparison of Correlations for Each Research Question

	Research Question 1: All 8th Graders	*Research Question 2: By Census Region*	*Research Question 3: Race/ Ethnicity and Census Region*	*Research Question 4: ELL*	*Research Question 5: FRL*	*Research Question 6: SWD*
R	.392	.748	.777	.865	.572	.587
R^2	.154	.560	.603	.749	.327	.345

When performance among census regions across subgroups are compared by examining coefficients (Table 24), inconsistencies surface. Results were not always statistically significant, and performance varied across subgroups. Eighth grade students in the Northeast most often demonstrated the highest proficiency percentages, but not

among ELL students. Eighth grade students in the South demonstrated the lowest proficiency percentages overall, but the highest proficiency percentages among ELL students and higher proficiency percentages for SWD students than the West census region. Altogether, the variance in assessment relationships demonstrated that eighth grade NAEP and state reading assessments are not ideally suited for relating, as stated in the findings of Kolen and Brennan (2004) among others.

Table 28. A Comparison of Coefficients for Each Research Question

	Research Question 1: All 8th Graders	*Research Question 2: By Census Region*	*Research Question 3: Race/ Ethnicity and Census Region*	*Research Question 4: ELL*	*Research Question 5: FRL*	*Research Question 6: SWD*
(Intercept)	15.665	24.813	9.080*	1.642	18.878	8.541
State	.213	.184	.679	.057	.018*	.003*
South		-11.484	-7.355	2.419	-5.014	-4.276
Midwest		-4.551	-4.583	1.106*	-.851*	-2.244
West		-8.456	-4.666*	-1.161*	-3.643	-4.322
American Indian			-.099			
Asian			.072			
Black			-.374			

* denotes not results were not significant (p >.05)

Discussion of the Findings

Previous researchers (Bandeira de Mello et al., 2009; Carnoy & Loeb, 2002; Dorans, 2004; Ercikan, 1997; Feuer, 1999; Gordon, 2009; Hombo, 2003; Kolen & Brennan, 2004; Linn, 1993; Linn & Kiplinger, 1995; Mislevy, 1992; Pommerich et al.,

2004; Prowker & Camilli, 2007; Taylor & Gordon, in press; Waltman, 1997) have extensively analyzed the issues surrounding relating NAEP and state assessments. This study relied on their findings to inform its approach to the research questions and in the interpretation of the following results. The goal of this study was to examine each population subgroup to fully explore the extent to which eighth grade state reading assessment proficiency percentages can predict eighth grade NAEP reading assessment proficiency percentages. This section discusses the results and implications of the findings for each of the seven research questions.

Kolen and Brennan (2004) referenced the Mislevy/Linn framework, which provided a conceptual model for relating assessments based on four methods: equating, calibration, statistical moderation, and projection. Mislevy (1992) and Linn (1993) discussed use of regression methodology to project, or predict proficiency percentages from one test by using proficiency percentages from another assessment. Specifically, Linn concluded the degree to which one assessment is comparable to another depends on how similar are the tests' questions, conditions, and cognitive complexity. Both Mislevy and Linn emphasized that the results from projection studies were situation, time, and group dependent. For that reason, this study investigated the relationship between assessment results for each subgroup, rather than relying on the results of the aggregated population.

Kolen and Brennan (2004), who studied the methodology of relating distinct assessments, asked four questions to determine whether test results can be related and

how results should be used:

1. Are the two tests "used to draw similar inferences?"
2. "Do the two tests measure the same constructs?"
3. "Are the two tests designed to be used with the same population?"
4. "Do the tests share common measurement conditions: …test length, test format, administration conditions, etc." (p. 224)

Using these criteria, Kolen and Brennan (2004) concluded that NAEP and state assessments were dissimilar in all areas. Despite their caution, numerous researchers (Bandeira de Mello et al., 2009; Carnoy & Loeb, 2002; Dorans, 2004; Ercikan, 1997; Feuer, 1999; Gordon, 2009; Hombo, 2003; Kolen & Brennan, 2004; Linn, 1993; Linn & Kiplinger, 1995; Mislevy, 1992; Pommerich et al., 2004; Prowker & Camilli, 2007; Taylor & Gordon, 2010; Waltman, 1997) have continued to explore relationships between NAEP and state assessment proficiency percentages. The following research questions are explored below with respect to links between results and relevant literature, as well as with respect to Kolen and Brennan's criteria for relating assessments.

Research Question One

To what extent can the percentage of eighth grade students demonstrating proficiency on the 2009 NAEP reading assessment be predicted by the percentage of eighth grade students demonstrating proficiency on the 2009 state reading assessment?

There is a significant relationship between percent proficient on eighth grade state reading assessments and percent proficient on eighth grade NAEP reading assessments. The correlation between the percent proficient on these two assessments is moderate

(Cohen, 1988). In every state, the percent proficient on eighth grade NAEP reading assessments is lower than the percent proficient on state assessments.

The findings resulting from Research Question One indicated a positive and significant relationship between the percentage of eighth grade students demonstrating proficiency on the 2009 NAEP reading assessment and the percentage of eighth grade students demonstrating proficiency on 2009 state reading assessments. With a correlation of .392, the relationship between the two assessment proficiency percentages was moderate. Approximately 15% of eighth grade NAEP reading assessment proficiency percentage could be explained by state assessment proficiency percentage value. This finding indicated there were many other variables contributing to the NAEP proficiency percentage.

The correlation resulting from this research question was strong enough to construct a predictive model, which indicated 16% as the average NAEP assessment proficiency percentage controlling for state assessment proficiency percentage. The regression coefficient for *x* (.213) indicated that for every one percent increase on the state assessment, the NAEP increases by only about 1/5 of 1% or approximately .21%. The very small number here indicates a great deal of difference between NAEP and state assessment proficiency percentages. In other words, this equation reveals that NAEP proficiency percentages are, on average, much lower than state proficiency percentages. A moderate correlation between these tests suggests there may not be the similarity between NAEP and state assessment proficiency scales that NCLB creators envisioned

when they required that states use NAEP performance standards as a guide when they created their own (Bourque, 2009). The correlation value between the assessments is not strong enough to suggest convergent validity between NAEP and state assessments.

The disparity between proficiency scales hearkens back to Kolen and Brennan's (2004) four questions. Disparity between assessment constructs violates their second requirement for assessment-relating validity that the two tests to be related should measure the same constructs (Kolen and Brennan, 2004). If NAEP and state assessments are significantly different, what is the value in conducting a study to relate assessments? Mislevy (1992) found value in using assessment-relating studies to make assessment performance projections; however, he cautioned that results could be unstable and might shift over time. Thus, Mislevy recommended that relating analyses be conducted regularly to strengthen findings. To that end, it is of interest to note that Gordon (2009) found a .327 correlation when he asked the same question posed in Research Question One of this study for 2007 results. The similarity of the correlation findings between this study (.392) and his (.327) corroborates his findings.

Another possible cause of variation seen in this study of 2009 eighth NAEP and state reading assessment proficiency percentages lies in different testing conditions. Differences in this area violate Kolen and Brennan's fourth requirement for common measurement conditions (2004). Whereas NAEP is a voluntary, low-stakes test, many state assessments are mandatory, high-stakes tests. Students tend to perform more

commensurately with their ability when the stakes are higher; for some students, there is less motivation to exert true effort on a voluntary assessment (Waltman, 1997).

Unlike most studies attempting to relate NAEP and state assessments that have used eighth grade populations, Waltman (1997) used fourth grade students because she believed they would be less vulnerable to differences in motivation. She concluded that her results were more stable than those reported in her colleagues' (Ercikan, 1997; Linn & Kiplinger, 1995) studies, perhaps because fourth graders demonstrate less variability in motivation from test to test (Waltman).

Research Question Two

To what extent can the percentage of eighth grade students demonstrating proficiency on the 2009 NAEP reading assessment be predicted by the percentage of eighth grade students demonstrating proficiency on the 2009 state reading assessment, controlling for census regions defined by NAEP?

When controlling for census regions, there is a significant relationship between percent proficient on eighth grade state reading assessments and percent proficient on eighth grade NAEP reading assessments. The correlation between the percent proficient on these two assessments is strong (Cohen, 1988). The relationship between percent proficient is significant in each of the four census regions. The Northeast demonstrated the highest proficiency percentages, followed by the Midwest, West, and South, in that order.

The findings resulting from Research Question Two indicated a positive and significant relationship between the percentage of eighth grade students demonstrating

proficiency on the 2009 NAEP reading assessment and the percentage of eighth grade students demonstrating proficiency on the 2009 state reading assessment, controlling for census regions defined by NAEP. With a correlation of .748, the relationship between the two assessment proficiency percentages controlling for census regions was strong. Approximately 56% of eighth grade NAEP reading assessment proficiency percentage could be explained by state assessment proficiency percentage value and census region. The regression coefficient for *x* (.184) indicated that for every one percent increase on the state assessment, the NAEP increases by only about 1/5 of 1% or less than .20%.

Research Question Two echoed Research Question One, but added the element of controlling for census region. When census region was controlled for, the predictive strength of this correlation was stronger than Research Question One. This finding indicated that census region is a very significant contributor to NAEP proficiency percentage. In addition, this finding indicated that the correlation between state and NAEP proficiency percentages is consistent nationwide.

The correlation resulting from this research question was sufficiently strong to construct a predictive model, which indicated 25% as the average NAEP assessment proficiency percentage before adjusting for state assessment proficiency percentage and census region. Interestingly, controlling for census region allowed the analysis of differences in proficiency percentage trends across the United States. The model constructed predicted that NAEP proficiency percentages in the South census region would be about 11% below those in the Northeast region; proficiency percentages in the

Midwest region would be nearly 5% below the Northeast region; and proficiency percentages in the West region would be approximately 8% below the Northeast region. Such wide variations in proficiency percentage by census region indicate striking differences in state assessment proficiency percentages and/or student achievement.

In 1984, the National Assessment Policy Committee, the committee charged with overseeing NAEP before NAGB, agreed to help states compare their state level assessments to NAEP as a way of determining validity (Vinovskis, 1998). When achievement levels were addressed in NCLB, the policy required states to set standards for their state assessments using NAEP achievement levels (Basic, Proficient, and Advanced (NCLB, 2002). Despite these guidelines, the results of Research Question Two demonstrate significant differences in the proficiency scales states use.

Bourque's (2009) review of performance standards among states found that states use proficiency scales with a different number of levels: "12 states use a 5-level system, 29 use a 4-level system, 10 use a 3-level system, and 1 uses a 6-level system" (p. 23). These differences in scale caused differences in where states position the "proficient" level. While some positioned proficient at the second highest level of a 3-level scale, others positioned proficient at the third highest level on a 5- or 6-level scale. Bourque noted that the difference in levels among assessments proficiency placement has "the likely effect of depressing the definition of Proficient…. [and] the definition of Proficient can vary from state to state" (p. 23).

The results of this research question indicate that state assessment proficiency percentages in the Northeast region have a higher degree of shared variance, or construct validity, with NAEP proficiency percentages than any other census region. In their study of NAEP proficiency percentages, Carnoy and Loeb (2002) found that students in high-accountability states significantly outperformed students in states with less rigorous accountability standards. Carnoy and Loeb's results suggest that one will find a positive relationship between NAEP proficiency percentages and assessment proficiency percentages in states that have higher accountability standards and, correspondingly, more difficult tests. Conversely, their study suggests that lower accountability standards and less rigorous assessments would lead to lower NAEP proficiency percentages. Accordingly, Carnoy and Loeb's conclusions suggest that more states in the Northeast census region have higher accountability standards and more rigorous tests. At the very least, the Northeast's high performance on NAEP tests suggest that the standards and rigor of the Northeast state tests more closely mirror NAEP tests than do the state assessments from any other region.

Research Question Three

To what extent can the percentage of eighth grade students demonstrating proficiency on the 2009 NAEP reading assessment be predicted by the percentage of eighth grade students demonstrating proficiency on the 2009 state reading assessment, controlling for each of the five major racial/ethnic groups identified as subgroups by NCLB: American Indian, Asian, Hispanic, Black, White; and controlling for census regions defined by NAEP?

When controlling for race, ethnicity, and census regions, there was a significant relationship between percent proficient on eighth grade state reading assessments and percent proficient on eighth grade NAEP reading assessments. The correlation between the percent proficient on these two assessments was strong (Cohen, 1988). The relationship between percent proficient was determined to be significant for percent proficient in the Black subgroup, as well as in the South census region (where percent proficient was lower than all other regions).

The findings resulting from Research Question Three indicated a positive, significant overall relationship between the total percentage of eighth grade students demonstrating proficiency on the 2009 NAEP reading assessment and the percentage of total eighth grade students demonstrating proficiency on the 2009 state reading assessment, controlling for census regions defined by NAEP and percent proficient by race/ethnicity. Results were significant for percent proficient in the Black subgroup, as well as the South census region. With an overall correlation of .777, the relationship between the two assessment proficiency percentages, controlling for census regions defined by NAEP and percent proficient by race/ethnicity, seemed strong. Approximately 60% of eighth grade NAEP reading assessment proficiency percentage could be explained by state assessment proficiency percentage value, census region, and percent proficient by race/ethnicity; however, the apparent strength of this finding may be weakened once issues with multicollinearity are explored. The regression coefficient for *x*

(.679) indicated that for every one percent increase on the state assessment, the NAEP increases by about 1%.

Research Question Three echoed Research Question Two, but supplied additional independent variables to control for census region and percent proficient by race/ethnicity. The resulting regression contained nine independent variables. Problems with multicollinearity prevented a complete answer to Research Question Three, which sought to determine whether 2009 eighth grade state reading assessment proficiency percentages could be used to predict 2009 eighth grade NAEP reading proficiency percentages for every race/ethnicity in each of the four census regions.

The correlation resulting from this research question constructed a predictive model, which indicated 9% as the average NAEP assessment proficiency percentage controlling for state assessment proficiency percentage, census region, and percent proficient by race/ethnicity. Given problems with multicollinearity, caution should be used in drawing conclusions here.

Small populations in some of the race/ethnicities could also contribute to confounding results for Research Question Three. Because NCLB allowed states to set differing minimum values for subgroups and results from subgroups without that minimum value are not reported, subgroup performance on NAEP and state assessments may not have been reported in the same proportions. This consideration is especially relevant for subgroups that have relatively small populations nationwide (American Indian and Asian, for example); as a result, data for small subgroups are very sensitive to

small changes. When multiple eigenvalues are close to zero, this means that small changes in data values can lead to large changes in coefficient estimates. In other words, when subgroups are relatively small, small changes in data can make a big difference. When it comes to discerning whether NAEP and state assessments are administered to the same population, AYP rules that affect which subgroups are reported for state assessments do not apply for NAEP testing. Therefore, the results of this study suggest that subgroup performance among states will show small but meaningful differences between NAEP and state assessments.

If one returns to Kolen and Brennan's (2004) four requirements for relating assessments, it seems that Research Question Three revealed a potential problem with the requirement that the two tests assess the same population. AYP reporting rules create differences in the racial and ethnic makeup of eighth grade populations sampled between NAEP and state assessments and may cause a violation of Kolen and Brennan's third rule.

Research Question Four

To what extent can the percentage of eighth grade ELL students demonstrating proficiency on the 2009 NAEP reading assessment be predicted by the percentage of eighth grade ELL students demonstrating proficiency on the 2009 state reading assessment, controlling for census regions defined by NAEP?

When controlling for census regions, there was a significant relationship between the percent of ELL eighth grade students proficient on state reading assessments and the percent of ELL eighth grade students proficient on NAEP reading assessments. The

correlation between the percent of ELL students proficient on these two assessments, controlling for census region, was strong (Cohen, 1988). It was determined that percent proficient for eighth grade ELL students in the South census region was significantly higher than in the Northeast, as well as higher than the other regions.

Findings resulting from Research Question Four indicated a positive and significant relationship between the percentage of eighth grade ELL students demonstrating proficiency on the 2009 NAEP reading assessment and the percentage of eighth grade ELL students demonstrating proficiency on the 2009 state reading assessment. The relationship was significant for aggregate scores, but the only census region that tested significant was the South. With a correlation of .865, the relationship between the two assessment proficiency percentages, controlling for census region, seemed strong. Approximately 75% of eighth grade NAEP reading assessment proficiency percentage for ELL students could be explained by state assessment proficiency percentage value and census region. The regression coefficient for *x* (.057) indicated that for every one percent increase on the state assessment, the NAEP increases by only about 1/20 of 1% or approximately .06%.

Research Question Four echoed Research Question Two, but asked the question for the ELL subgroup rather than the total sampled population. This finding indicated that the strong positive correlation between state and NAEP proficiency percentages was consistent throughout the nation with the ELL subgroup.

The correlation resulting from this research question constructed a predictive model, which indicated 2% as the average NAEP assessment proficiency percentage for ELL students controlling for state assessment proficiency percentage and census region. The model constructed predicted that NAEP proficiency percentages for ELL students in the South census region would be 2% above those in the Northeast region.

When the performance of different subgroups is analyzed, variations in proficiency percentage by census region show similar trends. Based on the results of this and Gordon's (2009) study, proficiency percentages in the Northeast census region are highest, followed by the Midwest, West, and, lastly, the South. This trend held true for total population, percent proficient by race/ethnicity for the most part, low SES students, and students with disabilities for the most part. However, results for Research Question Four showed that ELL students do not follow the same regional trend.

With the performance of ELL students, the South and Midwest census regions reported better proficiency percentages than the Northeast, and the West was only slightly behind Northeast proficiency percentages. This phenomenon could be due to statistical problems related to the small variation in scores when dealing with low proficiency percentages, AYP subgroup reporting issues, or differences in ELL student achievement.

The intercept value (1.642) represents the average NAEP assessment proficiency percentage for ELL students before adjusting for state assessment proficiency percentage and census region. This intercept value means that NAEP proficiency percentages for

ELL students are very low, hovering in the single digits. Given that ELL students are new to the English language, low reading assessment proficiency percentages are understandable. Nevertheless, these low proficiency percentage values can be problematic for statistical research because the relatively small range of values leads to minimum variation in proficiency percentages, which can sometimes create problems with producing valid results. In this vein, Linn and Kiplinger (1995) expressed reservations about the usefulness of comparing NAEP and state assessments for very high or very low proficiency percentages. Instead, they found assessment-relating results to be most useful for making estimates about average state performance (Linn & Kiplinger).

Bearing in mind Kolen and Brennan's (2004) four requirements for relating assessments, subgroup population issues again were likely to have violated their third requirement: that the tests be given to the same populations. The same issues that affect racial and ethnic subgroups with small populations are especially present with an even smaller population of ELL students. It is likely that a state would not have enough ELL students to report as a subgroup. However, since NAEP is not bound by NCLB's AYP subgroup reporting guidelines, proficiency percentages for ELL students would not be reported in the same way (NCES, 2009a). This reporting detail would cause different proportions of ELL students to be reported in some states. As discussed in Research Question Three, when the numbers in a sampled population are small, small changes—such as small variations in reporting—can create validity issues in an assessment-relating study.

Nevertheless, the variation of ELL performance from typical census region patterns—especially a departure from the trend of Northeast dominance—was noteworthy and could indicate differences in support systems for ELL students among states. This issue is discussed further in the conclusion.

Research Question Five

To what extent can the percentage of eighth grade students qualifying for free and reduced lunch who demonstrate proficiency on the 2009 NAEP reading assessment be predicted by the percentage of eighth grade students qualifying for free and reduced lunch who demonstrate proficiency on 2009 state reading assessments, controlling for census regions defined by NAEP?

When controlling for census regions, there was a significant relationship between percent of FRL eighth grade students proficient on state reading assessments and percent of FRL eighth grade students proficient on NAEP reading assessments. The correlation between the percent of FRL students proficient on these two assessments, controlling for census region, was strong (Cohen, 1988). The relationship between percent of FRL students proficient on NAEP was significant in only the South and West census regions. Percent proficient on NAEP for eighth grade FRL students in the South census region was significantly lower than in the Northeast. In addition, percent proficient for eighth grade students in the West census region was significantly lower than the Northeast, but higher than the South.

The findings resulting from Research Question Five indicated a positive and significant relationship between the percentage of eighth grade low SES students demonstrating proficiency on the 2009 NAEP reading assessment and the percentage of

eighth grade low SES students demonstrating proficiency on the 2009 state reading assessment, controlling for census regions defined by NAEP. The relationship was statistically lower in only the South and West census regions. With a correlation of .532, the relationship between the two assessment proficiency percentages (controlling for census region) appeared strong. Approximately 28% of eighth grade NAEP reading assessment proficiency percentage for low SES students could be explained by state assessment proficiency percentage value and census region. The regression coefficient for *x* (.018) indicated that for every one percent increase on the state assessment, the NAEP increases by only about 1/50 of 1% or approximately .02%.

Research Question Five echoed Research Question Two, but asked the question for the low SES subgroup rather than the total sampled population. This finding indicated that the strong positive correlation between state and NAEP proficiency percentages is consistent throughout the nation with the low SES subgroup.

The correlation resulting from this research question constructed a predictive model, which indicated 19% as the average NAEP assessment proficiency percentage for low SES students controlling for state assessment proficiency percentage and census region. The model constructed predicted that NAEP proficiency percentages for low SES students in the South census region would be about 5% below those in the Northeast region, and proficiency percentages in the West region would be nearly 4% below the Northeast region.

These findings also indicated that the correlation between NAEP and state assessments is weakest for the low SES subgroup when controlling for census region. This result is similar to that which Gordon (2009) found when he studied the same question. His results indicated that the proficiency percentages of FRL students could be predicted for only one of the four census regions tested by NAEP (Gordon). One possible reason for this weaker correlation could be due to the fact that low SES students are identified through those who qualify for free and reduced meal services. Because there can be a stigma to receiving free and reduced meals, not all eligible students were classified as low SES at the time of testing. This could cause the low SES population to be incomplete and the non-SES population to consist partially of low SES students.

Research Question Six

To what extent can the percentage of eighth grade students with disabilities demonstrating proficiency on the 2009 NAEP reading assessment be predicted by the percentage of eighth grade students with disabilities demonstrating proficiency on the 2009 state reading assessment, controlling for census regions defined by NAEP?

When controlling for census regions, there was a significant relationship between percent of SWD eighth grade students proficient on state reading assessments and percent of SWD eighth grade students proficient on NAEP reading assessments. The correlation between the percent of SWD students proficient on these two assessments was strong (Cohen, 1988). The relationship between percent of SWD students proficient was significant in all census regions. Percent proficient for eighth grade SWD students in the South census region was significantly lower than in the Northeast. Speaking

descriptively, percent proficient for eighth grade students in the West census region was significantly lower than the Northeast, but higher than the South. Eighth grade SWD students in the Northeast demonstrated the highest proficiency percentages, followed by the Midwest, South, and West, in that order.

The findings resulting from Research Question Six indicated a positive and significant relationship between the percentage of eighth grade SWD demonstrating proficiency on the 2009 NAEP reading assessment and the percentage of eighth grade SWD demonstrating proficiency on 2009 state reading assessments; this relationship remains significant in each of the census regions defined by NAEP. With a correlation of .587, the relationship between the two assessment proficiency percentages appeared strong. Approximately 35% of eighth grade NAEP reading assessment proficiency percentage for SWD could be explained by state assessment proficiency percentage value and census region. The regression coefficient for *x* (.003) indicated that for every one percent increase on the state assessment, the NAEP increases by only about 1/100 of 1% or less than .01%.

Research Question Six echoed Research Question Two, but asked the question for the SWD subgroup rather than the total sampled population. This finding indicated that the strong positive correlation between state and NAEP proficiency percentages controlling for census region is consistent throughout the nation with the SWD subgroup.

The correlation resulting from this research question constructed a predictive model, which indicated 9% as the average NAEP assessment proficiency percentage for

SWD controlling for state assessment proficiency percentage and census region. The model constructed predicted that NAEP proficiency percentages for SWD in the South census region would be 4% below those in the Northeast region; proficiency percentages in the Midwest region would be 2% below the Northeast region; and proficiency percentages in the West region would be 4% below the Northeast region.

The same caution attended to in Research Question Four regarding very low proficiency percentages should be observed when considering the proficiency percentages of the SWD subgroup, which had a intercept value (average NAEP proficiency percentage proficiency percentage before adjustments are made) of 8.541. Proficiency percentages in the single- and low double-digits are potentially problematic because the relatively small range of values leads to minimum variation in proficiency percentages. Linn and Kiplinger (1995) warned against comparing very low proficiency percentages when relating assessments.

Similarly, the same problem with tsubgroup sample makeup expressed in Research Question Five (related to those designated as FRL and, thus, low SES) could be at work in these results. Because not all students who have disabilities are known or wish to be classified as such, the makeup of this subgroup sample is incomplete and the non-SWD sample would have some students who would qualify for SWD services within its sample.

Research Question Seven

On average does the difference between the percentage of eighth grade students demonstrating proficiency on state and NAEP reading assessments change from 2007 to 2009?

The findings resulting from Research Question Seven indicated an increase in the mean difference between the percentage of eighth grade students demonstrating proficiency on state and NAEP reading assessments from 2007 to 2009. The mean difference between the two assessments was 39.5 in 2007; in two years, the mean difference increased to 41.5. This increase demonstrates that student achievement on state assessments gives a different perception of student proficiency than does student achievement on NAEP. For instance, parents and educators in Texas might be confused to learn that the Texas state assessment deems 94% of the state's eighth grade students proficient in reading, whereas NAEP finds only 27% of the same student population proficient in reading. It is unclear whether the discrepancy between the Texas NAEP and state assessment proficiency percentages was a result of different standards, different proficiency scales, different conditions, differences in rigor, other differences, or (likely) a combination of several of these four. From an NCLB standpoint, more schools in Texas will be likely to make AYP; from a NAEP standpoint, where proficiency percentages rank Texas as 36th among all states, such performance is cause for some discussion, if not alarm. One parents' guide to NAEP offers the following comment on discrepancies between NAEP and state proficiency reporting:

"NAEP data will highlight the rigor of standards and tests for individual states: If

> there is a large discrepancy between children's proficiency on a state's tests and their performance on NAEP, that would suggest that the state needs to take a closer look at its standards and assessments and consider making improvements" (U.S. Department of Education, 2003, p. 14).

In light of such a discrepancy—and the fact that such discrepancies are the norm—these words from the parent guide seem sound.

Hess (2005) noted that Peterson and Hess (2005) "documented the immense disparity in the rigor of state accountability systems, and the perverse reality that NCLB's AYP requirements make school performance look worse in states with more demanding accountability systems" (p. 53). States with especially rigorous assessments who set a high cut score for proficiency would understandably classify many more of their schools as in need of improvement; conversely, states with less rigorous assessments with a low cut score for proficiency would classify fewer schools in need of improvement (Hess).

Gordon (2009) determined there were significant disparities between the percent of students performing at proficiency on NAEP and state assessments, and his findings from studying 2007 data were confirmed through this analysis of 2009 proficiency percentage difference data.

If these gaps were consistent among states, the difference in scale might be overcome by creating concordant scores. However, Research Questions One through Six demonstrate that proficiency percentages between NAEP and state assessments vary somewhat predictably, but not nearly predictably enough to meet the rigorous standards

required to create concordant scores (Dorans, 2004). Moreover, when one compares the lists of states with top proficiency percentages on NAEP with the list of states with top proficiency percentages on state assessments (Tables 22 and 23), the complete lack of correspondence between top achieving states on both lists is most troublesome. Furthermore, differences in state assessments are creating a situation where schools that would make AYP according to some states' standards are failing AYP in others (Casserly, 2004). This situation does not seem in keeping with the spirit of equity behind NCLB.

Data from 2006 showed 27% of school districts in the U.S. failed to make AYP for two or more consecutive years; in Florida, that same number was a staggering 72% (McLester, p. 20). Such a difference among states, as compared to the national average, begs the question whether current assessments afford a fair comparison of schools making AYP.

Limitations

Inherent to this study are several limitations. As demonstrated through an exploration of Feuer's (1999) cautions and Kolen and Brennan's (2004) requirements for relating assessments, NAEP and state assessments are not easily suited for comparison. Their results are not used for the same purposes, do not measure the same standards, are not given to consistent populations, and are not administered under the same conditions. Despite these limitations, studies relating NAEP and state assessments are of interest to

those in the education field.

In addition, NAEP and state assessments are vulnerable to variations in testing motivation as demonstrated by Waltman (1997) in her study that showed eighth-graders are more susceptible to testing motivation issues than fourth-graders. Unlike many state assessments that are high-stakes exams, NAEP is a voluntary, low-stakes assessment. Scores from more savvy eighth-grade students will be more vulnerable to variations in motivation.

Another key limitation is that the data set included data aggregated to the state level. Since the data are aggregated, interpretations can be made at only the state level, not the individual student or school level. In addition, not all school types are included in this study. Also, each state has different rules for which students are included in its test. The sample of students used for NAEP testing is often small and may not represent a school district (or state) accurately. Moreover, differences in the difficulty of each state's test may affect the percentage of students demonstrating proficiency.

Pommerich et al. (2004) saw a cause for concern when assessment-relating studies were conducted between tests whose scores could not be equated. For scores to be equated, the two assessments must measure the same construct and must be expressed using the same metric (Dorans, 2004). Because of variations in assessment standards and proficiency scale differences, it is not appropriate to equate NAEP and state assessment scores (Kolen & Brennan, 2004). Pommerich et al. suggested that attempting to relate scores from tests that cannot be equated leads to a greater likelihood that related scores

could be misused or misinterpreted. They cautioned that it would be harder to use the results in a clear-cut way. Despite these reservations, Pommerich et al. do see merit in relating test results for distinct tests—provided that caution is used. The results of this study are useful only if those who read and interpret them understand the limitations.

Implications for Practice

Certainly, NCLB has transformed the way people talk about schools, measure educational achievement, and approach challenges (Hess, 2005). However, it is debatable whether NCLB is currently accomplishing its purpose to "ensure that all children have a fair, equal, and significant opportunity to obtain a high-quality education and reach, at a minimum, proficiency on challenging state academic achievement standards and state academic assessments" (U.S. Department of Education, 2001, Section 1001). Sanctions have increased, but inequalities in standards among states are allowing consistent differences in regional performance to persist (as shown in this study of NAEP and state assessment data and exemplified by the inequalities among census regions present in Research Question Two). Although Tennessee has adopted proficiency levels in accordance with NCLB legislation that are the same as NAEP's proficiency levels, there is clearly a difference in rigor between the two assessments' definition of what skills a "proficient" student can perform. Accordingly, "because of the variation in assessments and where proficiency is set, state to state comparisons are not meaningful" (Taylor & Gordon, in press, p. 3).

Since the inception of mandated accountability with the IASA in 1994, state responses have been uneven:

> Although all states developed assessments, standards, performance reporting, and, in most cases, consequences for performance, states have found different ways to define what it meant for schools to succeed, what indicators to include in their definition of success, and what the consequences would be for noncompliance (Goertz, 2005).

It was true in 1994, and the legacy of unevenness remains in the way states are carrying out NCLB guidelines. As long as states have freedom to design their own standards, assessments, and proficiency levels, there will be enough difference among results that regional inequalities persist and the spirit of NCLB will not be accomplished.

Competition for federal Race to the Top grants has pushed the issue of national assessments (Florida Department of Education [FDOE], 2010). To date, two consortia that will seek to develop national assessments among participating states have been funded. One such consortium, the Partnership for the Assessment of Readiness for College and Careers (PARCC) has been established, with Florida slated to serve as its fiscal agent (FDOE). Many education professionals have responded to the inequalities present among state assessments by lobbying for change through a system of national assessments with common standards, content, and proficiency scales.

A system composed of national common assessments, national common standards, or a common scale of designating proficiency would level the playing field.

NAEP could be used in this role, but not without changes to existing law. Additionally, recent methods developed by NCES (Bandeira de Mello et al., 2009) have made it possible for NAEP to serve as a metric against which state assessments can be measured and compared. In this area of comparing states, however, NCES must tread lightly due to prohibitions in NCLB legislation that limit the use of NAEP for comparing states. Data from NAEP are not to be used "to rank, compare, or otherwise evaluate individual students or teachers or to provide rewards or sanctions for individual students, teachers, schools or local educational agencies" (Public Law 107-110, Sec. 411 (B) (4) (A)). Nor are NAEP assessment results to be used "to establish, require, or influence the standards, assessments, curriculum, including lesson plans, textbooks, or classroom materials, or instructional practices of States or local educational agencies" (Sec. 411 (B) (4) (B)). Although this definition limits the role of NAEP to an extent, more than $112 million was earmarked to fund NAEP testing and its National Assessment Governing Board in 2003 (Public Law 107-279, Sec. 305 (A) (1) (A-B)).

The federal dollars appropriated to finance the NAEP assessment system underscore its importance, but NCLB guidelines have resulted in a system of state assessments that are not entirely useful for comparison to one another or to NAEP (Taylor & Gordon, in press). Consequently, this national test is not offering results in the most effective possible manner. It is Bourque's (2009) opinion that having a consistent definition and positioning of proficiency "would go a long way to resolving the disparity

between NAEP results for the states and the states' performance on their approved NCLB assessments" (p. 23).

In one sense, policymakers need to be more restrictive in compelling states to teach and assess common standards; ideally, NAEP would be redesigned to assess these common standards as well. Doing so would make state-to-state assessment comparisons, as well as state to NAEP assessment comparisons, more valid. In another sense, policymakers need to be less restrictive about the use of NAEP so that it can be used for the purpose of comparing states. NAEP could also be redesigned as a national common system of end-of-course assessments.

In order for these changes to take place, the public has to be aware of the nature of the existing disparities. Many people do not know what NAEP is, what it is designed to assess, how its results differ from their states' assessment results, or how the NCLB's AYP guidelines are preventing accurate state-to-state comparisons. Educational leaders have the ability to communicate these issues to the public. The cost of hiding inequitable assessment systems could be quite high when one considers the number of school shutdowns looming. Because 72% of Florida's school districts have failed to make AYP, the consequences could be major and widespread (McLester, 2006).

Recommendations for Further Research

The goal of this study was to investigate the extent to which state assessment proficiency percentages could be used to predict NAEP assessment proficiency

percentages. Data were collected to test seven research questions related to this goal. Many significant findings resulted from the exploration of these seven questions; however, there are several significant limitations to these findings that warrant further study.

1. One limitation is the fact that the results are situation, time, and group dependent (Linn, 1993; Mislevy, 1992). Results can vary over time and among subgroups. A repeated measures investigation of 2011 eighth grade NAEP and state reading assessment proficiency percentages could further substantiate the findings of this study, or reveal important contradictions.
2. This study was also limited by its design to study aggregate state data. An investigation of these same questions at the student level, with a model that allows the examination of students nested within schools or districts nested within states, is a possibility.
3. The research design of Research Question Three studying racial and ethnic subgroups could be studied in further detail for each subgroup and census region to see if subgroups' NAEP proficiency percentages can be predicted with similar patterns, or if different subgroups outperform one another in different census regions.
4. A detailed investigation into the standards tested on each eighth grade state reading assessment as compared to one another and to the NAEP reading assessment is also worthy of study to determine just how different (or similar)

these assessments are.

5. It is also of interest to examine which aspects of the Northeast census region's educational system most contribute to high NAEP proficiency percentages. A study of educational funding, professional development, and/or teacher experience as compared to NAEP proficiency percentages by state could reveal some possible answers behind high performance.
6. With more and more states leaning toward the creation of end-of-course assessments, additional studies comparing these state end-of-course exams to NAEP tests in the same subject area could be conducted.

Conclusions

This study expanded the work of researchers, most notably Gordon (2009), in the field of relating NAEP and state assessments. This investigation revealed that state assessment proficiency percentages could be used to predict NAEP proficiency percentages. This finding remained true across census regions, racial and ethnic groups, ELL students, low SES students, and students with disabilities. Throughout this study, significant gaps in achievement among census regions were consistently apparent. In addition, state proficiency percentages were consistently scaled higher than NAEP proficiency percentages.

NCLB (2002) guidelines seek to "ensure that all children have a fair, equal, and significant opportunity to obtain a high-quality education and reach, at a minimum, proficiency on challenging state academic achievement standards and state academic

assessments" However, the findings of this study revealed trends of inequality that warrant further study. Sanctions have increased for schools that do not measure up to NCLB's AYP guidelines. However, schools are not required to use a uniform system of measurement as they go about determining AYP. Consequently, AYP inequities result in NCLB sanctions being meted out inconsistently.

For parents, the variation between assessments is difficult at best. Members of the general public who listen to comparisons between different types of school assessments, between different states' assessments, and between NAEP and state assessments are rarely presented with the idea that the assessments under scrutiny should not really be compared side by side. When members of the press seek to present evidence of failing schools, NAEP scores are often used. However, it is rarely mentioned that states may not be teaching some of the NAEP standards upon which they are being measured.

The result is that many people too easily accept the premise that America's schools are failing. In truth, it is impossible for a citizen to compare the performance of students from state to state using state and/or NAEP assessments until one is ensured that each state is teaching the same standards. Instead, the public should demand common standards so they are able to determine what is really happening in schools and so that it is not so easy to be confounded by data that defies comparison.

In this study, the rules set out by each of the researchers who have sought to relate assessment results have been obeyed insofar as was possible. Whereas Dorans (2004) gives the goal of minimizing any error, it is not altogether possible to minimize the error

introduced due to variation assessments. Kolen and Brennan (2004) gave four logical questions to ask when considering whether assessments should be related that point to problems inherent to relating NAEP and state assessments. The discussion of the four state assessment instruments in Chapter Three, as well as the discussion of the NAEP instrument, reveals significant differences in difficulty, standards, and proficiency scale (to name a few).

Despite these significant variations, it is possible to spot trends in the data. For instance, in every subgroup but one (ELL students), the South census region posts proficiency percentages significantly behind peers in other regions. Conversely, the Northeast consistently achieves top proficiency percentages in all subgroups save one (ELL students). Trends such as these cannot and should not be used to impose sanctions on individual states, but they can be used to spot weak areas. For example, educators in the Northeast can look at the results of this study and ask themselves what they might do to increase the support systems they have in place for ELL students. On the other hand, educators of students with disabilities can seek direction from their peers in the Northeast to understand how they can use some common Northeastern SWD practices to get better achievement results with their own students.

Studies that relate assessment results are useful for demonstrating and documenting the persistent inequities between states' performance. As long as states can design assessments of unequal standards, rigor, and scale, inequities will go unnoticed and uncorrected. The U.S. founding fathers who established the groundwork for education

gave full responsibility to individual states and regarded any federal intervention as intrusive (Bourque, 2009). However, data from this study consistently revealed large and persistent inequalities among census regions. Giving states the freedom to design their own assessment systems is perpetuating a system of inequality that is clearly leaving some children behind.

APPENDIX A: CENSUS REGIONS DEFINED BY NAEP

Table 29. States within regions of the country defined by the U.S. Census Bureau

Northeast	South	Midwest	West
Connecticut□	Alabama□	Illinois*□	Alaska□
Maine□	Arkansas□	Indiana□	Arizona□
Massachusetts□	Delaware□	Iowa□	California*□
New Hampshire□	Washington, D.C.□	Kansas□	Colorado□
New Jersey□	Florida□	Michigan□	Hawaii□
New York*□	Georgia□	Minnesota□	Idaho□
Pennsylvania□	Kentucky□	Missouri□	Montana□
Rhode Island□	Louisiana□	Nebraska□	Nevada□
Vermont	Maryland□	North Dakota□	New Mexico□
	Mississippi□	Ohio□	Oregon□
	North Carolina□	South Dakota□	Utah□
	Oklahoma□	Wisconsin	Washington□
	South Carolina□		Wyoming
	Tennessee□		
	Texas*□		
	Virginia□		
	West Virginia		

SOURCE: U.S. Department of Commerce Economics and Statistics Administration.
*Indicates representative states in this study

APPENDIX B: 2007 NAEP AND STATE ASSESSMENT PROFICIENCY

Jurisdiction	State All 2007	NAEP All 2007
Alabama	71.8	21
Alaska	79.2	27
Arizona	63.2	24
Arkansas	62.5	25
California	42.2	21
Colorado	86.6	35
Connecticut	74.6	37
Delaware	78.1	31
Washington, D.C.	16.9	12
Florida	49.0	28
Georgia	88.9	26
Hawaii	60.2	20
Idaho	85.8	32
Illinois	80.9	30
Indiana	68.2	31
Iowa	72.5	36
Kansas	80.7	35
Kentucky	64.3	28
Louisiana	58.8	19
Maine	64.8	37
Maryland	68.7	33
Massachusetts	75.2	43
Michigan	71.8	28
Minnesota	63.6	37
Mississippi	51.6	17
Missouri	42.5	31
Montana	78.8	39
Nebraska	90.7	35
Nevada	56.9	22
New Hampshire	65.8	37
New Jersey	72.4	39
New Mexico	56.2	17
New York	57.3	32
North Carolina	87.9	28
North Dakota	75.7	32
Ohio	80.2	36
Oklahoma	77.7	26
Oregon	68.1	34
Pennsylvania	74.4	36
Rhode Island	58.1	27
South Carolina	34.5	25
South Dakota	78.0	37
Tennessee	92.1	26
Texas	87.5	28
Utah	80.8	30
Vermont	65.3	42
Virginia	79.5	34
Washington	66.6	34
West Virginia	80.2	23
Wisconsin	84.1	33
Wyoming	71.3	33

APPENDIX C: 2009 NAEP AND STATE ASSESSMENT PROFICIENCY

Jurisdiction	State All 2009	NAEP All 2009
Alabama	74.7	24
Alaska	81.7	27
Arizona	69.3	27
Arkansas	71.4	27
California	47.6	22
Colorado	88.5	32
Connecticut	76.6	43
Delaware	77.3	31
Washington, D.C.	46.4	14
Florida	54.2	32
Georgia	93.8	27
Hawaii	68.2	22
Idaho	91.5	33
Illinois	83.4	33
Indiana	68.5	32
Iowa	73.2	32
Kansas	85.4	33
Kentucky	68	33
Louisiana	61.3	20
Maine	71	35
Maryland	80.2	36
Massachusetts	78.7	43
Michigan	77	31
Minnesota	67.2	38
Mississippi	48.3	19
Missouri	50.2	34
Montana	80.8	38
Nebraska	95.2	35
Nevada	61.1	22
New Hampshire	69.8	39
New Jersey	81.6	42
New Mexico	62.5	22
New York	68.5	33
North Carolina	66.8	29
North Dakota	76.2	34
Ohio	72.4	37
Oklahoma	65.8	26
Oregon	69.5	33
Pennsylvania	79.7	40
Rhode Island	61.8	28
South Carolina	77.5	24
South Dakota	73.9	37
Tennessee	92.6	28
Texas	94.3	27
Utah	62.5	33
Vermont	68.9	41
Virginia	87.4	32
Washington	67.9	36
West Virginia	60.9	22
Wisconsin	85.2	34
Wyoming	64.9	34

APPENDIX D: 2007 NAEP AND STATE ASSESSMENTS DIFFERENCES

Jurisdiction	State All 2007	NAEP All 2007	2007 Difference
Alabama	71.8	21	50.8
Alaska	79.2	27	52.2
Arizona	63.2	24	39.2
Arkansas	62.5	25	37.5
California	42.2	21	21.2
Colorado	86.6	35	51.6
Connecticut	74.6	37	37.6
Delaware	78.1	31	47.1
Washington, D.C.	16.9	12	4.9
Florida	49	28	21
Georgia	88.9	26	62.9
Hawaii	60.2	20	40.2
Idaho	85.8	32	53.8
Illinois	80.9	30	50.9
Indiana	68.2	31	37.2
Iowa	72.5	36	36.5
Kansas	80.7	35	45.7
Kentucky	64.3	28	36.3
Louisiana	58.8	19	39.8
Maine	64.8	37	27.8
Maryland	68.7	33	35.7
Massachusetts	75.2	43	32.2
Michigan	71.8	28	43.8
Minnesota	63.6	37	26.6
Mississippi	51.6	17	34.6
Missouri	42.5	31	11.5
Montana	78.8	39	39.8
Nebraska	90.7	35	55.7
Nevada	56.9	22	34.9
New Hampshire	65.8	37	28.8
New Jersey	72.4	39	33.4
New Mexico	56.2	17	39.2
New York	57.3	32	25.3
North Carolina	87.9	28	59.9
North Dakota	75.7	32	43.7
Ohio	80.2	36	44.2
Oklahoma	77.7	26	51.7
Oregon	68.1	34	34.1
Pennsylvania	74.4	36	38.4
Rhode Island	58.1	27	31.1
South Carolina	34.5	25	9.5
South Dakota	78	37	41
Tennessee	92.1	26	66.1
Texas	87.5	28	59.5
Utah	80.8	30	50.8
Vermont	65.3	42	23.3
Virginia	79.5	34	45.5
Washington	66.6	34	32.6
West Virginia	80.2	23	57.2
Wisconsin	84.1	33	51.1
Wyoming	71.3	33	38.3

APPENDIX E: 2009 NAEP AND STATE ASSESSMENTS DIFFERENCES

Jurisdiction	State All 2009	NAEP All 2009	2009 Difference
Alabama	74.7	24	50.7
Alaska	81.7	27	54.7
Arizona	69.3	27	42.3
Arkansas	71.4	27	44.4
California	47.6	22	25.6
Colorado	88.5	32	56.5
Connecticut	76.6	43	33.6
Delaware	77.3	31	46.3
Washington, D.C.	46.4	14	32.4
Florida	54.2	32	22.2
Georgia	93.8	27	66.8
Hawaii	68.2	22	46.2
Idaho	91.5	33	58.5
Illinois	83.4	33	50.4
Indiana	68.5	32	36.5
Iowa	73.2	32	41.2
Kansas	85.4	33	52.4
Kentucky	68	33	35
Louisiana	61.3	20	41.3
Maine	71	35	36
Maryland	80.2	36	44.2
Massachusetts	78.7	43	35.7
Michigan	77	31	46
Minnesota	67.2	38	29.2
Mississippi	48.3	19	29.3
Missouri	50.2	34	16.2
Montana	80.8	38	42.8
Nebraska	95.2	35	60.2
Nevada	61.1	22	39.1
New Hampshire	69.8	39	30.8
New Jersey	81.6	42	39.6
New Mexico	62.5	22	40.5
New York	68.5	33	35.5
North Carolina	66.8	29	37.8
North Dakota	76.2	34	42.2
Ohio	72.4	37	35.4
Oklahoma	65.8	26	39.8
Oregon	69.5	33	36.5
Pennsylvania	79.7	40	39.7
Rhode Island	61.8	28	33.8
South Carolina	77.5	24	53.5
South Dakota	73.9	37	36.9
Tennessee	92.6	28	64.6
Texas	94.3	27	67.3
Utah	62.5	33	29.5
Vermont	68.9	41	27.9
Virginia	87.4	32	55.4
Washington	67.9	36	31.9
West Virginia	60.9	22	38.9
Wisconsin	85.2	34	51.2
Wyoming	64.9	34	30.9

APPENDIX F: GRAPHS RELATED TO TESTING ASSUMPTIONS

Figures: Research Question One

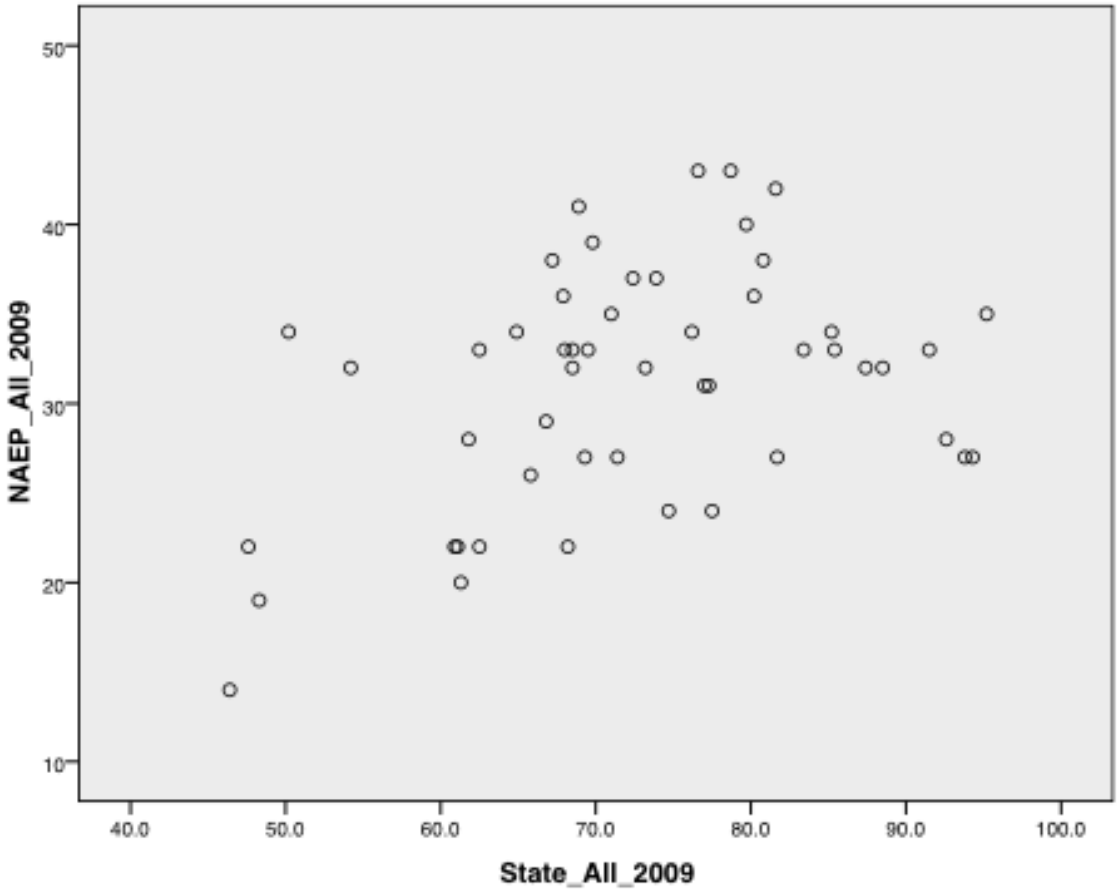

Figure 1. Scatterplot of 2009 NAEP and State Percent Proficient

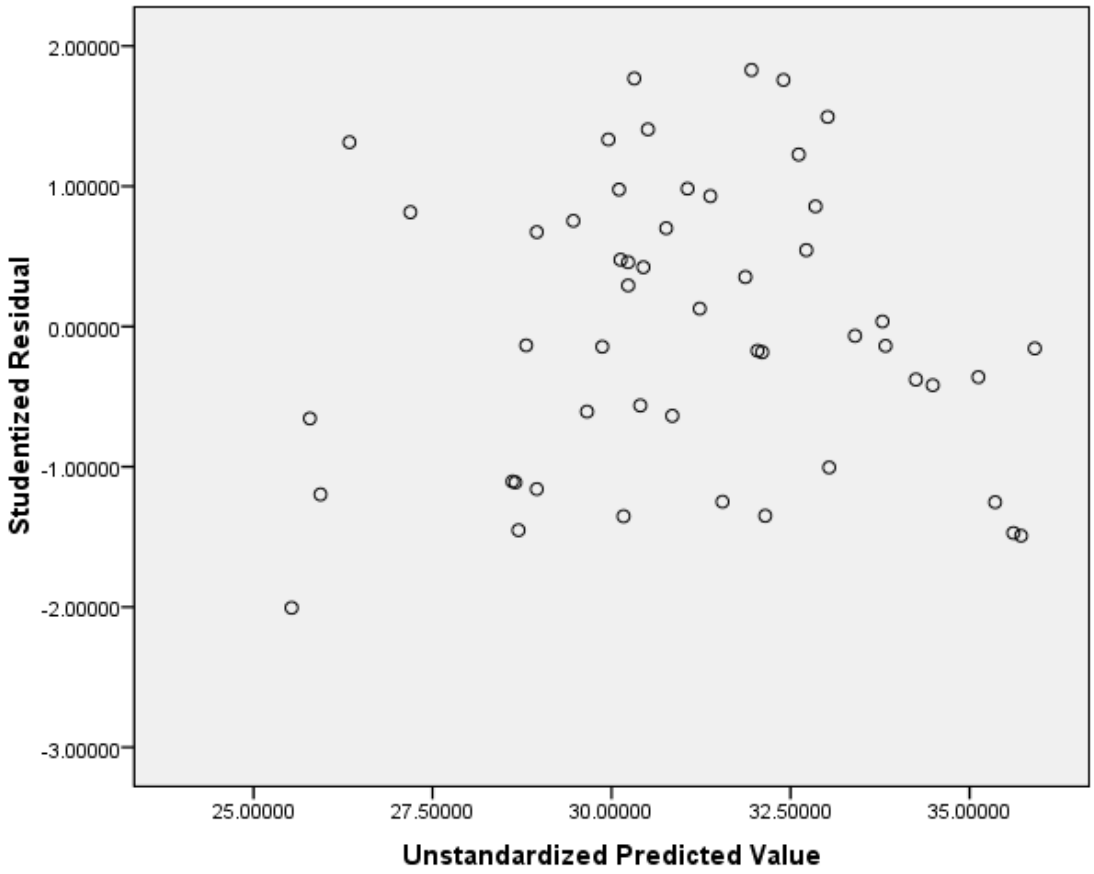

Figure 2. Scatterplot of Studentized Residuals to Unstandardized Predicted Values

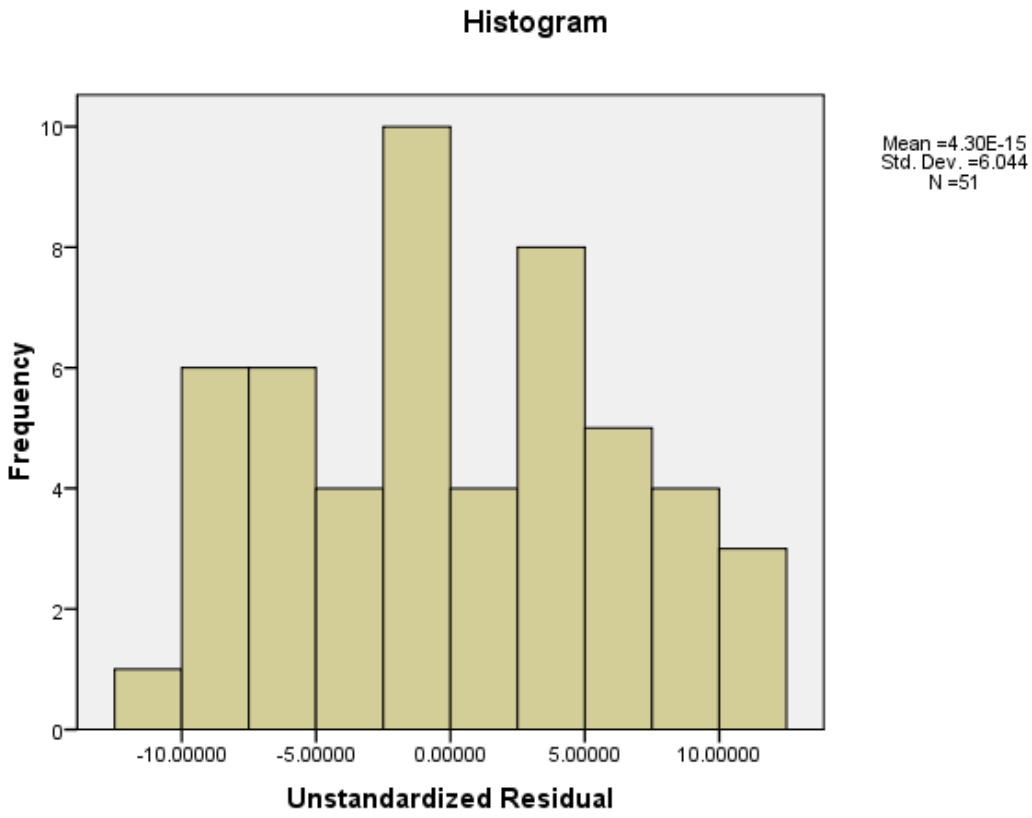

Figure 3. Histogram of Unstandardized Residuals

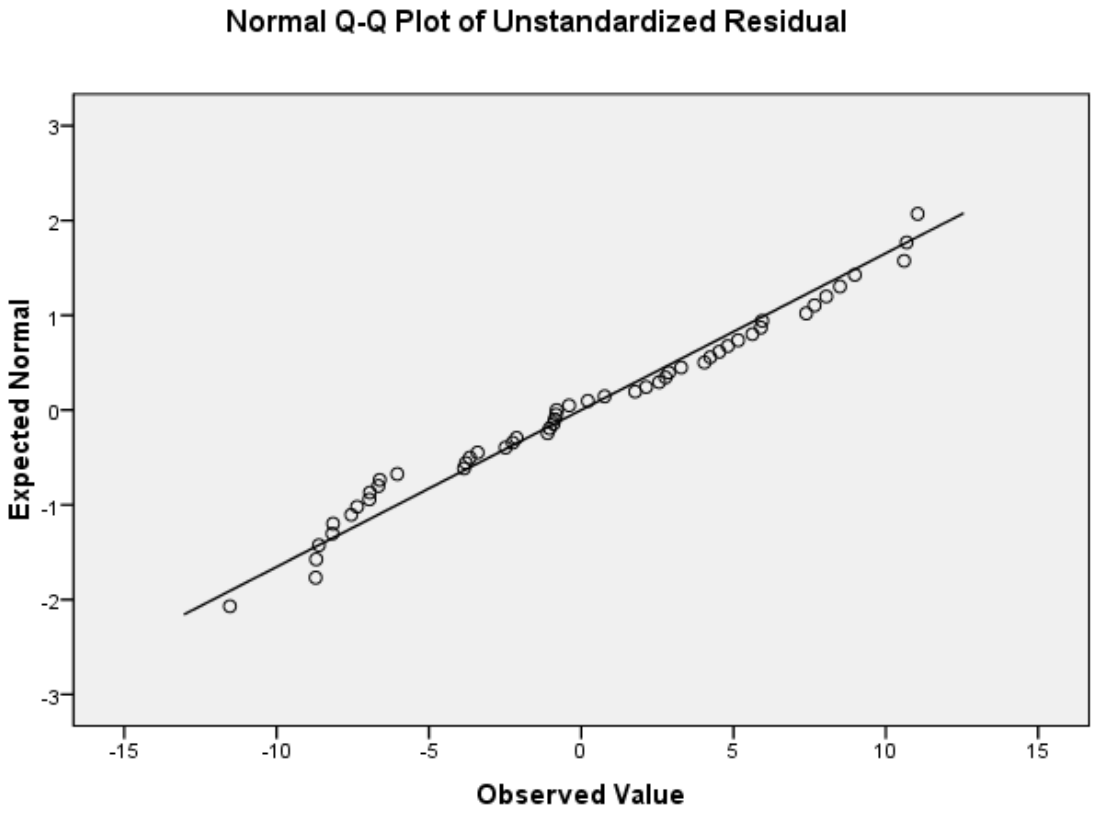

Figure 4. Q-Q Plot of Unstandardized Residuals

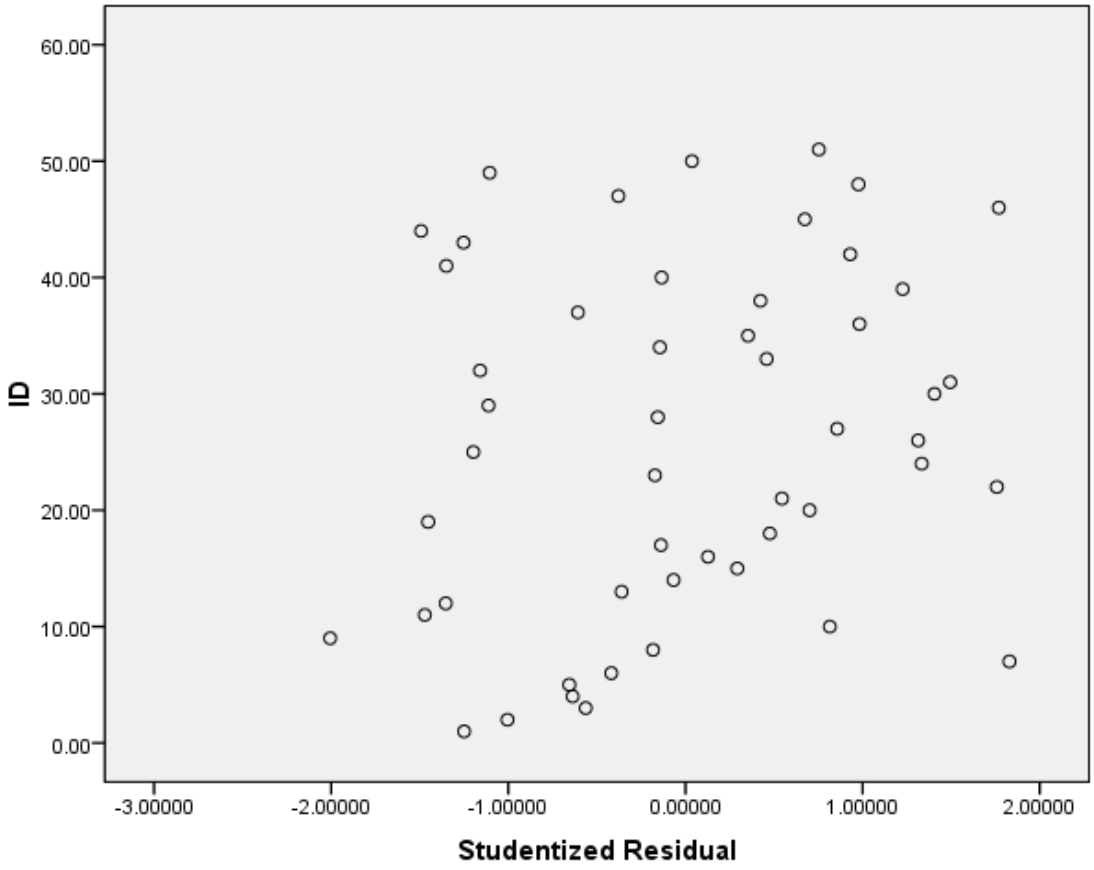

Figure 5. Scatterplot of Studentized Residuals to Case Number

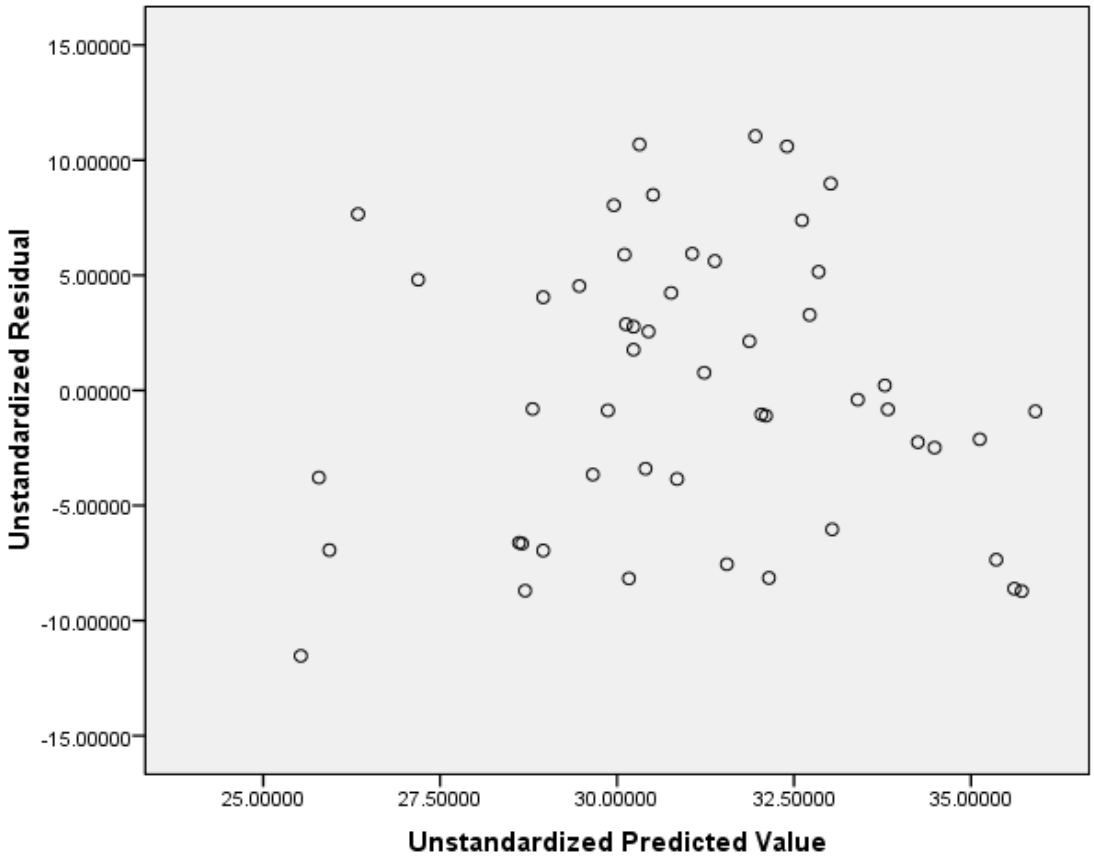

Figure 6. Scatterplot of Unstandardized Residuals to Unstandardized Predicted Values

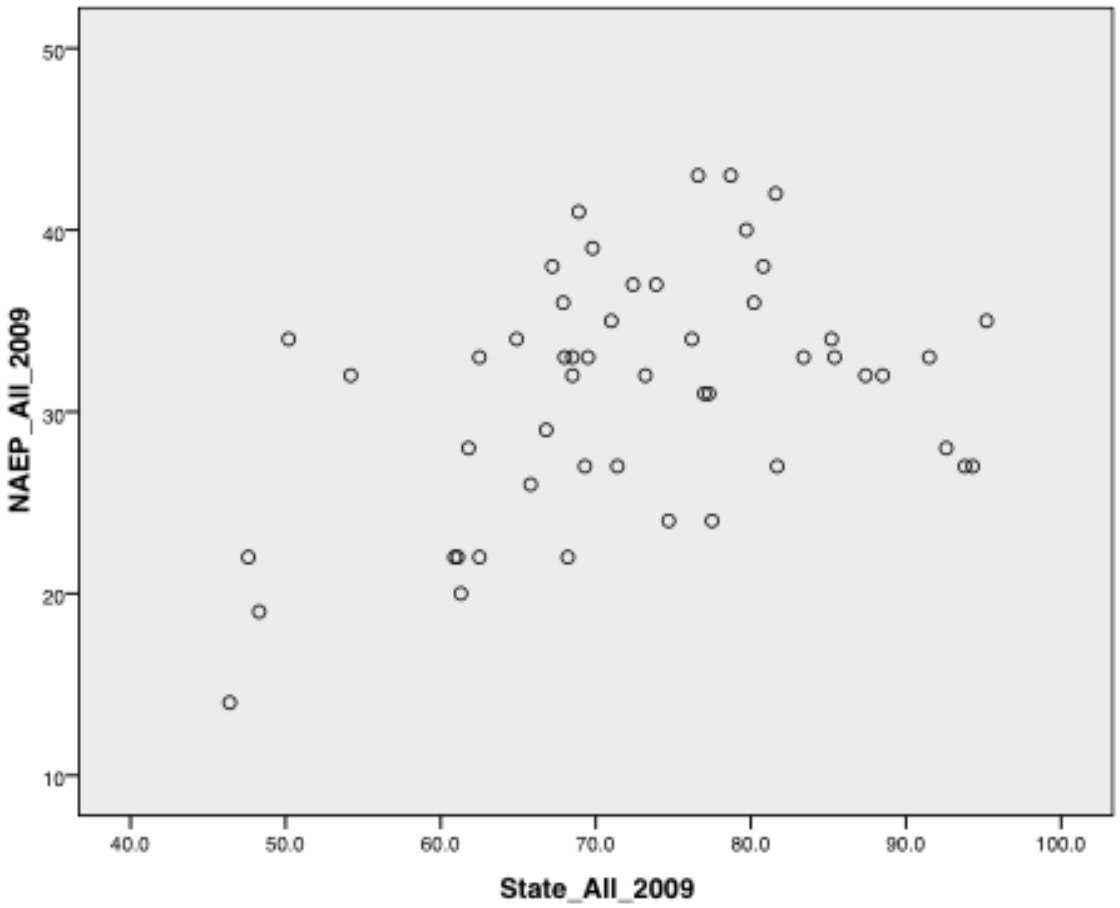

Figure 7. Scatterplot of Studentized Residuals to Unstandardized Predicted Values

Figures: Research Question Two

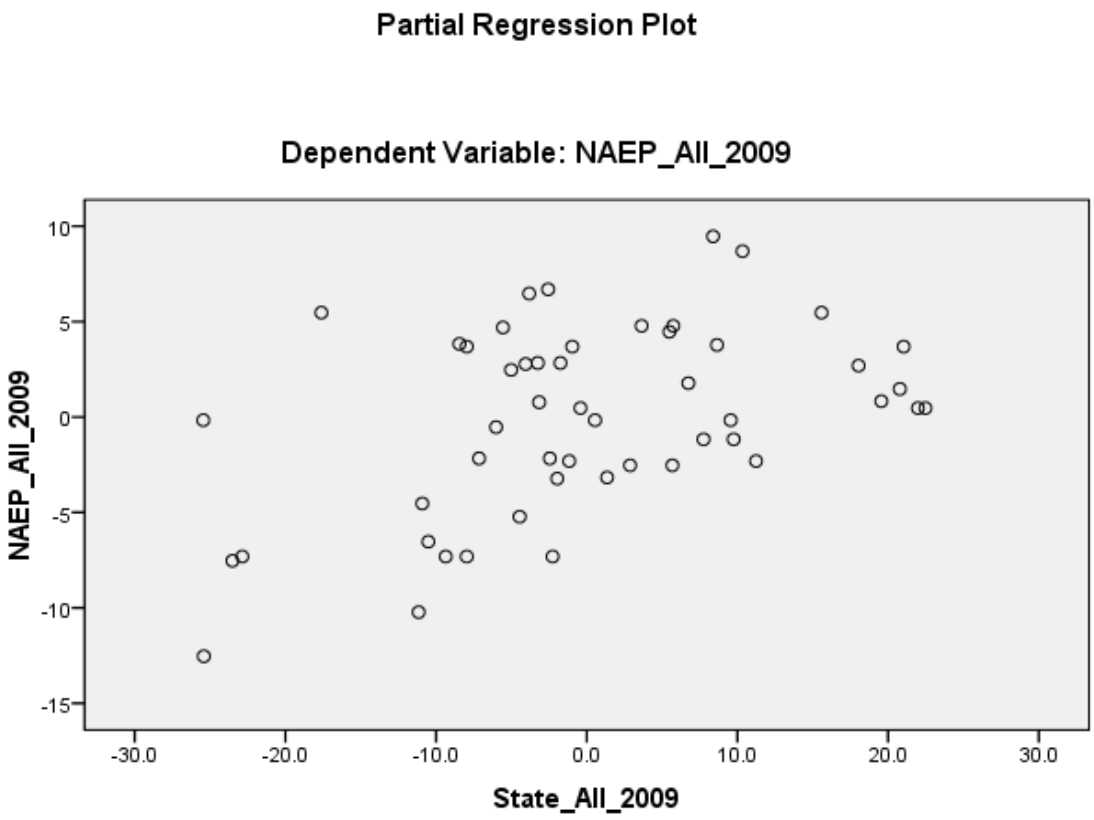

Figure 8. Partial Regression Plot of 2009 NAEP and State Percent Proficient

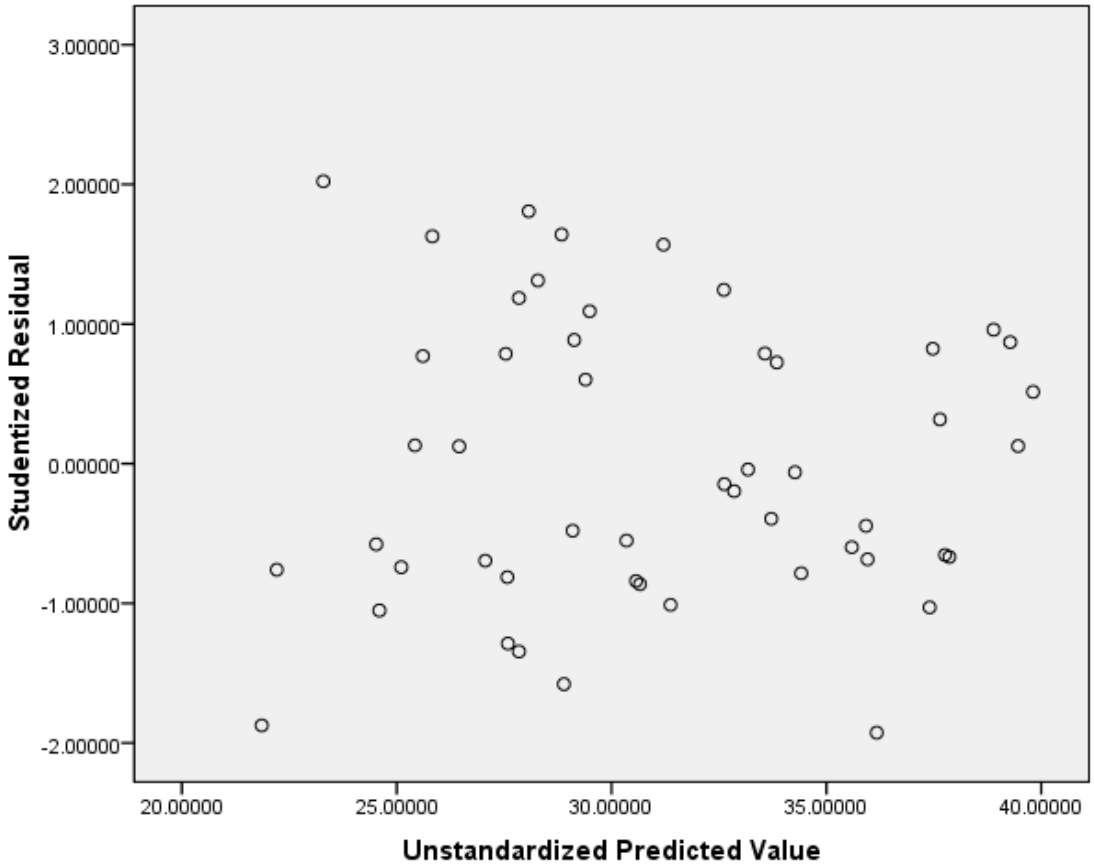

Figure 9. Scatterplot of Studentized Residuals to Unstandardized Predicted Values

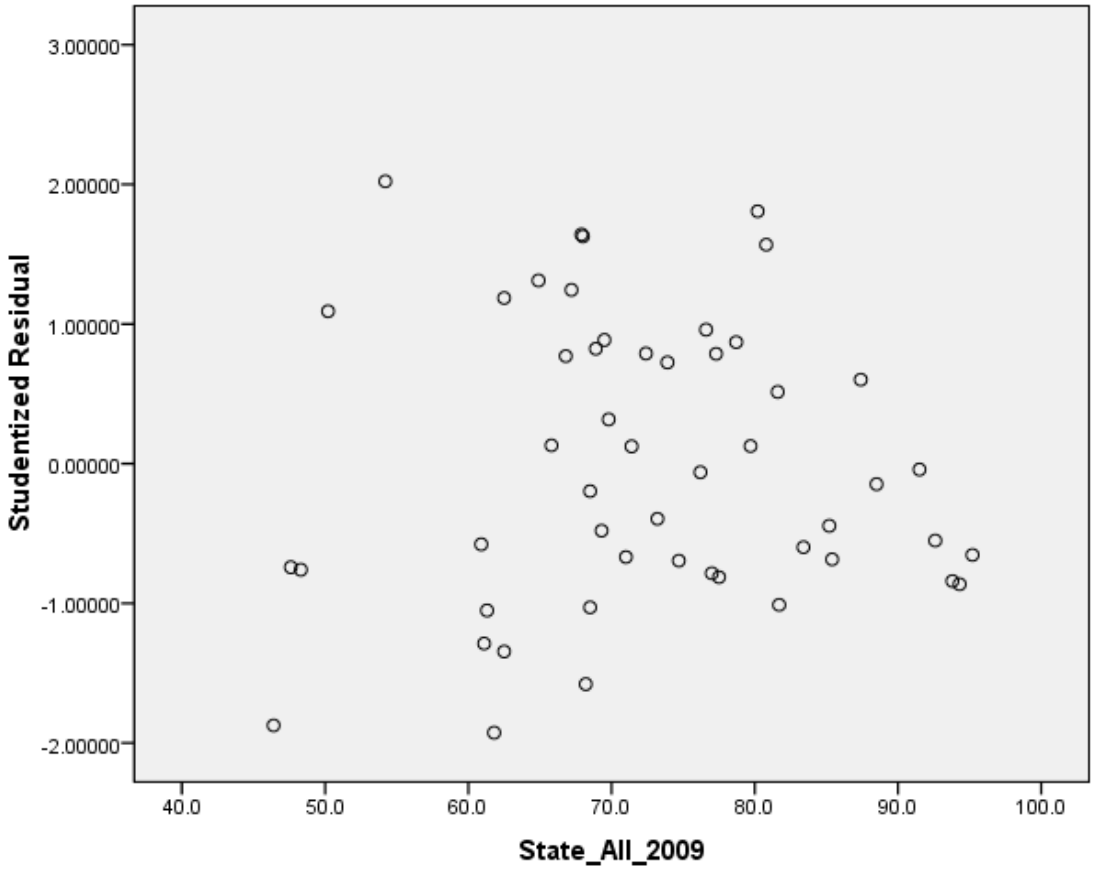

Figure 10. Scatterplot of Studentized Residuals to 2009 State Percent Proficient

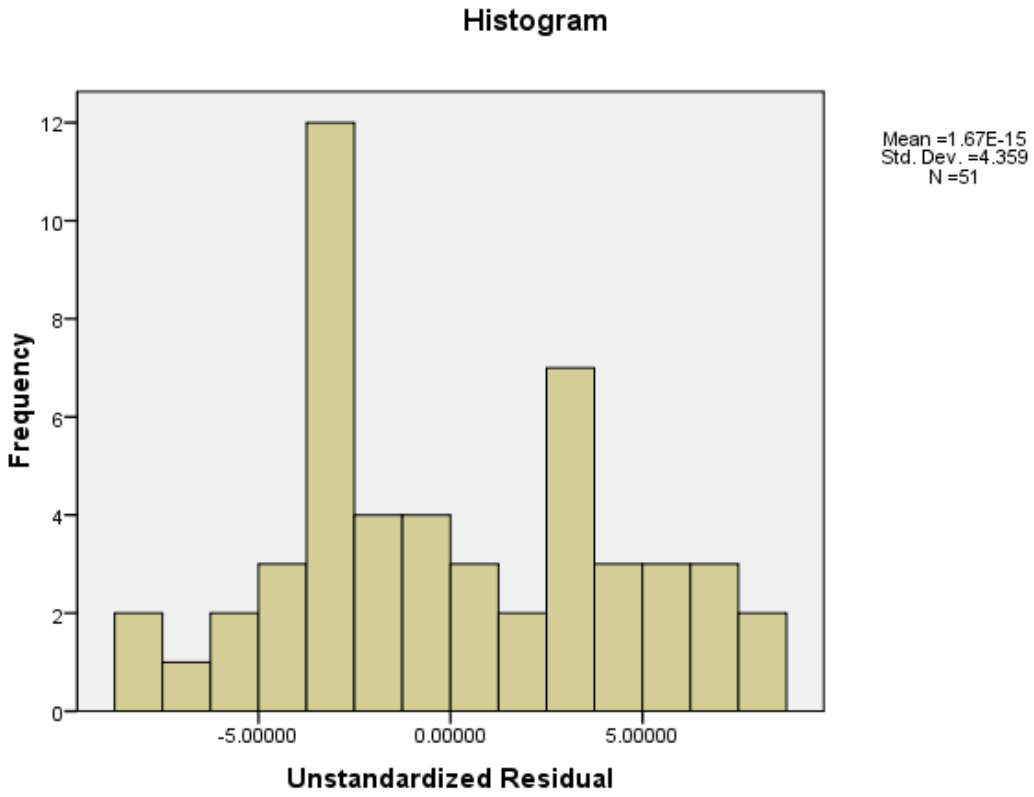

Figure 11. Histogram of Unstandardized Residuals

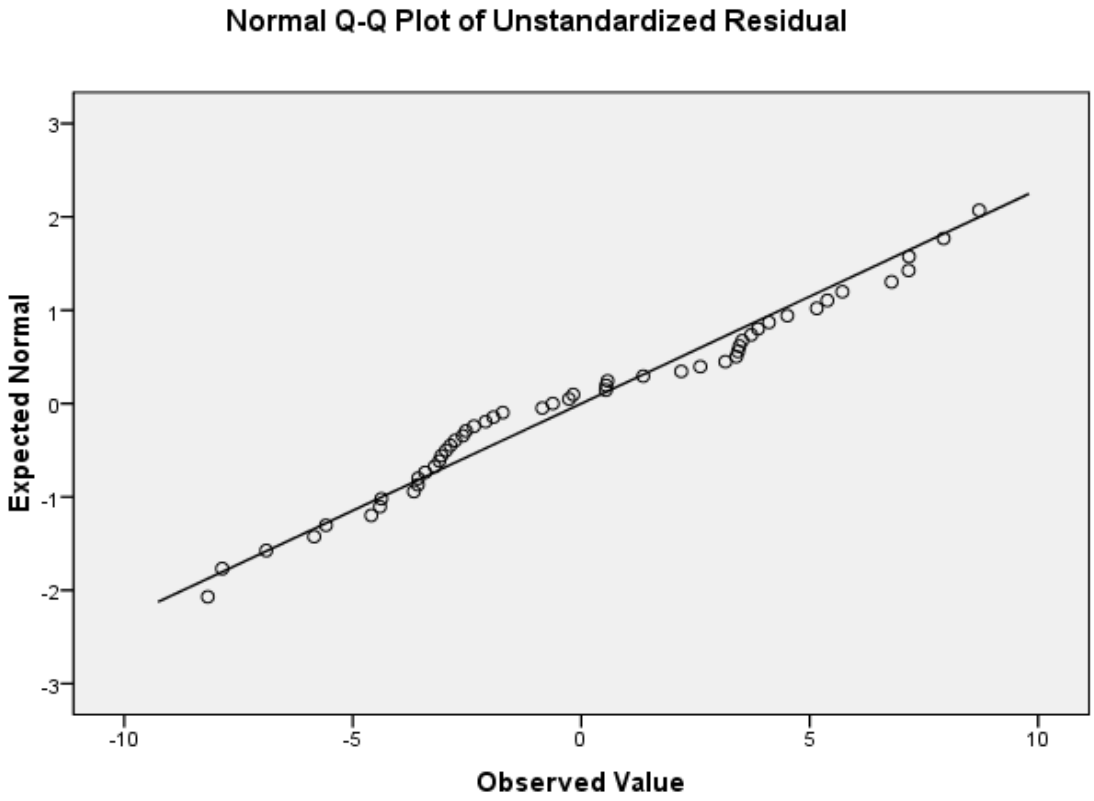

Figure 12. Q-Q Plot of Unstandardized Residuals

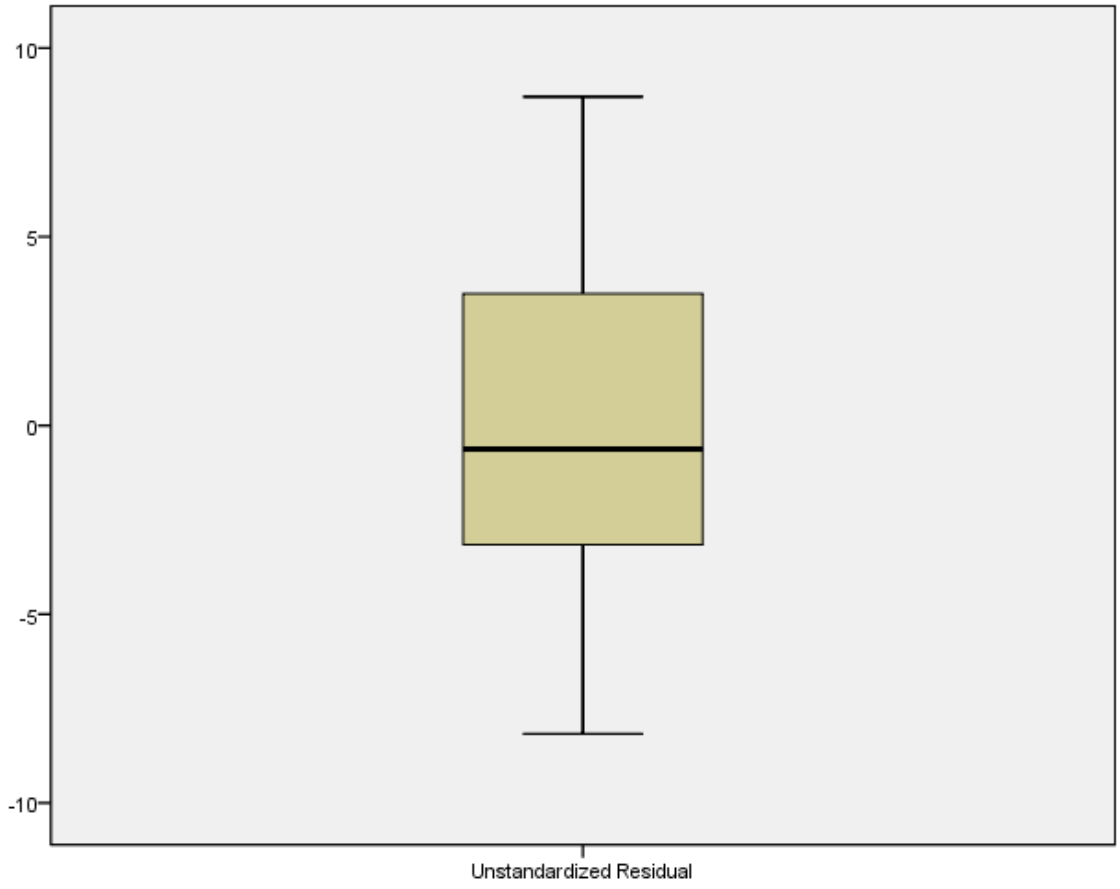

Figure 13. Boxplot of Unstandardized Residuals

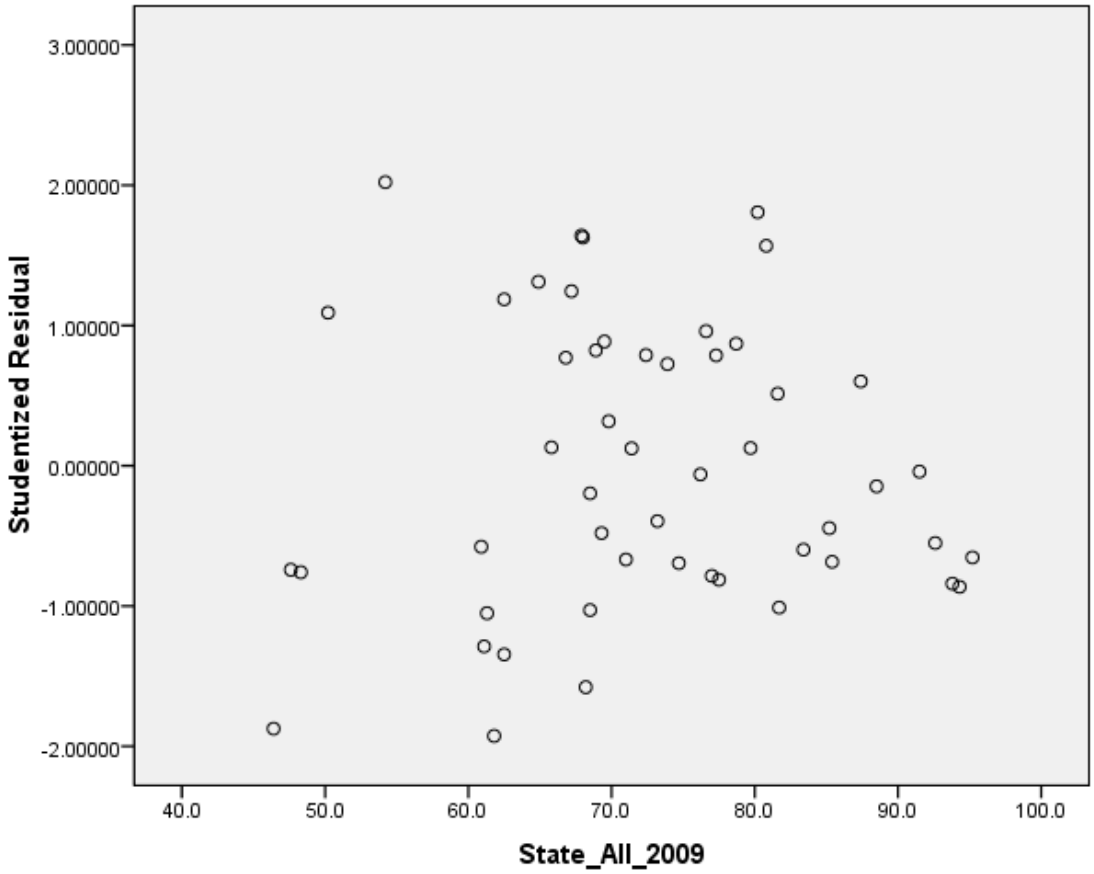

Figure 14. Scatterplot of Studentized Residuals to 2009 State Percent Proficient

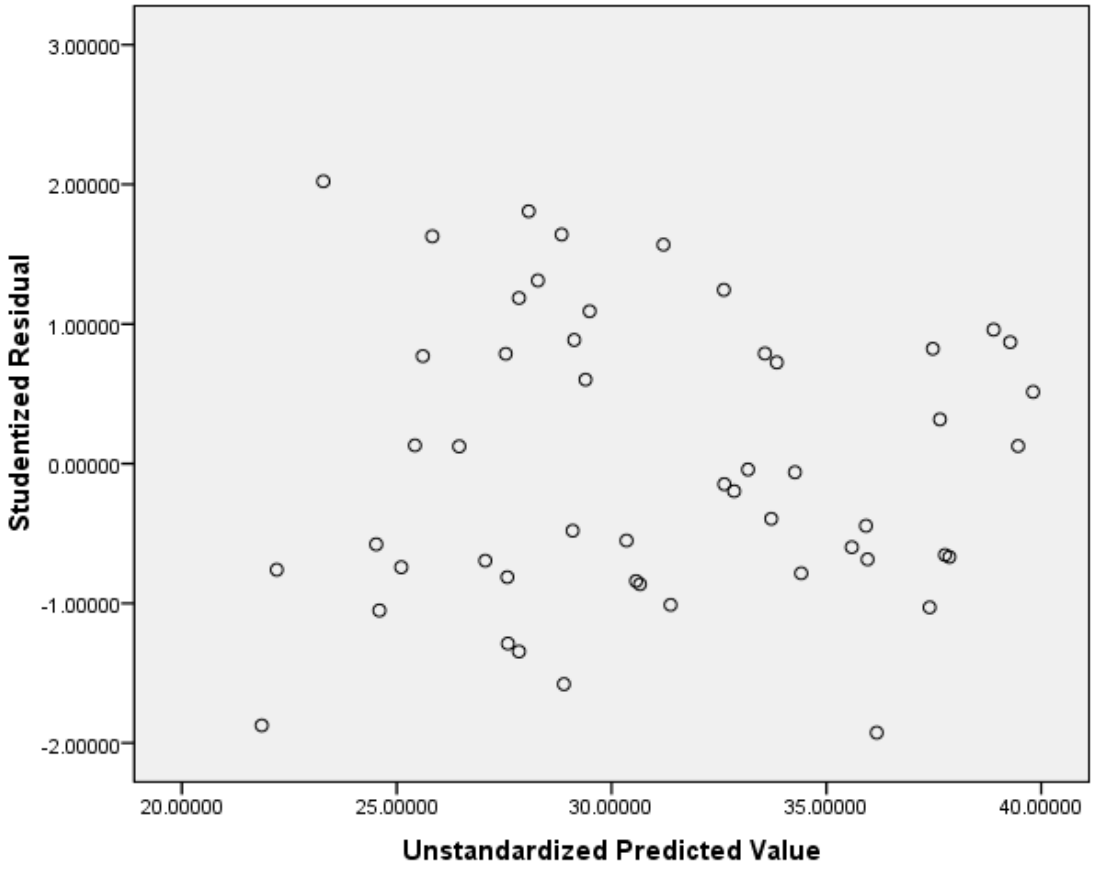

Figure 15. Scatterplot of Studentized Residuals to Unstandardized Predicted Values

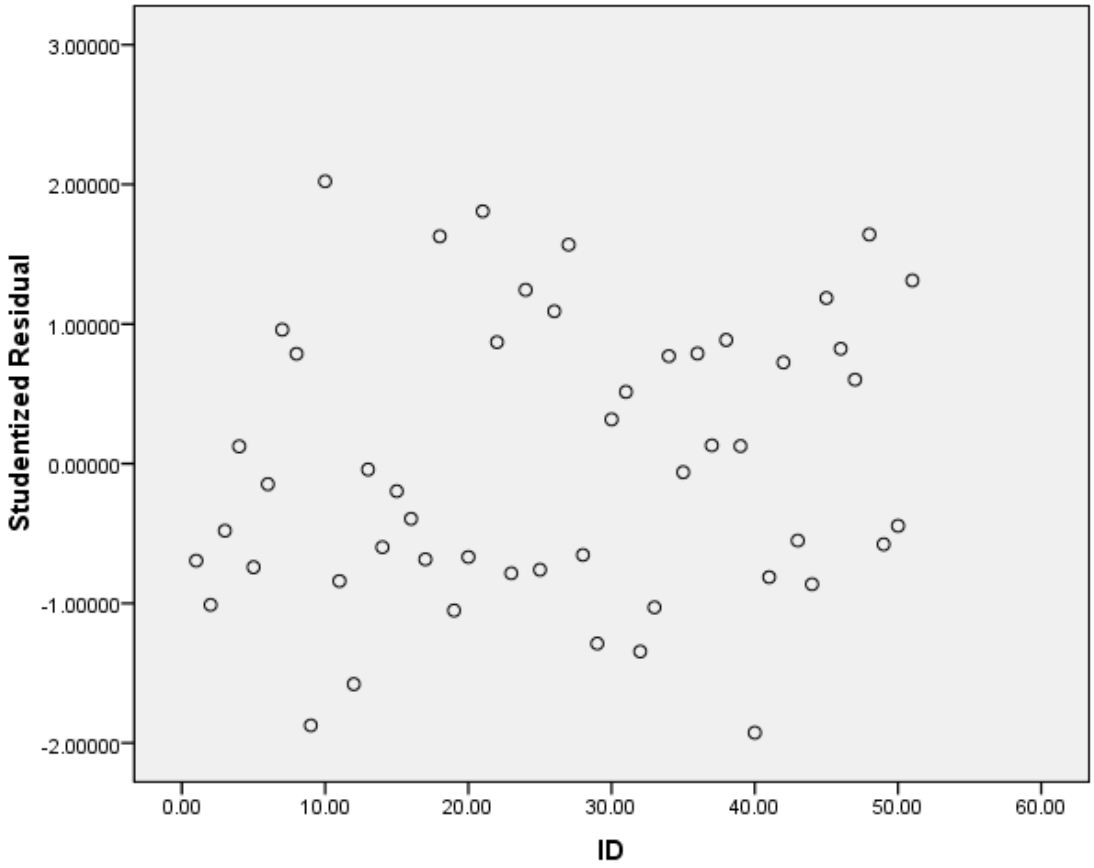

Figure 16. Scatterplot of Studentized Residuals to Case Number

Figures: Research Question Three

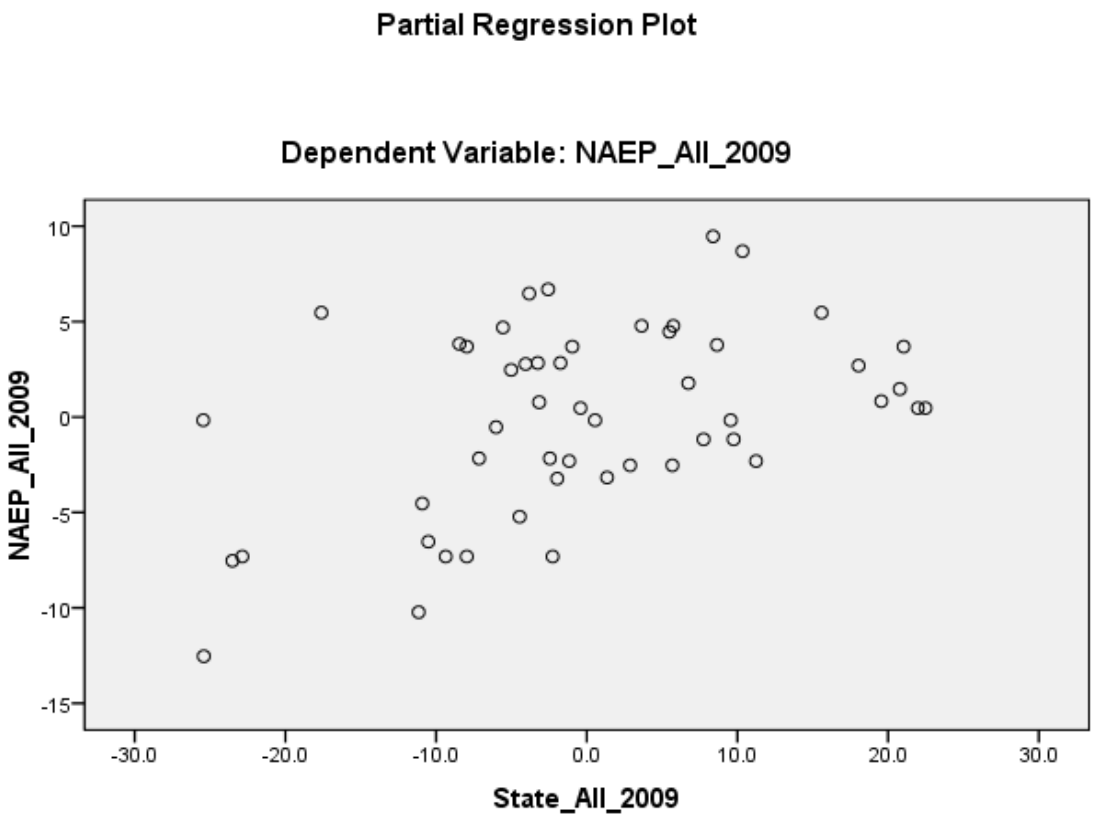

Figure 17. Partial Regression Plot of 2009 NAEP and State Percent Proficient

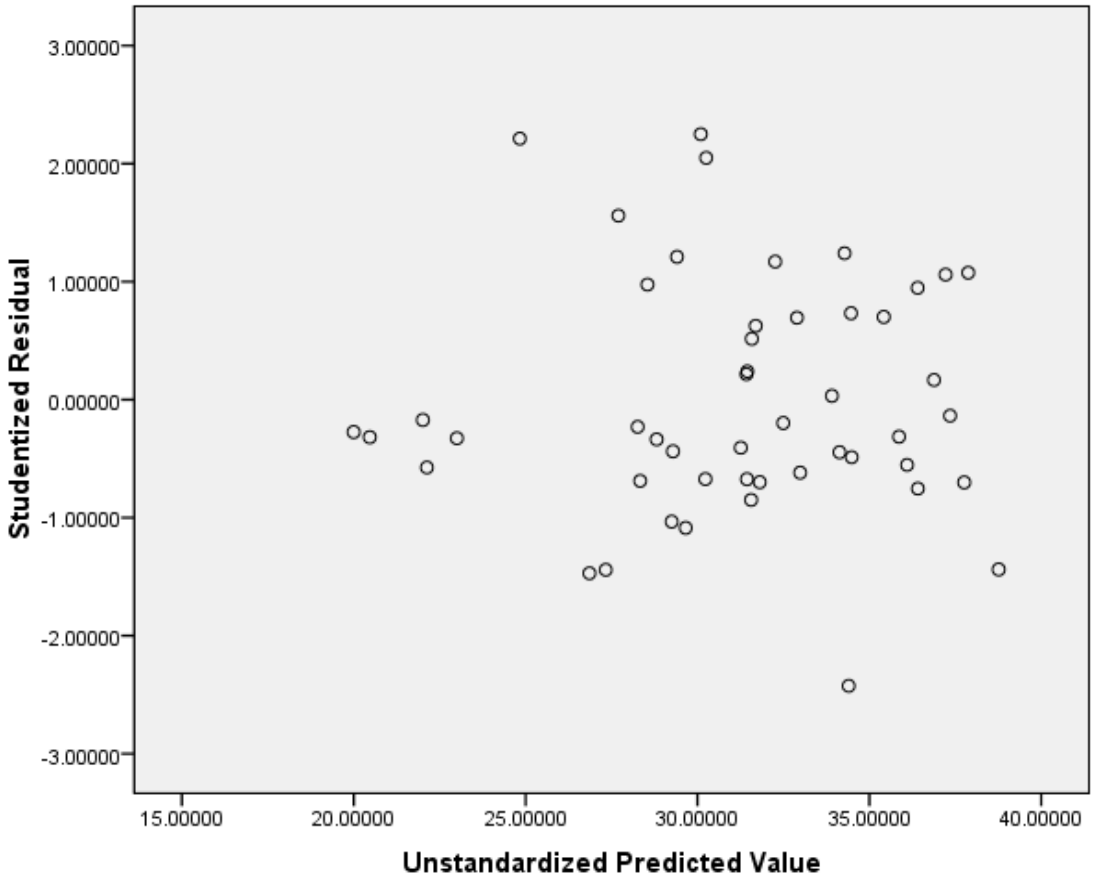

Figure 18. Scatterplot of Studentized Residuals to Unstandardized Predicted Values

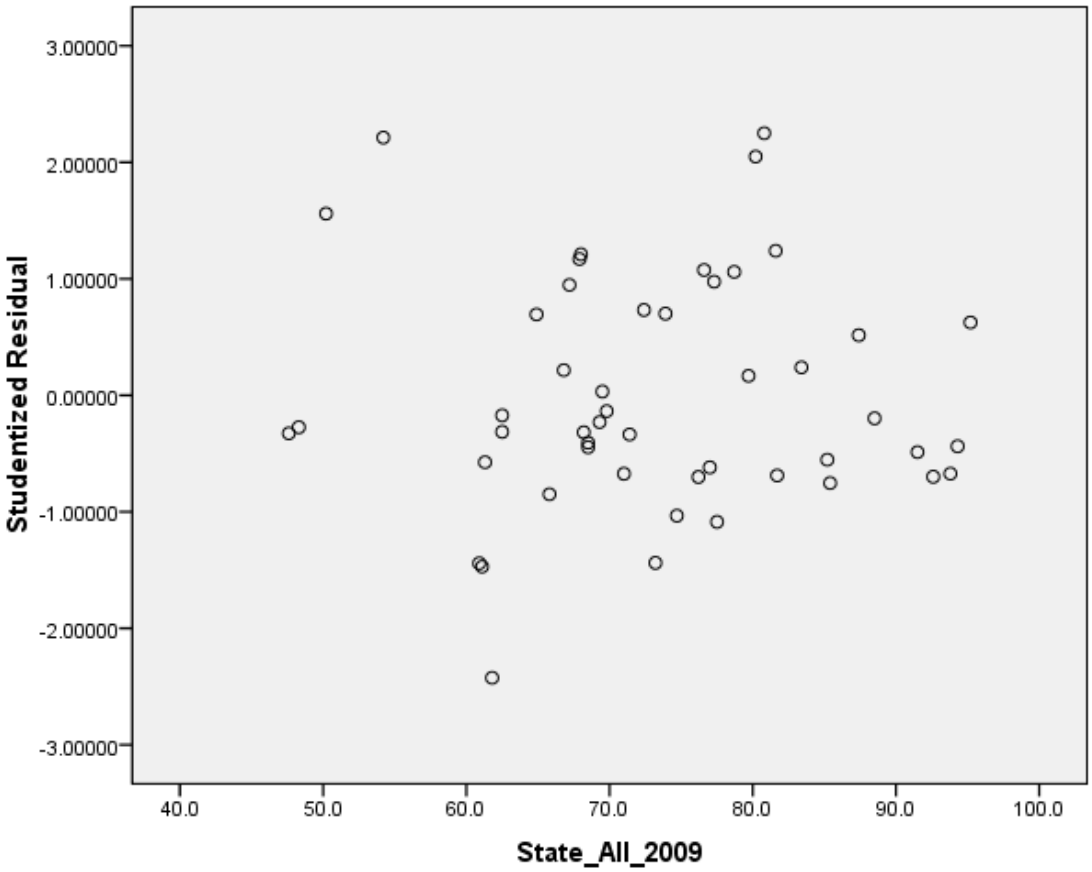

Figure 19. Scatterplot of Studentized Residuals to 2009 State Percent Proficient

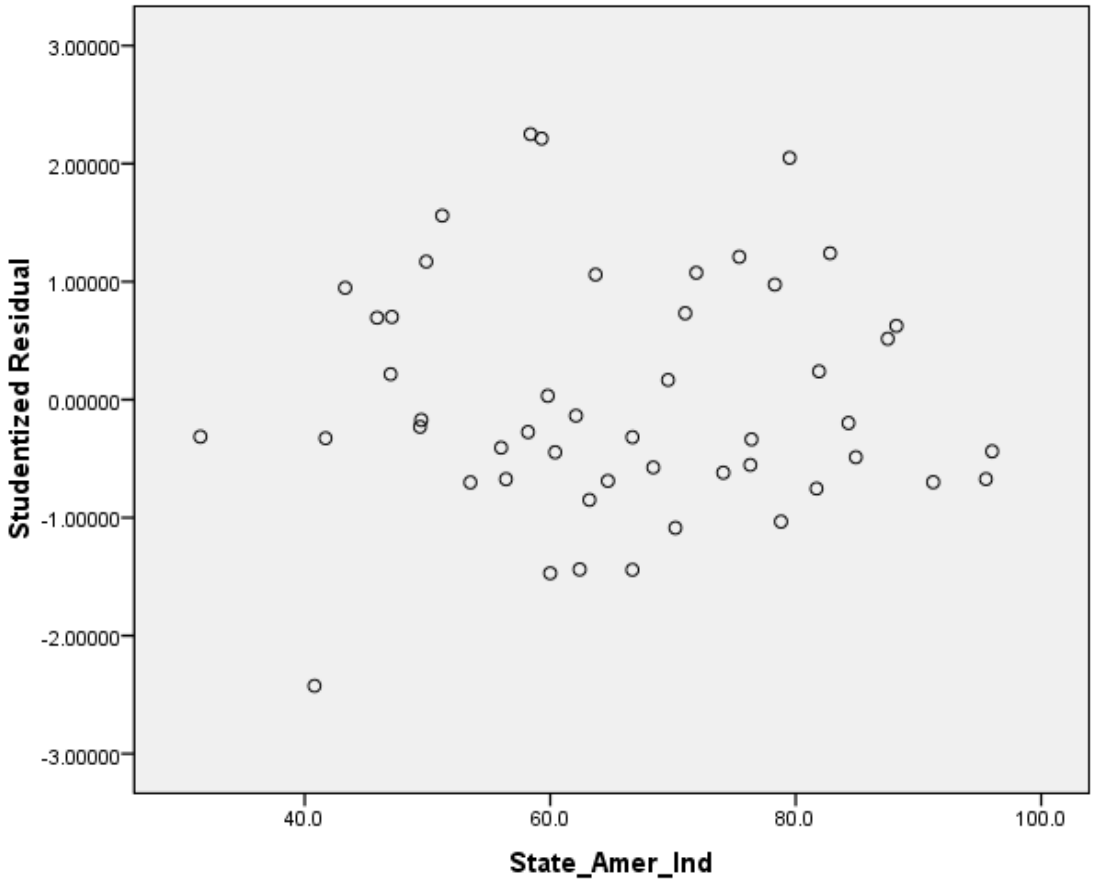

Figure 20. Scatterplot of Studentized Residuals to 2009 State American Indian Percent Proficient

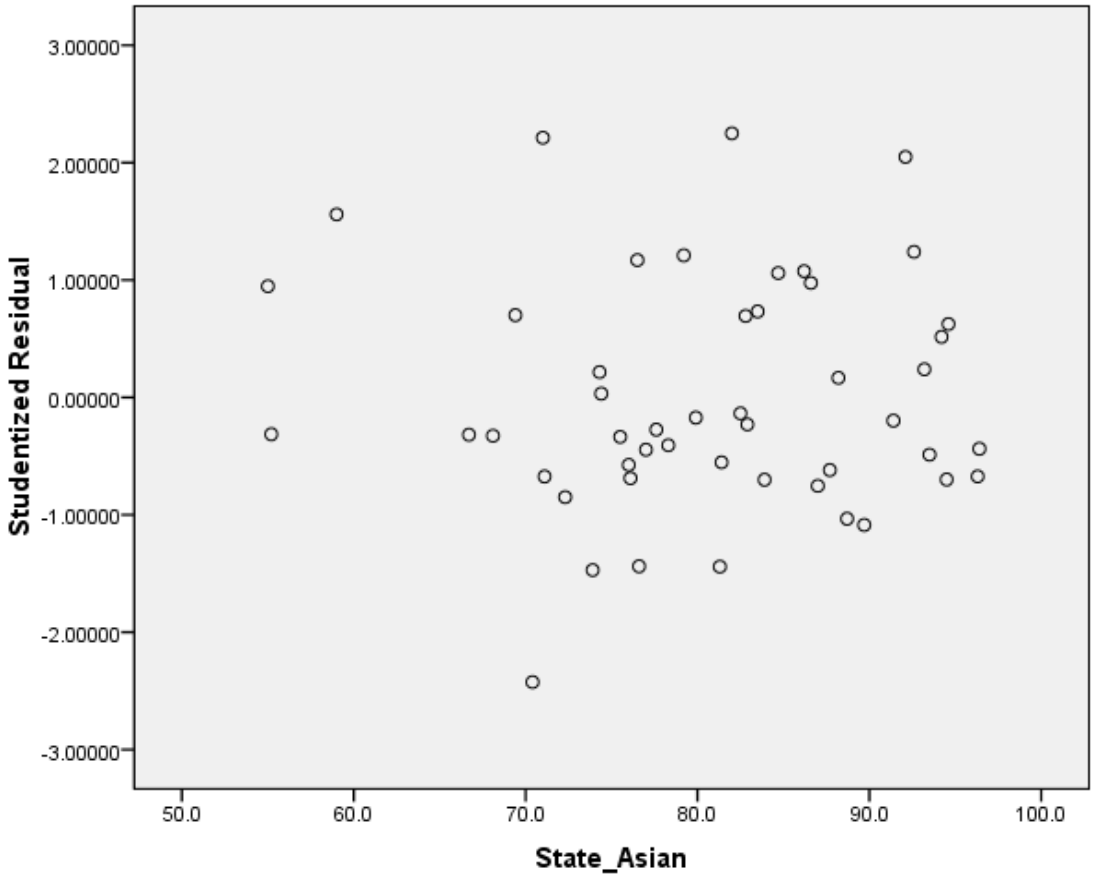

Figure 21. Scatterplot of Studentized Residuals to 2009 State Asian Percent Proficient

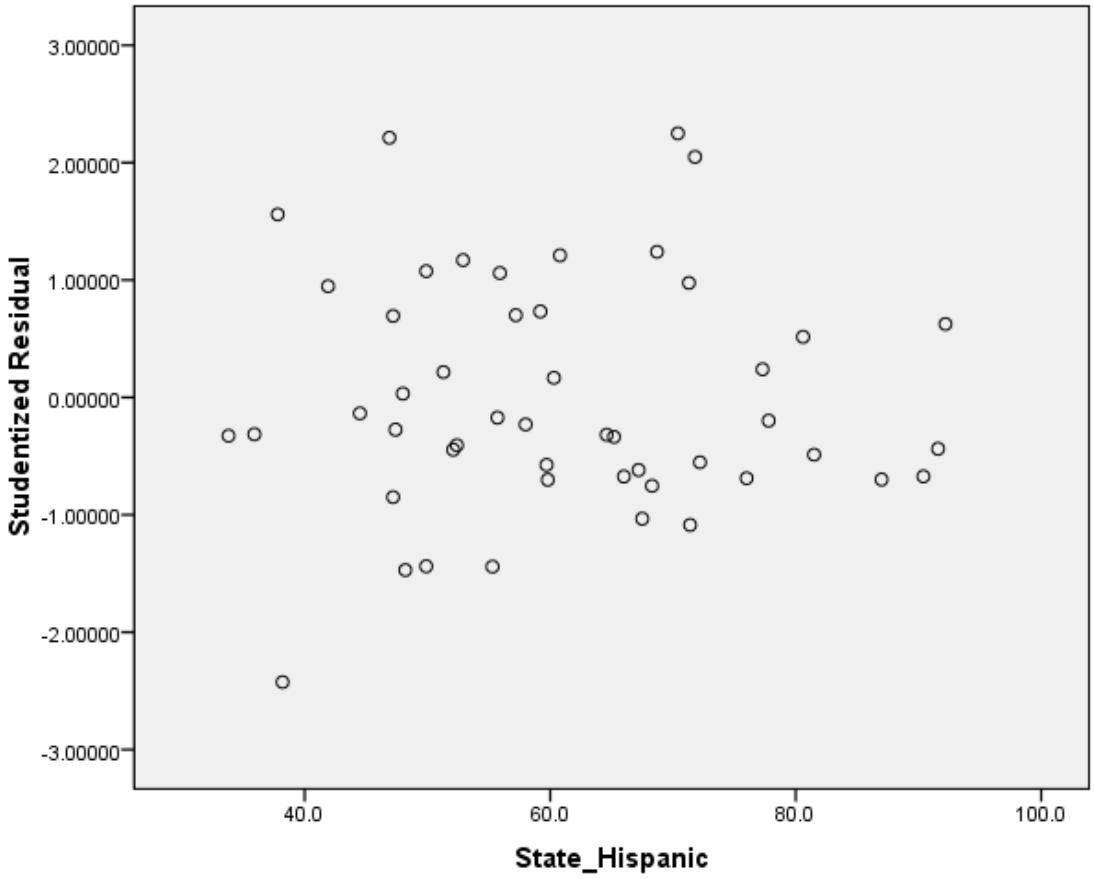

Figure 22. Scatterplot of Studentized Residuals to 2009 State Hispanic Percent Proficient

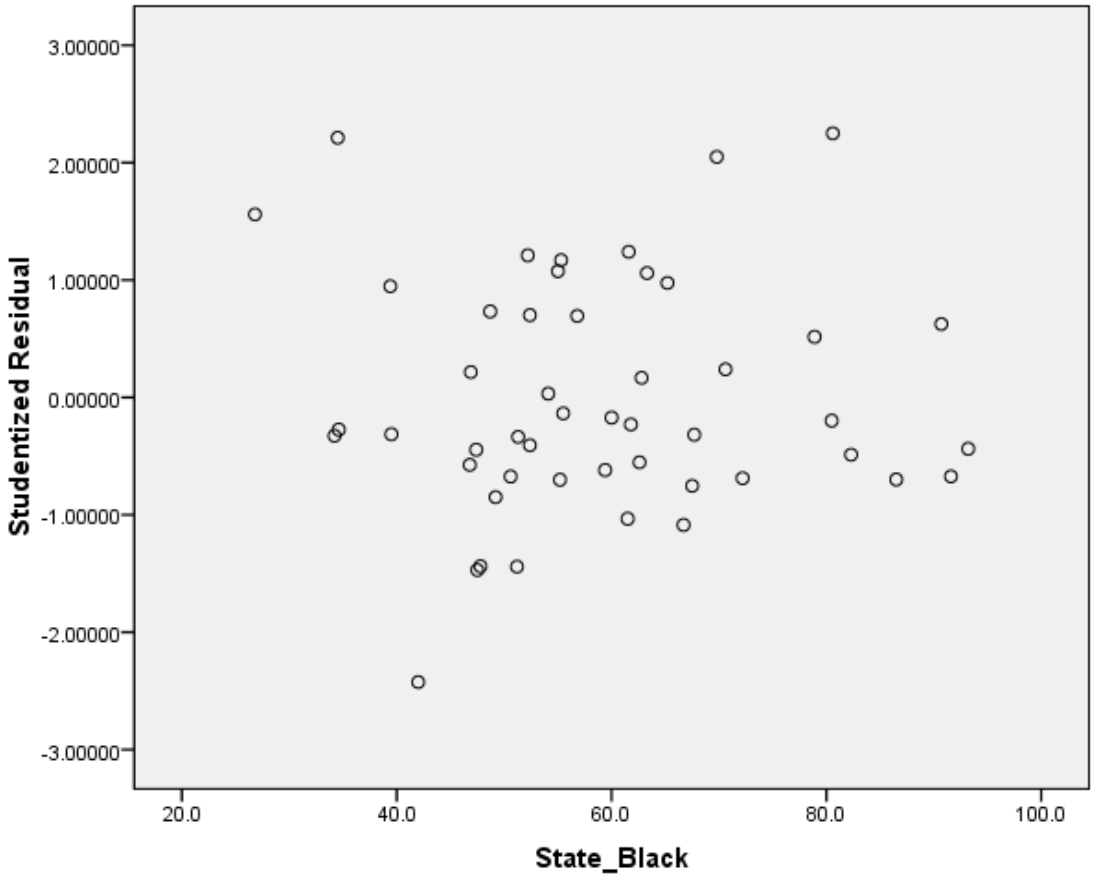

Figure 23. Scatterplot of Studentized Residuals to 2009 State Black Percent Proficient

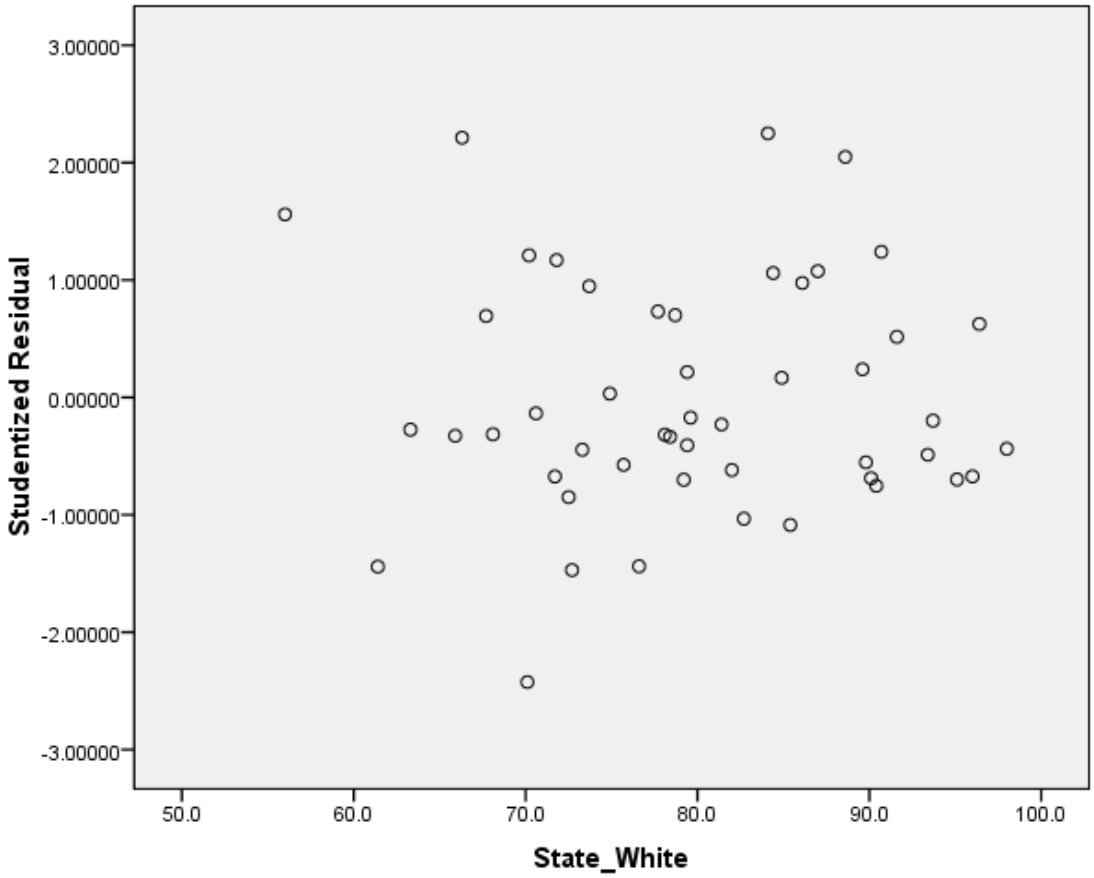

Figure 24. Scatterplot of Studentized Residuals to 2009 State White Percent Proficient

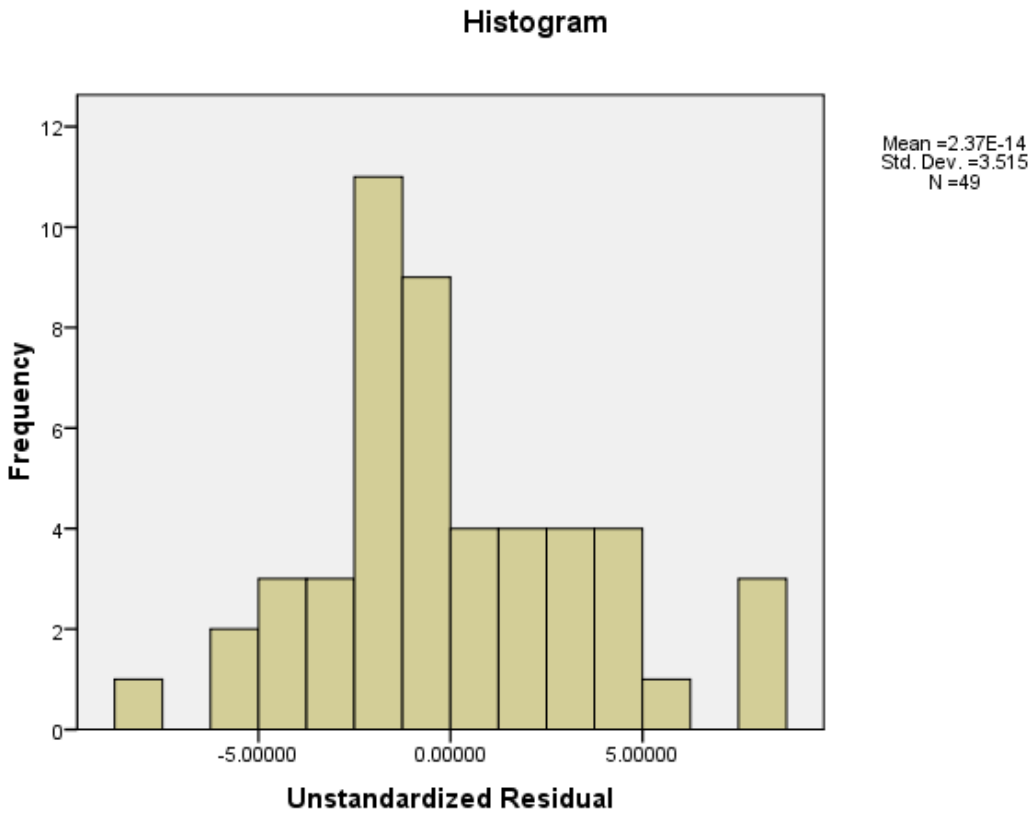

Figure 25. Histogram of Unstandardized Residuals

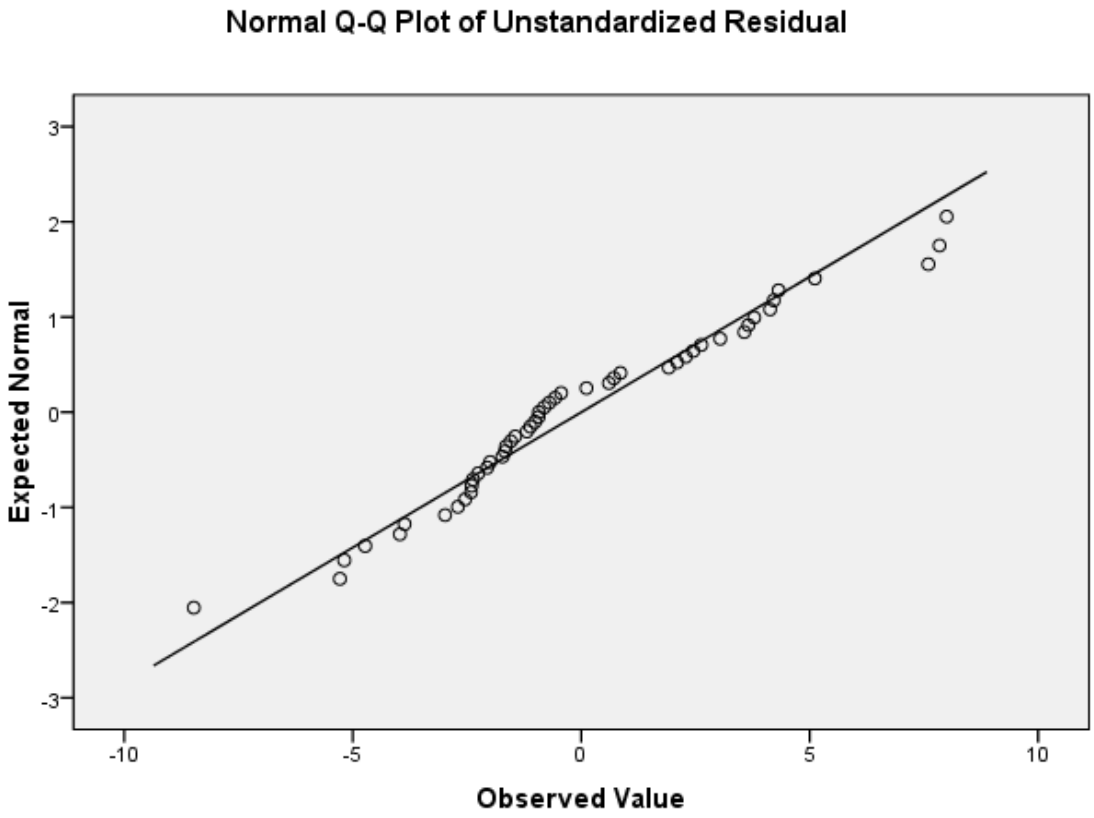

Figure 26. Q-Q Plot of Unstandardized Residuals

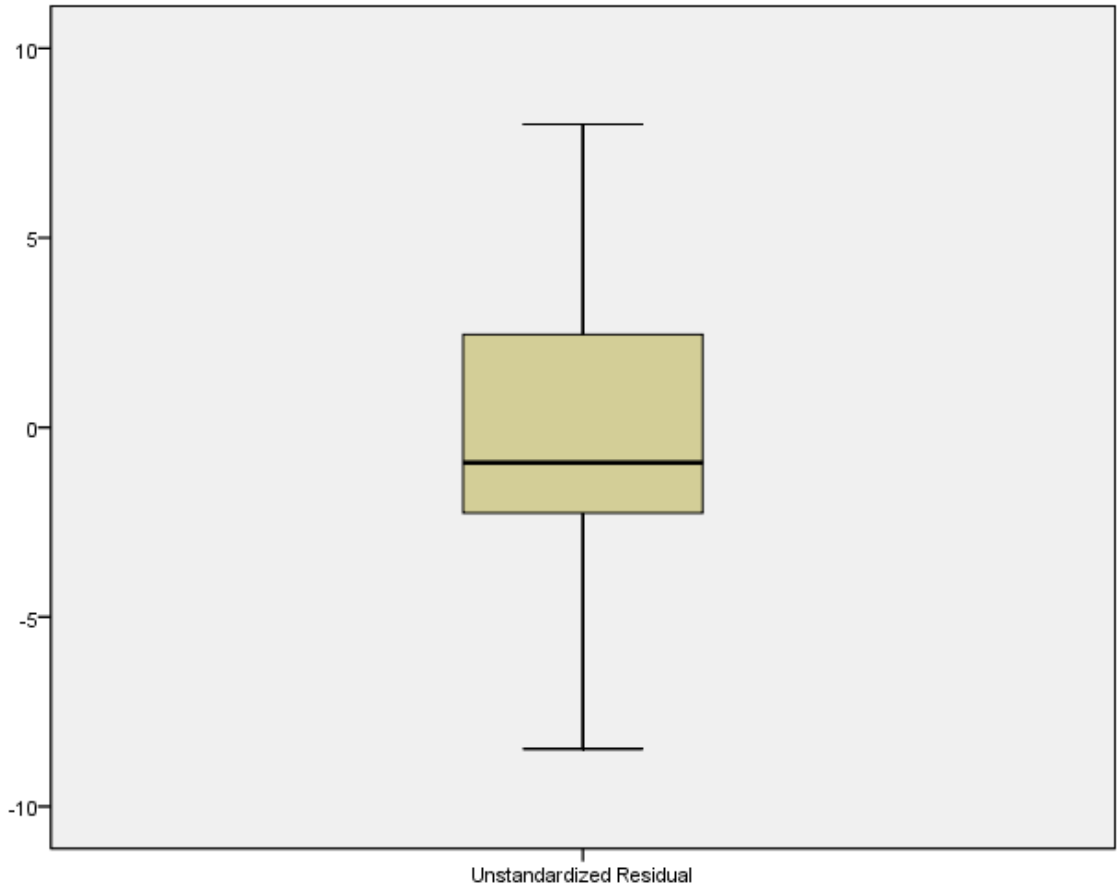

Figure 27. Boxplot of Unstandardized Residuals

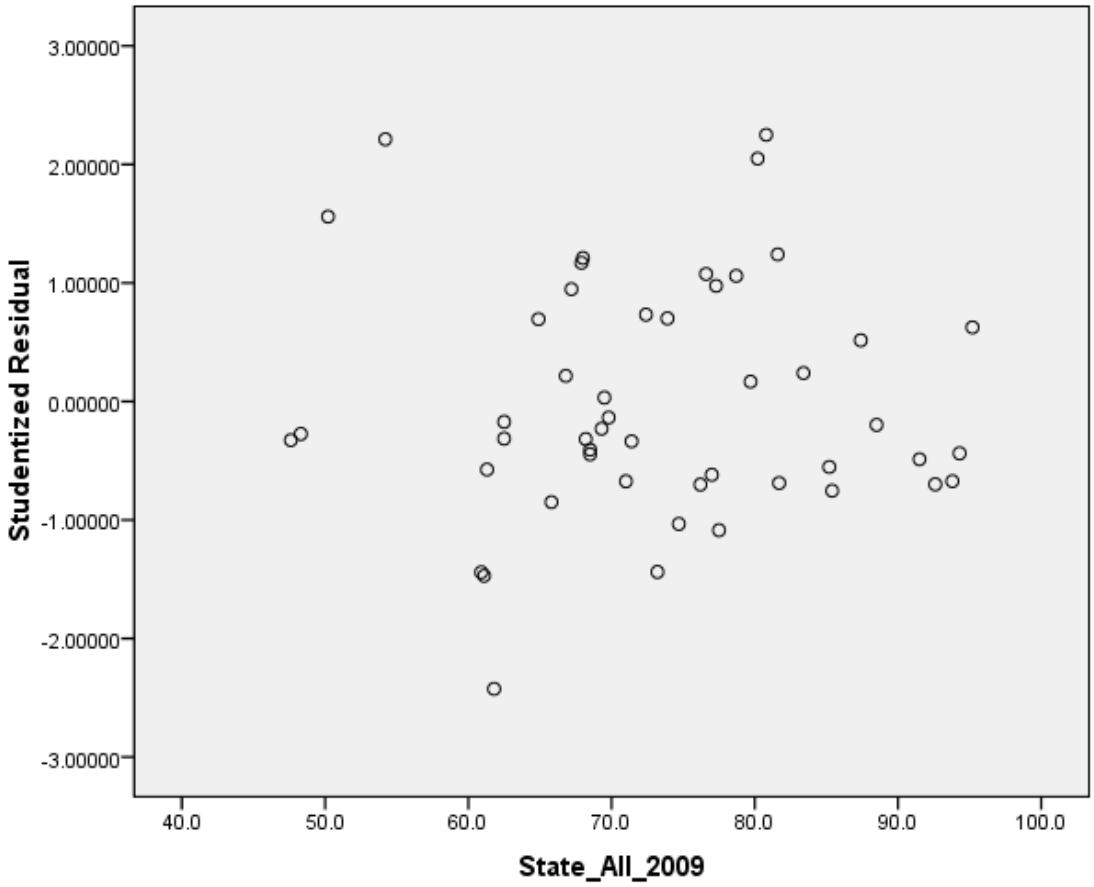

Figure 28. Scatterplot of Studentized Residuals to 2009 State Percent Proficient

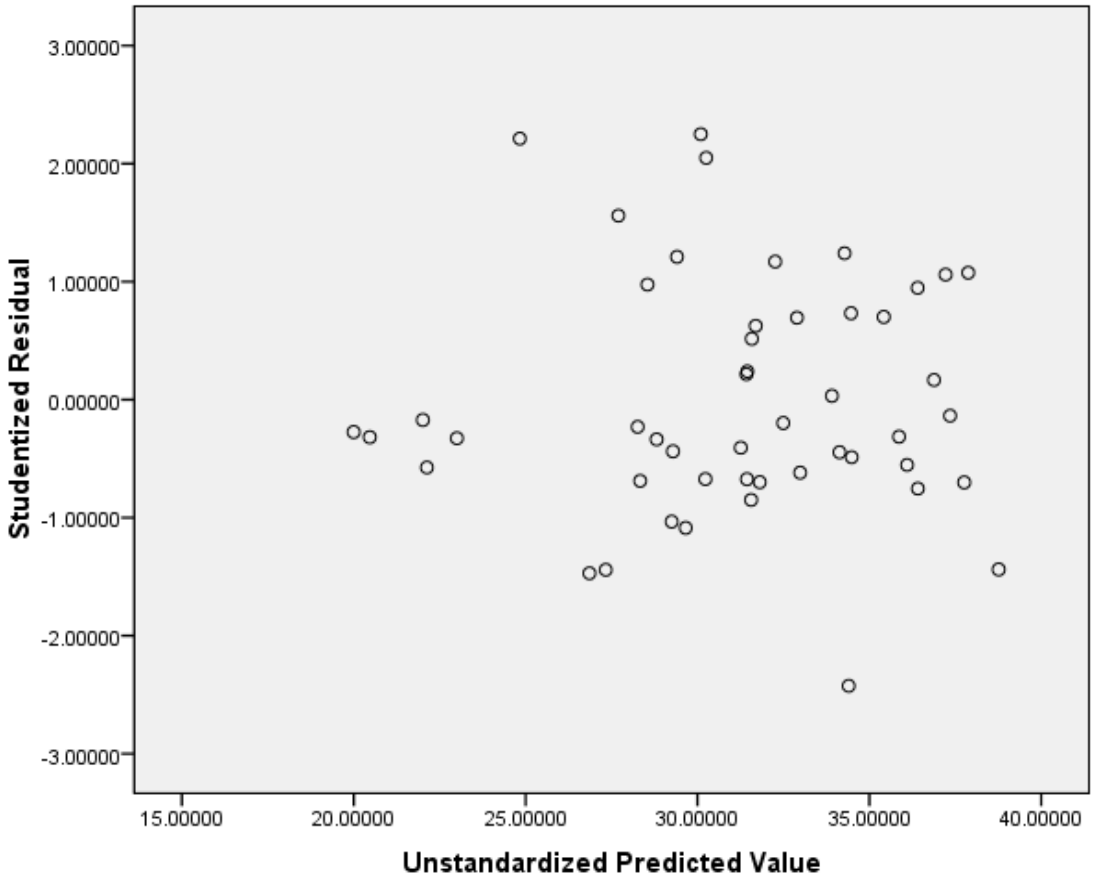

Figure 29. Scatterplot of Studentized Residuals to Unstandardized Predicted Values

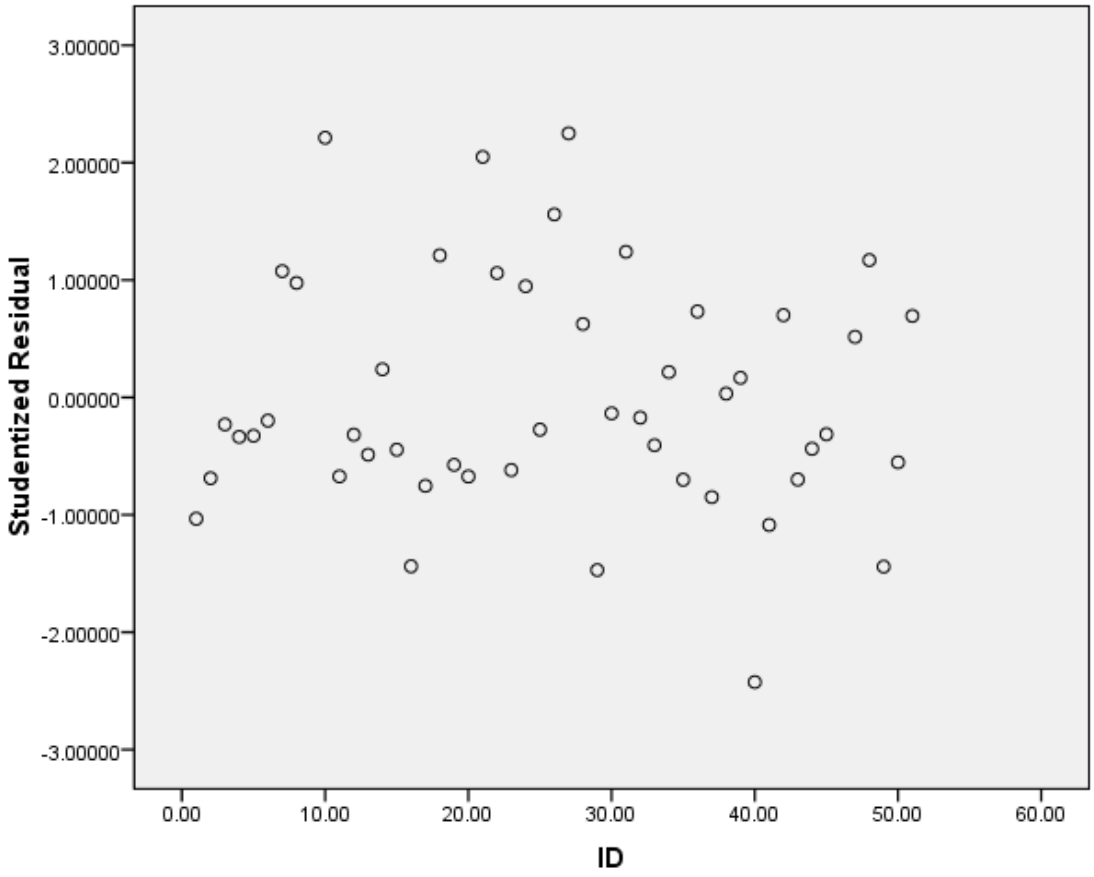

Figure 30. Scatterplot of Studentized Residuals to Case Number

Figures: Research Question Three with Variable Removed

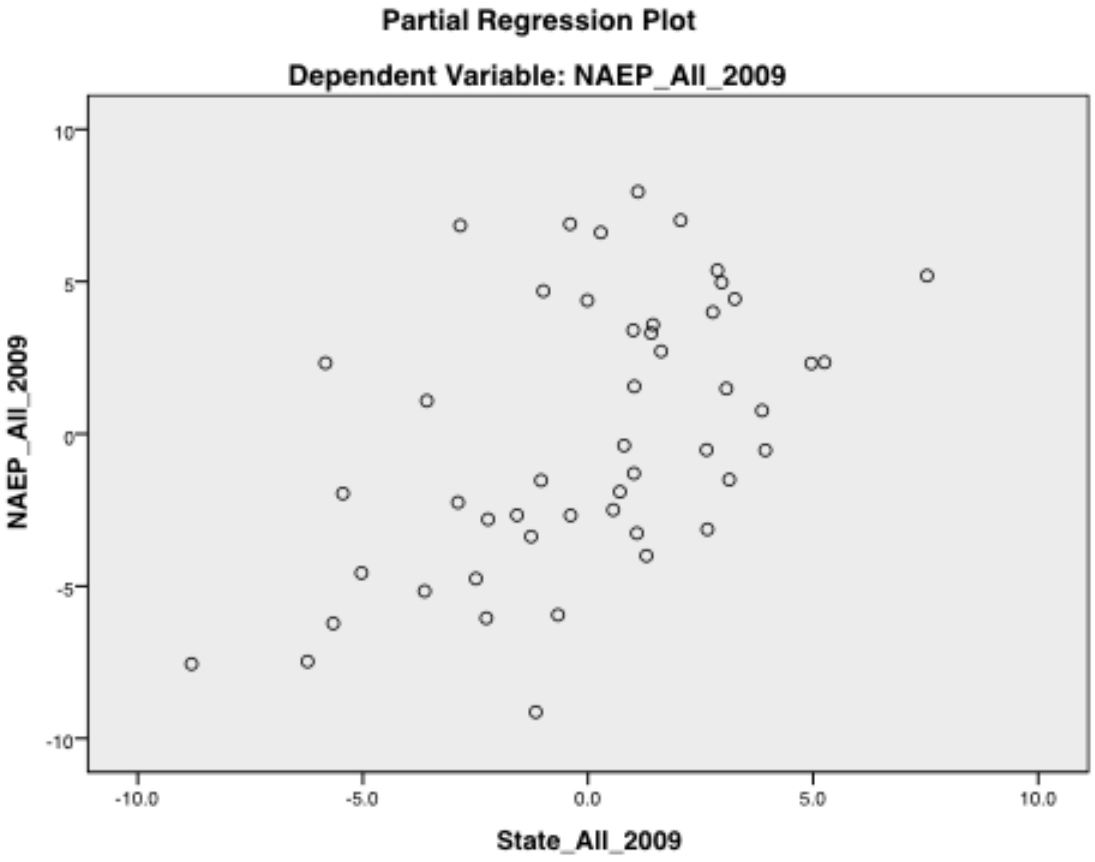

Figure 31. Partial Regression Plot of 2009 NAEP and State Percent Proficient

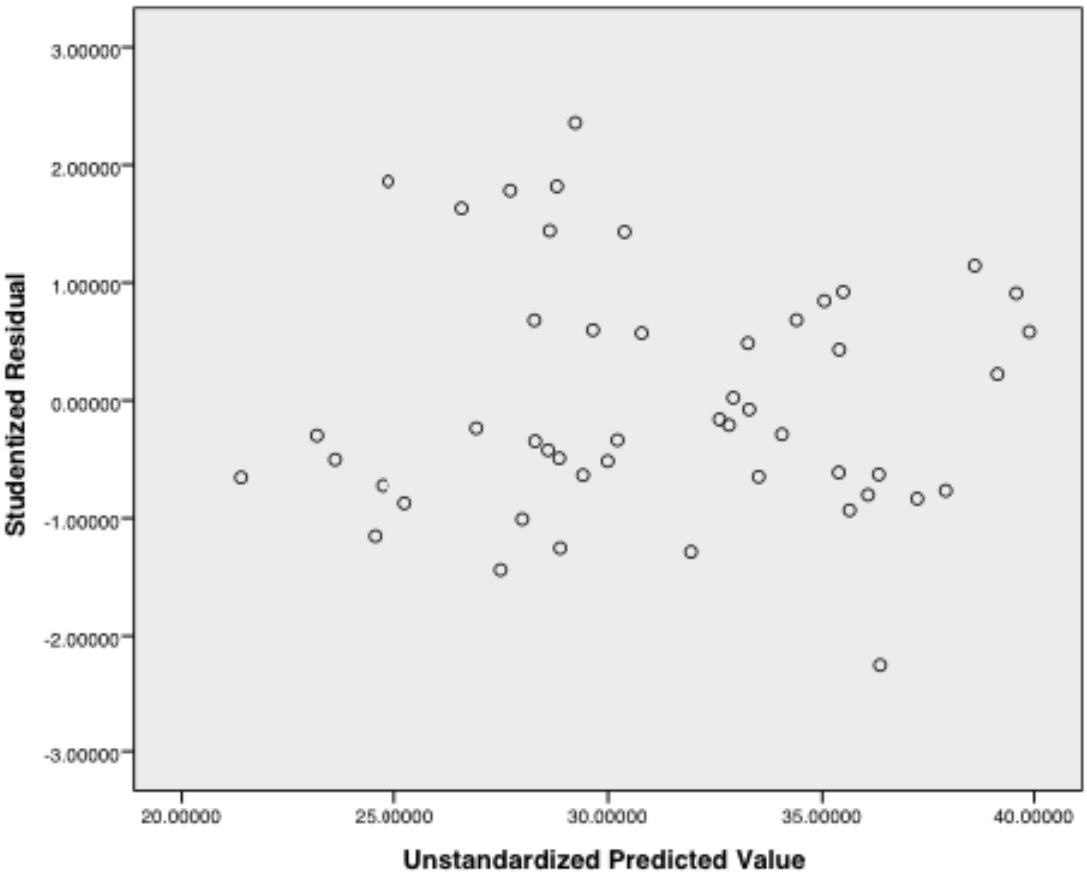

Figure 32. Scatterplot of Studentized Residuals to Unstandardized Predicted Values

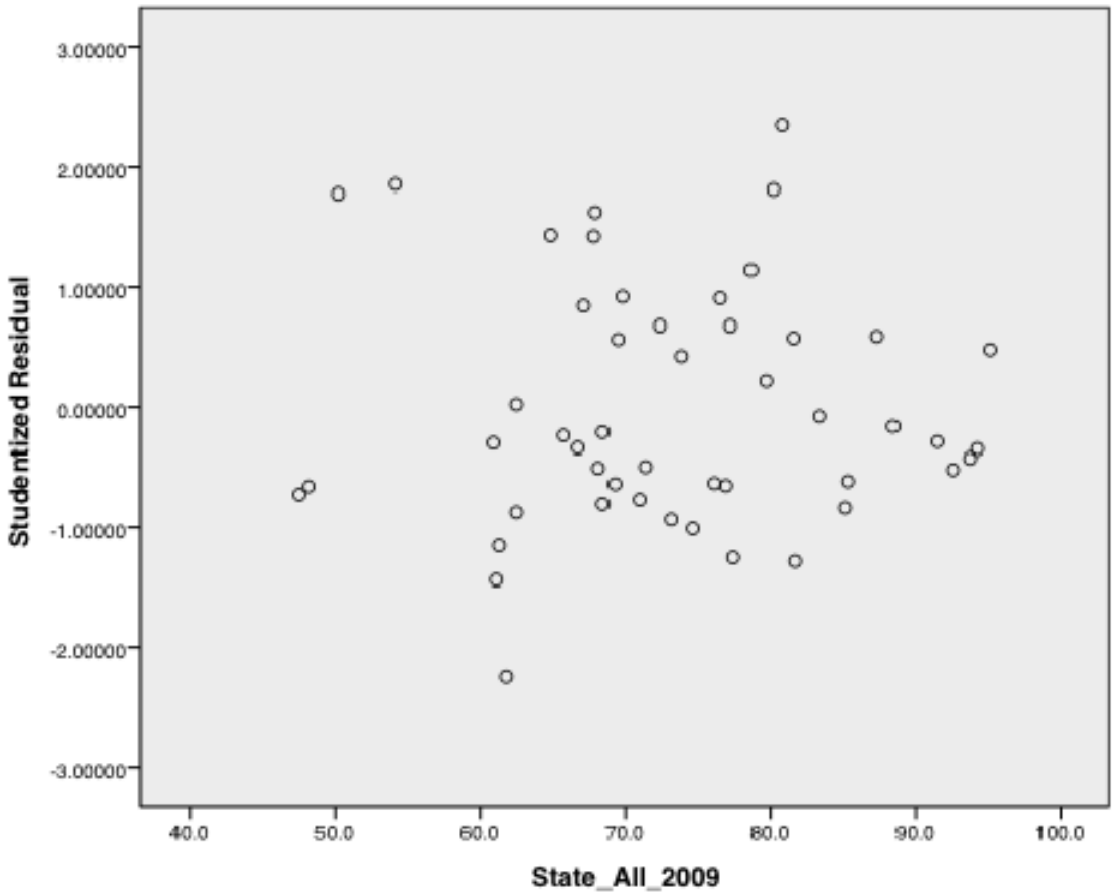

Figure 33. Scatterplot of Studentized Residuals to 2009 State Percent Proficient

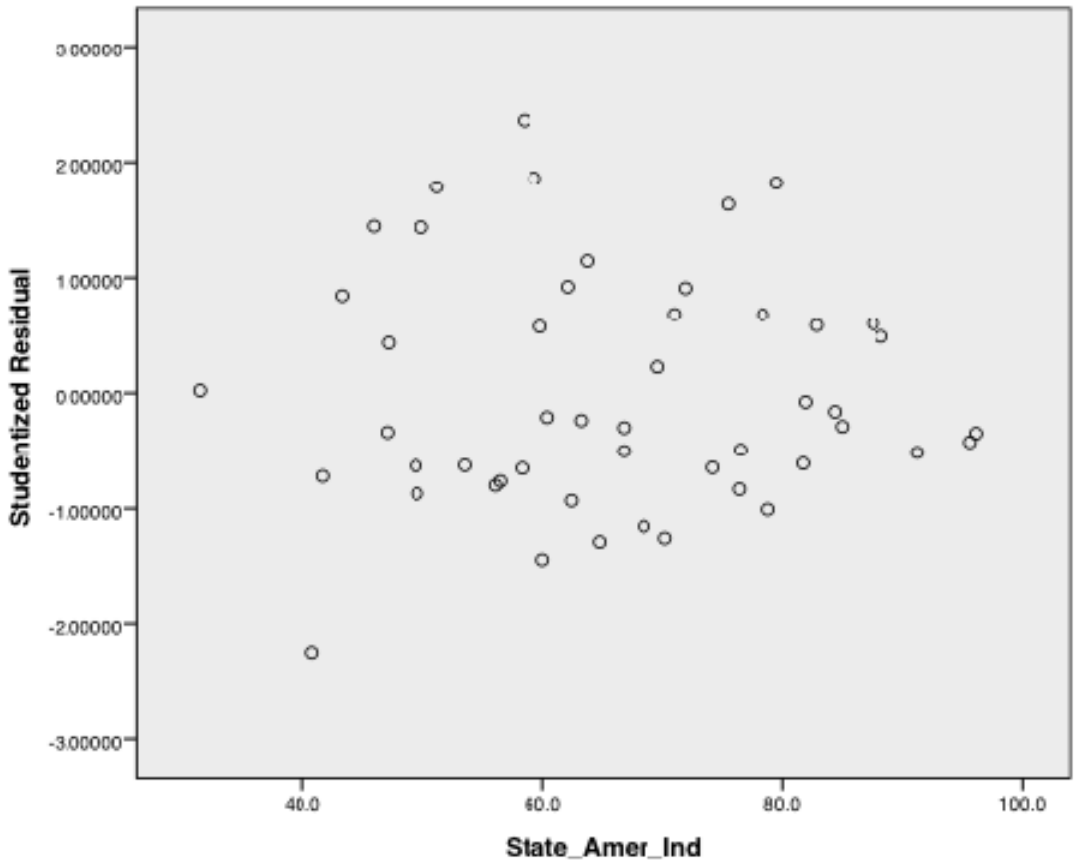

Figure 34. Scatterplot of Studentized Residuals to 2009 State American Indian Percent Proficient

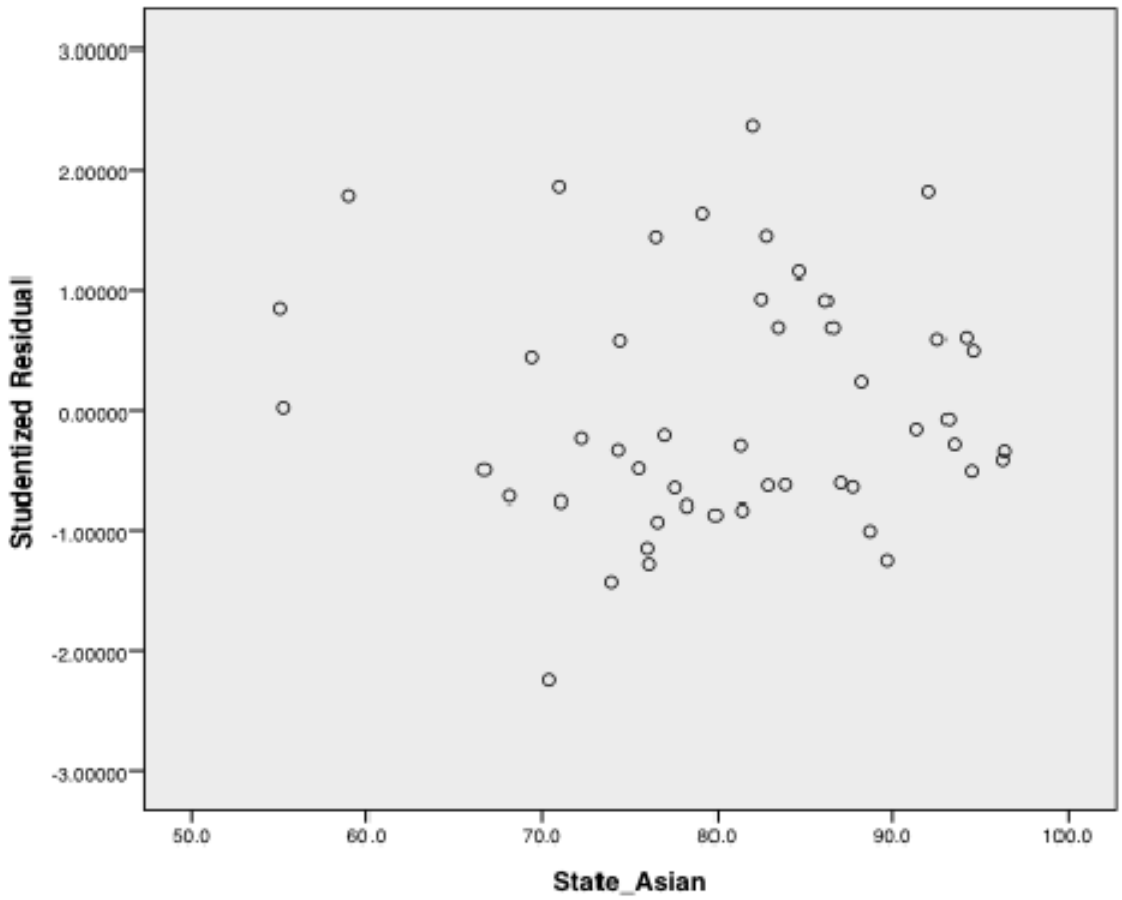

Figure 35. Scatterplot of Studentized Residuals to 2009 State Asian Percent Proficient

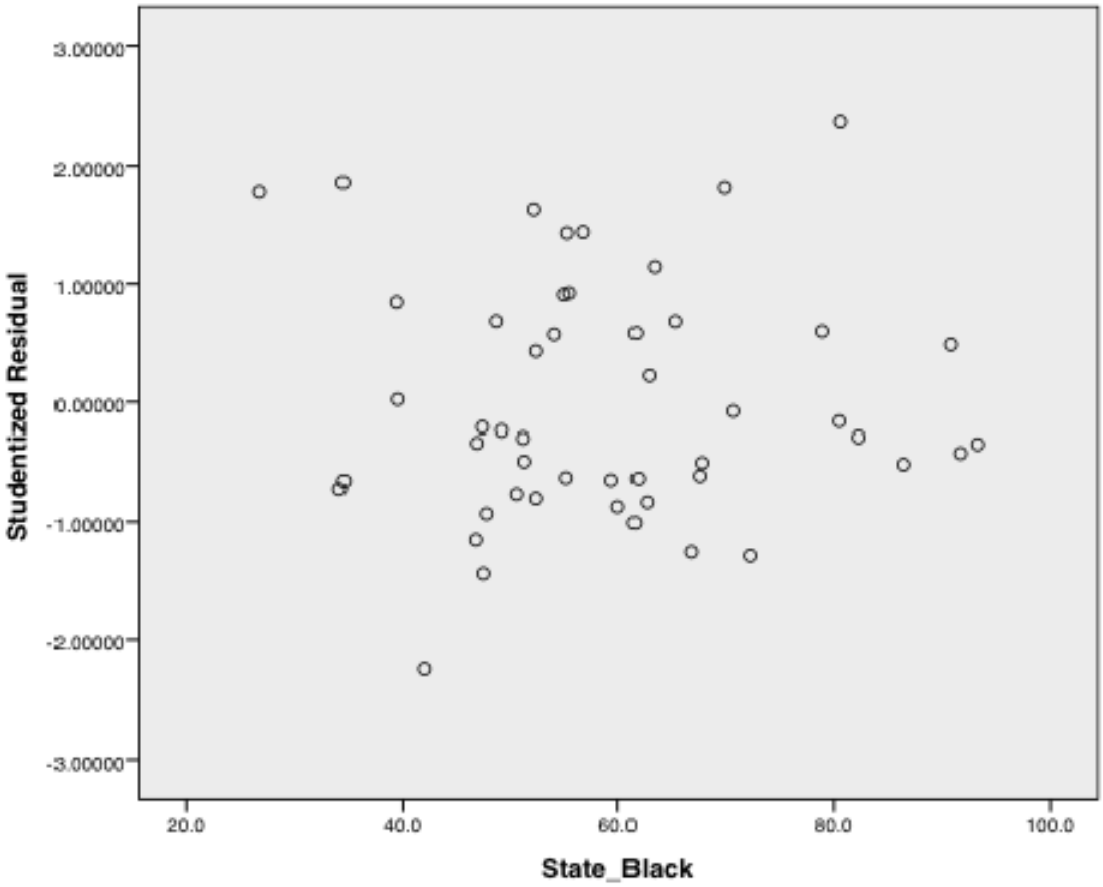

Figure 36. Scatterplot of Studentized Residuals to 2009 State Black Percent Proficient

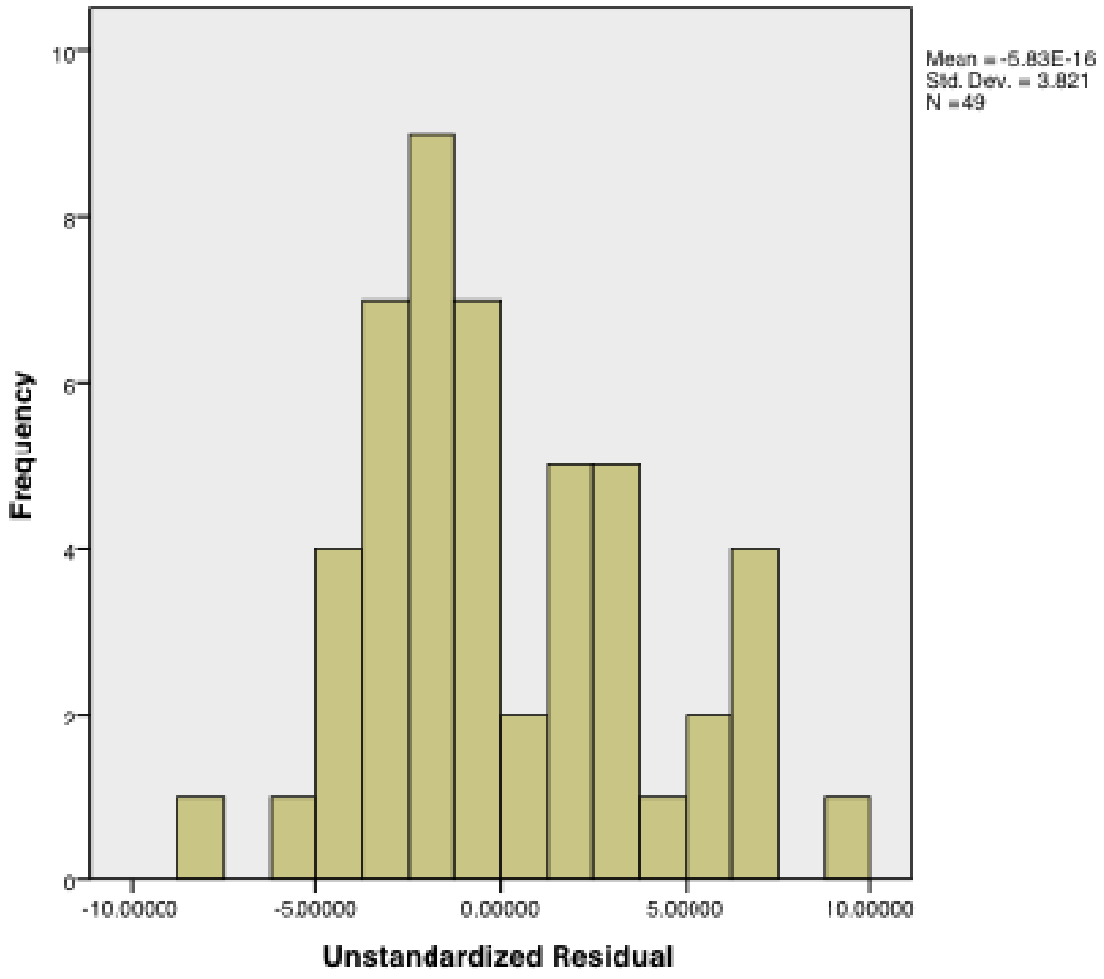

Figure 37. Histogram of Unstandardized Residuals

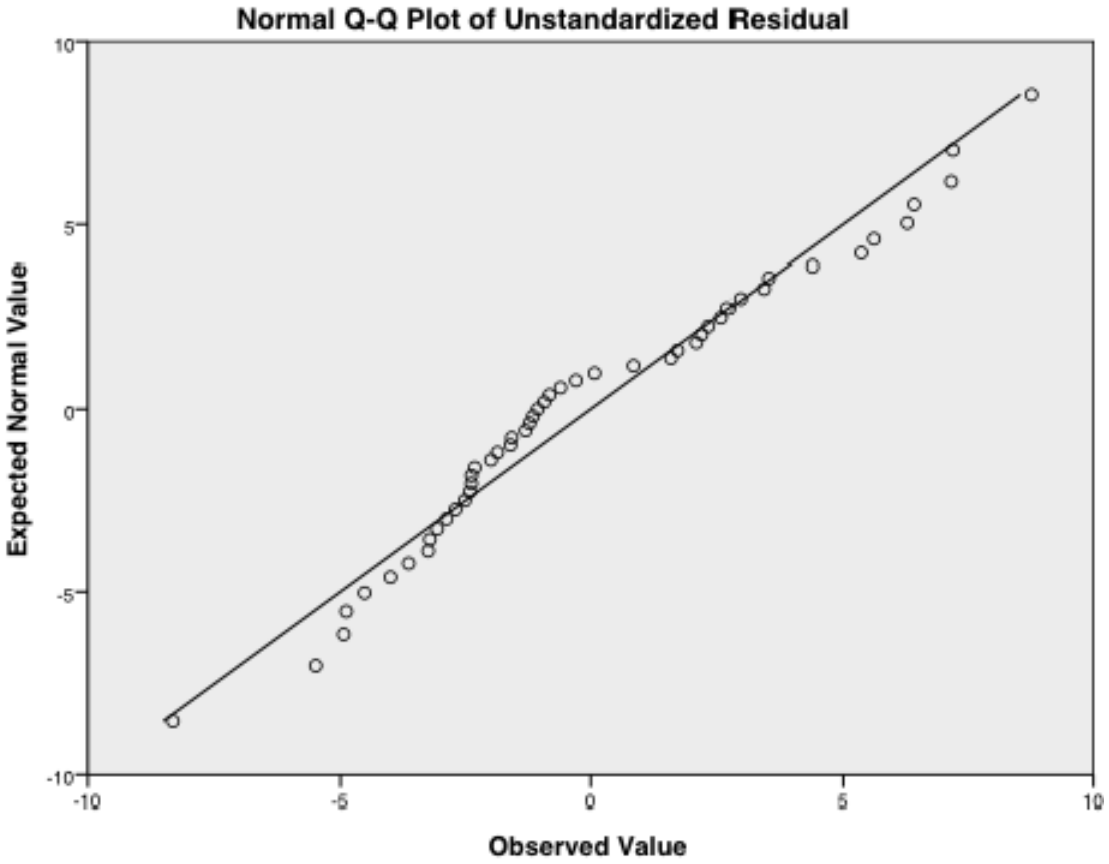

Figure 38. Q-Q Plot of Unstandardized Residuals

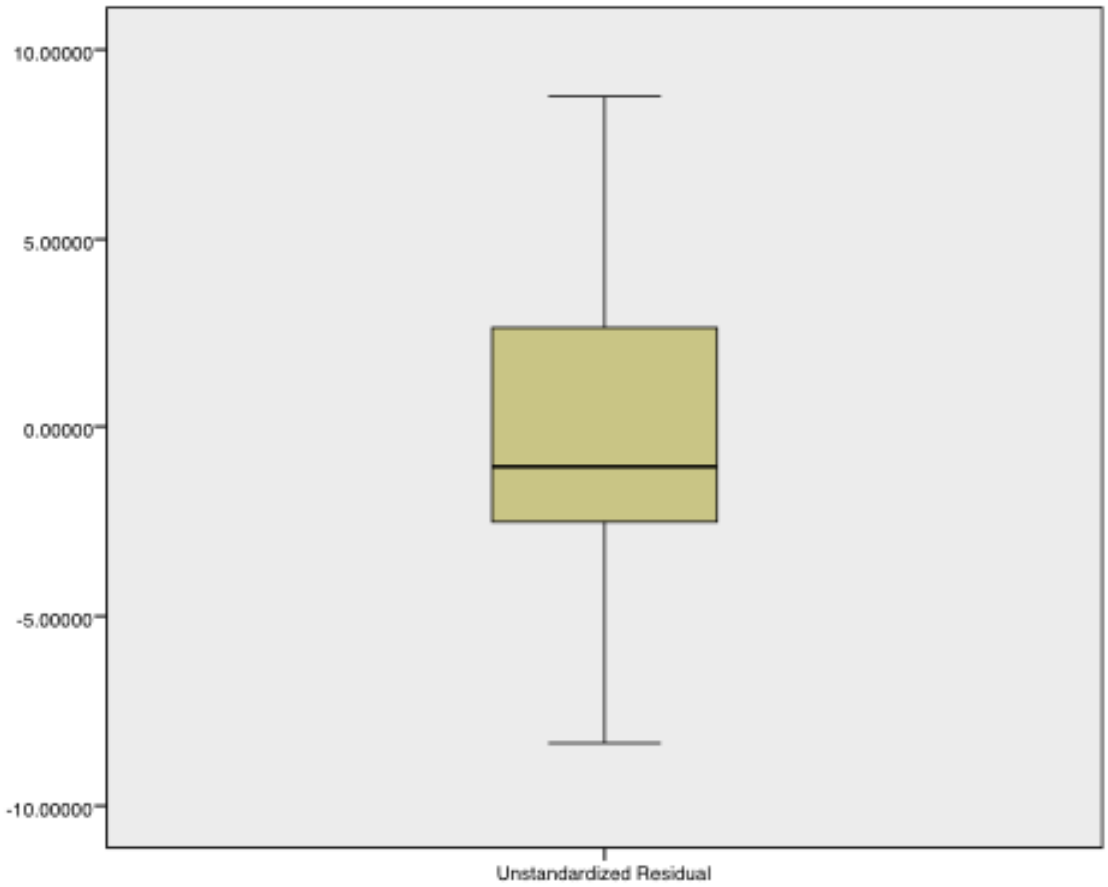

Figure 39. Boxplot of Unstandardized Residuals

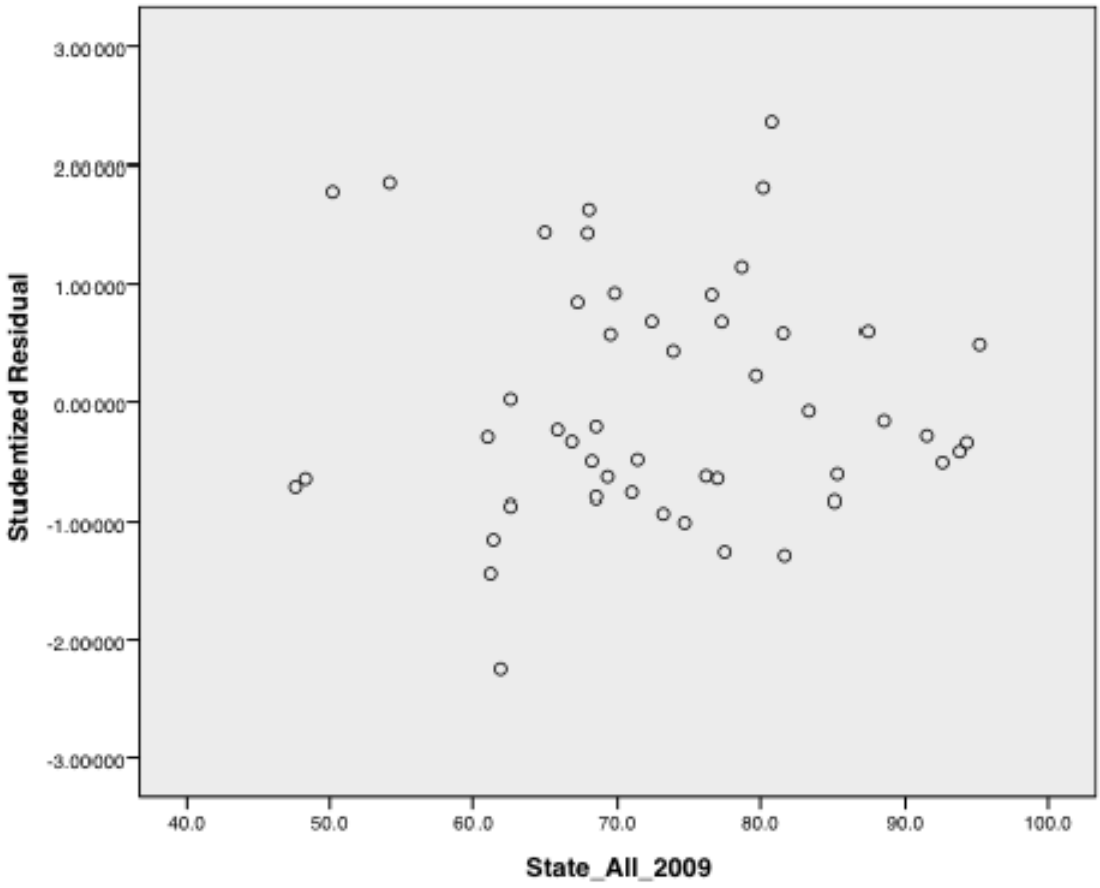

Figure 40. Scatterplot of Studentized Residuals to 2009 State Percent Proficient

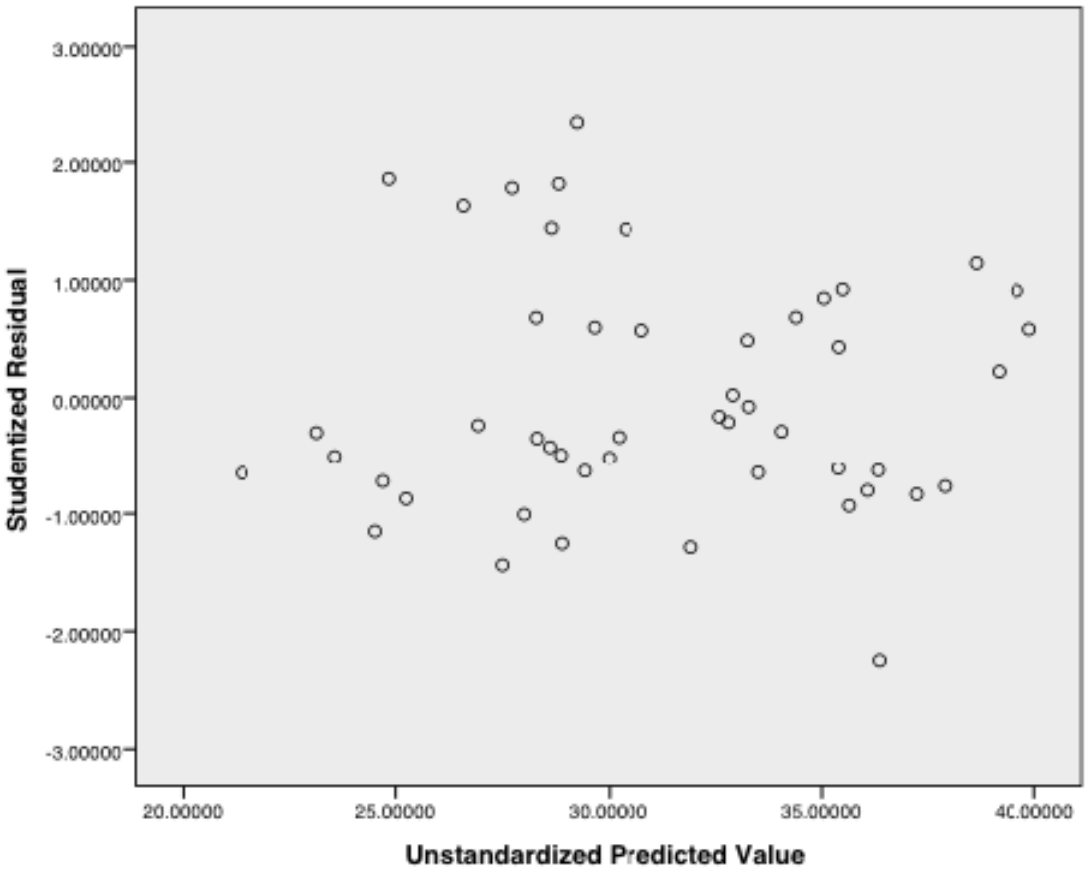

Figure 41. Scatterplot of Studentized Residuals to Unstandardized Predicted Values

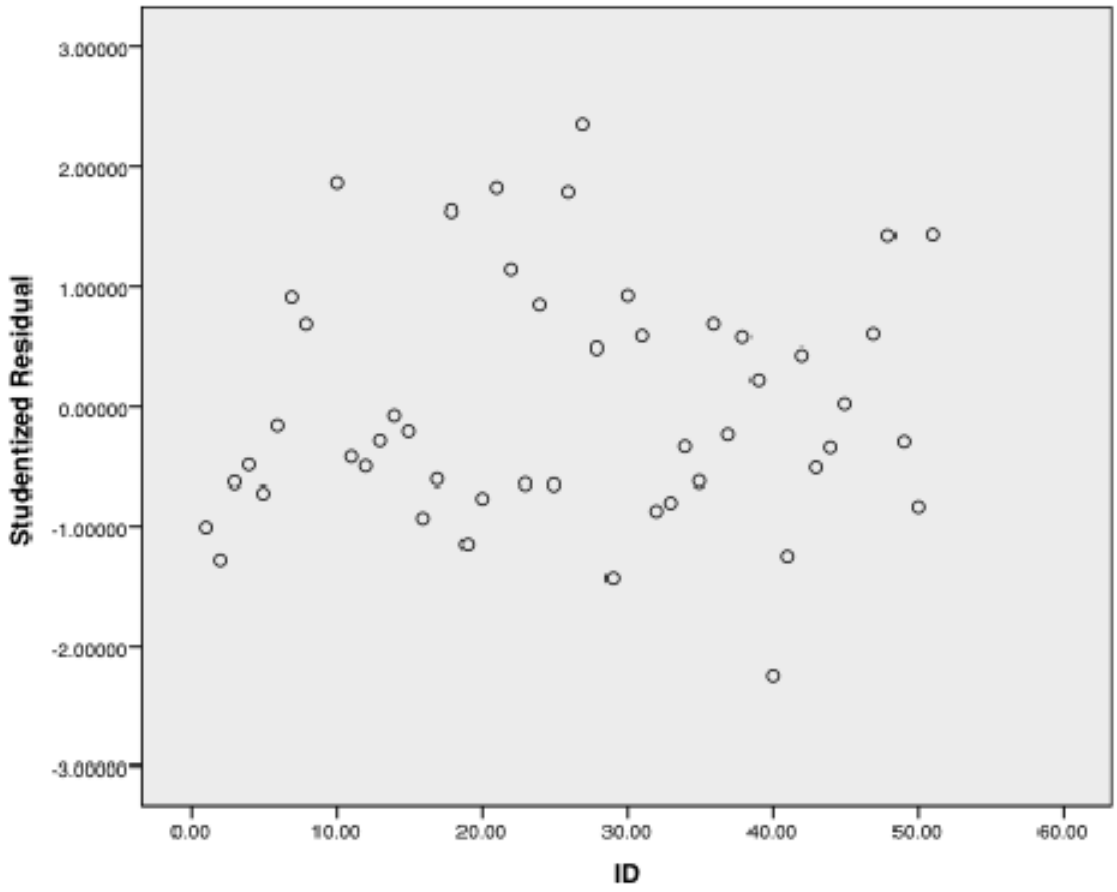

Figure 42. Scatterplot of Studentized Residuals to Case Number

Figures: Research Question Four

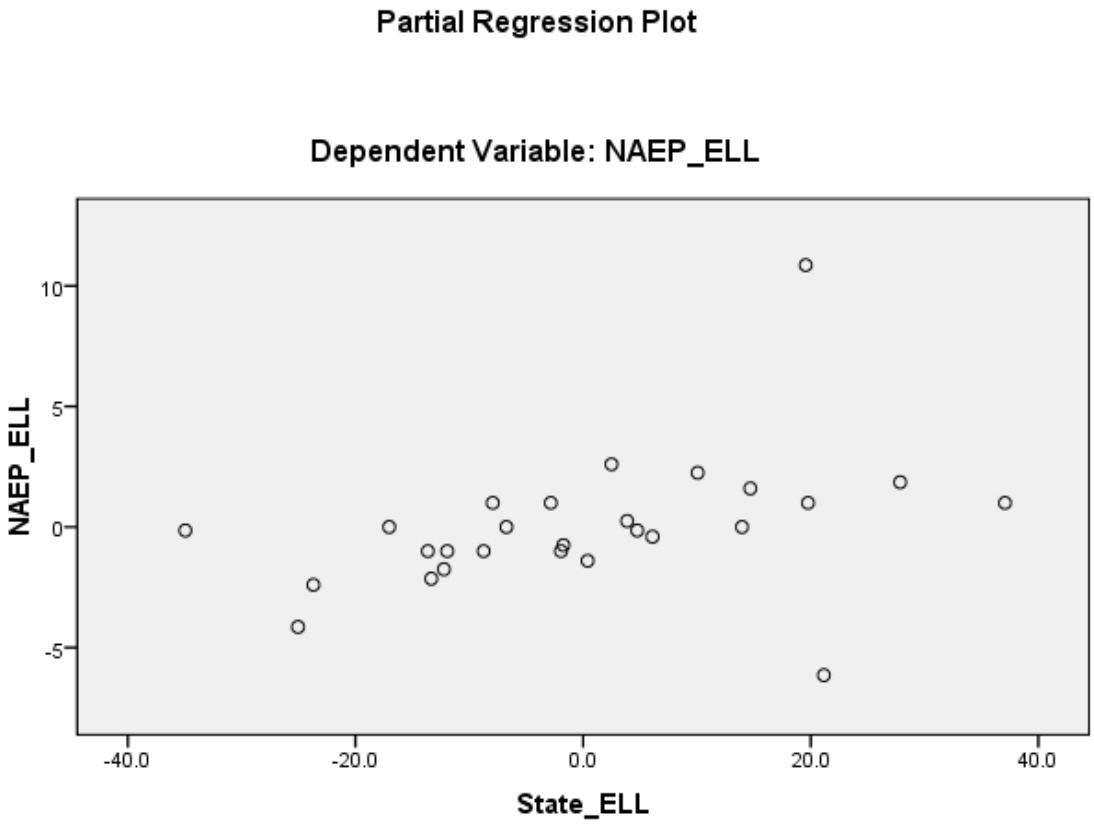

Figure 43. Partial Regression Plot of 2009 NAEP to State ELL Percent Proficient

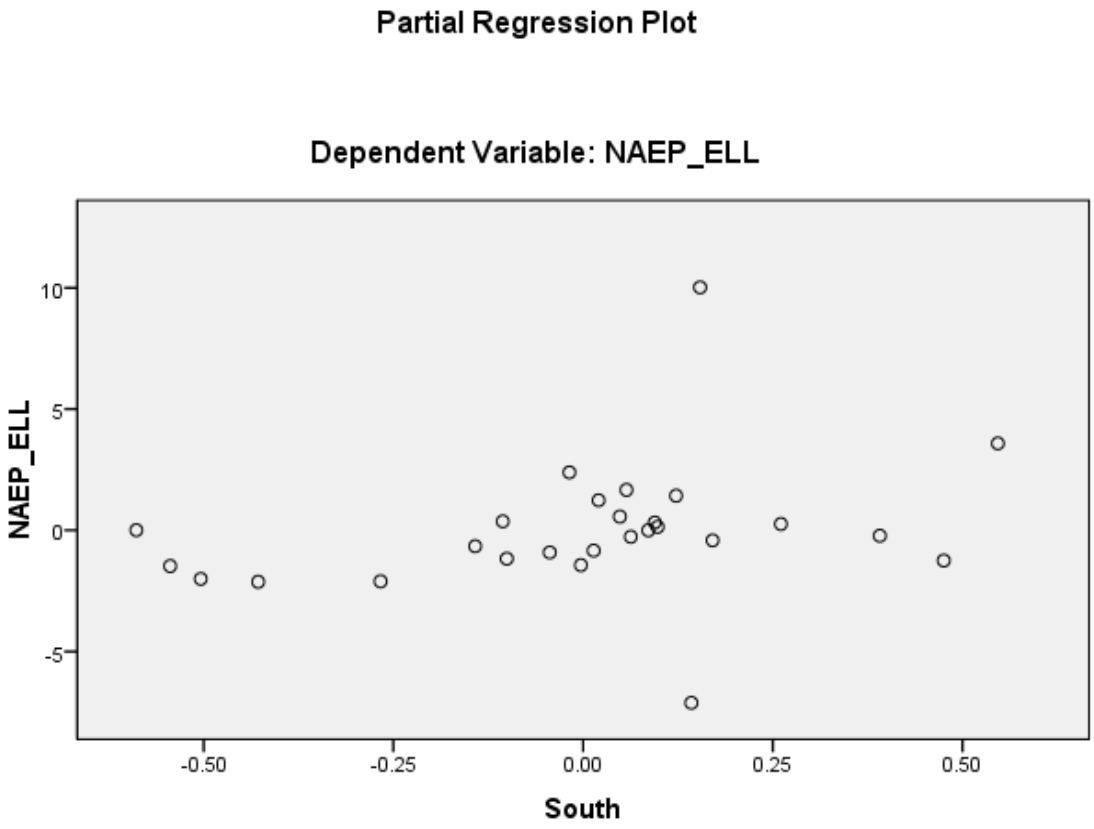

Figure 44. Partial Regression Plot of 2009 NAEP and State ELL Percent Proficient in the South

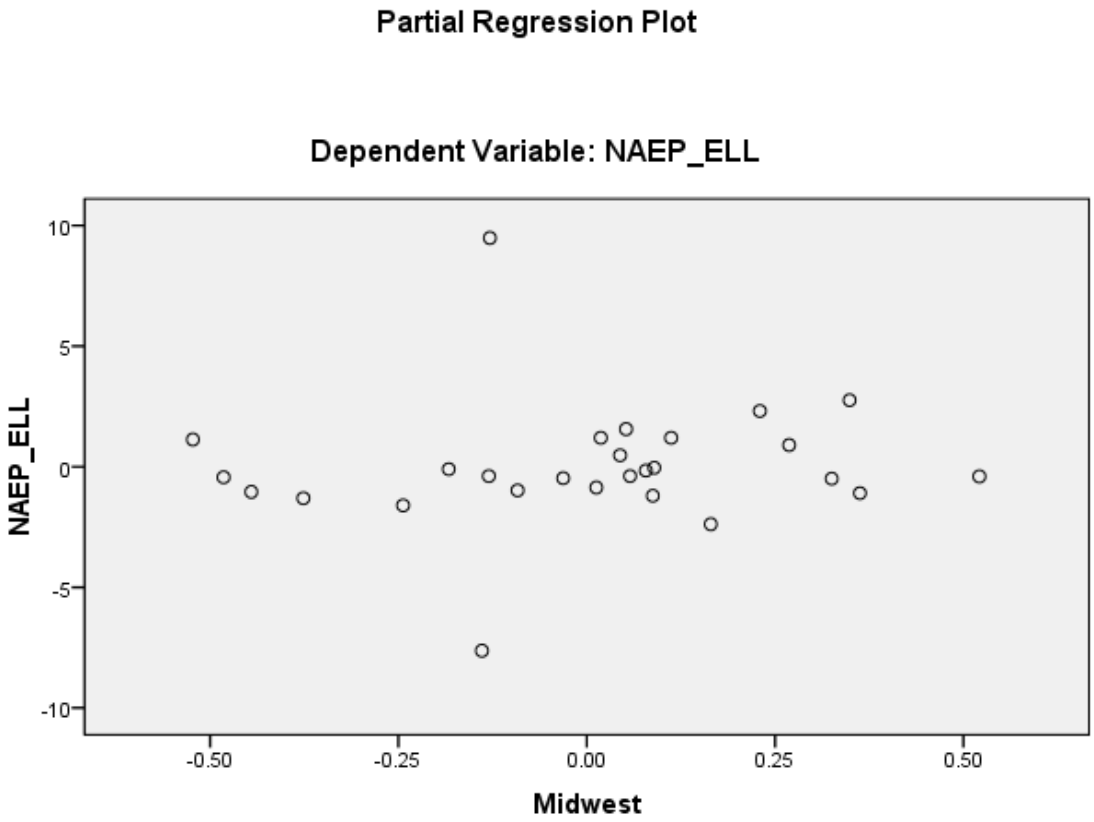

Figure 45. Partial Regression Plot of 2009 NAEP and State ELL Percent Proficient in the Midwest

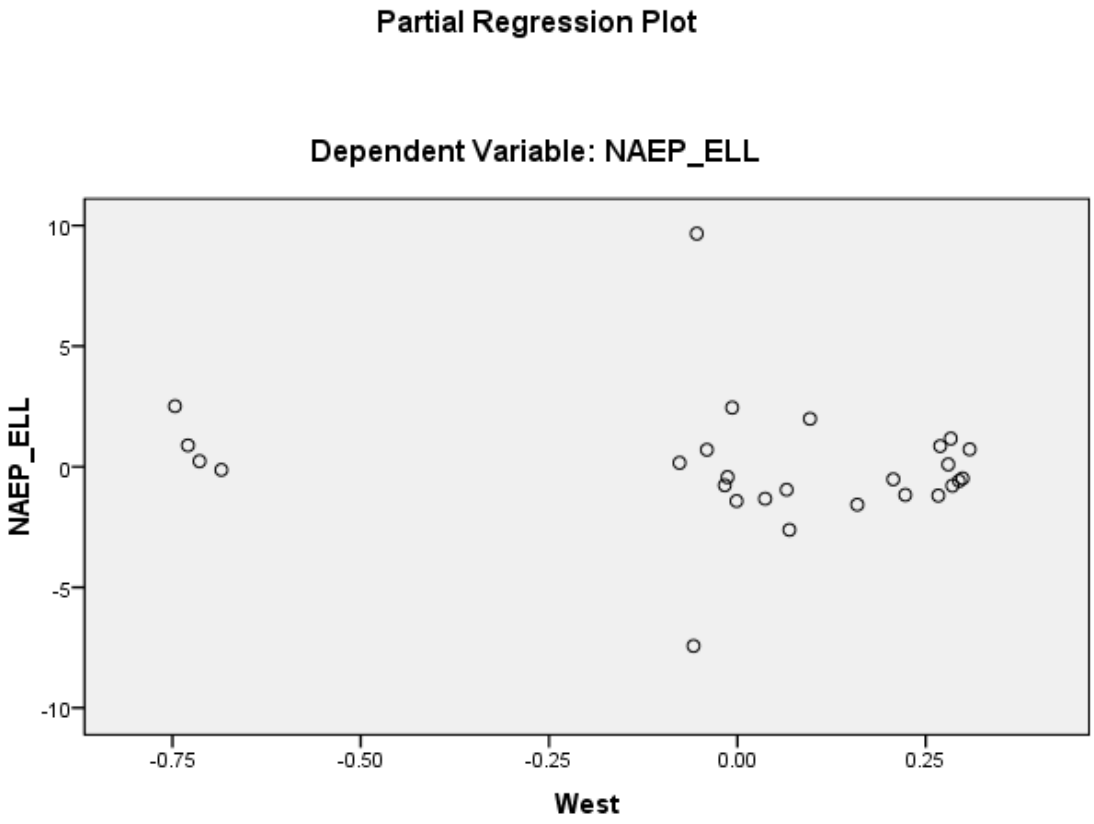

Figure 46. Partial Regression Plot of 2009 NAEP and State ELL Percent Proficient in the West

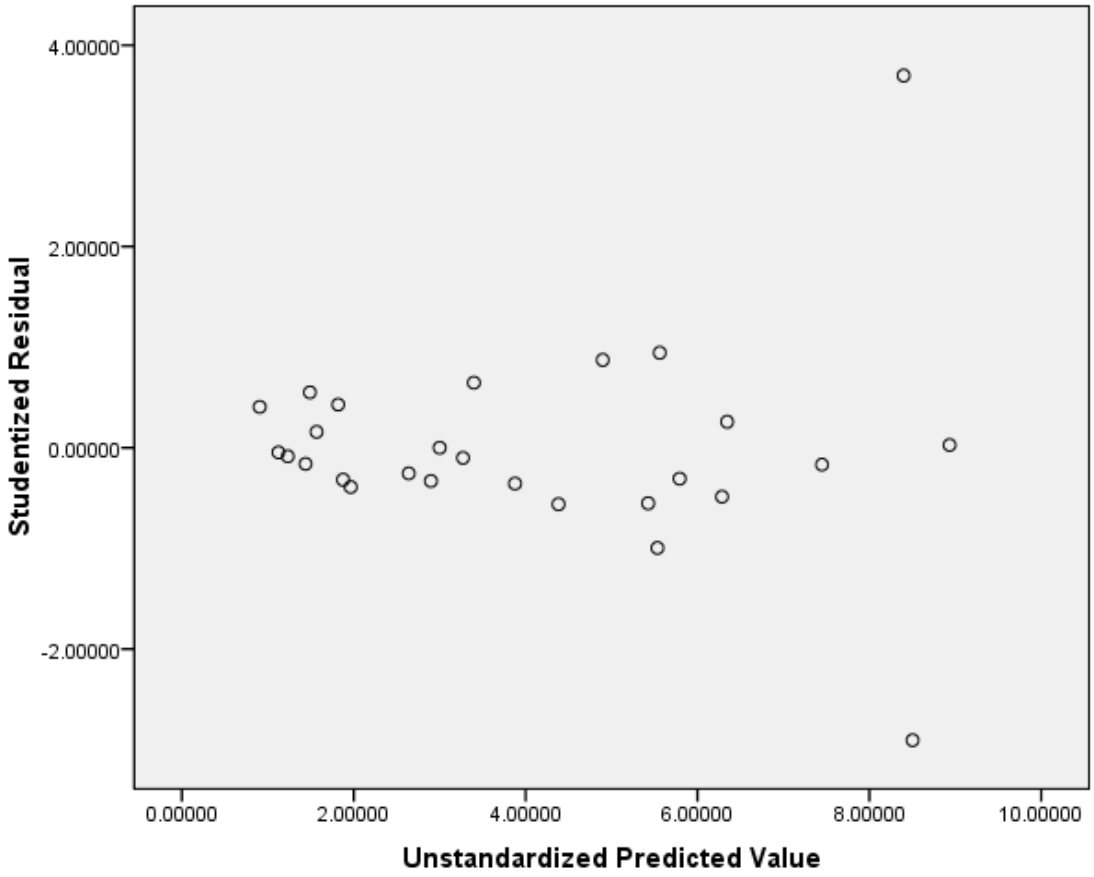

Figure 47. Scatterplot of Studentized Residuals to Unstandardized Predicted Values

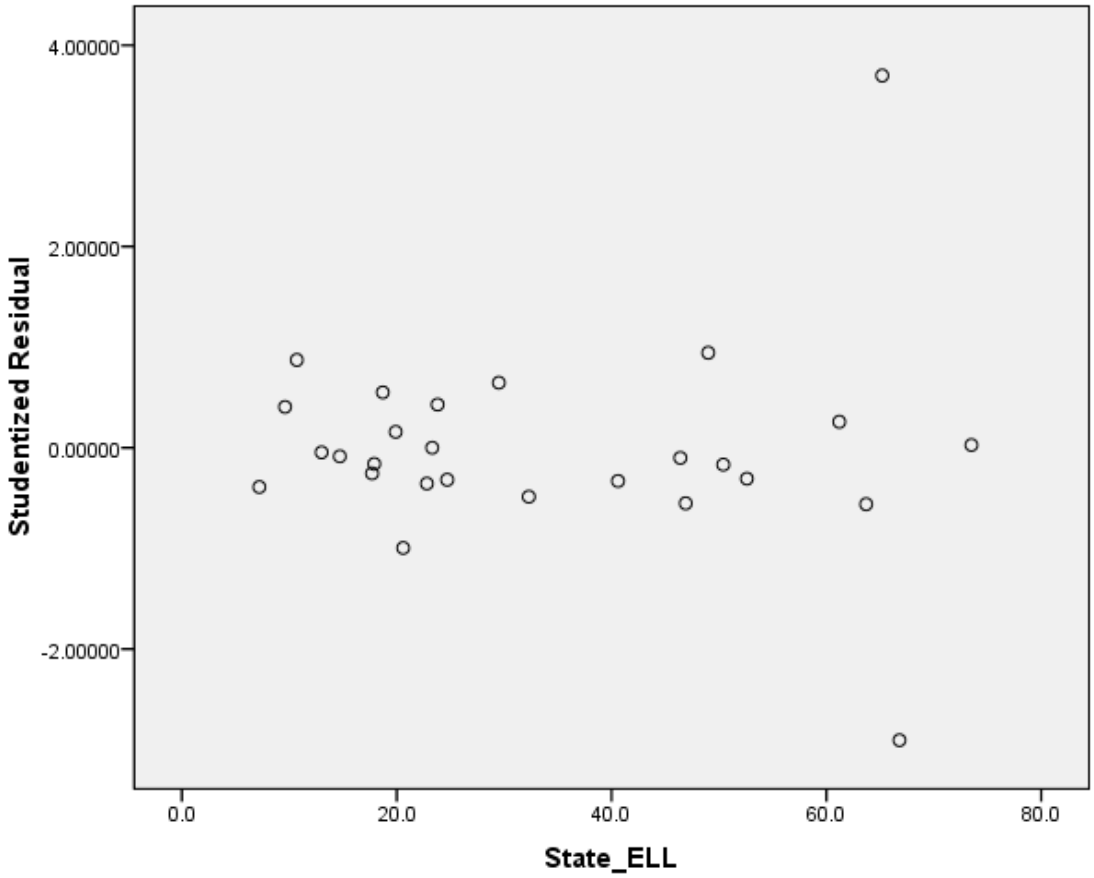

Figure 48. Scatterplot of Studentized Residuals to 2009 State ELL Percent Proficient

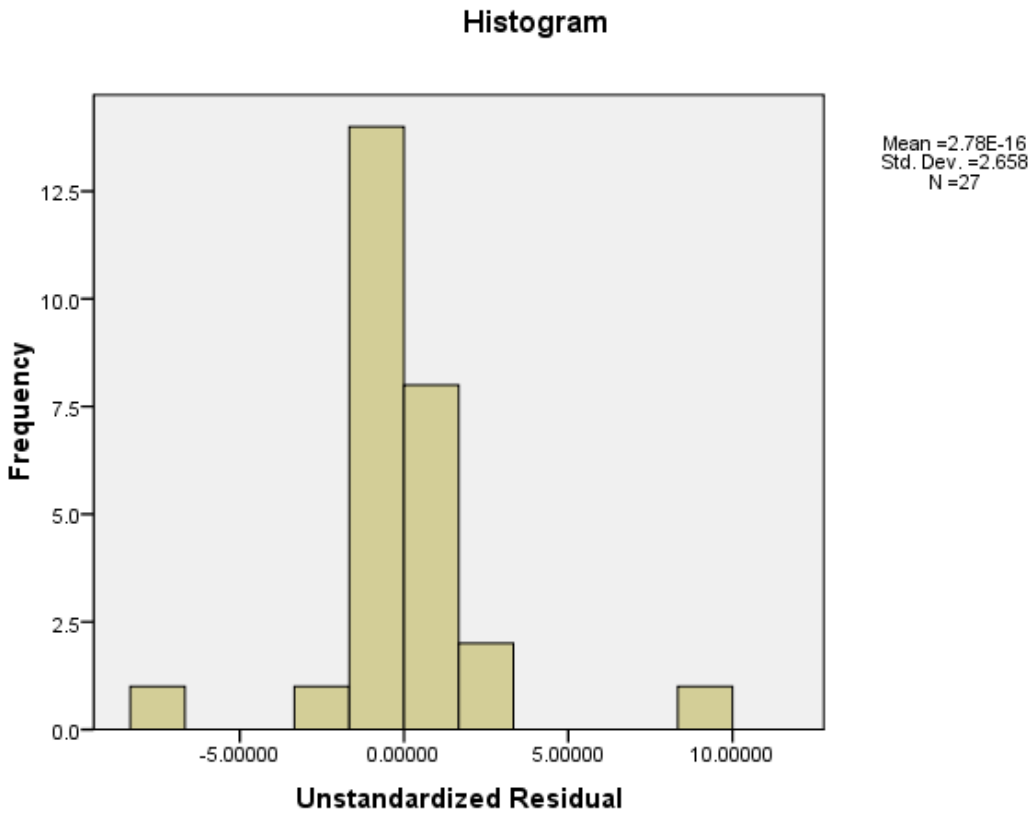

Figure 49. Histogram of Unstandardized Residuals

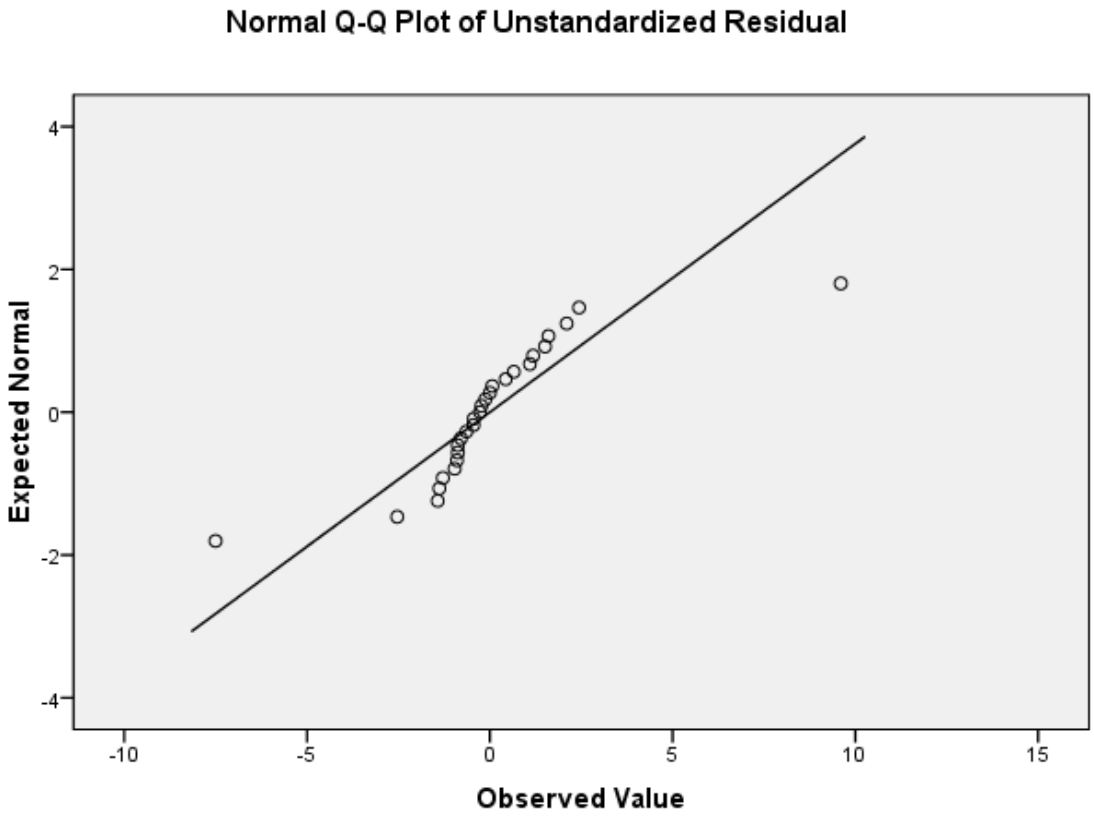

Figure 50. Q-Q Plot of Unstandardized Residuals

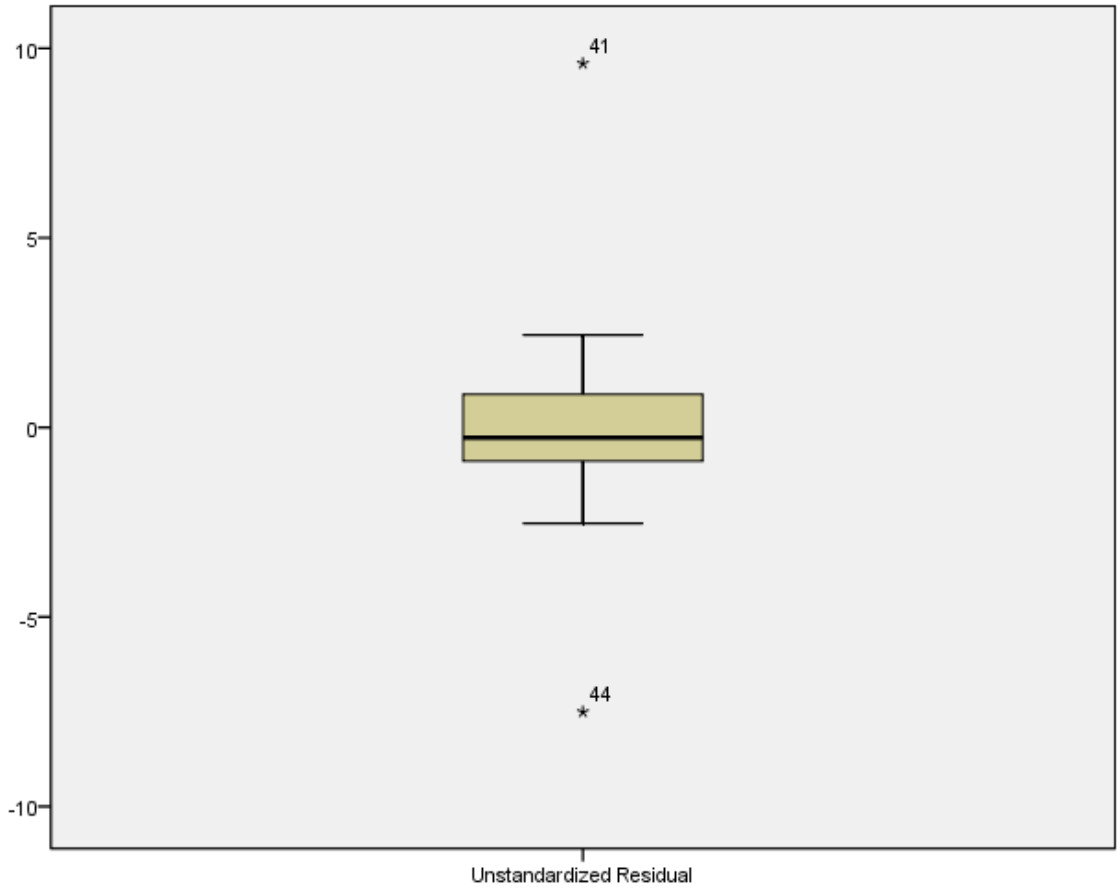

Figure 51. Boxplot of Unstandardized Residuals

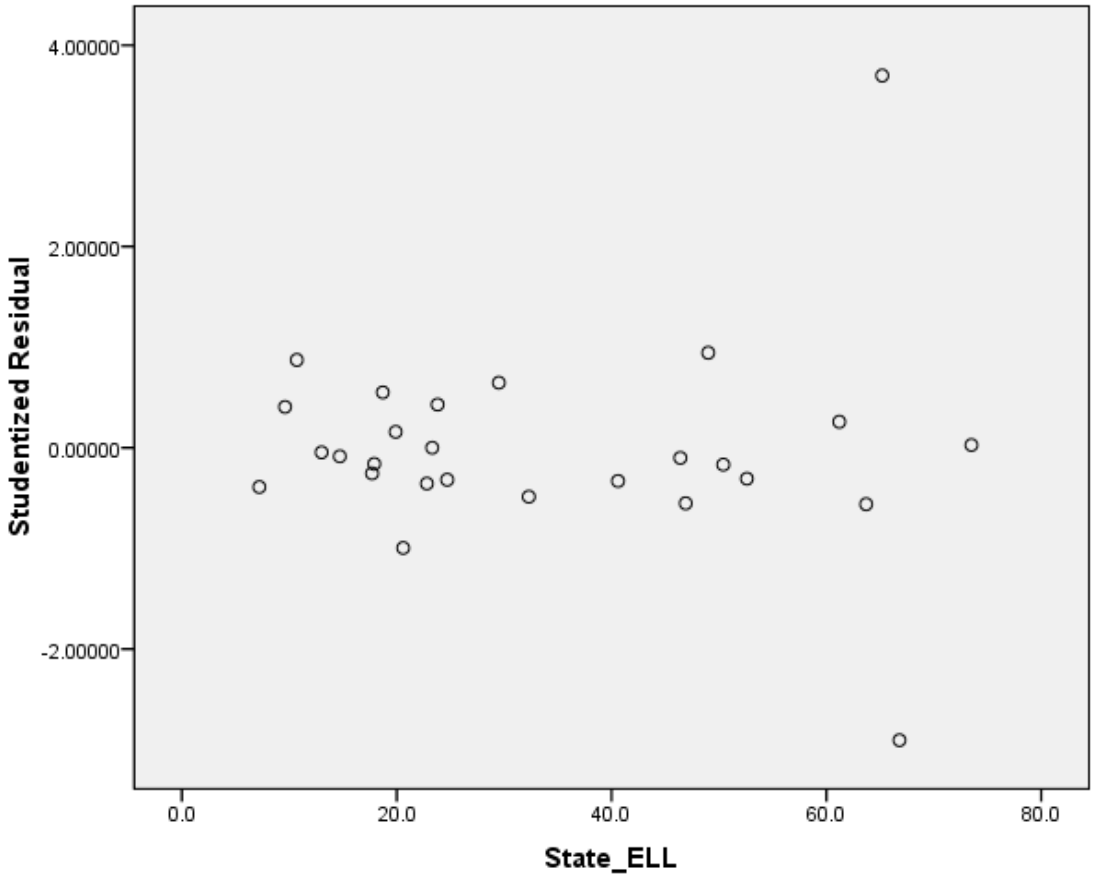

Figure 52. Scatterplot of Studentized Residuals to 2009 State ELL Percent Proficient

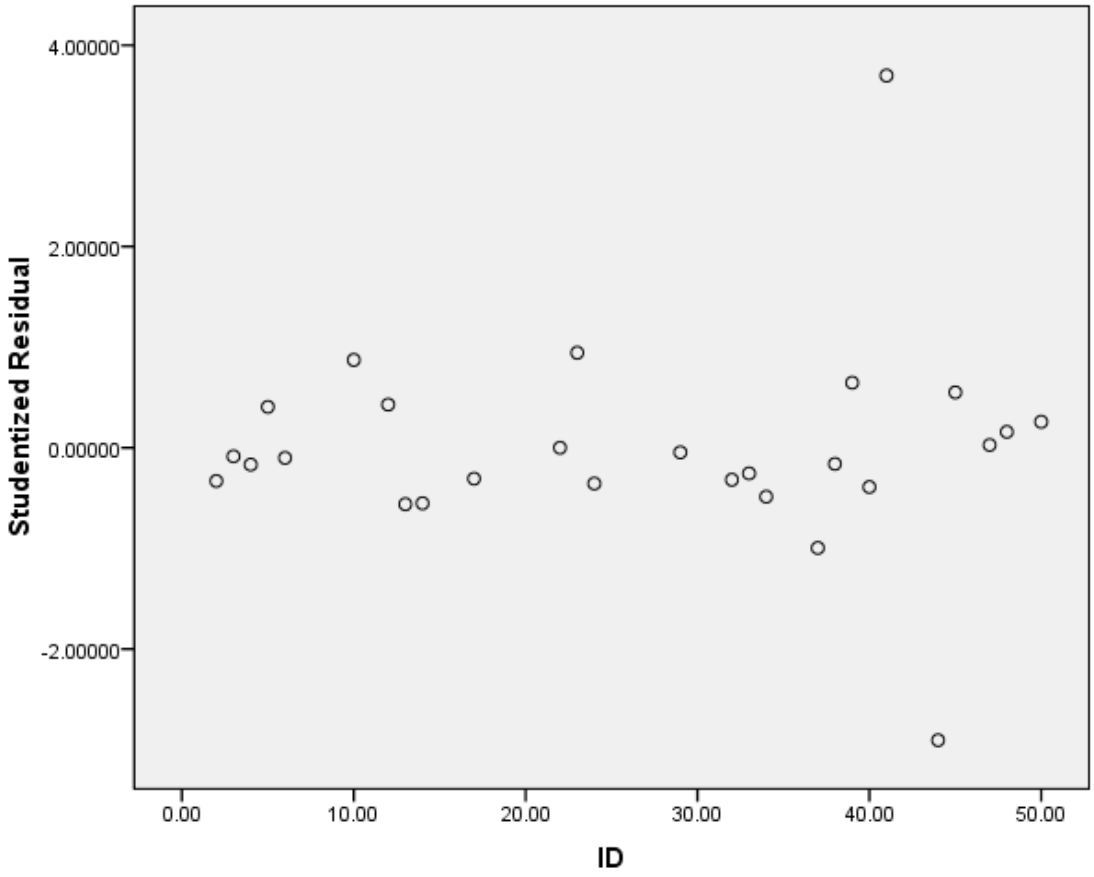

Figure 53. Scatterplot of Studentized Residuals to Case Number

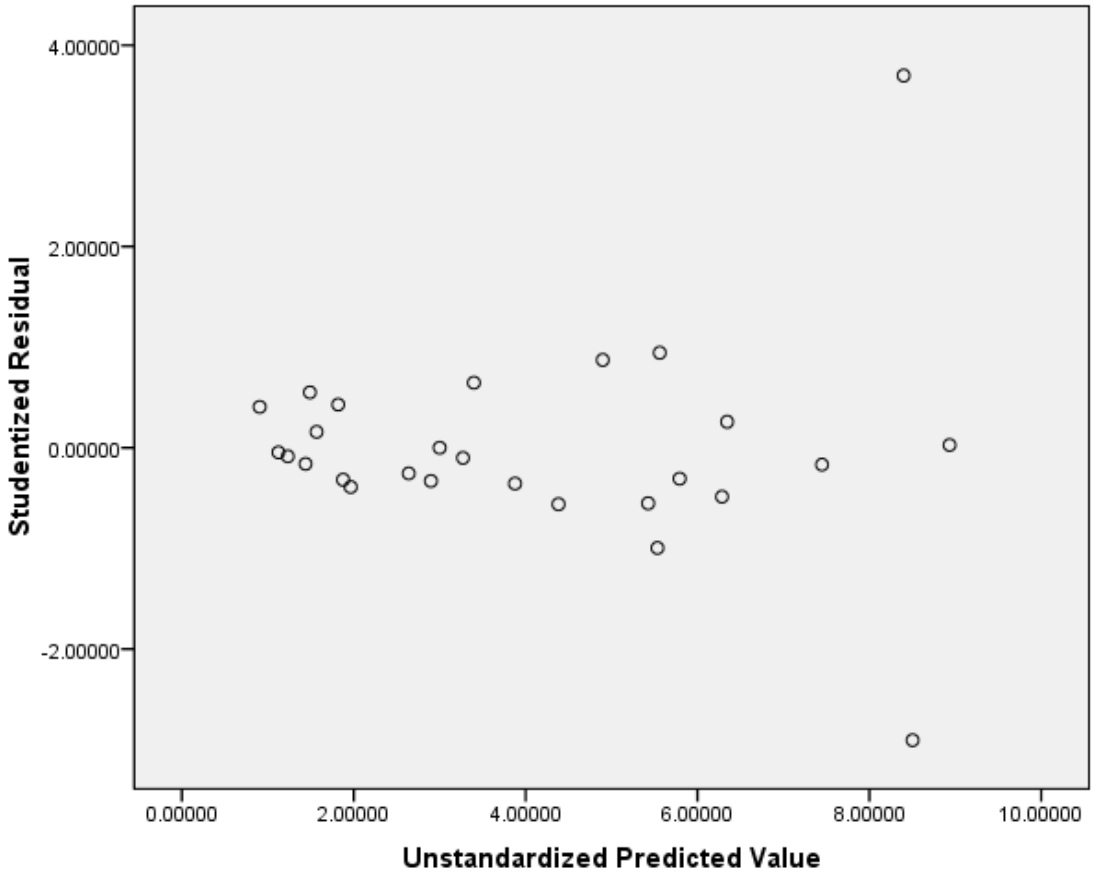

Figure 54. Scatterplot of Studentized Residuals to Unstandardized Predicted Values

Figures: Research Question Four With Outliers Removed

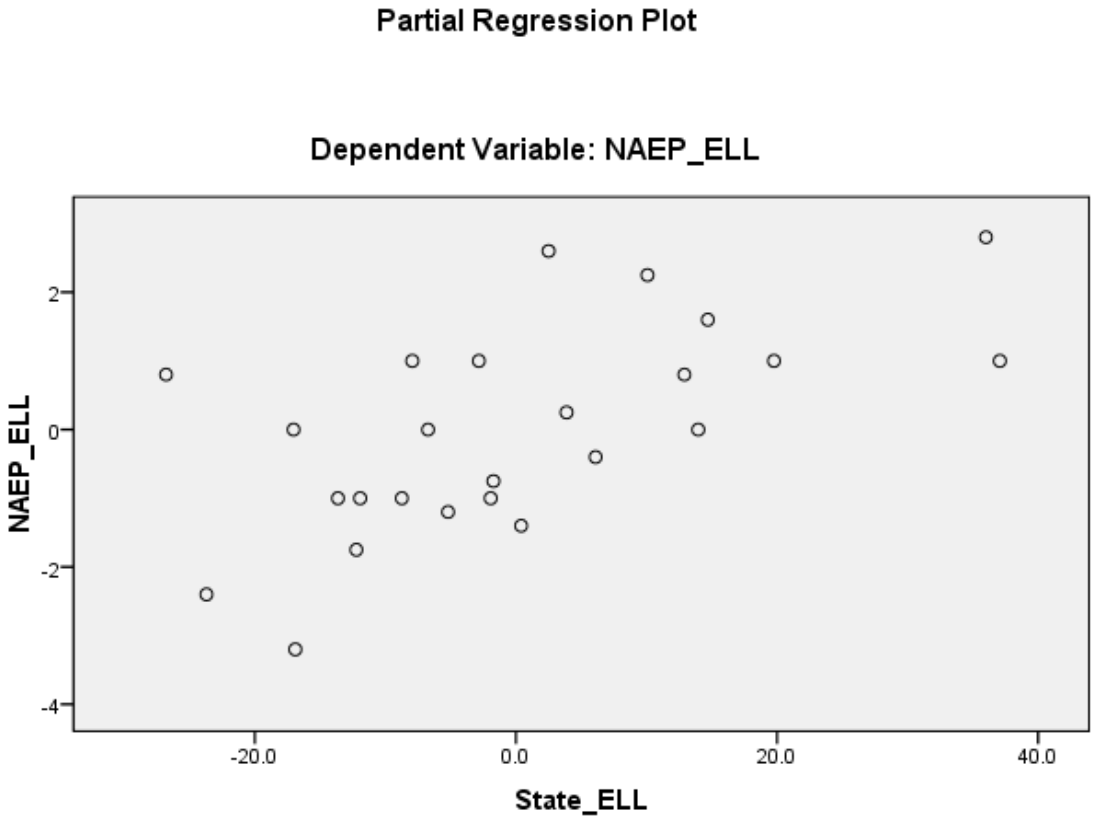

Figure 55. Partial Regression Plot of 2009 NAEP to State ELL Percent Proficient With Outliers Removed

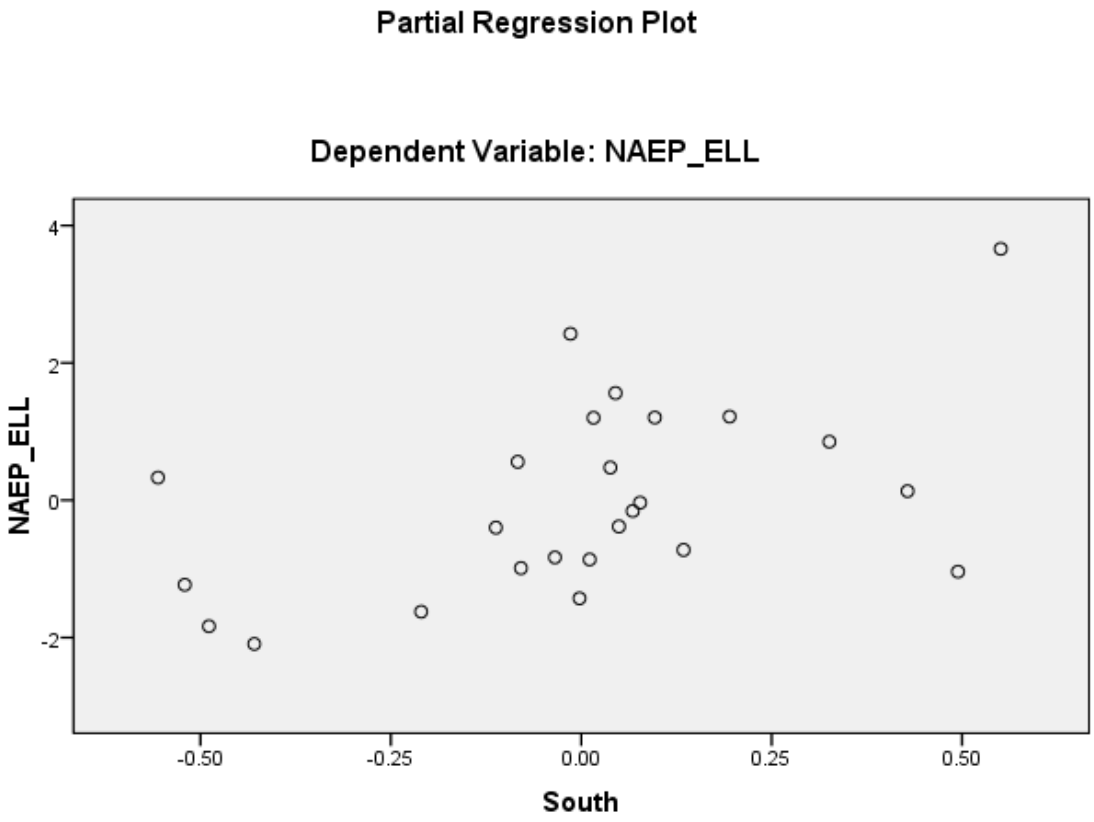

Figure 56. Partial Regression Plot of 2009 NAEP and State ELL Percent Proficient in the South With Outliers Removed

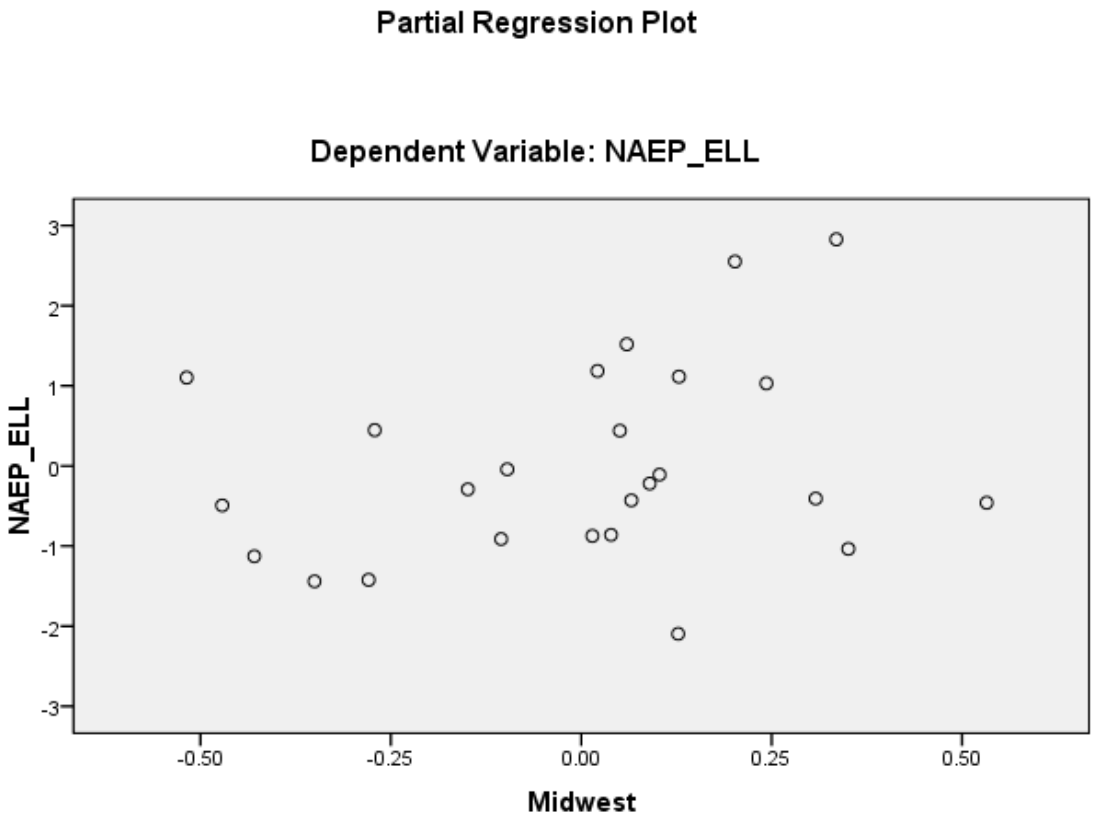

Figure 57. Partial Regression Plot of 2009 NAEP and State ELL Percent Proficient in the Midwest With Outliers Removed

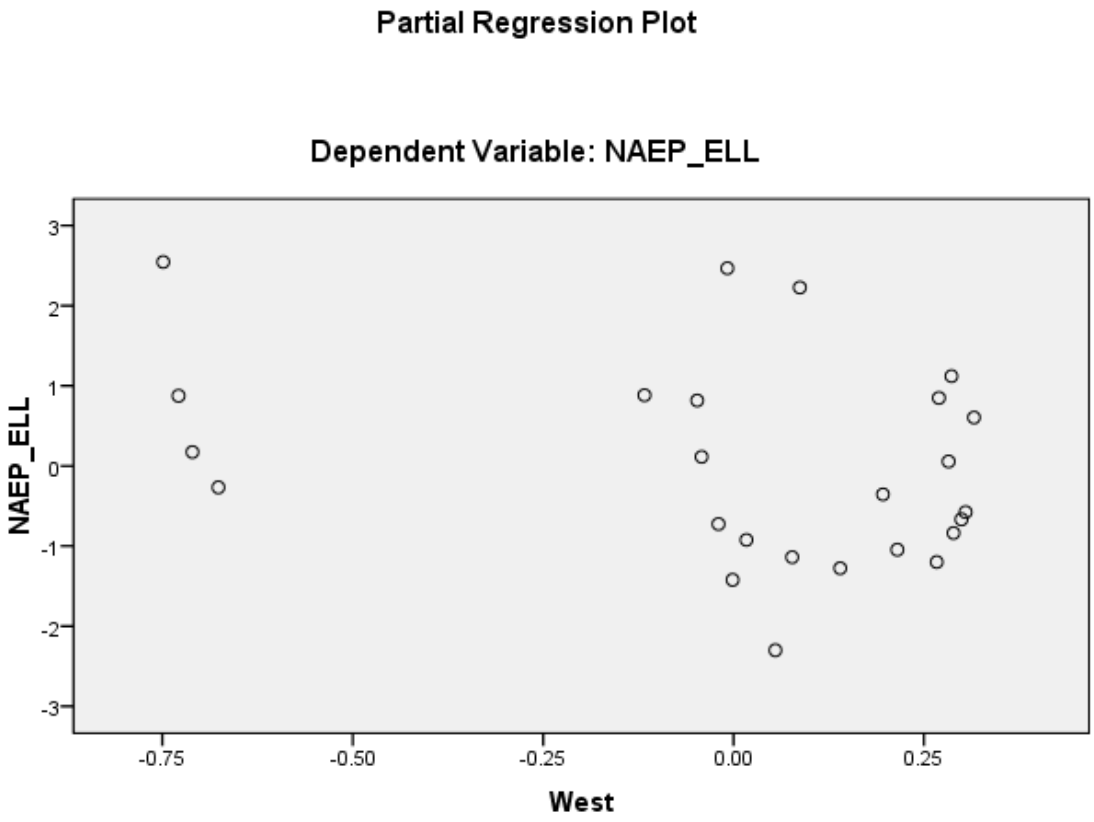

Figure 58. Partial Regression Plot of 2009 NAEP and State ELL Percent Proficient in the West With Outliers Removed

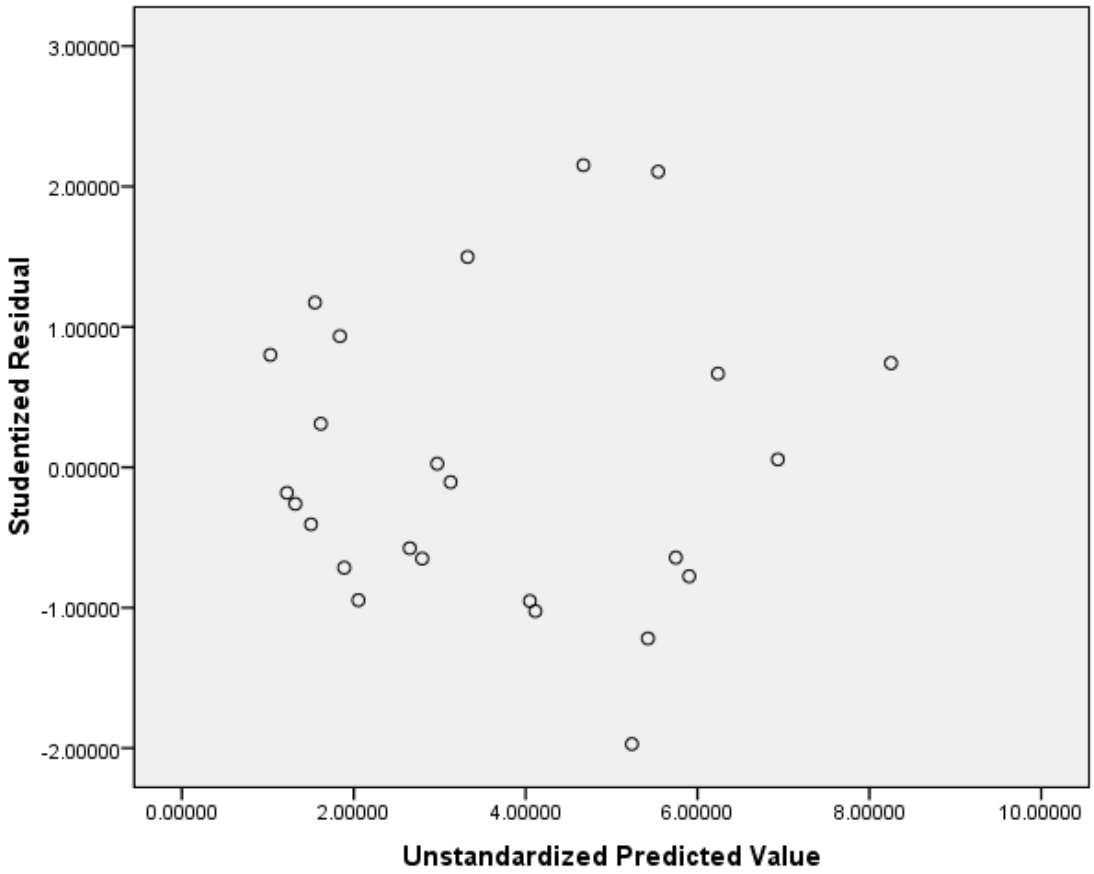

Figure 59. Scatterplot of Studentized Residuals to Unstandardized Predicted Values With Outliers Removed

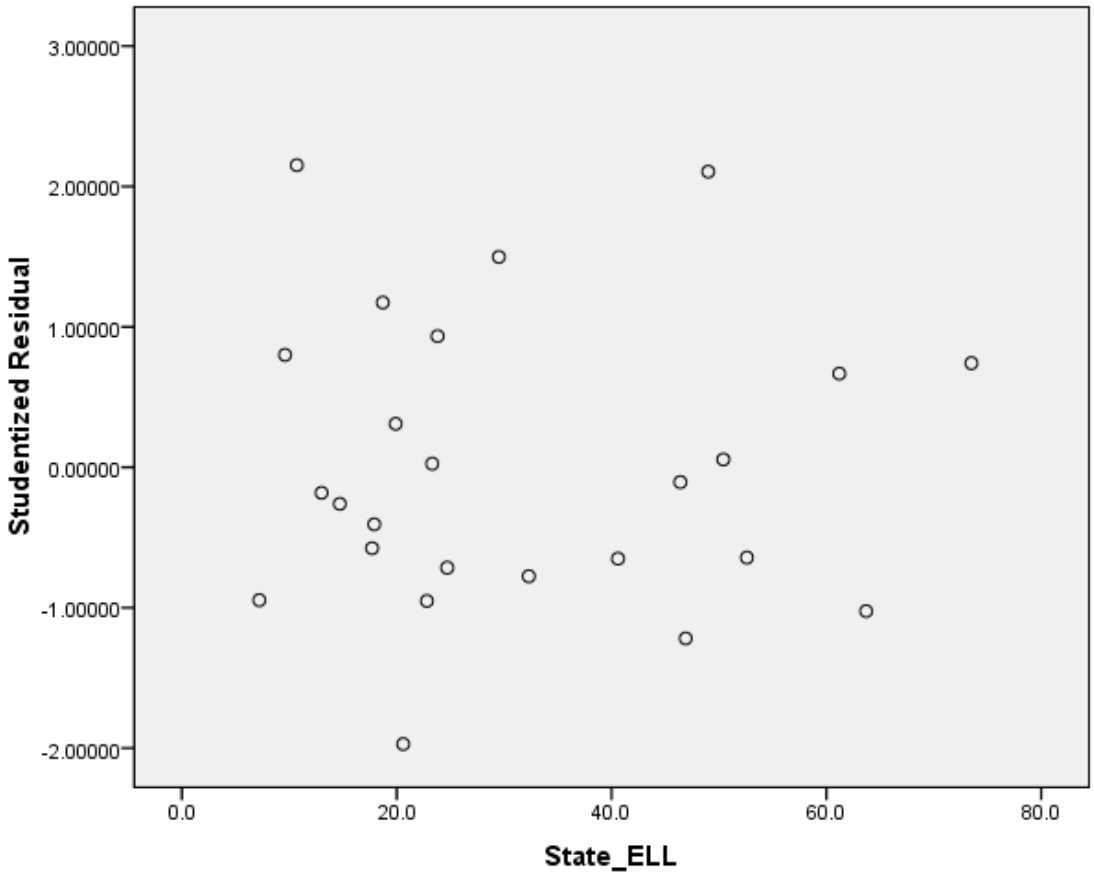

Figure 60. Scatterplot of Studentized Residuals to 2009 State ELL Percent Proficient With Outliers Removed

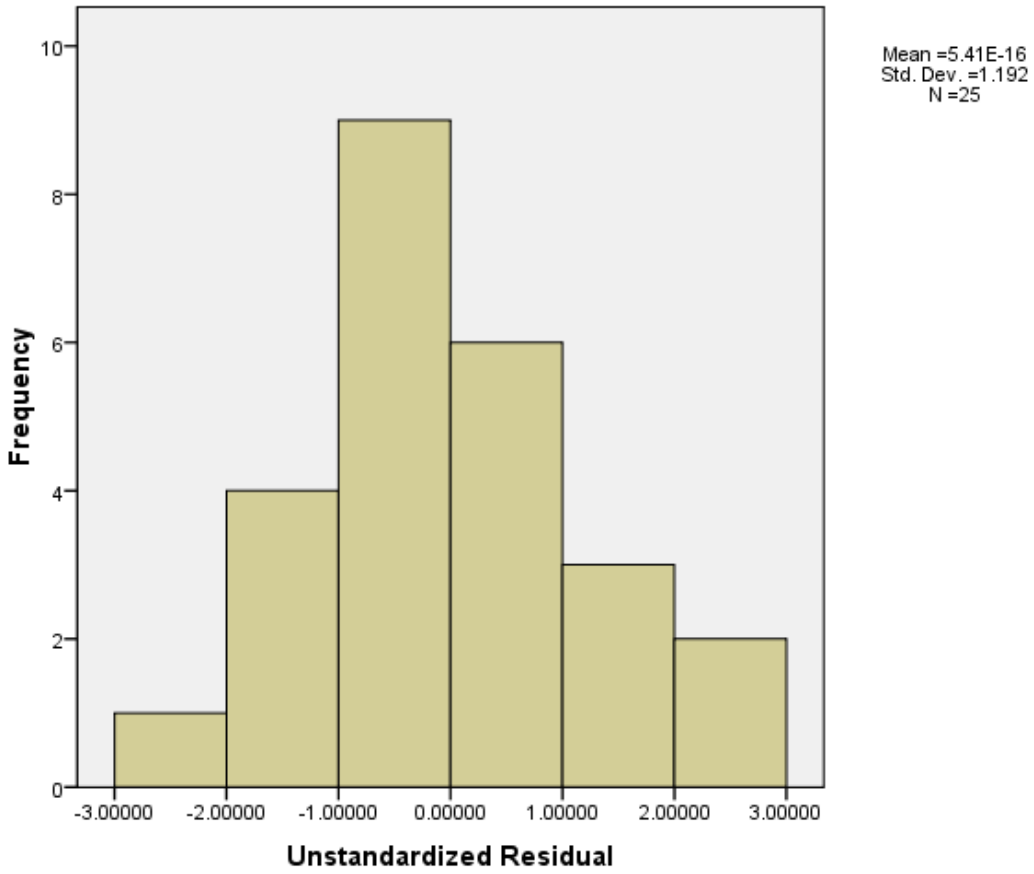

Figure 61. Histogram of Unstandardized Residuals With Outliers Removed

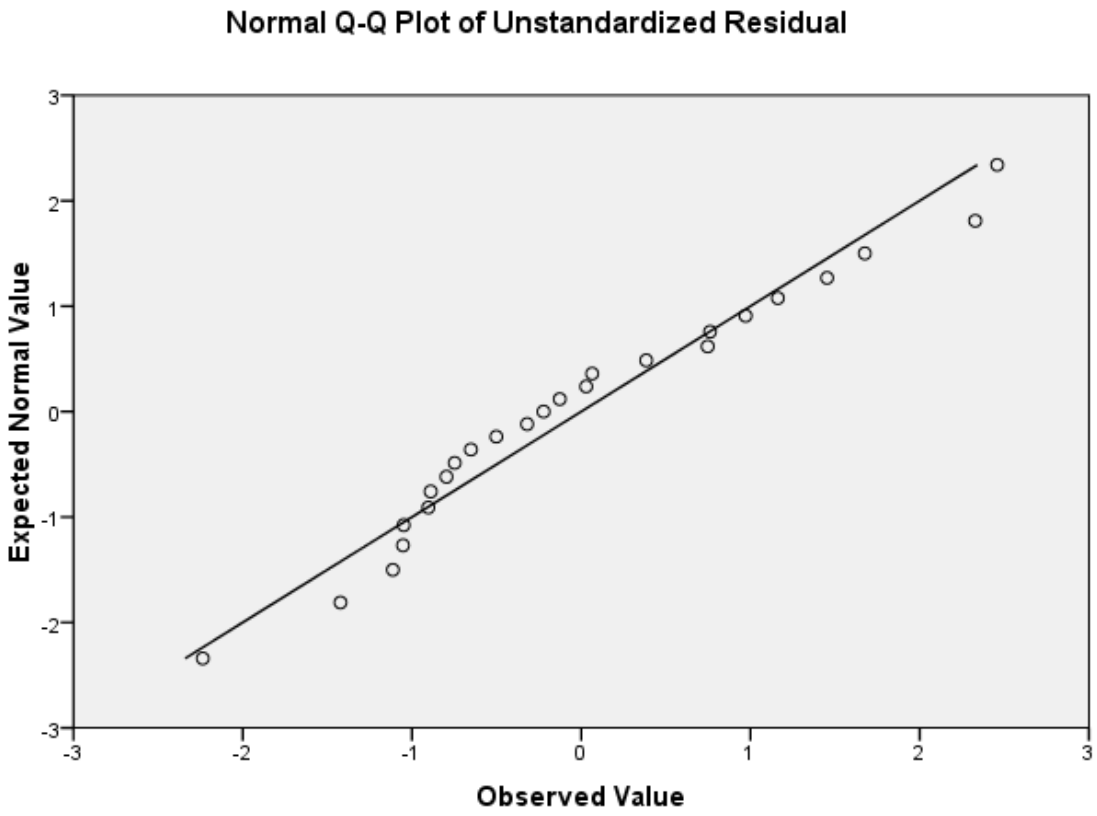

Figure 62. Q-Q Plot of Unstandardized Residuals With Outliers Removed

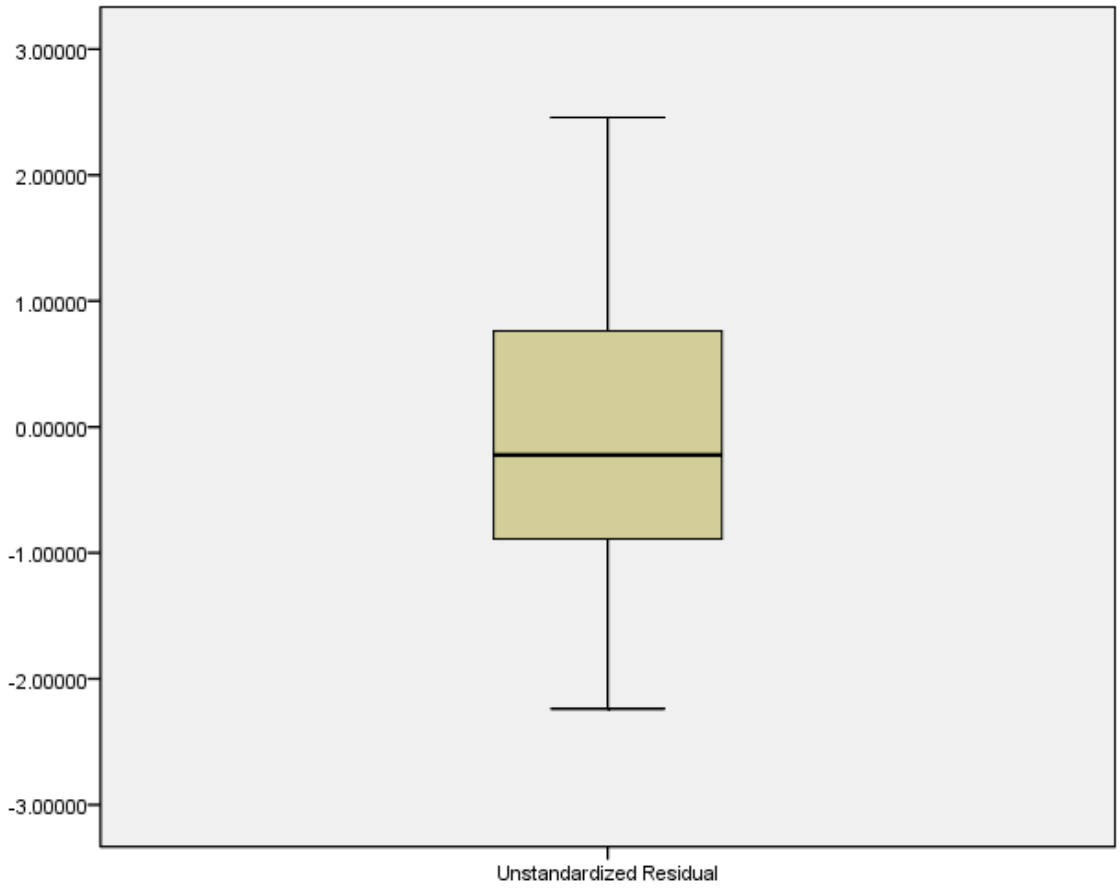

Figure 63. Boxplot of Unstandardized Residuals With Outliers Removed

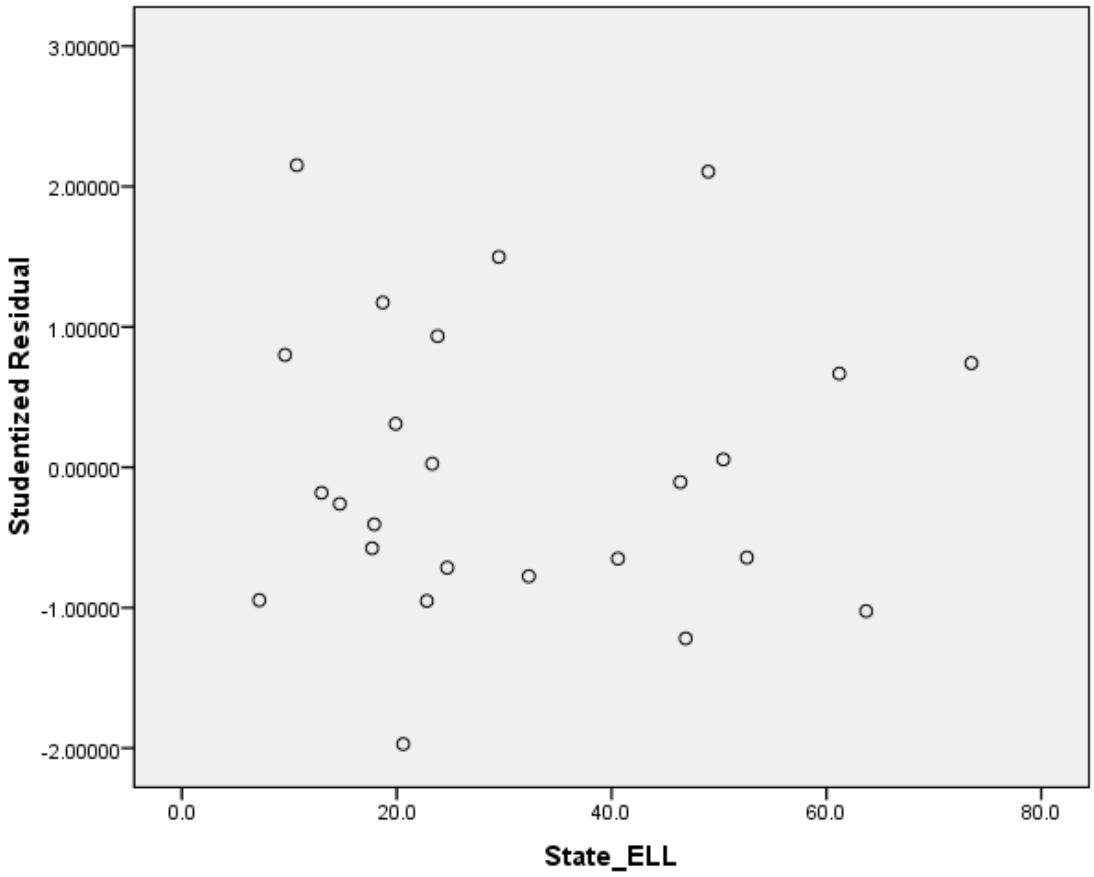

Figure 64. Scatterplot of Studentized Residuals to 2009 State ELL Percent Proficient With Outliers Removed

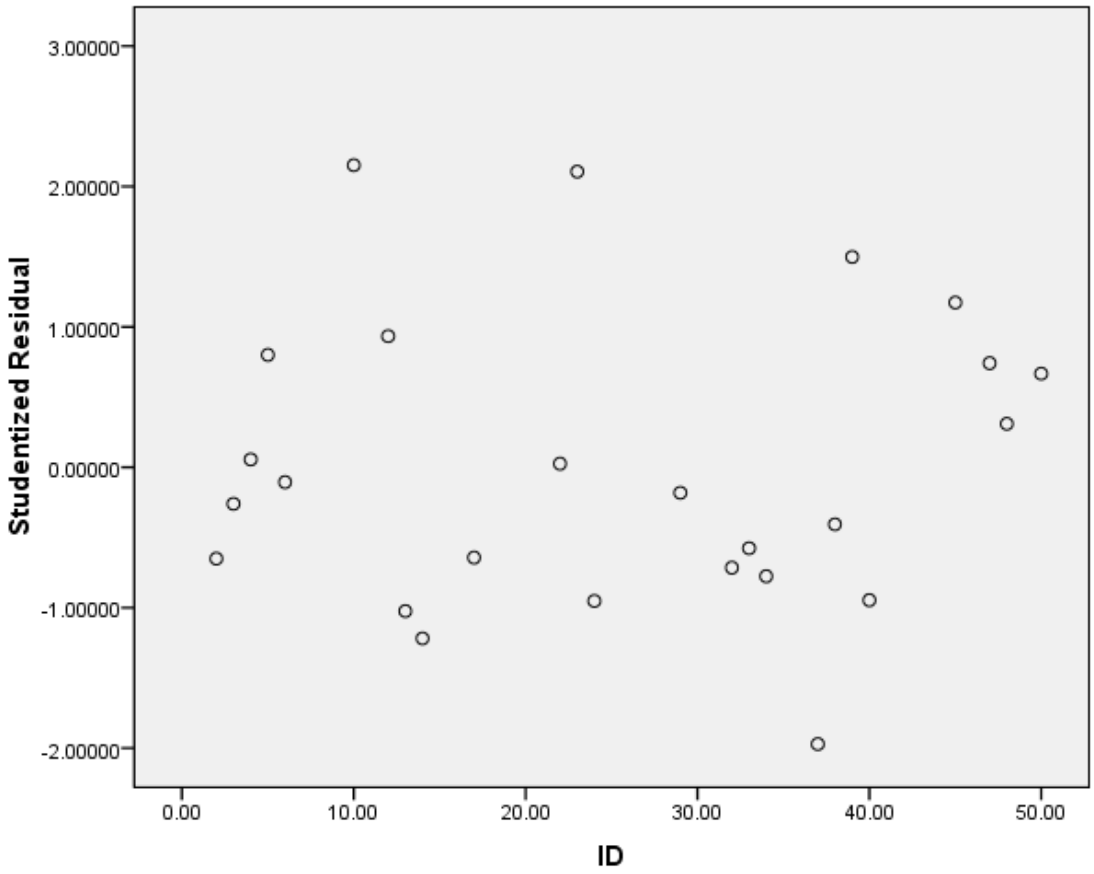

Figure 65. Scatterplot of Studentized Residuals to Case Number With Outliers Removed

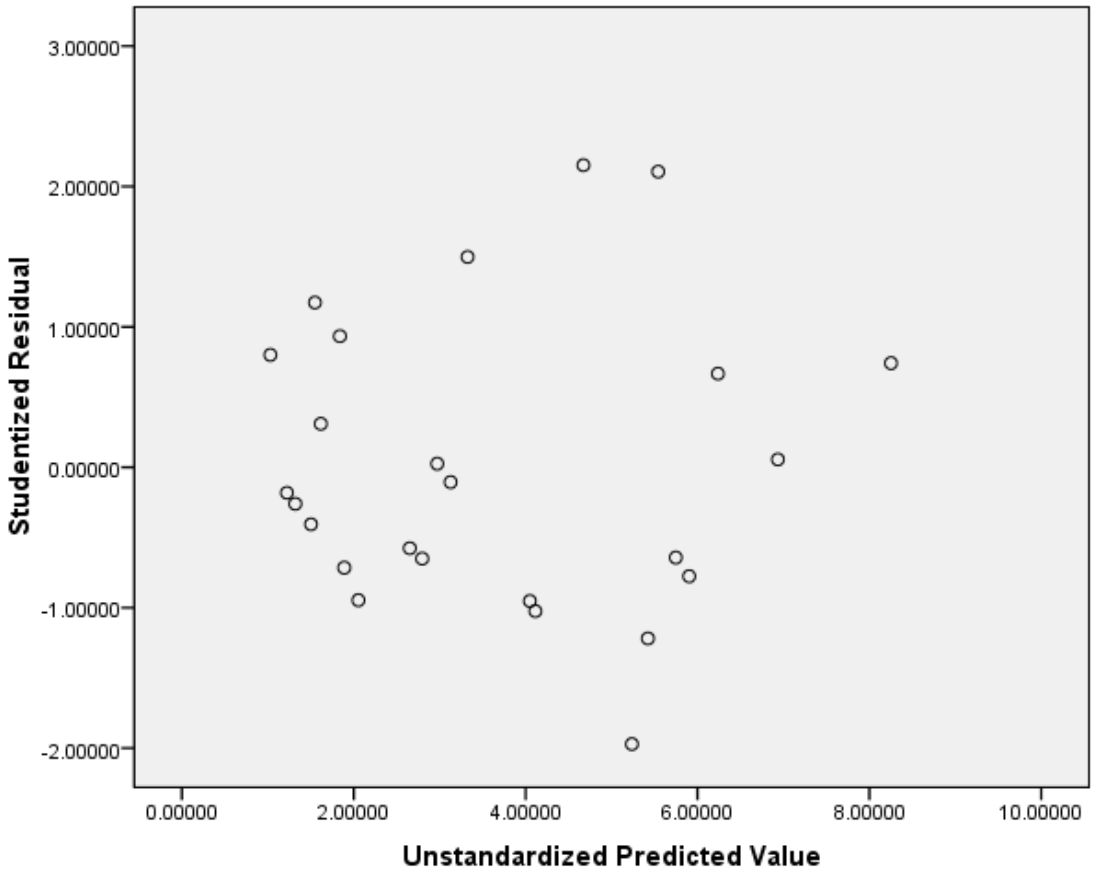

Figure 66. Scatterplot of Studentized Residuals to Unstandardized Predicted Values With Outliers Removed

Figures: Research Question Five

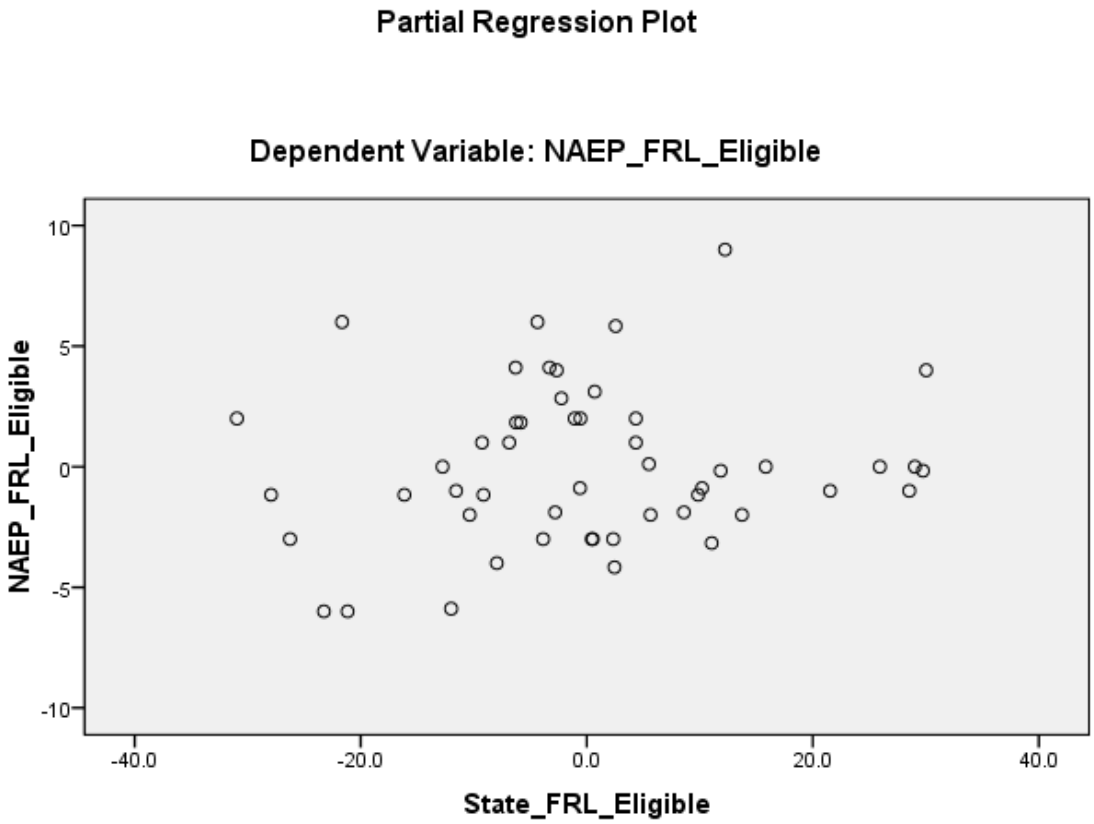

Figure 67. Partial Regression Plot of 2009 NAEP to State FRL Percent Proficient

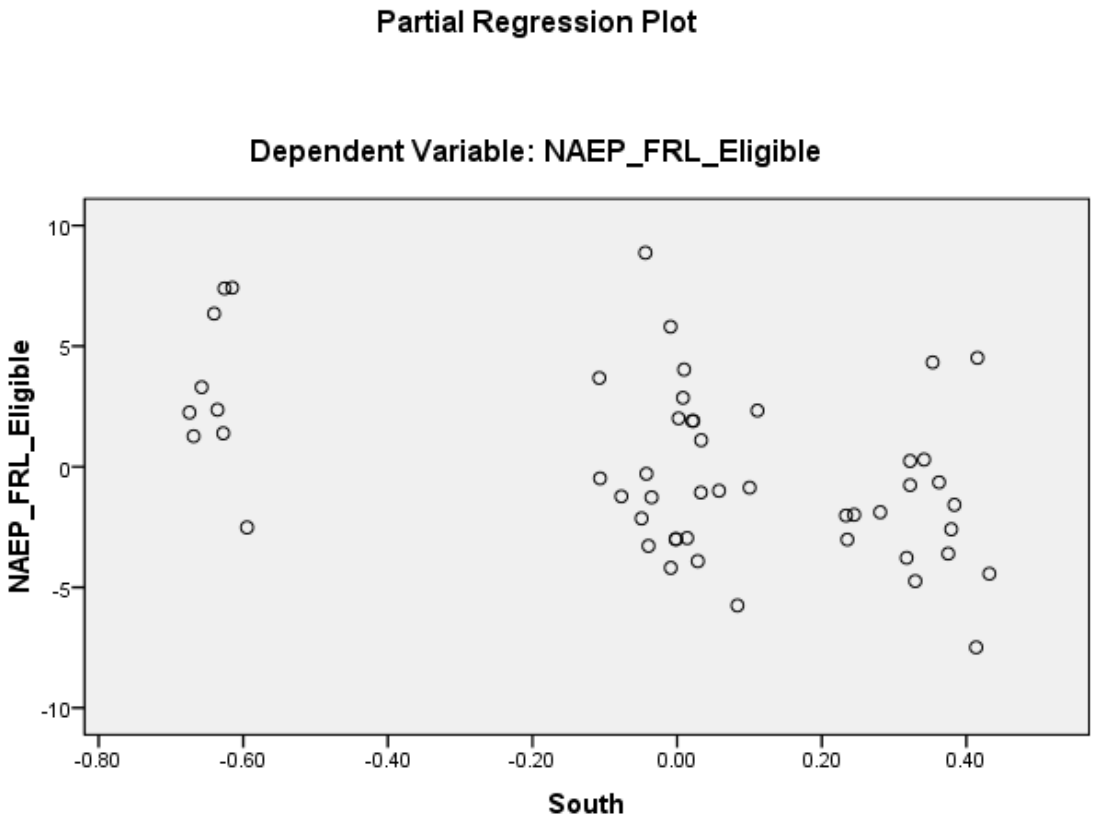

Figure 68. Partial Regression Plot of 2009 NAEP to State FRL Percent Proficient in the South

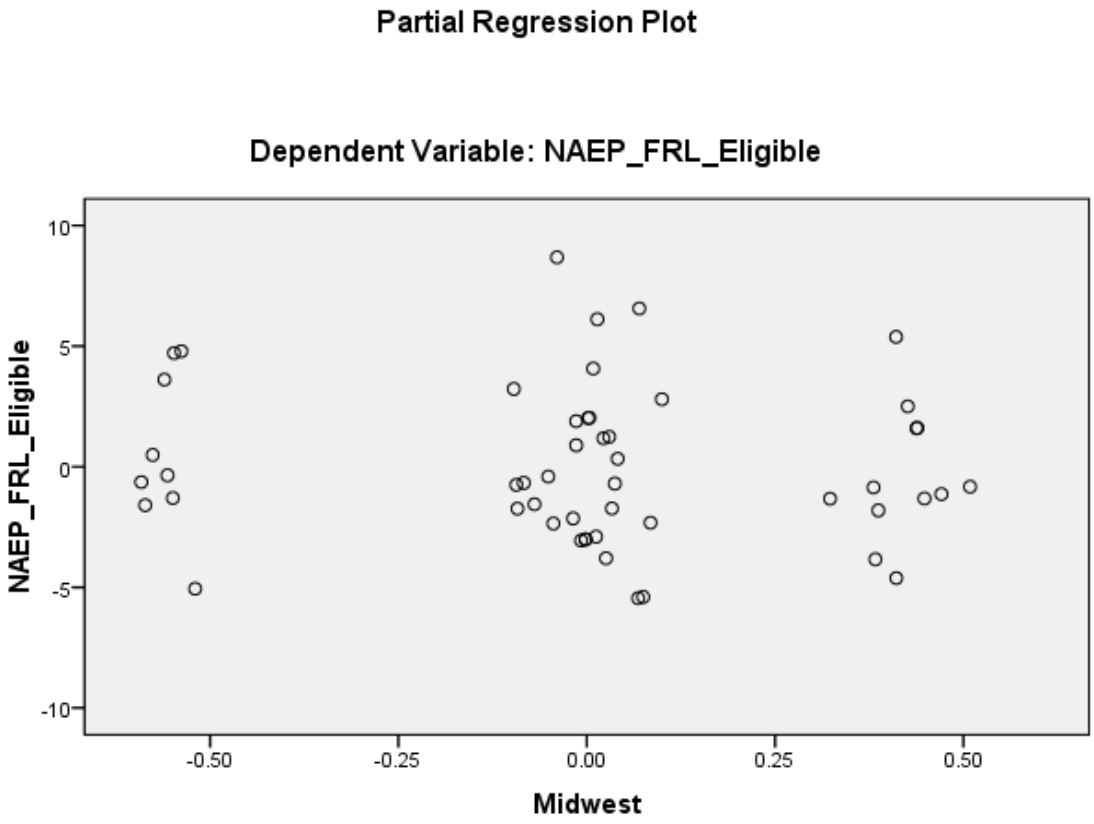

Figure 69. Partial Regression Plot of 2009 NAEP to State FRL Percent Proficient in the Midwest

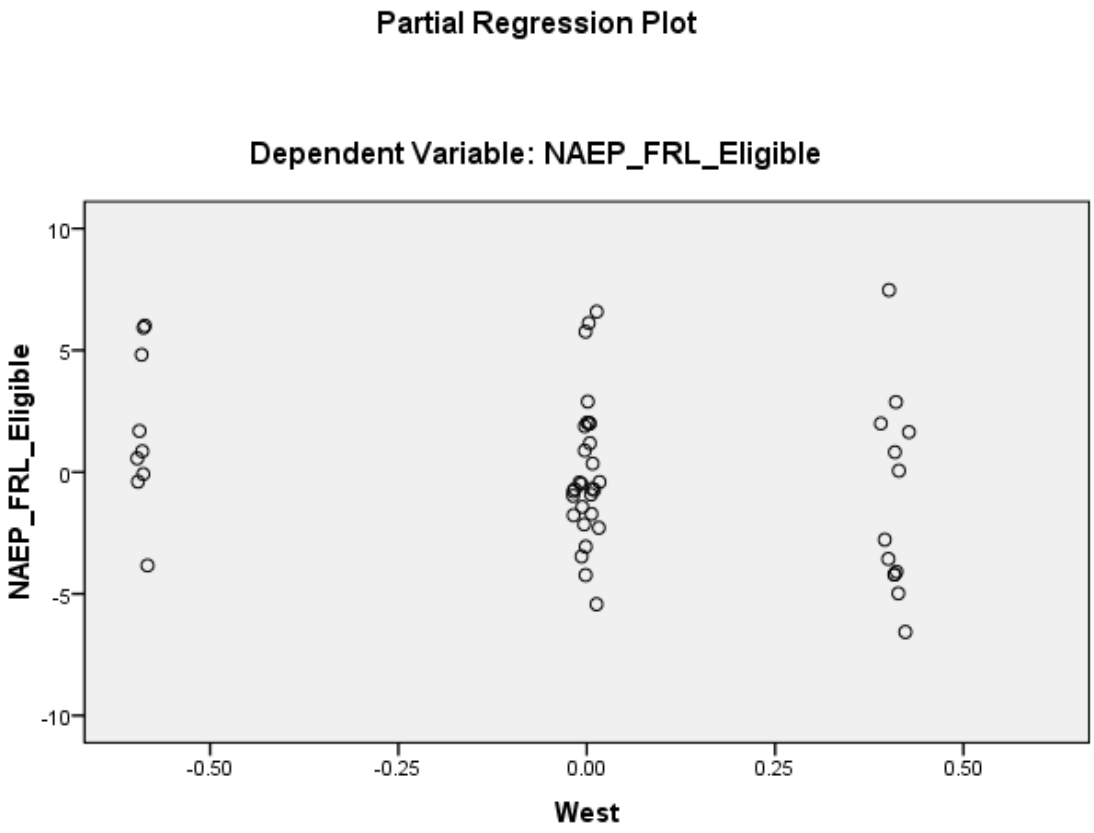

Figure 70. Partial Regression Plot of 2009 NAEP to State FRL Percent Proficient in the West

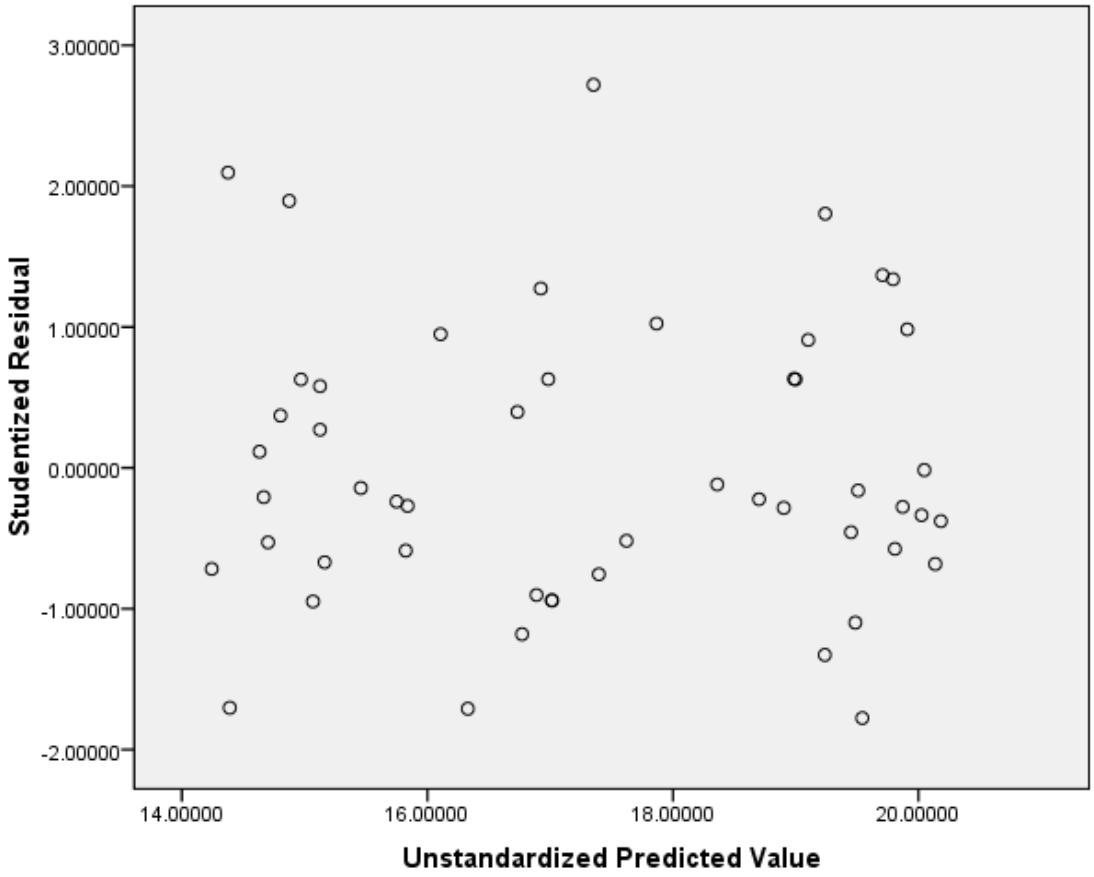

Figure 71. Scatterplot of Studentized Residuals to Unstandardized Predicted Values

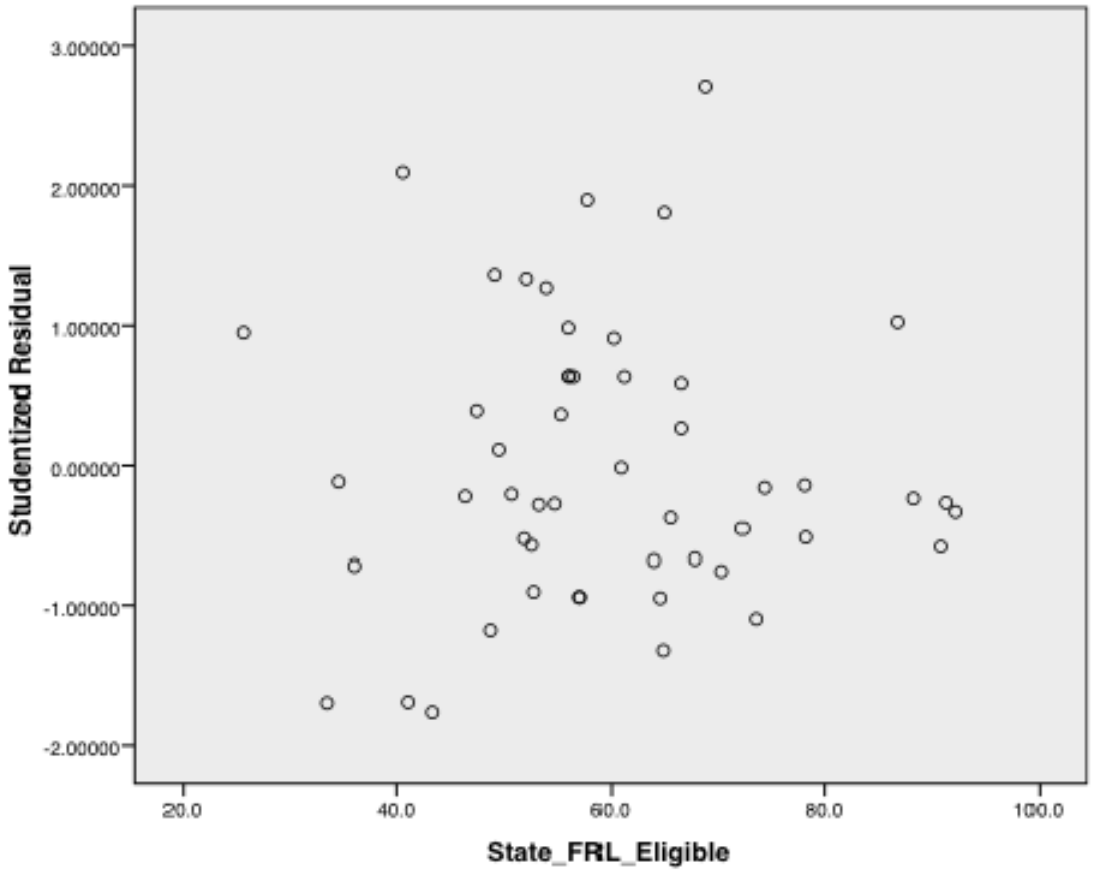

Figure 72. Scatterplot of Studentized Residuals to 2009 Low SES Percent Proficient

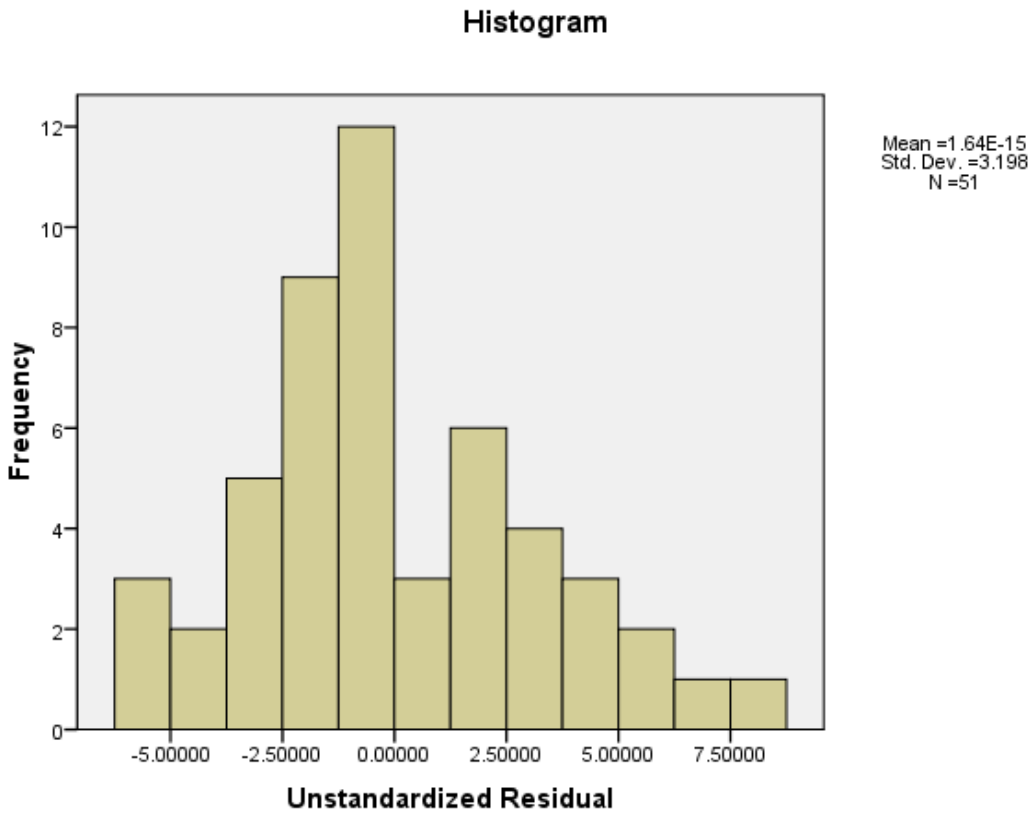

Figure 73. Histogram of Unstandardized Residuals

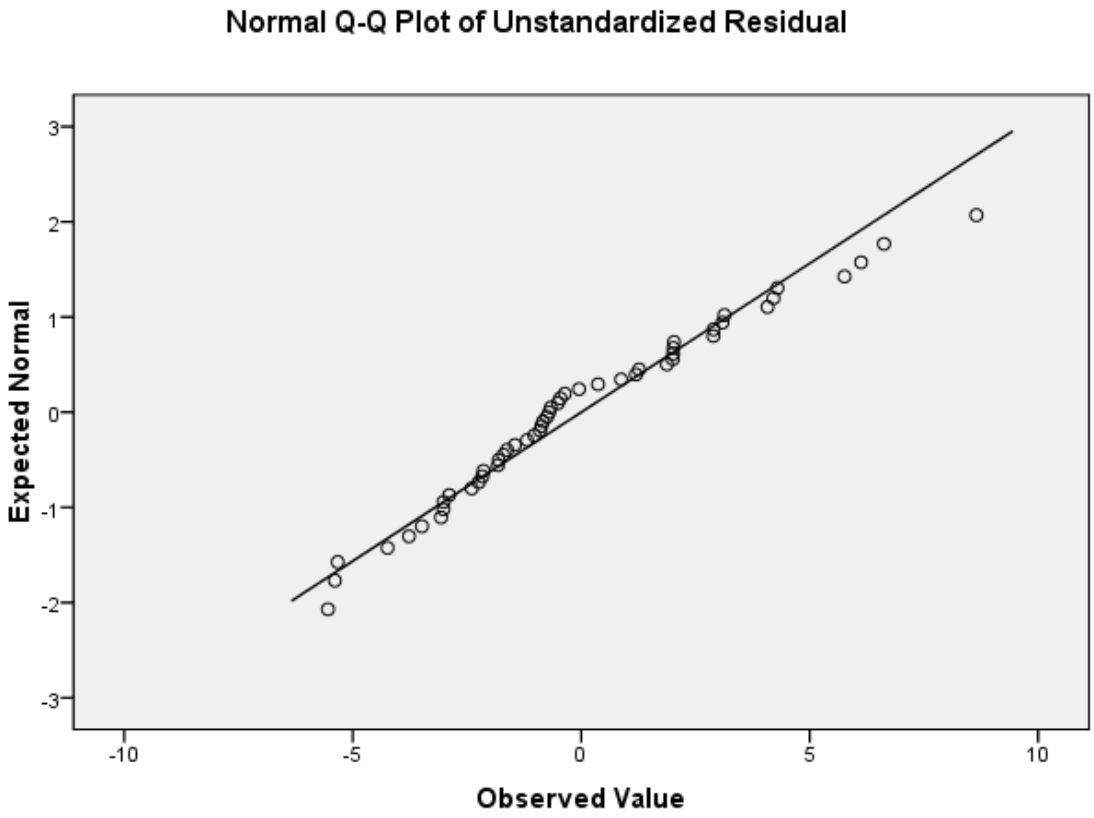

Figure 74. Q-Q Plot of Unstandardized Residuals

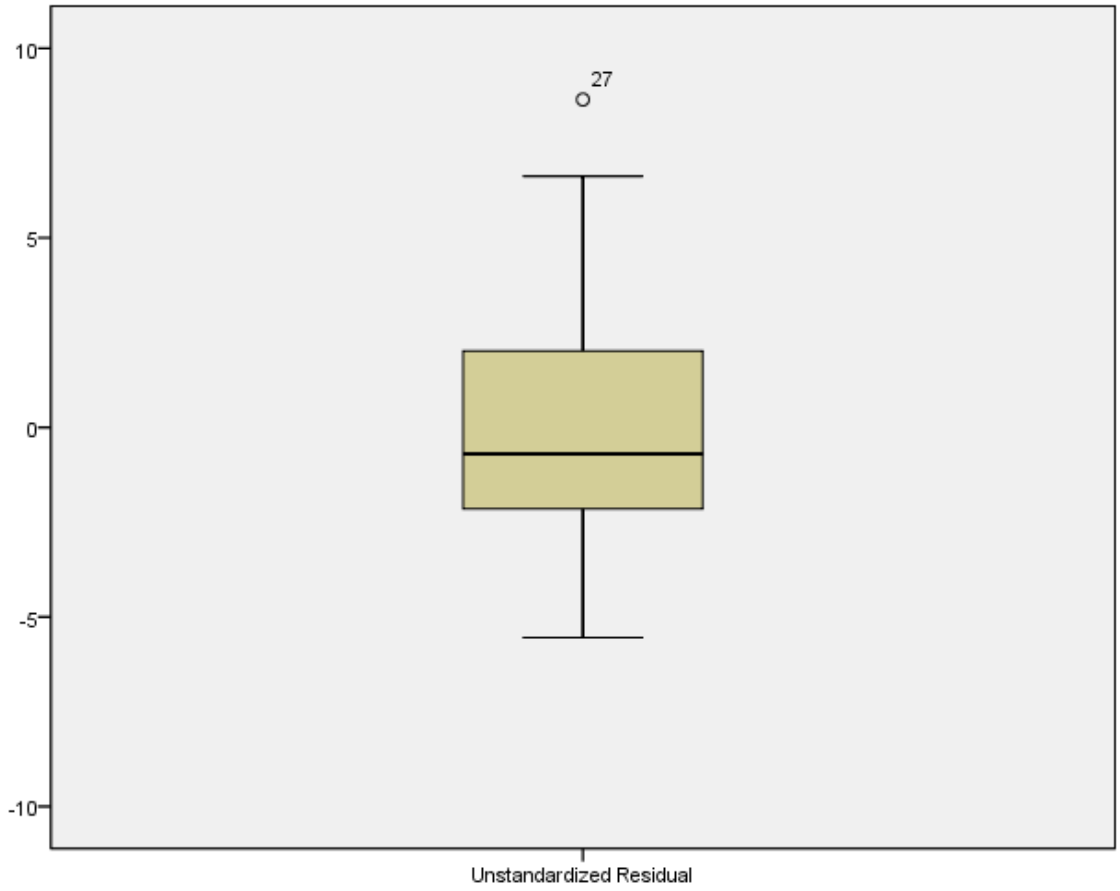

Figure 75. Boxplot of Unstandardized Residuals

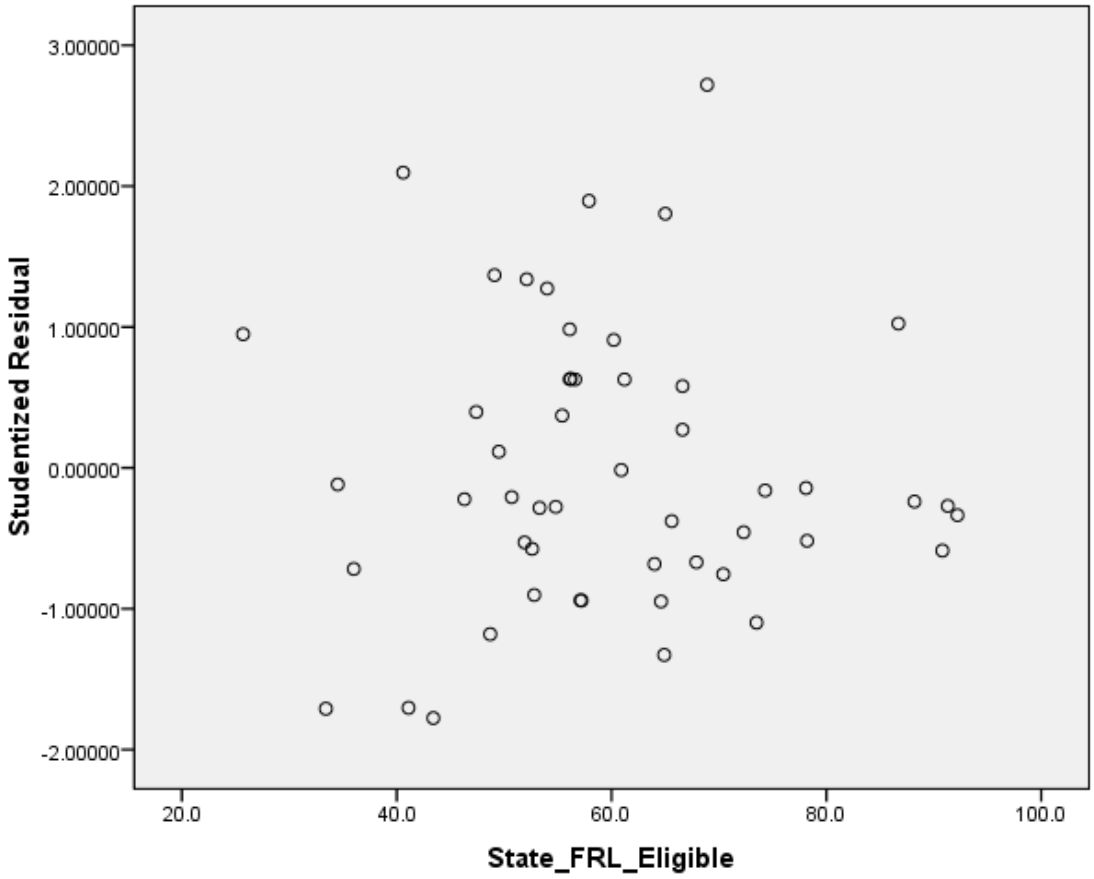

Figure 76. Scatterplot of Studentized Residuals to 2009 State Low SES Percent Proficient

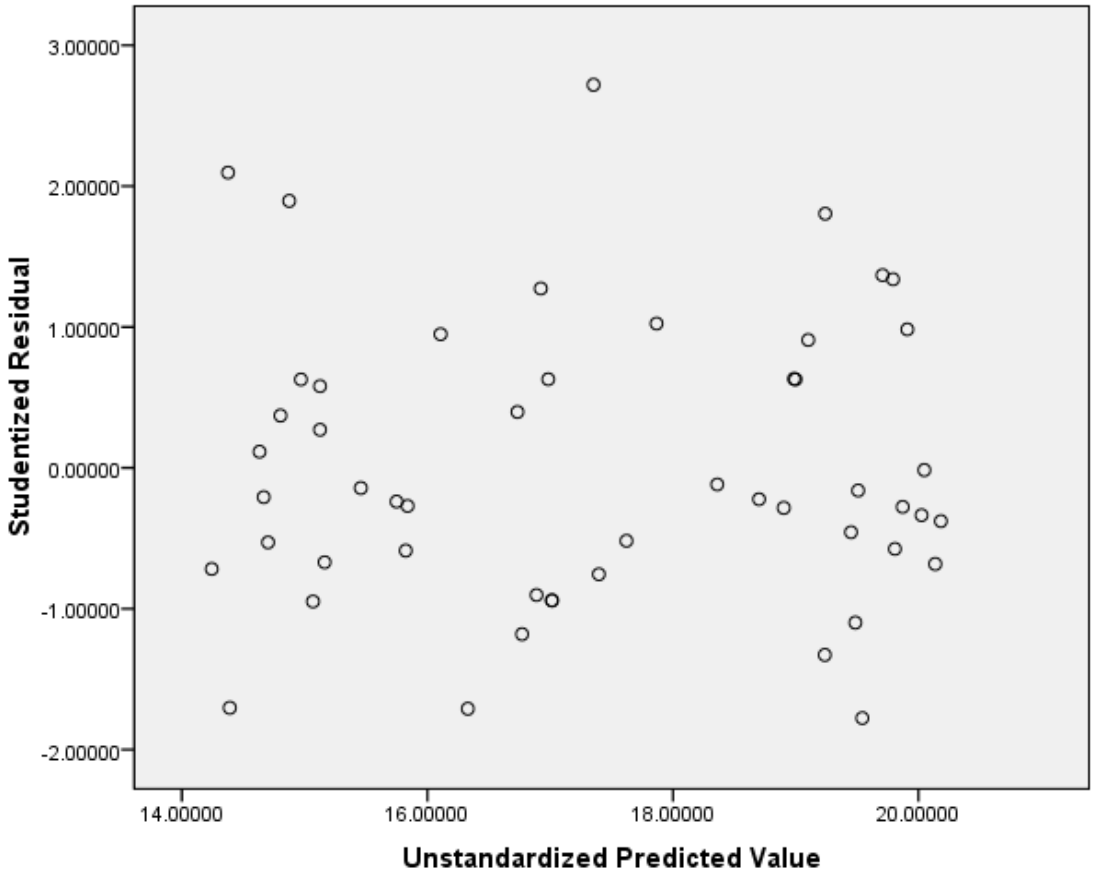

Figure 77. Scatterplot of Studentized Residuals to Unstandardized Predicted Values

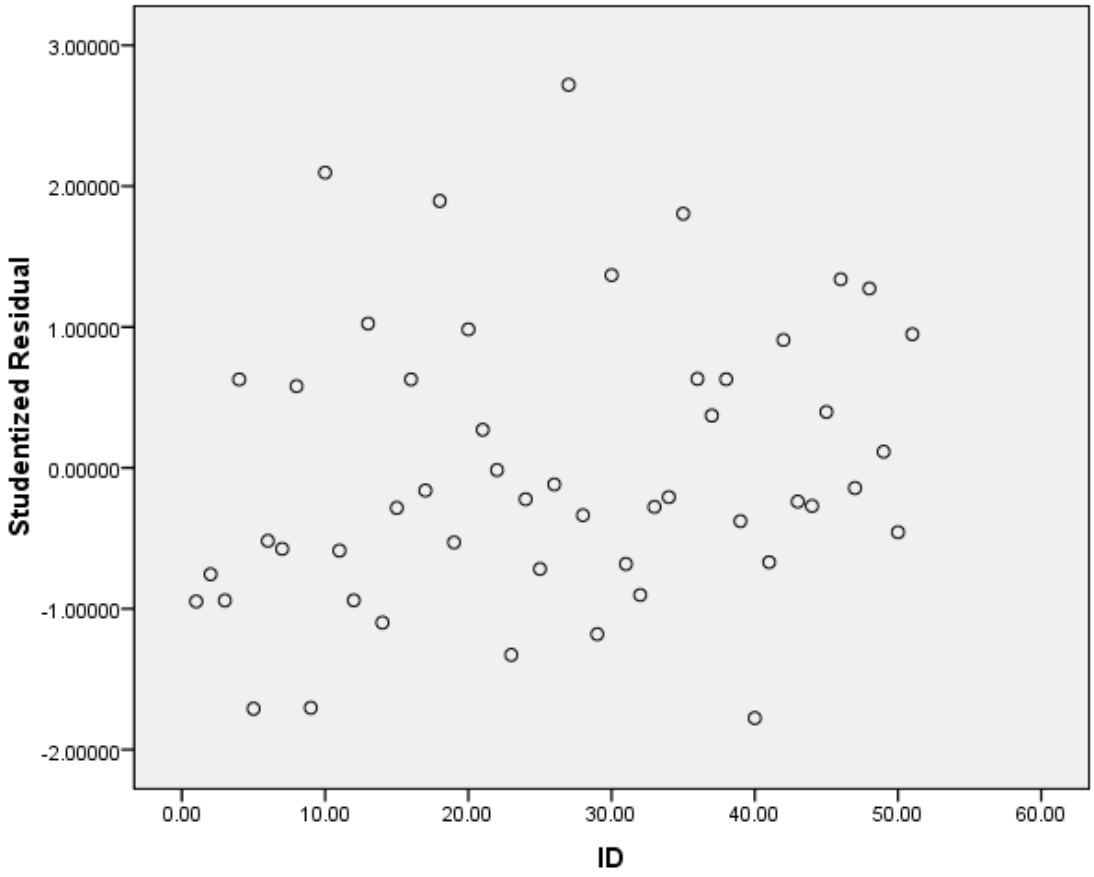

Figure 78. Scatterplot of Studentized Residuals to Case Number

Figures: Research Question Five With Outliers Removed

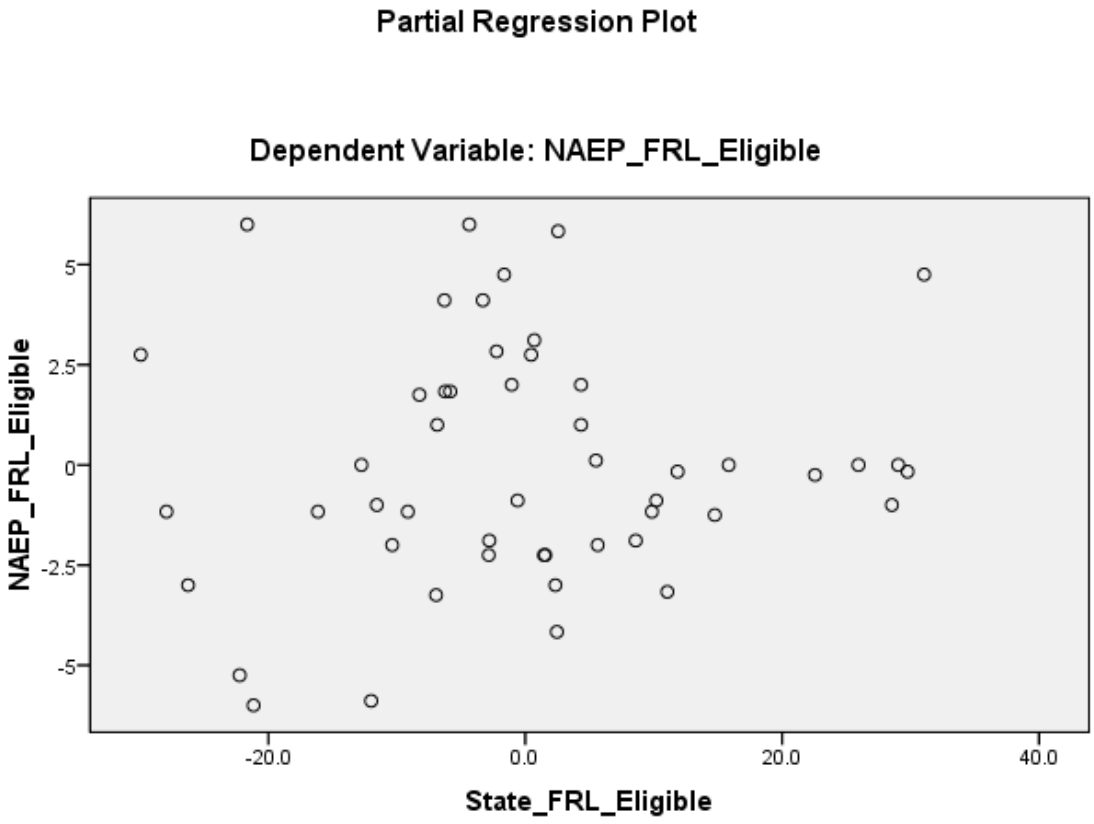

Figure 79. Partial Regression Plot of 2009 NAEP to State FRL Percent Proficient With Outliers Removed

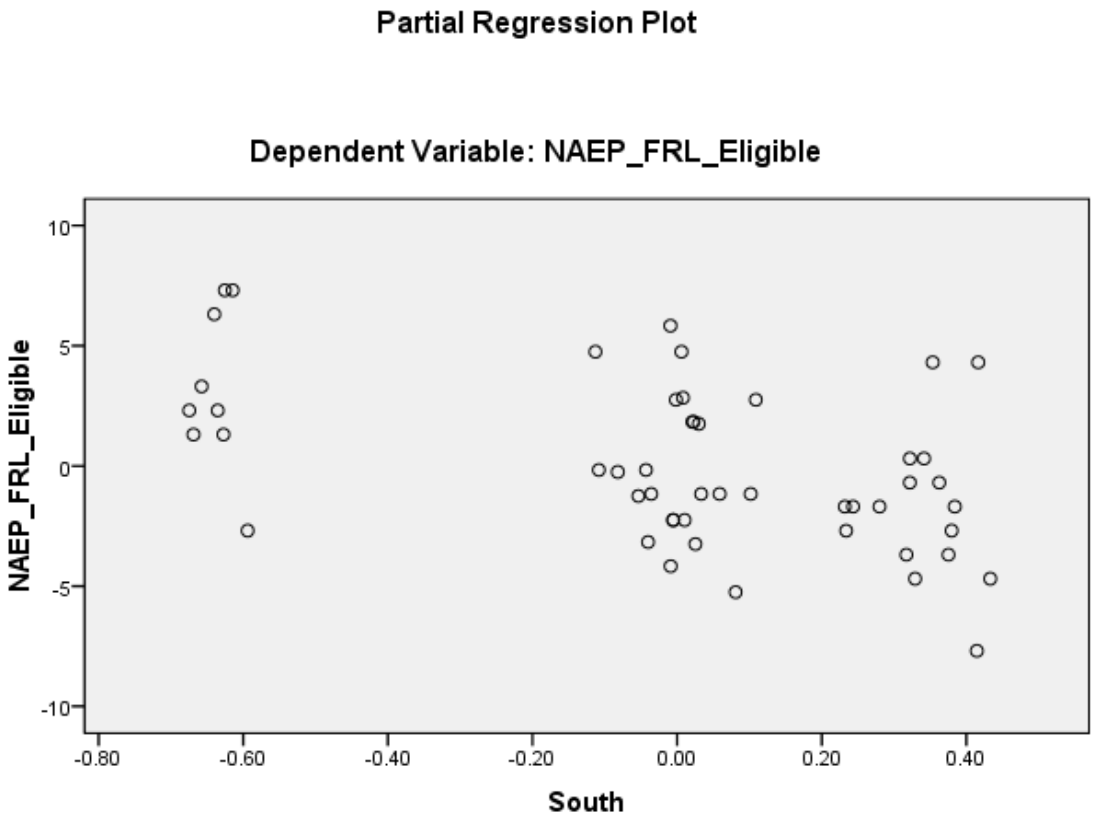

Figure 80. Partial Regression Plot of 2009 NAEP to State FRL Percent Proficient in the South With Outliers Removed

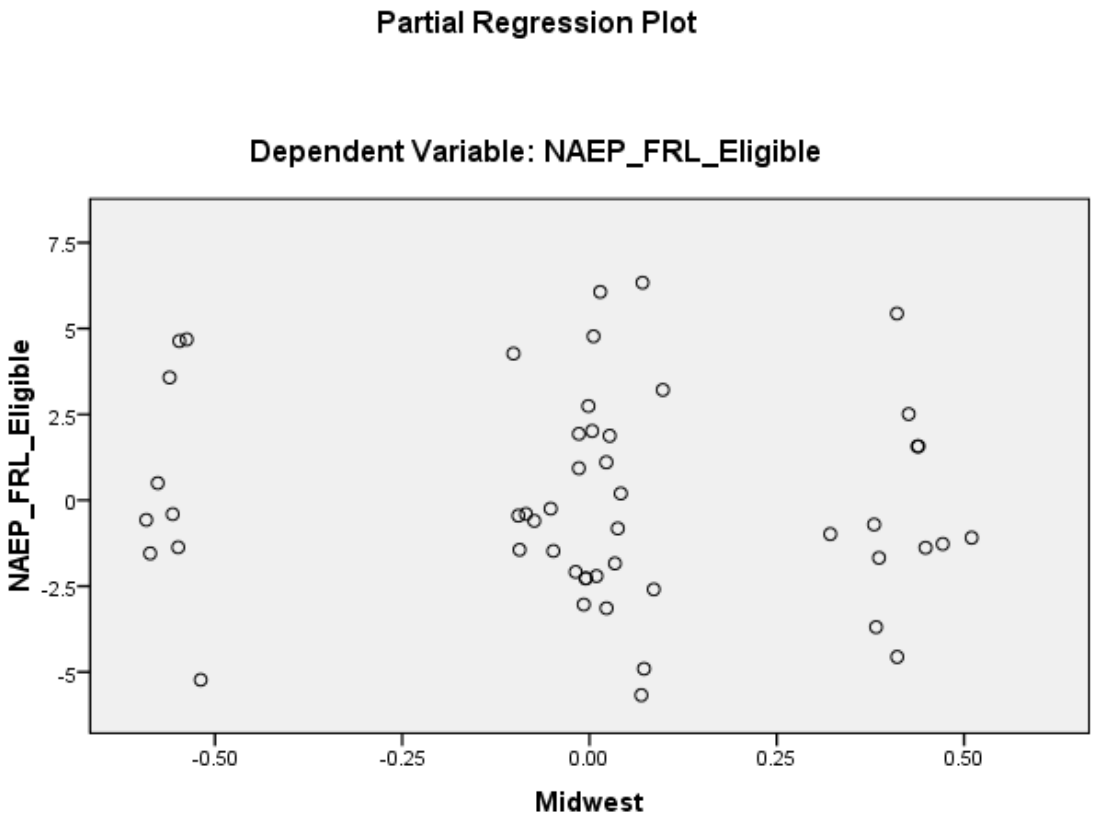

Figure 81. Partial Regression Plot of 2009 NAEP to State FRL Percent Proficient in the Midwest With Outliers Removed

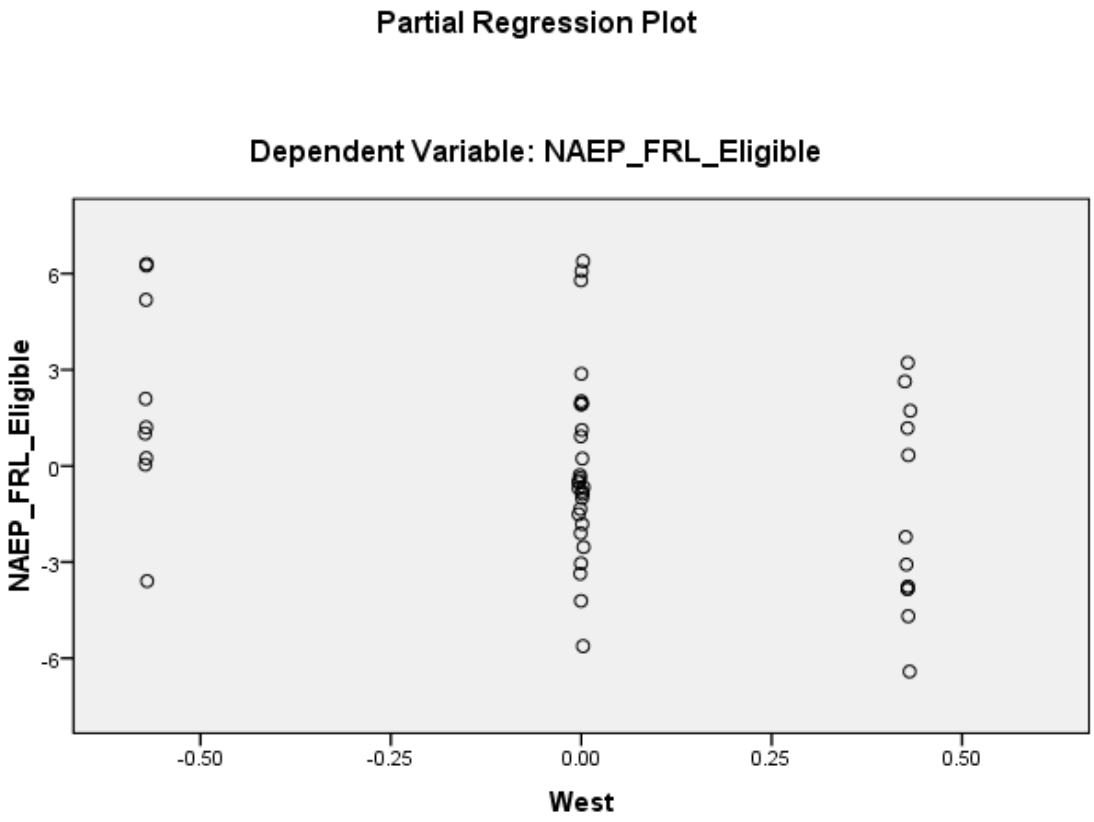

Figure 82. Partial Regression Plot of 2009 NAEP to State FRL Percent Proficient in the West With Outliers Removed

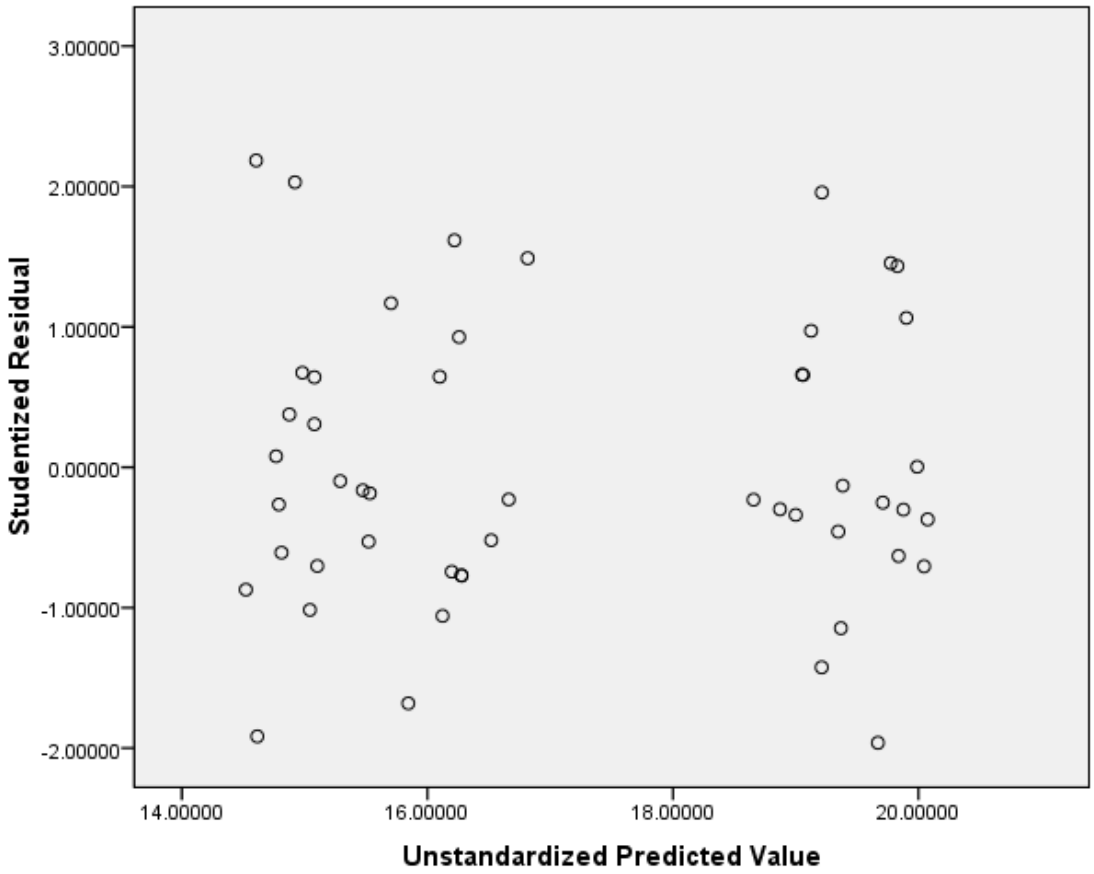

Figure 83. Scatterplot of Studentized Residuals to Unstandardized Predicted Values With Outliers Removed

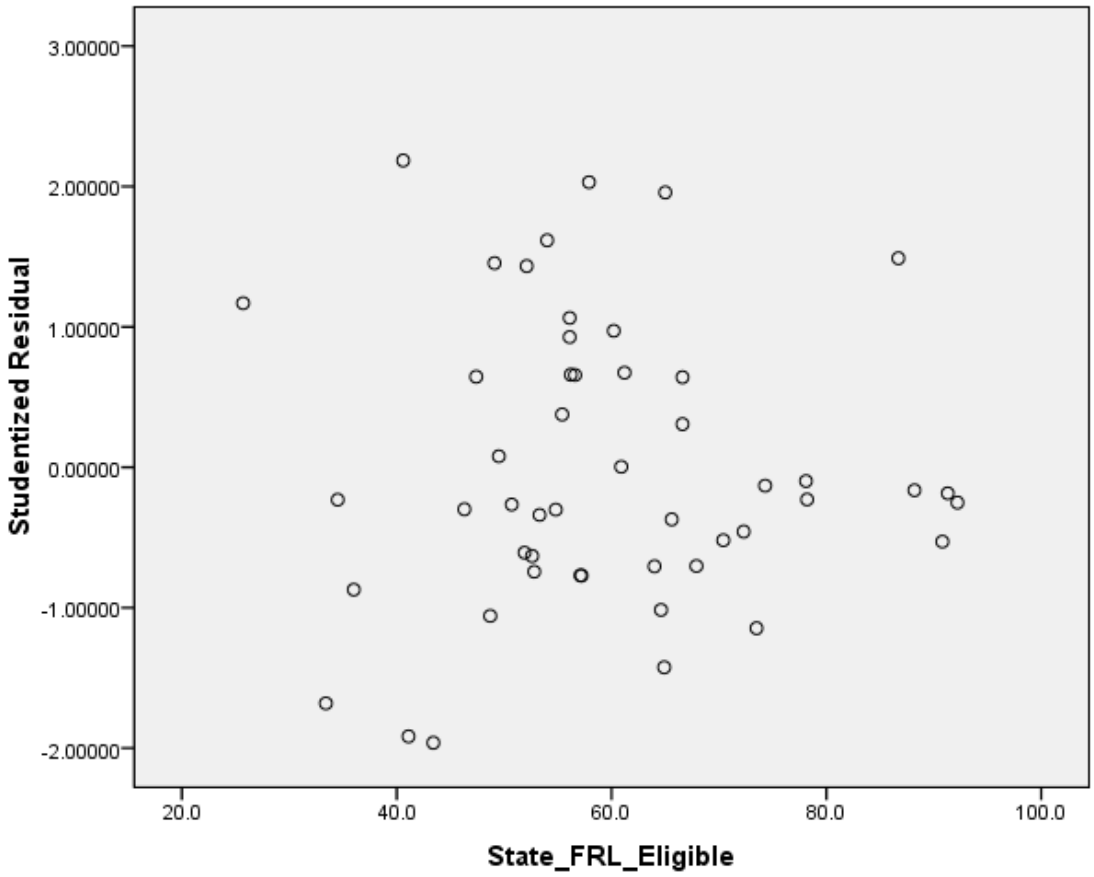

Figure 84. Scatterplot of Studentized Residuals to 2009 Low SES Percent Proficient With Outliers Removed

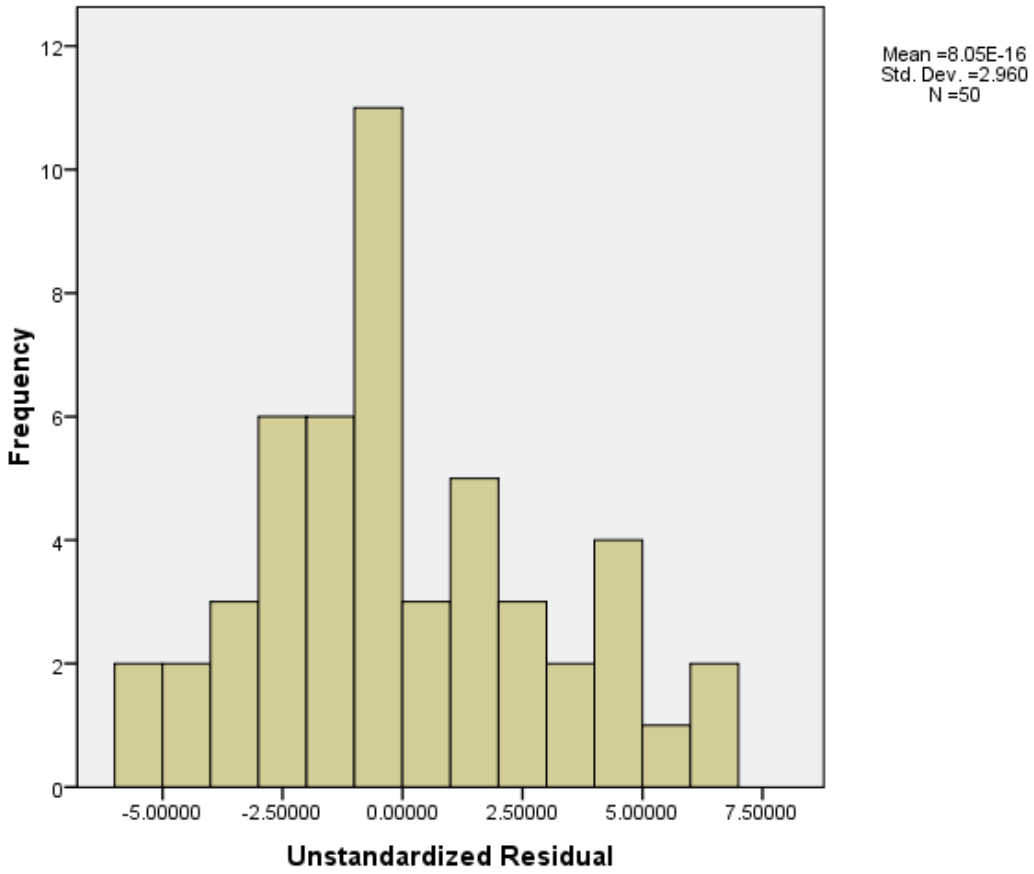

Figure 85. Histogram of Unstandardized Residuals With Outliers Removed

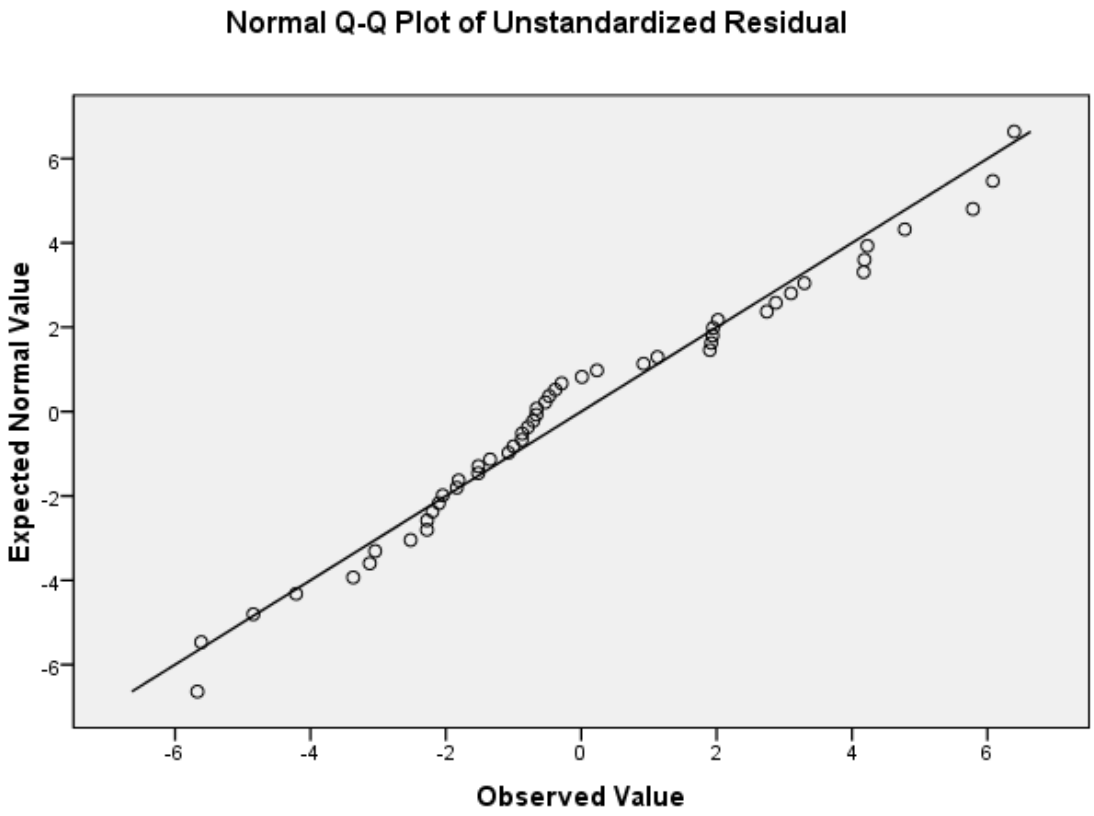

Figure 86. Q-Q Plot of Unstandardized Residuals With Outliers Removed

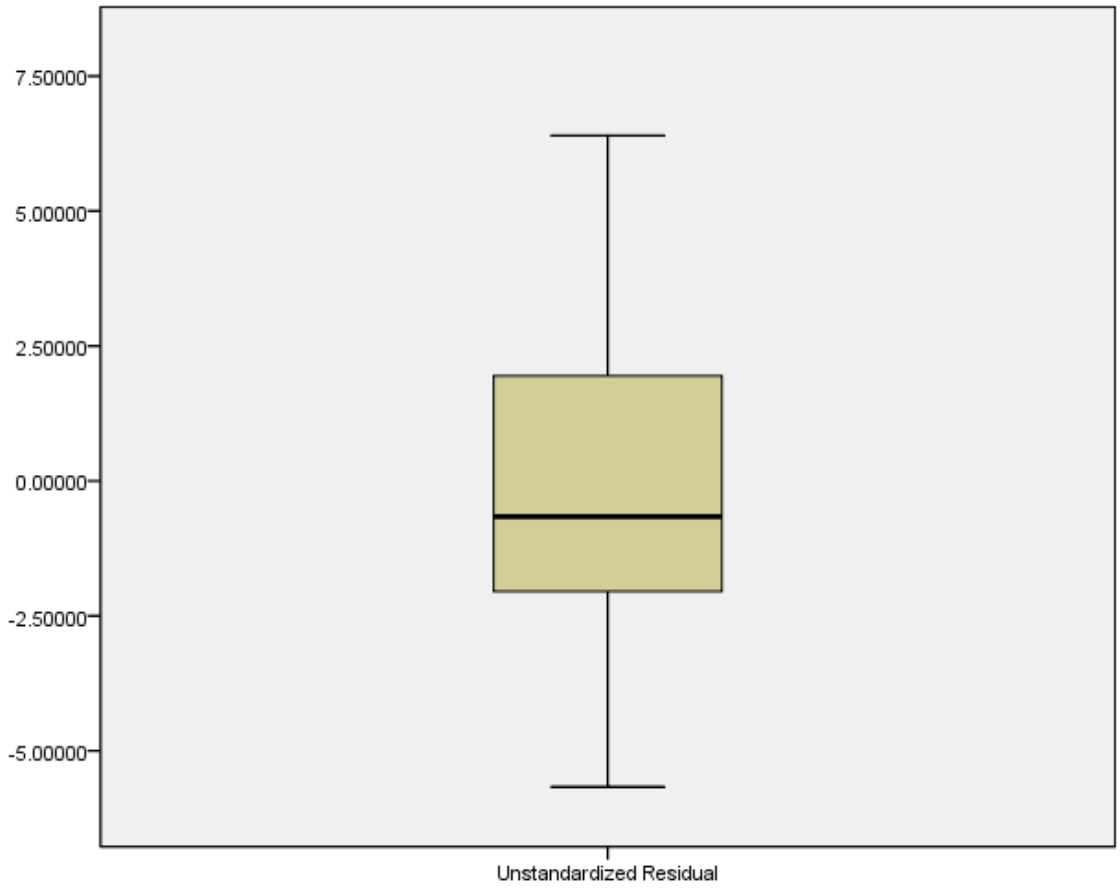

Figure 87. Boxplot of Unstandardized Residuals With Outliers Removed

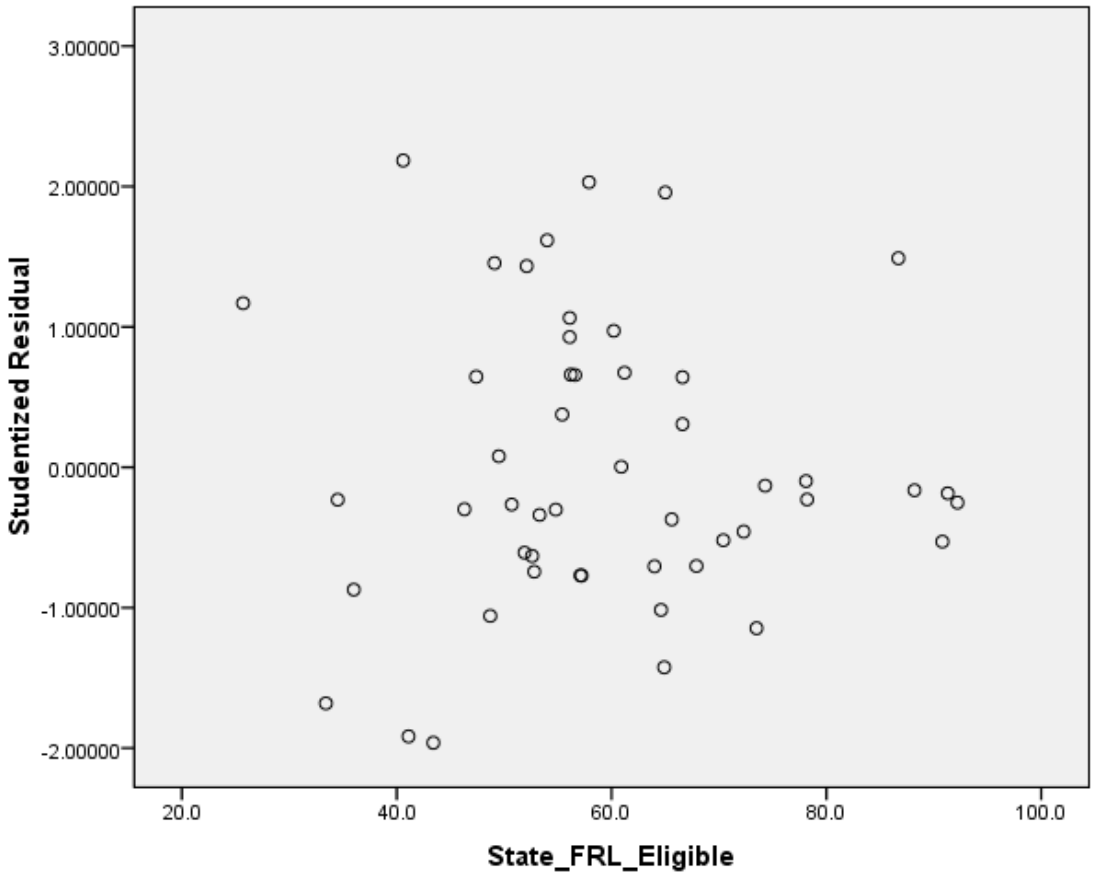

Figure 88. Scatterplot of Studentized Residuals to 2009 State Low SES Percent Proficient With Outliers Removed

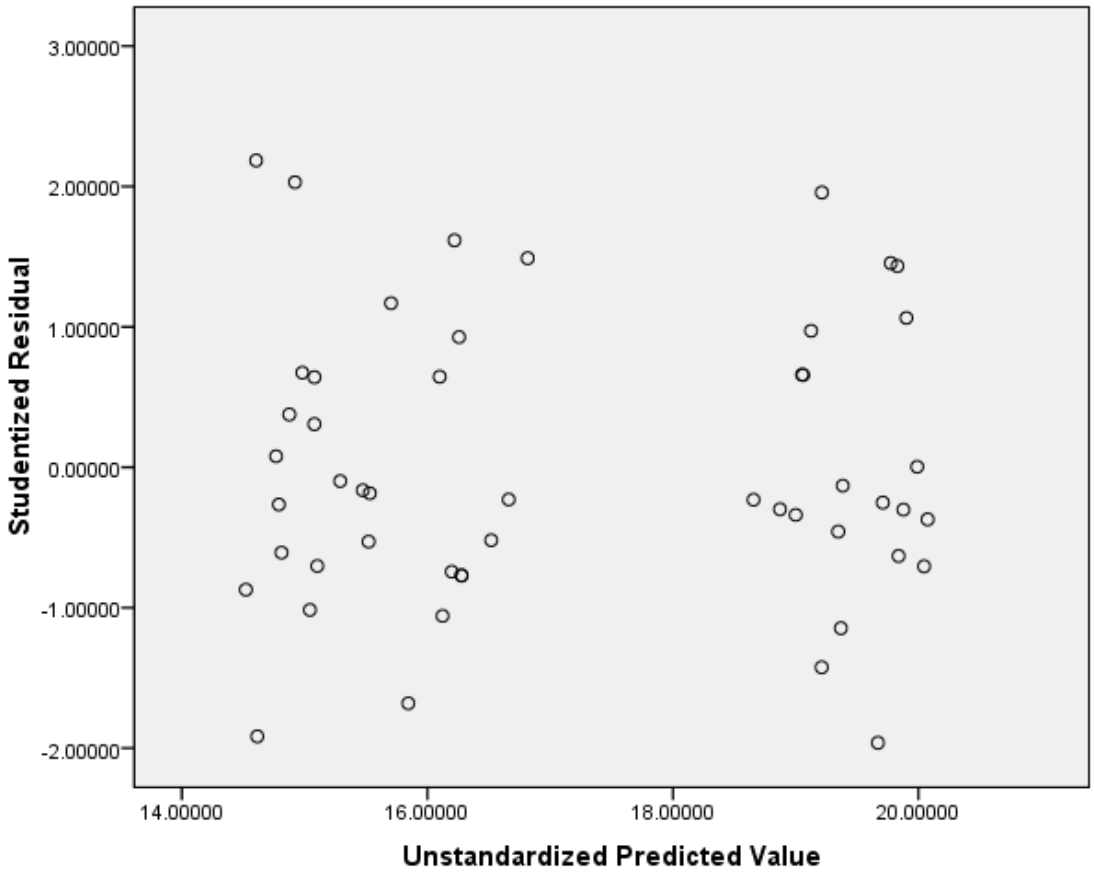

Figure 89. Scatterplot of Studentized Residuals to Unstandardized Predicted Values With Outliers Removed

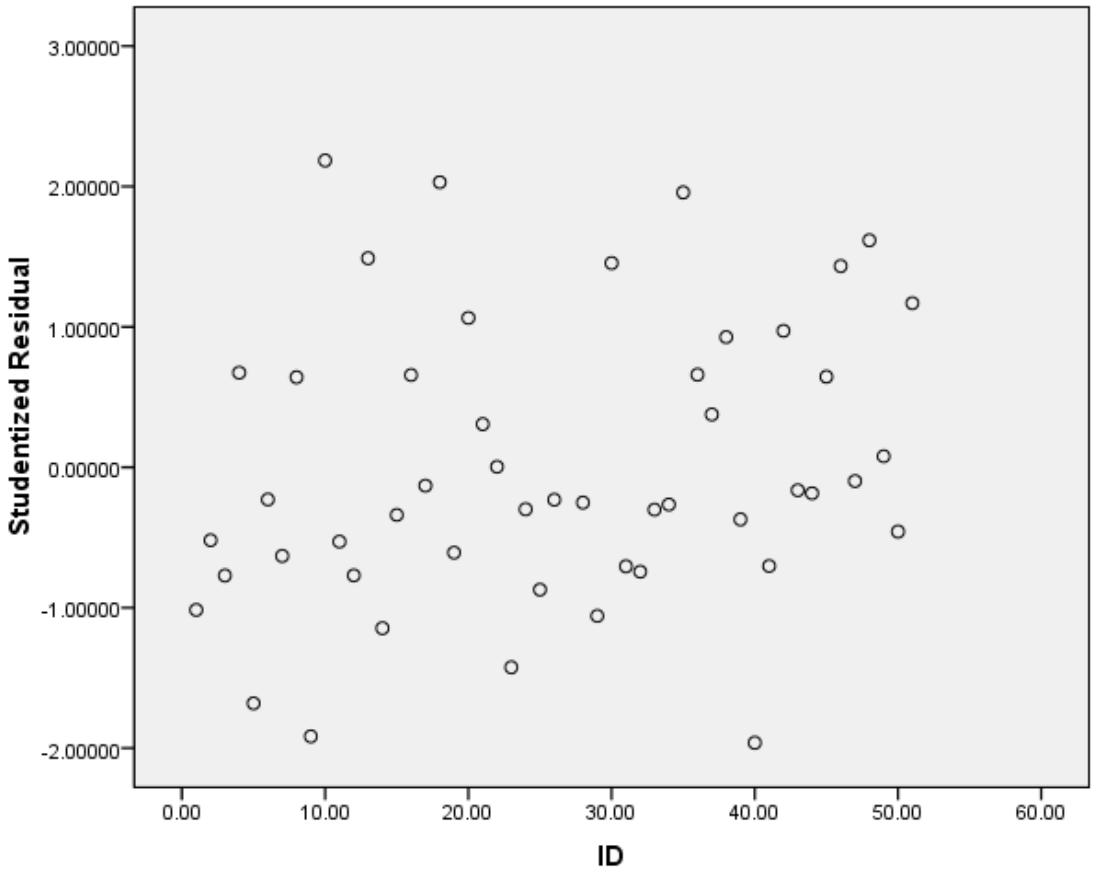

Figure 90. Scatterplot of Studentized Residuals to Case Number With Outliers Removed

Figures: Research Question Six

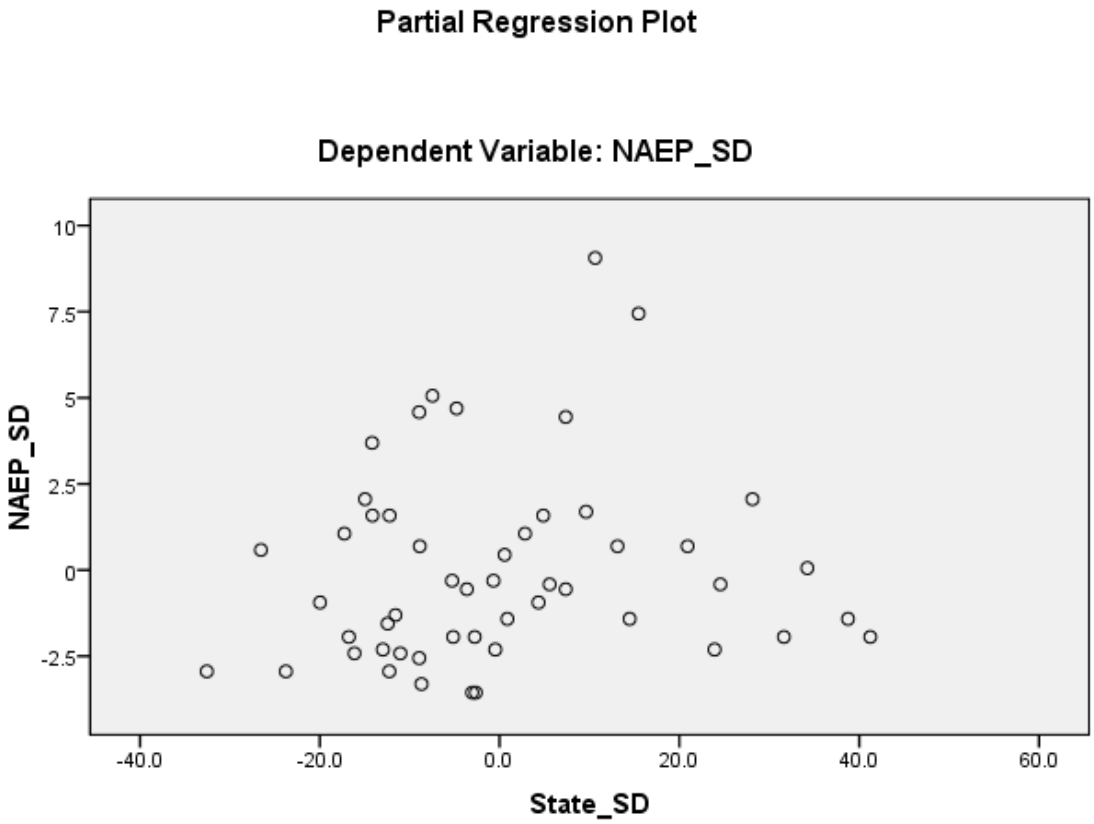

Figure 91. Partial Regression Plot of 2009 NAEP to State SWD Percent Proficient

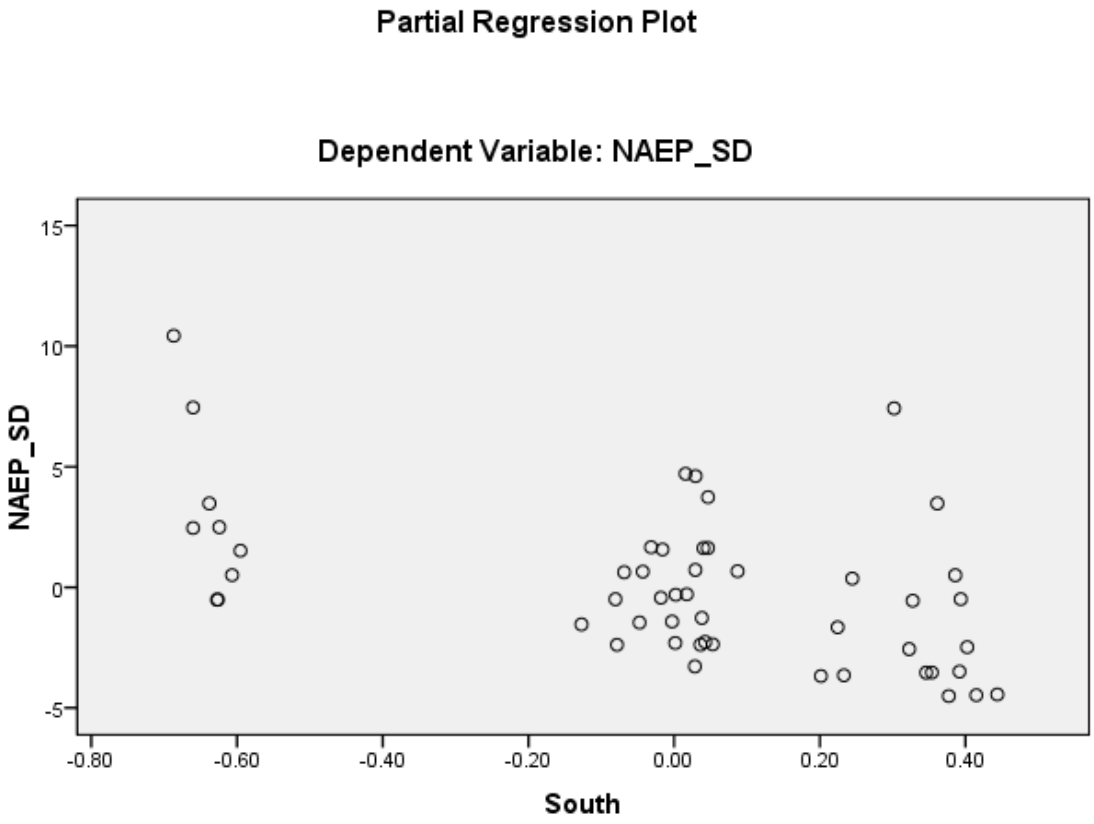

Figure 92. Partial Regression Plot of 2009 NAEP to State SWD Percent Proficient in the South

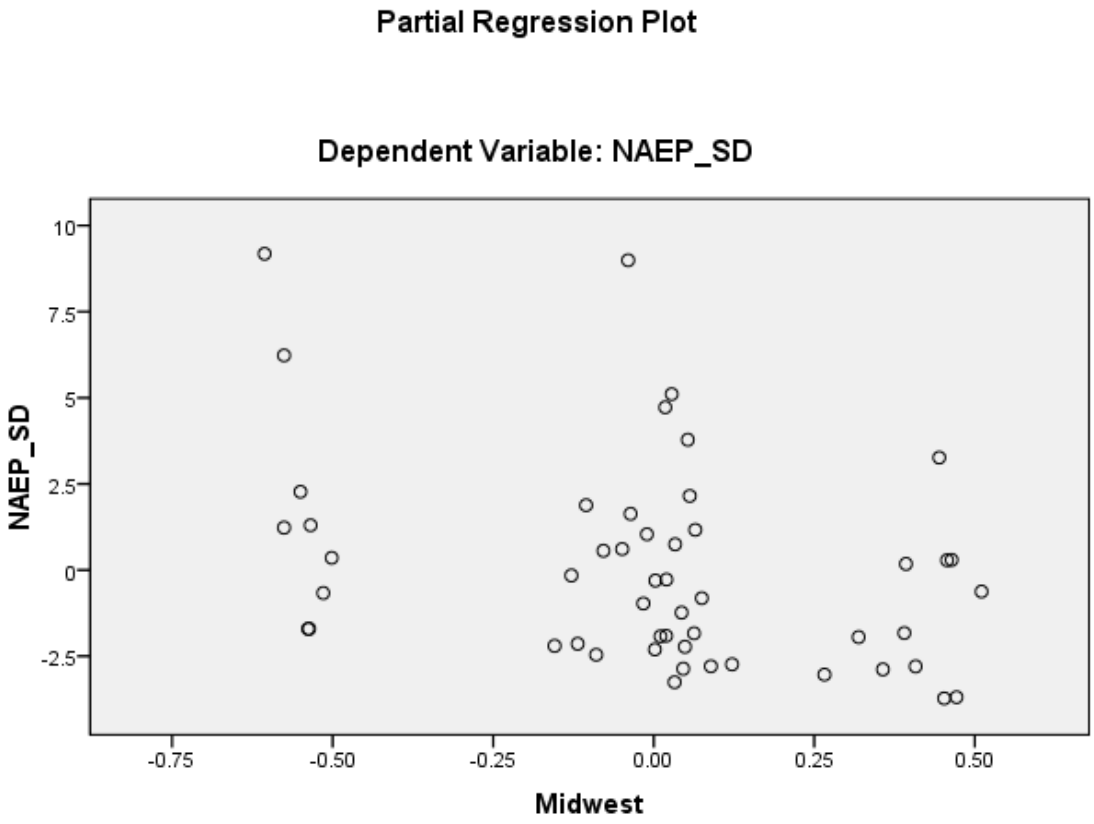

Figure 93. Partial Regression Plot of 2009 NAEP to State SWD Percent Proficient in the Midwest

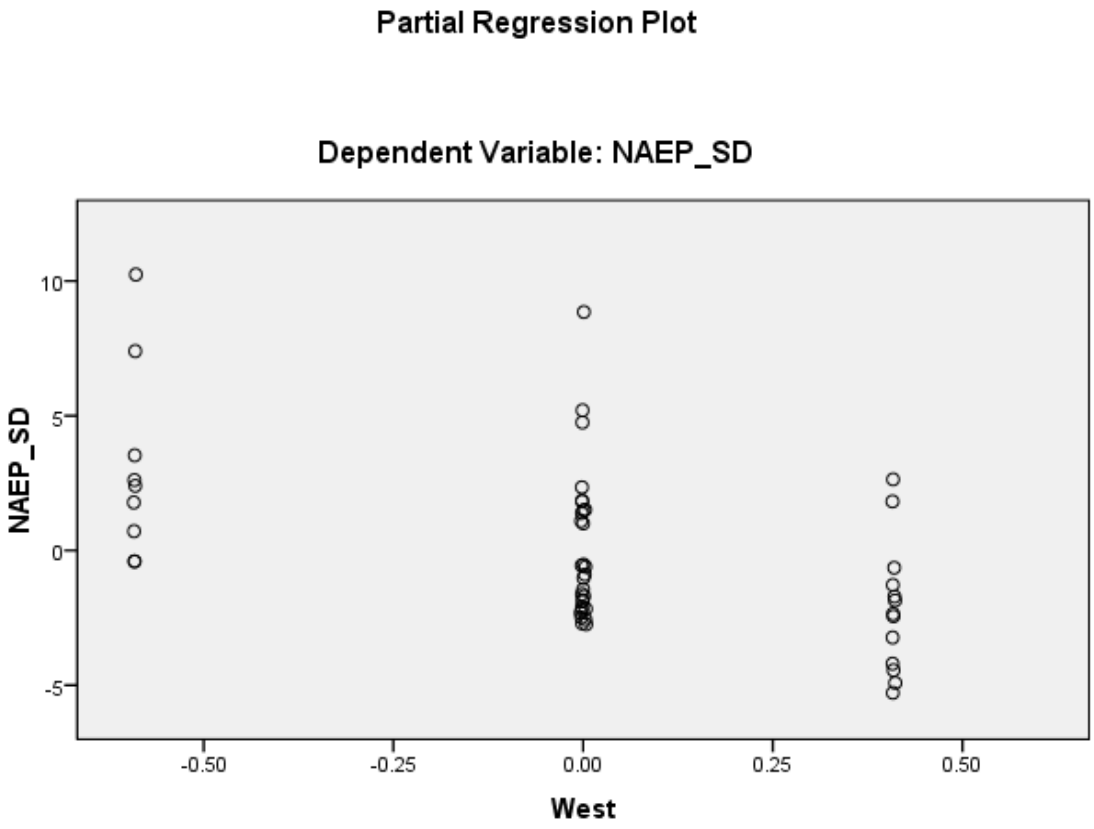

Figure 94. Partial Regression Plot of 2009 NAEP to State SWD Percent Proficient in the West

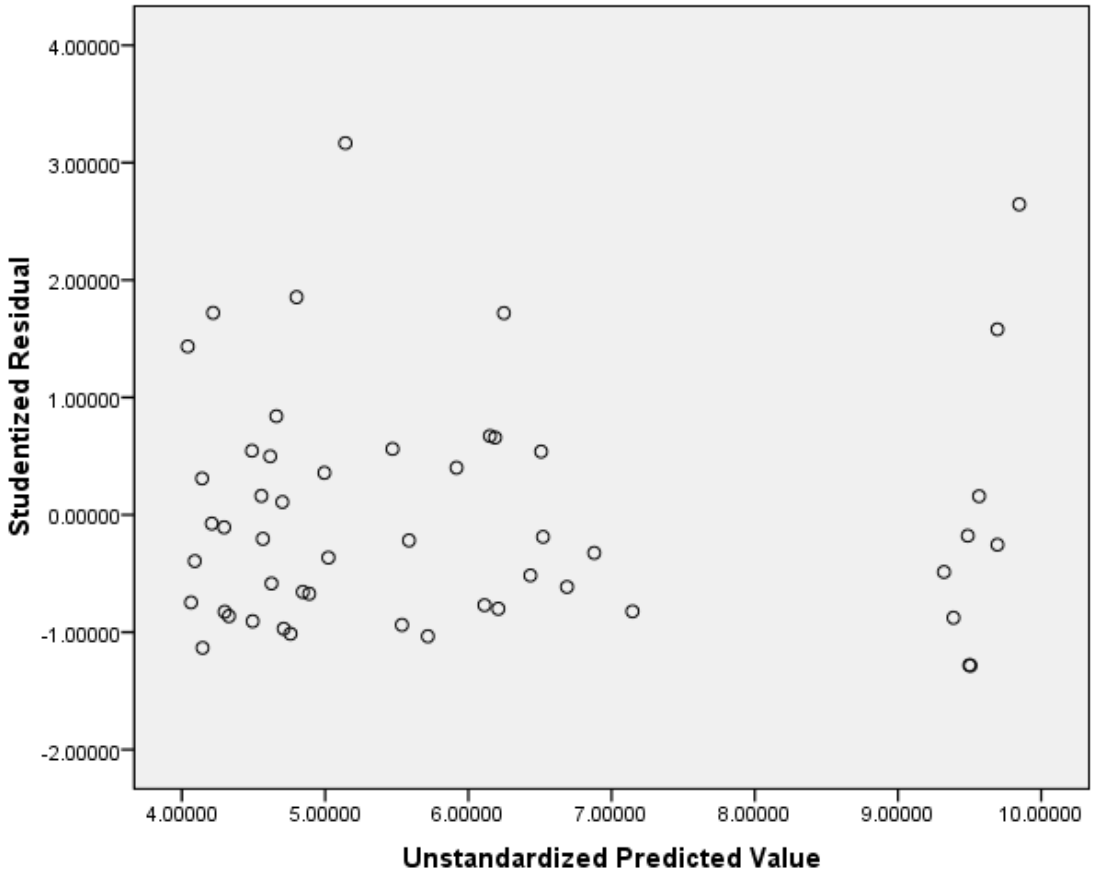

Figure 95. Scatterplot of Studentized Residuals to Unstandardized Predicted Values

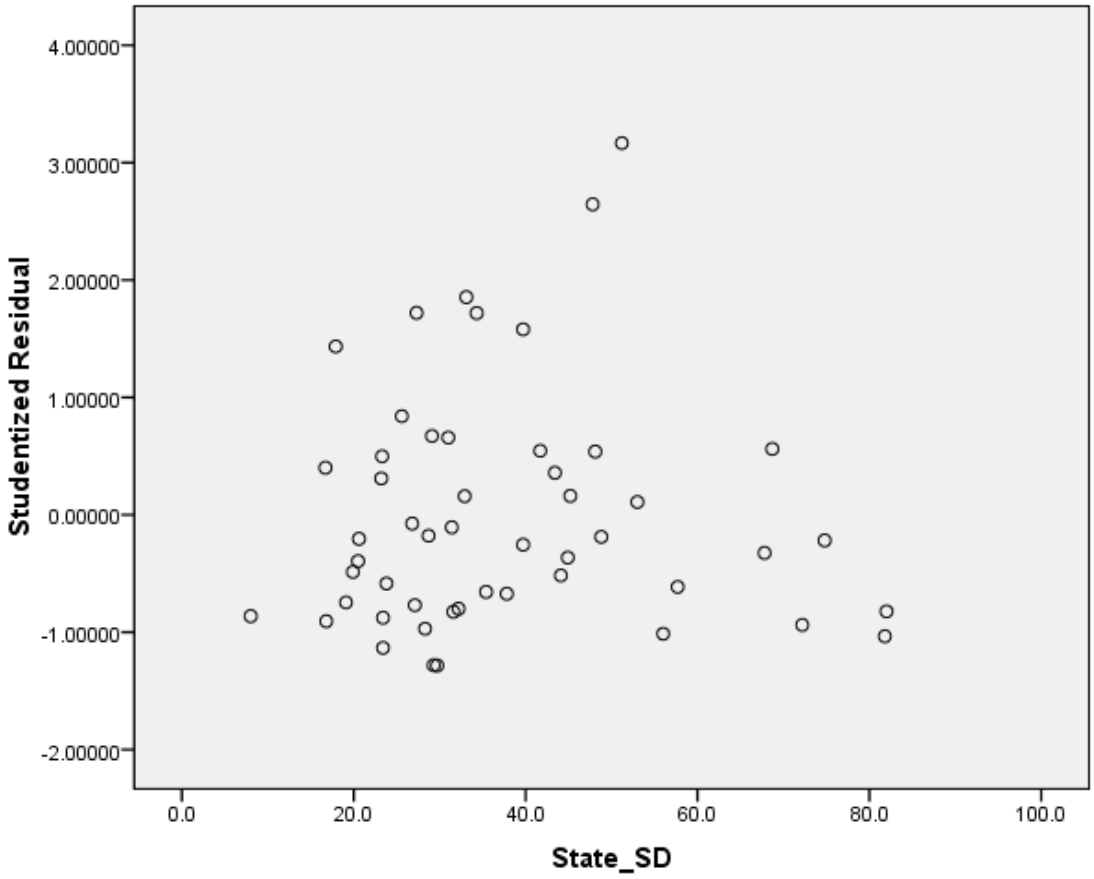

Figure 96. Scatterplot of Studentized Residuals to 2009 State SWD Percent Proficient

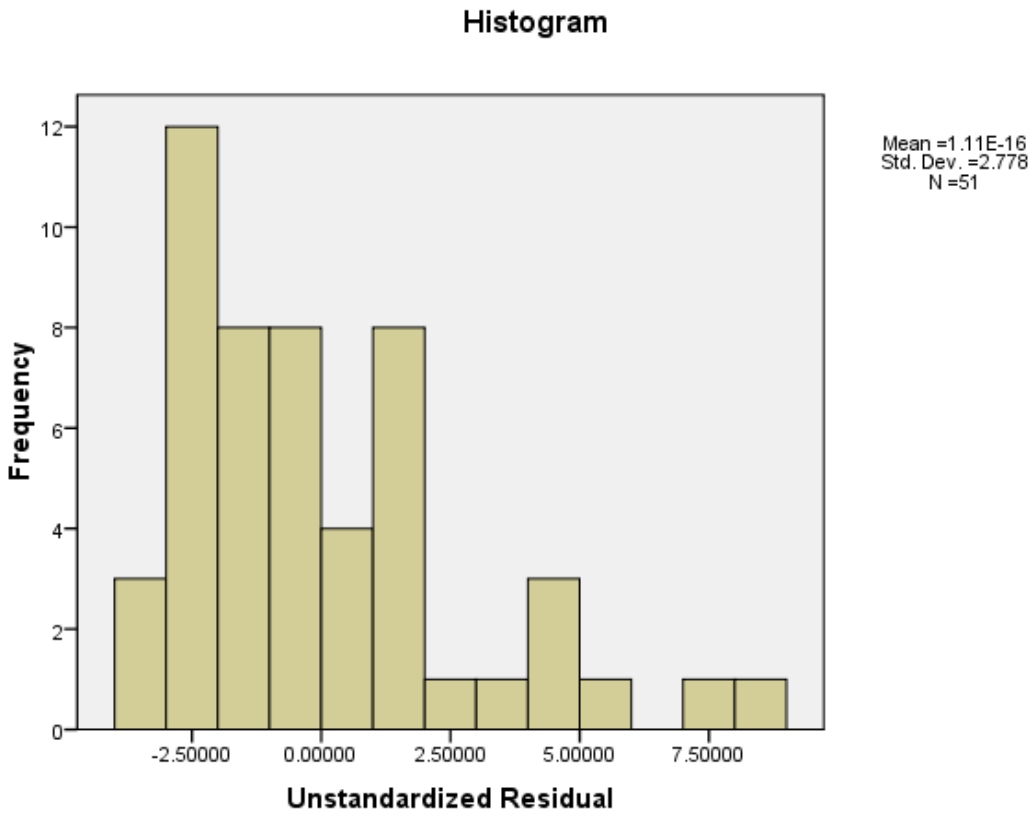

Figure 97. Histogram of Unstandardized Residuals

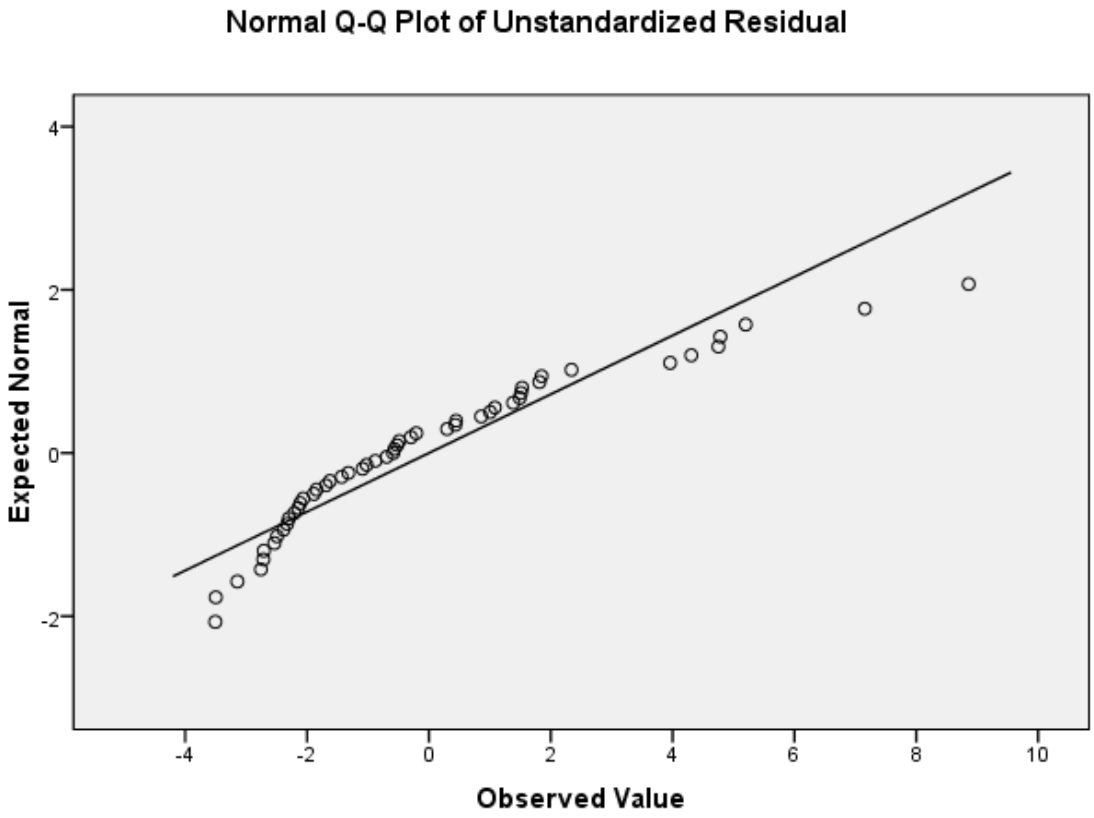

Figure 98. Q-Q Plot of Unstandardized Residuals

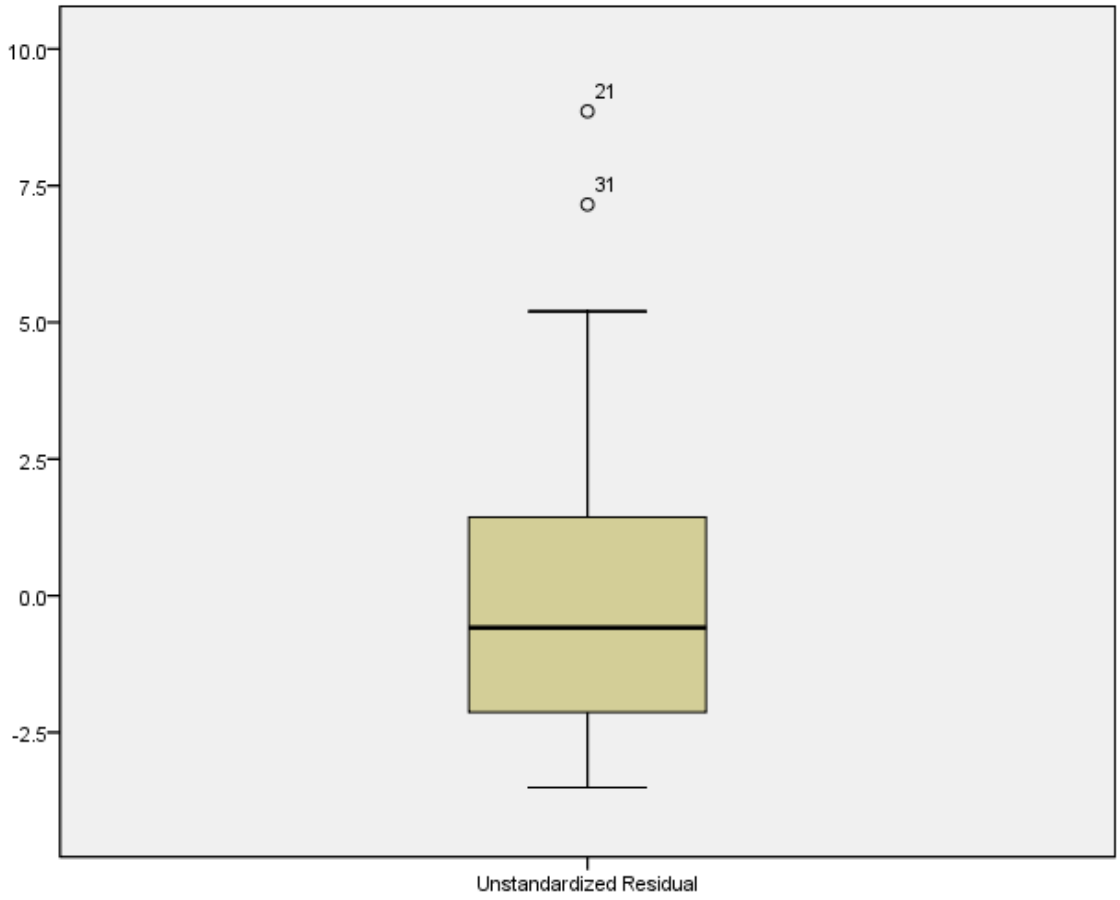

Figure 99. Boxplot of Unstandardized Residuals

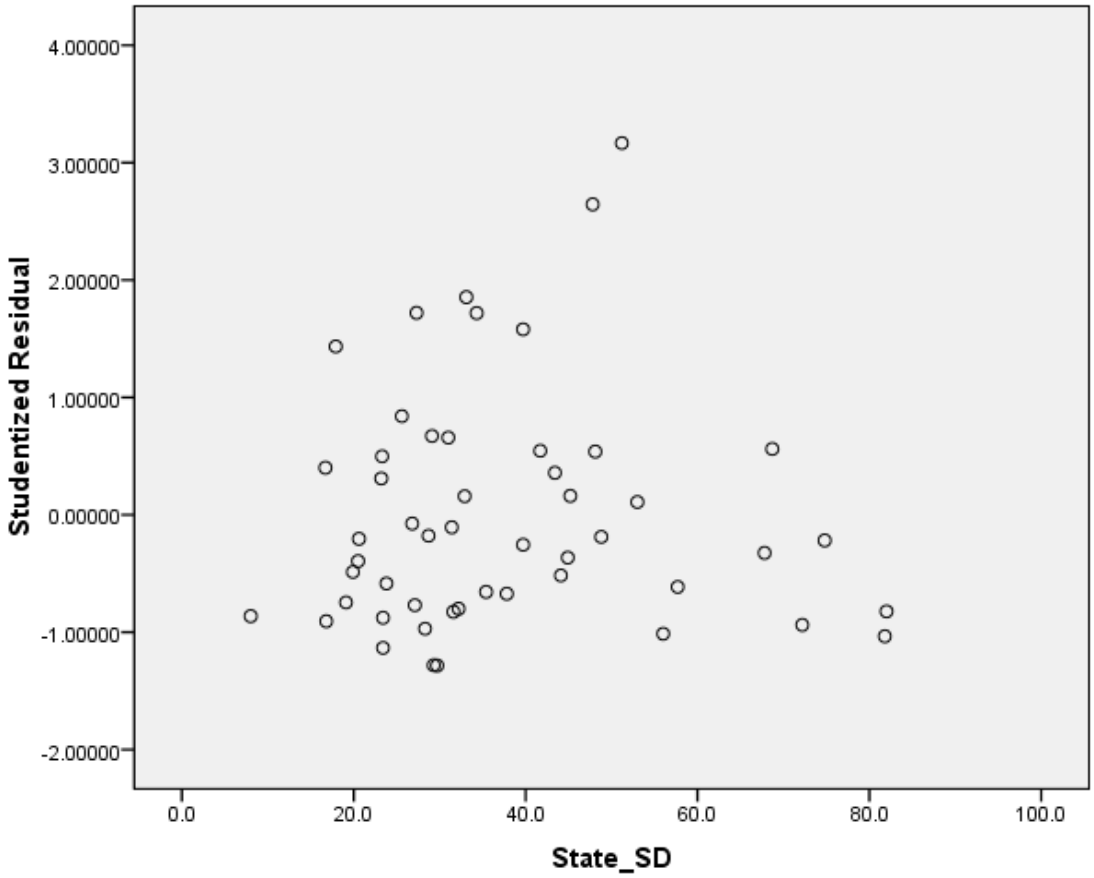

Figure 100. Scatterplot of Studentized Residuals to 2009 State SWD Percent Proficient

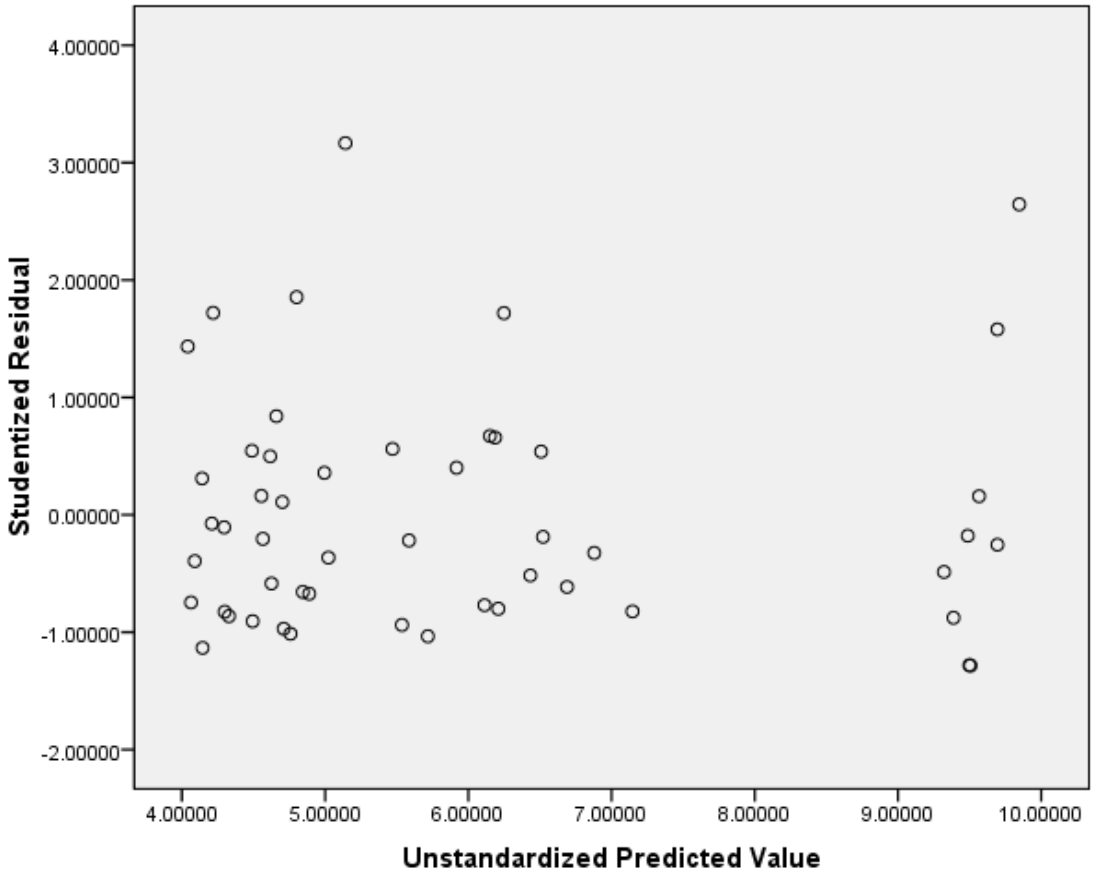

Figure 101. Scatterplot of Studentized Residuals to Unstandardized Predicted Values

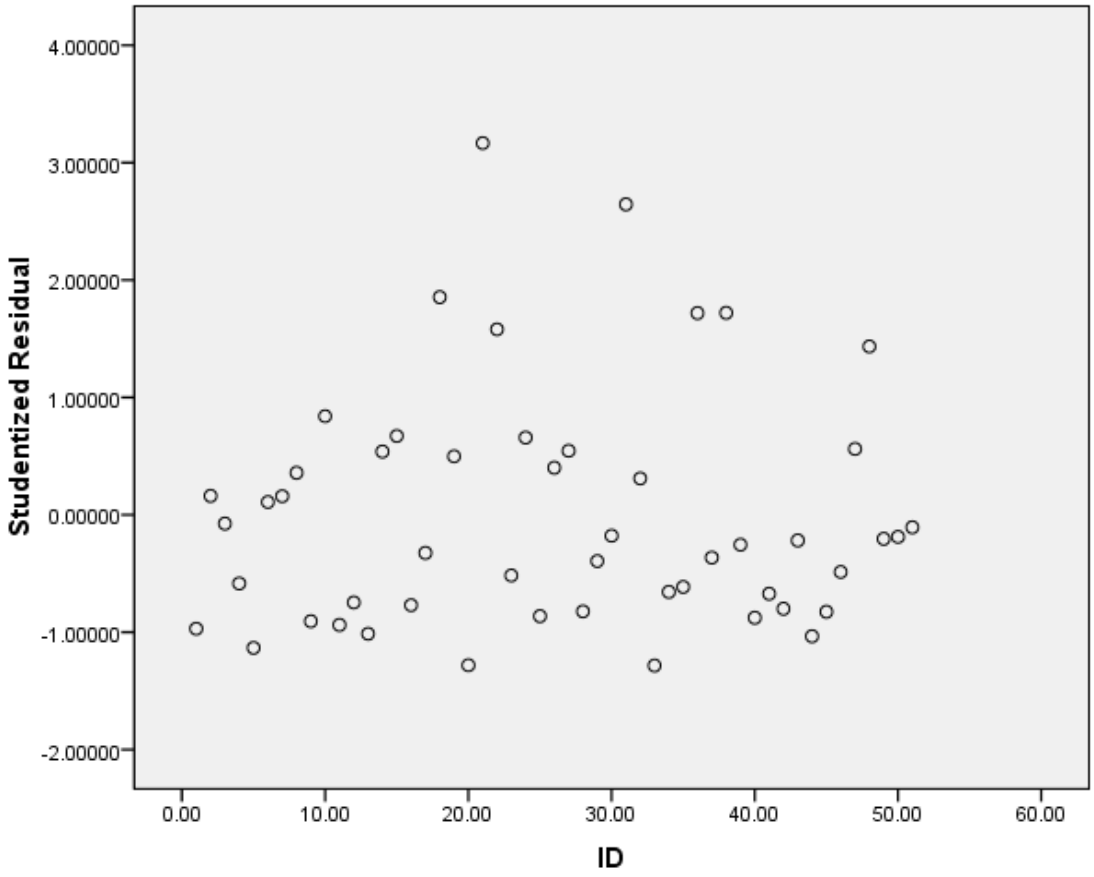

Figure 102. Scatterplot of Studentized Residuals to Case Number

Figures: Research Question Six with Outliers Removed

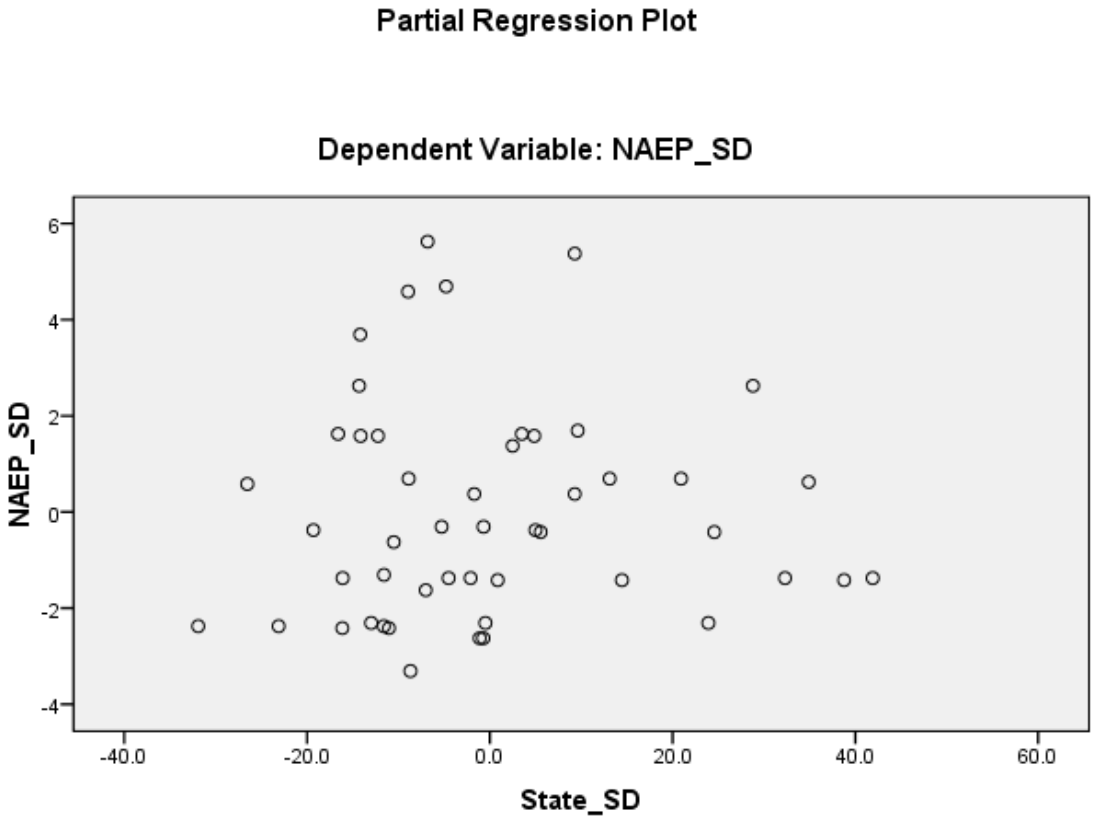

Figure 103. Partial Regression Plot of 2009 NAEP to State SWD Percent Proficient with Outliers Removed

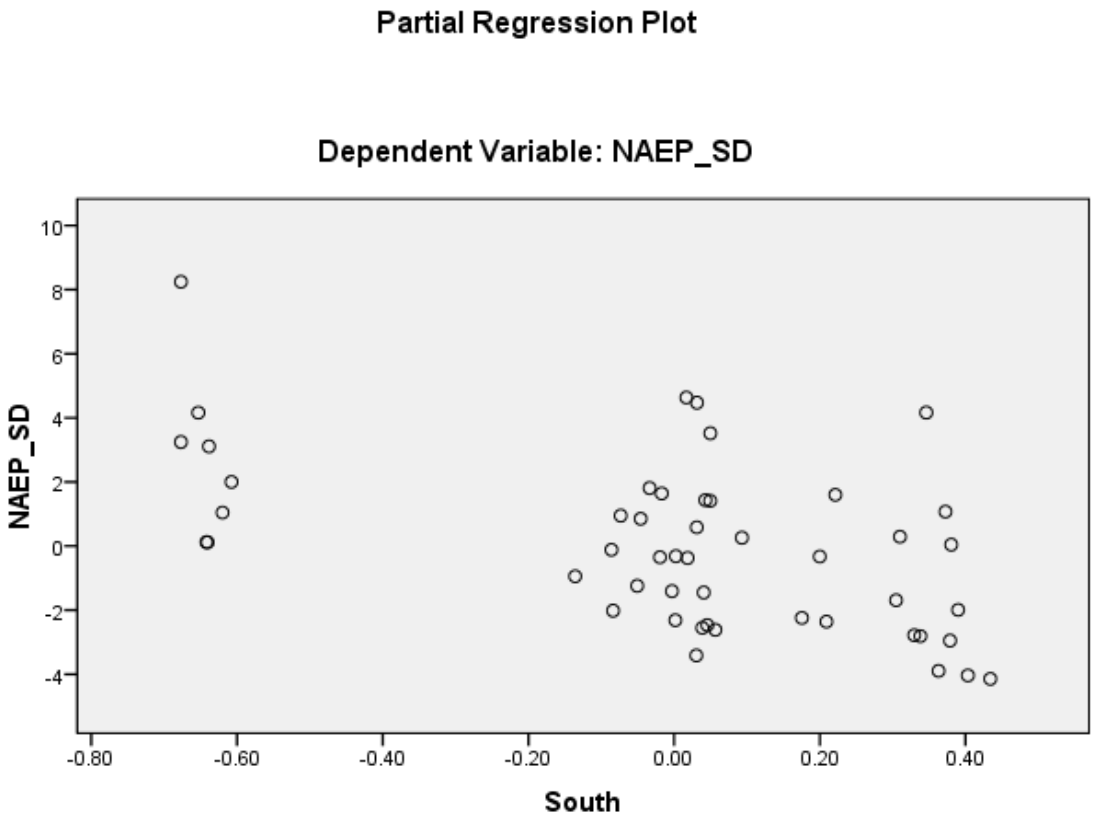

Figure 104. Partial Regression Plot of 2009 NAEP to State SWD Percent Proficient in the South with Outliers Removed

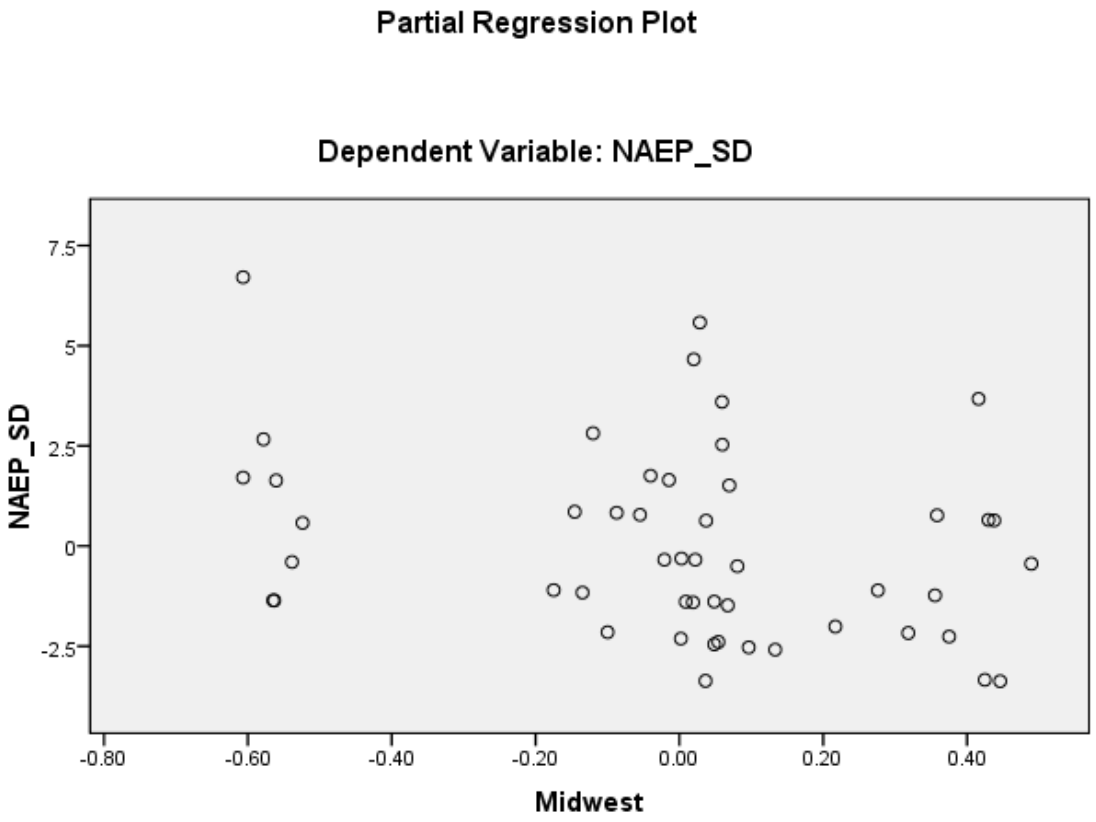

Figure 105. Partial Regression Plot of 2009 NAEP to State SWD Percent Proficient in the Midwest with Outliers Removed

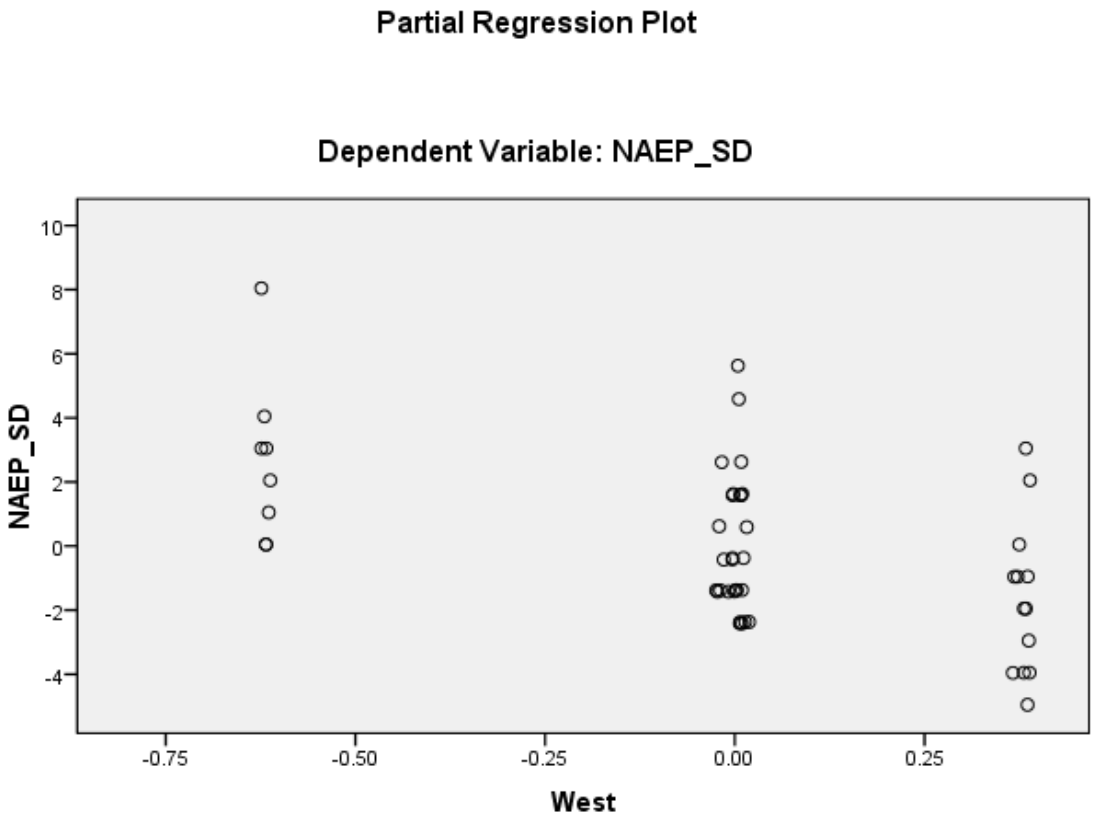

Figure 106. Partial Regression Plot of 2009 NAEP to State SWD Percent Proficient in the West with Outliers Removed

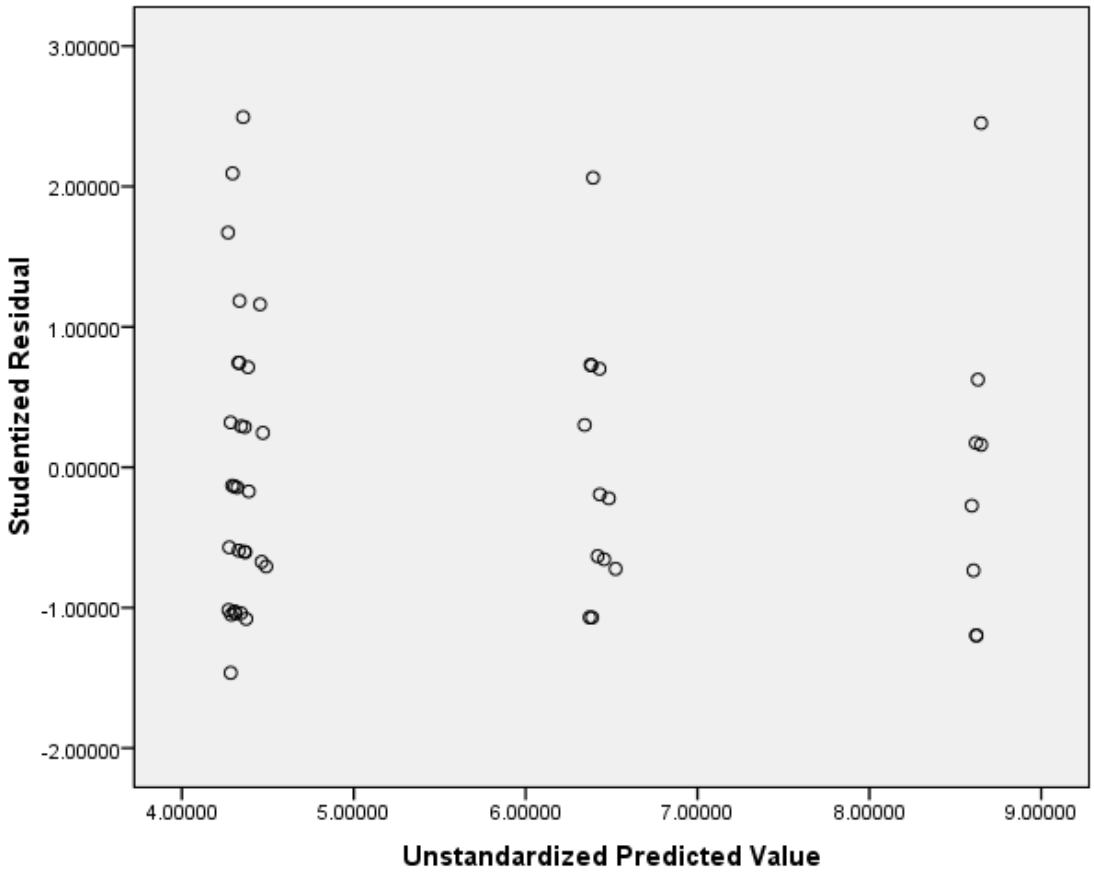

Figure 107. Scatterplot of Studentized Residuals to Unstandardized Predicted Values with Outliers Removed

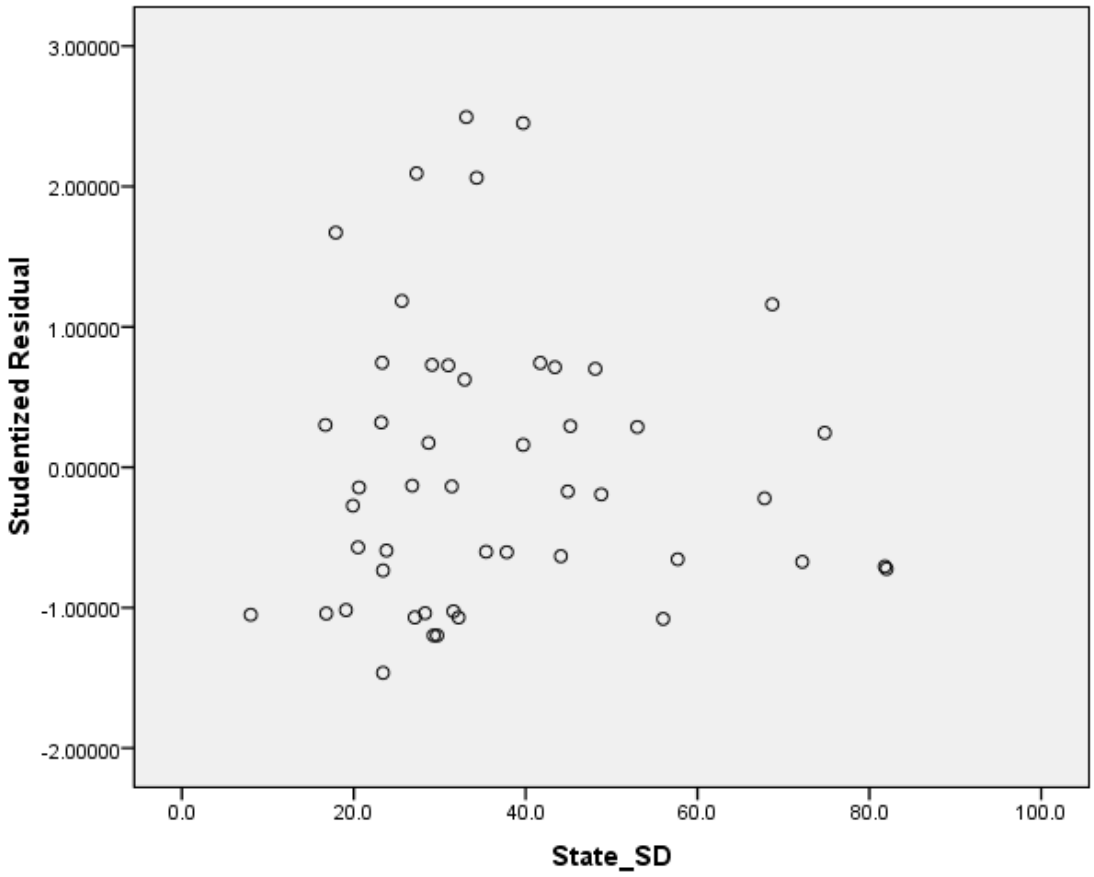

Figure 108. Scatterplot of Studentized Residuals to 2009 State SWD Percent Proficient with Outliers Removed

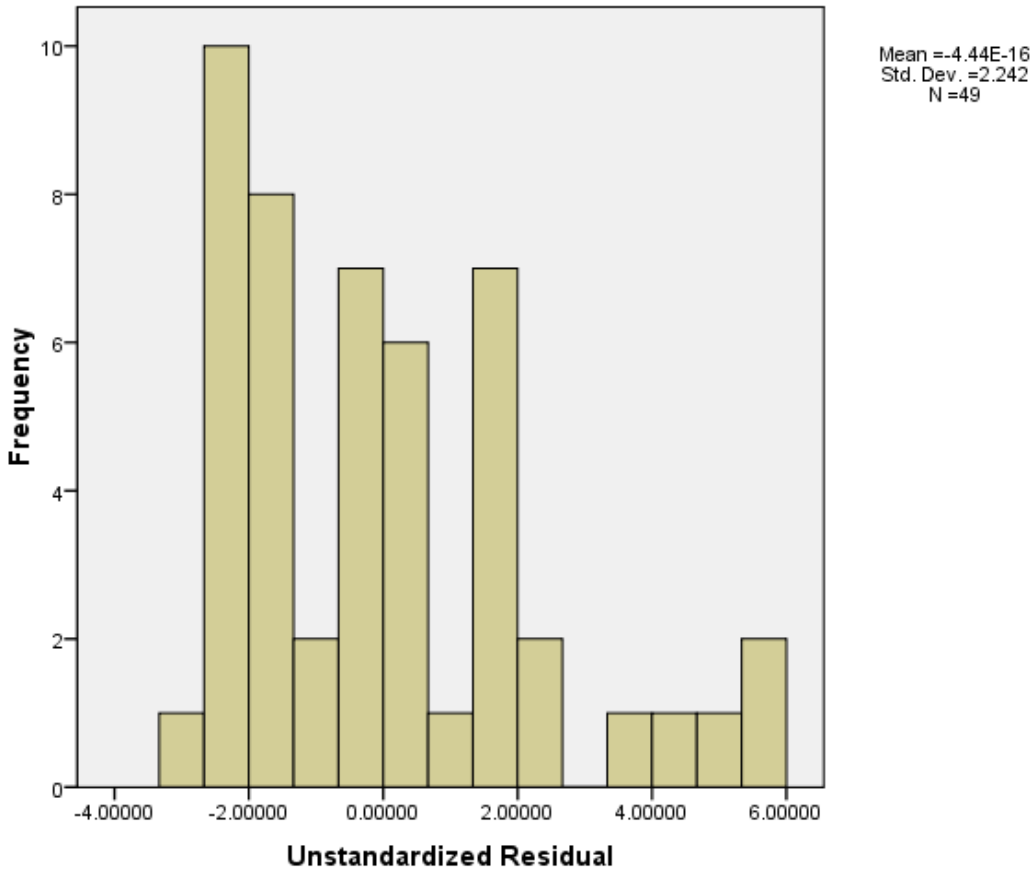

Figure 109. Histogram of Unstandardized Residuals with Outliers Removed

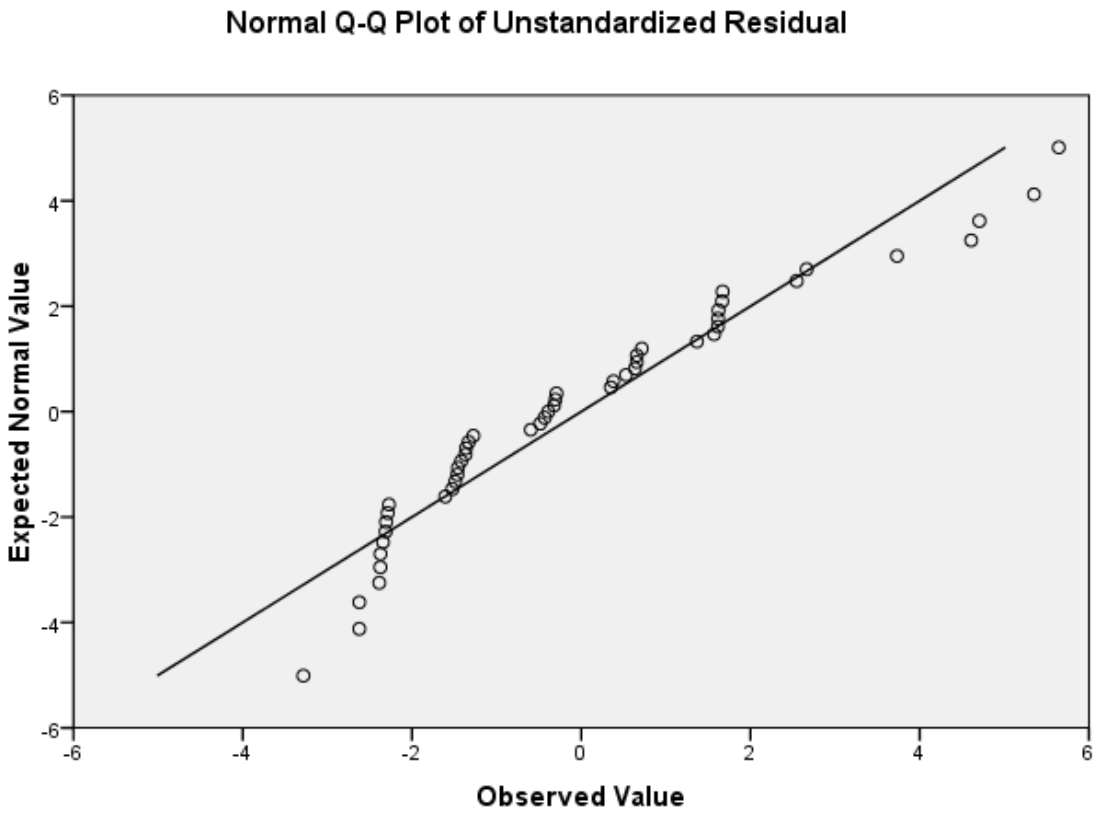

Figure 110. Q-Q Plot of Unstandardized Residuals with Outliers Removed

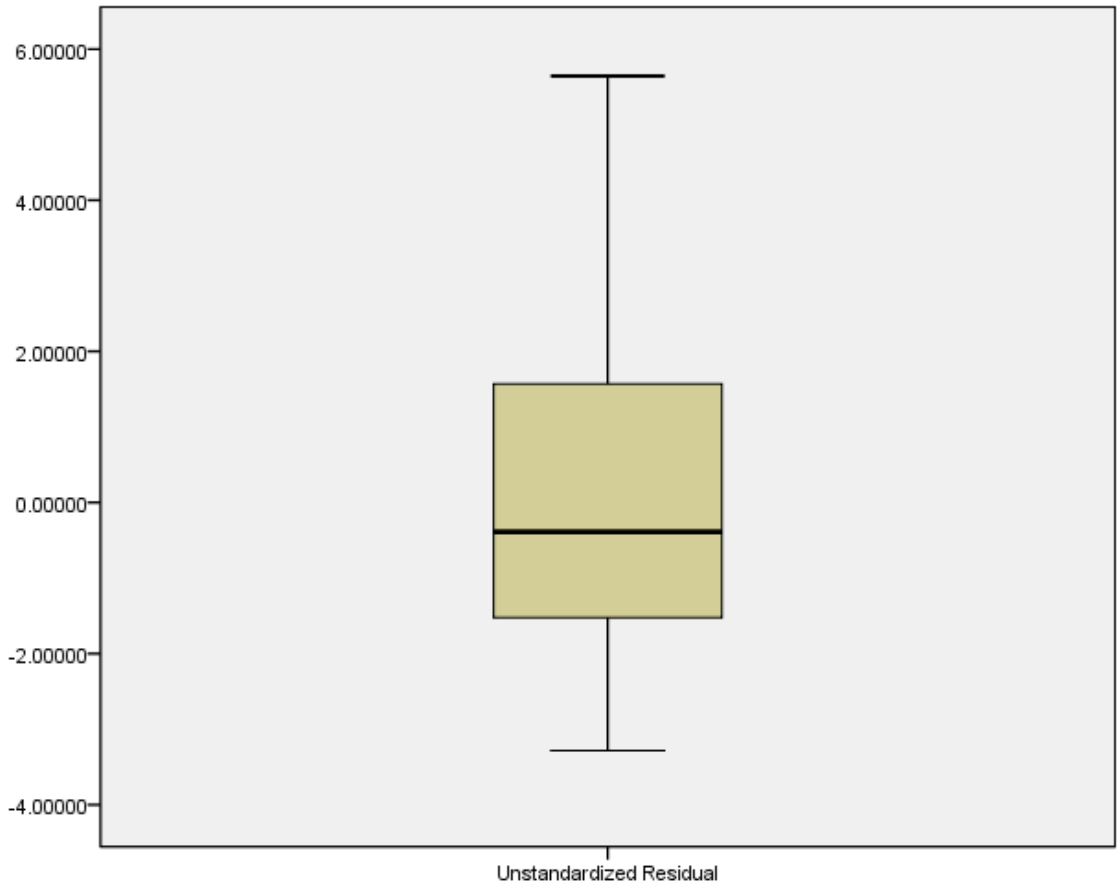

Figure 111. Boxplot of Unstandardized Residuals with Outliers Removed

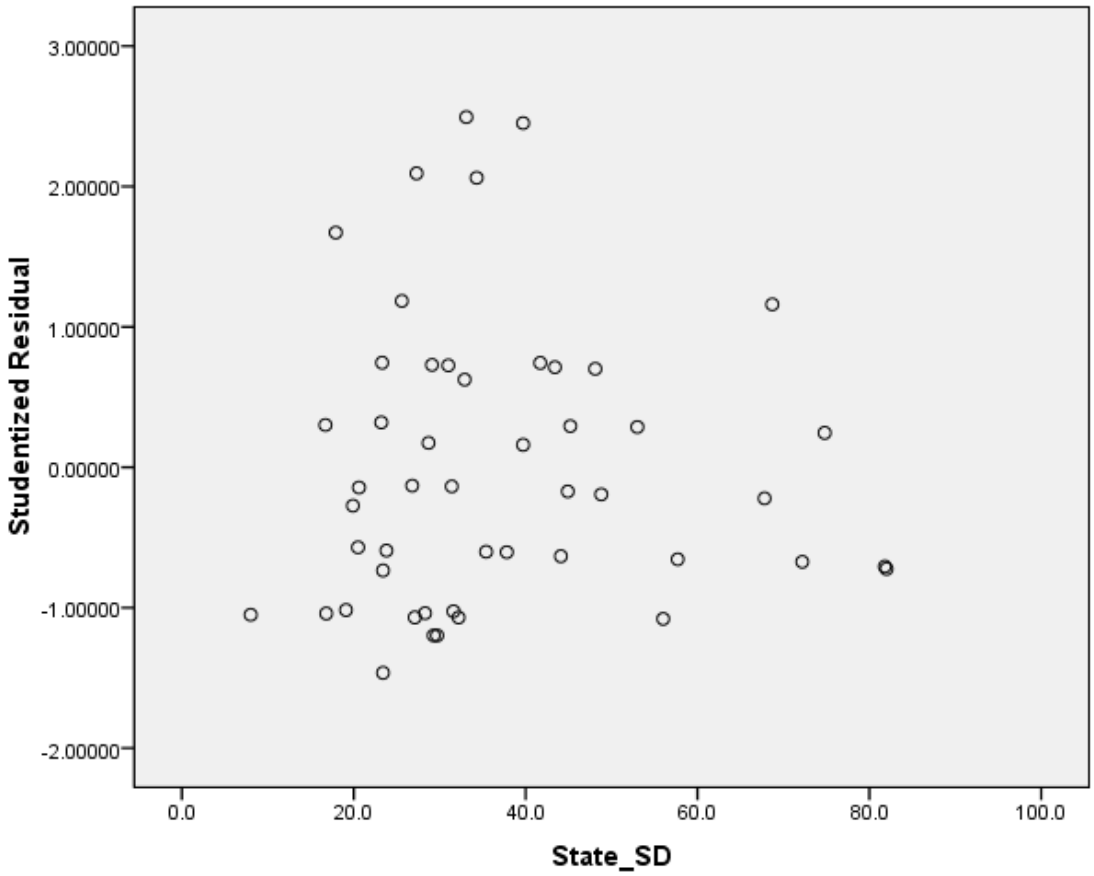

Figure 112. Scatterplot of Studentized Residuals to 2009 State SWD Percent Proficient with Outliers Removed

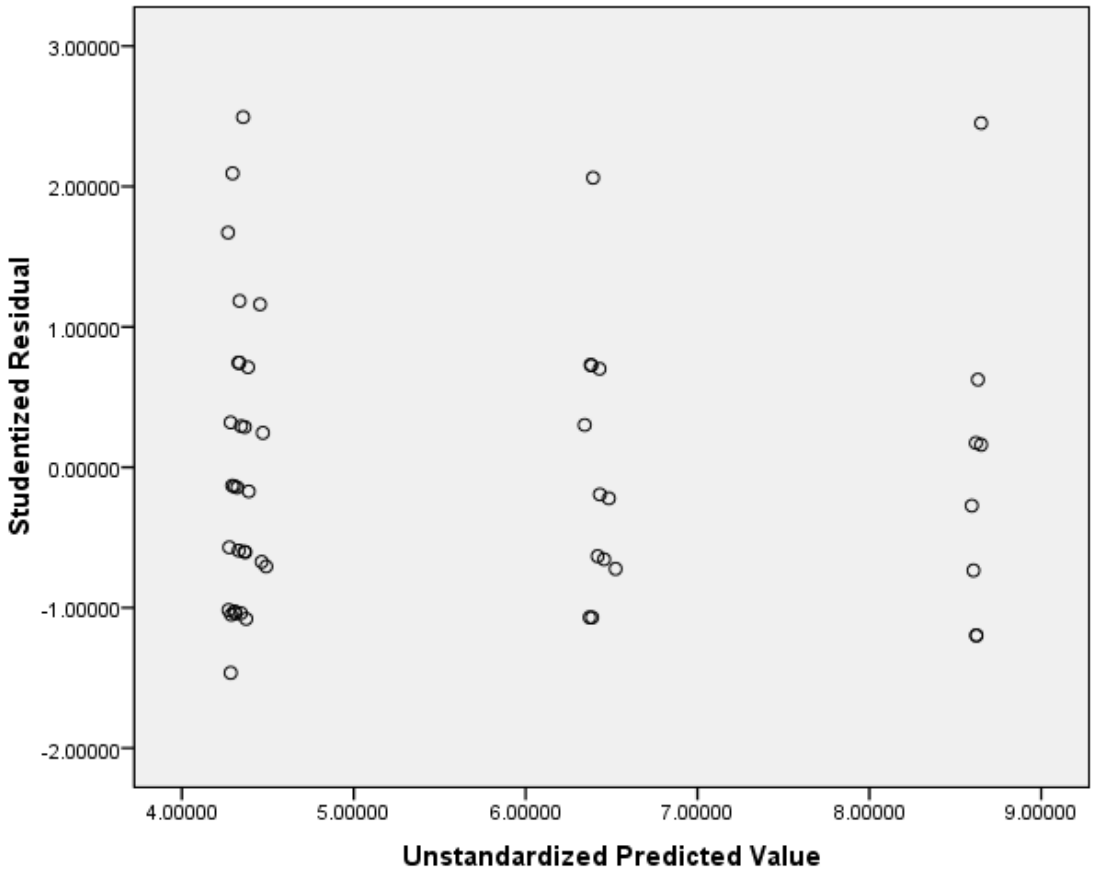

Figure 113. Scatterplot of Studentized Residuals to Unstandardized Predicted Values with Outliers Removed

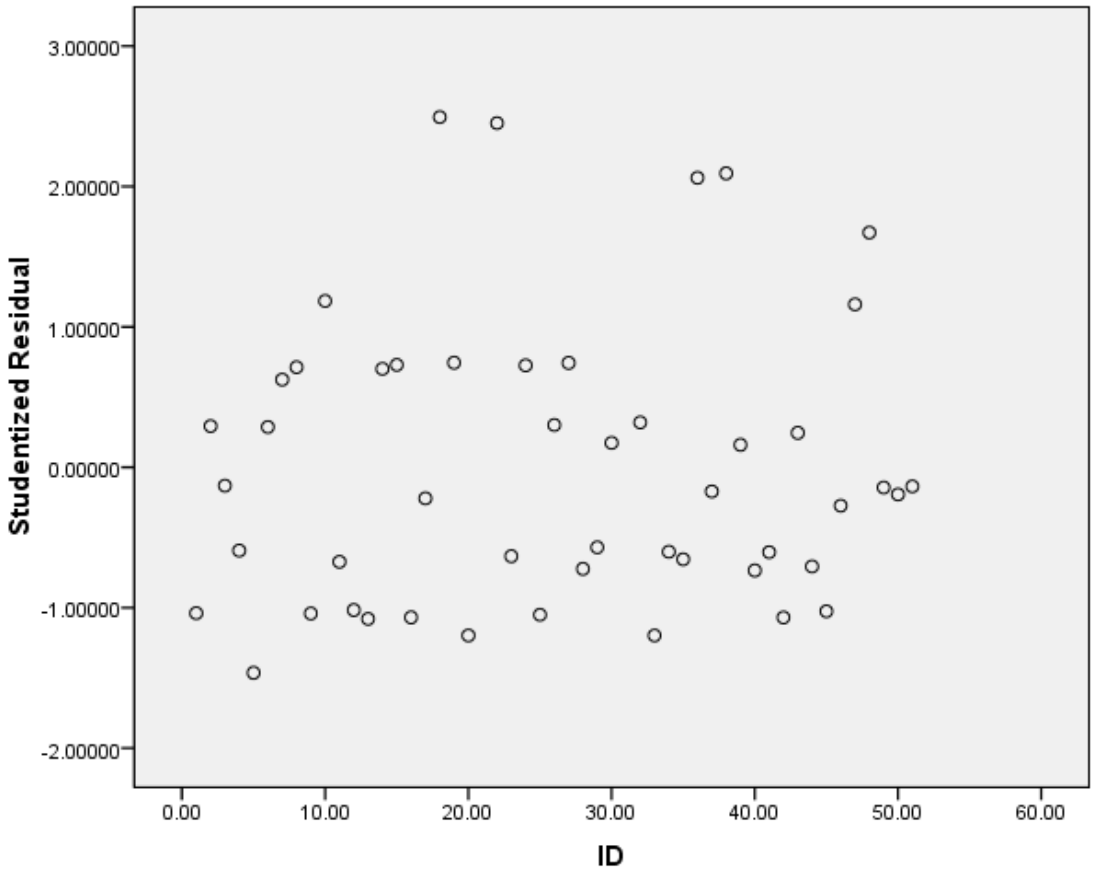

Figure 114. Scatterplot of Studentized Residuals to Case Number with Outliers Removed

APPENDIX G: IRB APPROVAL

University of Central Florida Institutional Review Board
Office of Research & Commercialization
12201 Research Parkway, Suite 501
Orlando, Florida 32826-3246
Telephone: 407-823-2901, 407-882-2012 or 407-882-2276
www.research.ucf.edu/compliance/irb.html

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

Type of Review:	[redacted]
Project Title:	[redacted]
Investigator:	[redacted]
IRB #:	[redacted]
Funding Agency:	[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

LIST OF REFERENCES

Anchorage School District. (2010). *NCLB requirements: A brief summary*. Retrieved from http://www.asdk12.org/nclb/everyone/summary.asp

Bandeira de Mello, V., Blankenship, C., & McLaughlin, D. H. (2009). *Mapping State Proficiency Standards Onto NAEP Scales: 2005-2007 (NCES 2010-456).* National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.

Bourque, M., & Hambleton, R. (1993). Setting performance standards on the national assessment of educational progress. *Measurement & Evaluation in Counseling & Development (American Counseling Association)*, *26*(1), 41. Retrieved from Professional Development Collection database.

Bourque, M. L. (2009). *A history of NAEP achievement levels: Issues, implementation, and impact 1989-2009.* Retrieved from http://www.nagb.org/who-we-are/20-anniversary/bourque-achievement-levels-formatted.pdf

Burke, J. (1996). Reviews. *Educational Leadership*, *54*(3), 88.

California Department of Education. (2004). *Key Elements of Testing*. Retrieved from http://www.cde.ca.gov/ta/tg/sa/documents/keyelements0504.pdf

California Department of Education. (2009a). *2009 California Standardized Testing and Reporting Post-test guide: Technical information for STAR district and test site coordinators and research specialists.* Retrieved from http://www.startest.org/pdfs/STAR.posttest_guide.2009.pdf

California Department of Education. (2009b). *About STAR 2009*. Retrieved from http://star.cde.ca.gov/star2009/aboutSTAR.asp

California Department of Education. (2009c). *Standardized Testing and Reporting – STAR Sample test questions.* Retrieved from http://starsamplequestions.org/starRTQ/search.jsp

California Department of Education. (2010a). *California Standardized Testing and Reporting: California Standards Test scores 2009.* Retrieved from http://star.cde.ca.gov/star2009/viewreport.asp

California Department of Education. (2010b). *California Standards Test Technical Report: Spring 2009 Administration.* Retrieved from http://www.cde.ca.gov/ta/tg/sr/documents/csttechrpt09.pdf

California State Board of Education. (2001). *California State Board of Education Policy #01-09.* Retrieved from http://www.cde.ca.gov/be/ms/po/policy01-09-dec2001.asp

Carnoy, M., & Loeb, S. (2002). Does external accountability affect student outcomes: A cross-state analysis. *Educational Evaluation and Policy Analysis*, *24*(4), 305-331.

Casserly, M. (2004). Driving change. *Education Next*, *4*(3), 32-37.

Center on Education Policy. (2008). Many states have taken a 'backloaded' approach to No Child Left Behind goal of all students scoring 'proficient.' Retrieved from http://blog.news-record.com/staff/chalkboard/May2008%20nclb%20report.pdf

Cohen, J. (1988). Statistical power analysis (2nd ed.). Hillsdale, NJ: Erlbaum.

Decker, D., & Bolt, S. (2008). Challenges and opportunities for promoting student achievement through large-scale assessment results: Research, reflections, and future directions. *Assessment for Effective Intervention*, *34*(1), 43-51.

Dorans, N. J. (2004). Equating, concordance, and expectation. *Applied Psychological Measurement*, *28*(4), 227-246.

Easton, J., Rosenkranz, T., Bryk, A., & Consortium on Chicago School Research. (2001). Annual CPS test trend review, 2000: Research data brief. Academic Productivity Series, 2000 Results. Retrieved from ERIC database.

Ercikan, K. (1997). Linking statewide tests to the National Assessment of Educational Progress: Accuracy of combining test results across states. *Applied Measurement in Education*, *10*(2), 145-159.

Feuer, M. J., National Research Council (1999). *Uncommon measures: Equivalence and linkage among educational tests.* Washington, DC: National Academy Press.

Florida Department of Education. (2007). *Assessment and accountability briefing book: FCAT, school accountability, teacher certification tests.* Retrieved from http://fcat.fldoe.org/pdf/BriefingBook07web.pdf

Florida Department of Education. (2010). *PARRC submits application for Race to the Top assessment funds.* Retrieved from http://www.fldoe.org/news/2010/2010_06_23.asp

Goertz, M. (2005). Implementing the No Child Left Behind Act: Challenges for the states. *Peabody Journal of Education*, *80*(2), 73-89.

Gordon, W. R. (2009). A comparison of eighth grade reading scores by state and by the four census defined regions identified by NAEP. Orlando, FL: University of Central Florida. Retrieved from Dissertation & Theses: Full Text. (CFE0002536.)

Hansen, J. (1993). Is educational reform through mandated accountability an oxymoron? *Measurement and Evaluation in Counseling and Development*, *26*(1), 11-21.

Hess, F. (2005). Commentary: Accountability policy and scholarly research. *Educational Measurement: Issues & Practice*, *24*(4), 53-57.

Hoff, D. (2007). Turnarounds central issue under NCLB. *Education Week*, *26*(42), 1.

Hoff, D. (2008). Steep climb to NCLB goal for 23 states. *Education Week*, *27*(39), 1.

Hombo, C. H. (2003). NAEP and No Child Left Behind: Technical challenges and practical solutions. *Theory into Practice*, *42*(1), 59-65.

Illinois State Board of Education Division of Assessment. (2009). *Illinois Standards Achievement Test 2009 Technical Manual.* Illinois: NCS Pearson. Retrieved from http://www.isbe.state.il.us/assessment/pdfs/2009_ISAT_Tech_Manual.pdf

Illinois State Board of Education. (2009). *ISAT sample book 2009: Grade 8*. Illinois: NCS Pearson. Retrieved from http://www.isbe.state.il.us/assessment/pdfs/2009/ISAT_Sample_Book_gr8.pdf

Improving America's Schools Act of 1994 Pub. L. No. 103-382, § 2, 108 Stat. 3518 (1994).

Jones, L., & Olkin, I. (2004). The nation's report card: Evolution and perspectives. Bloomington, IN: Phi Delta Kappa Educational Foundation.

Kolen, M. J. (2004). Linking assessments: Concept and history. *Applied Psychological Measurement*, *28*(4), 219-226.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking, 2nd ed*. New York, NY: Springer.

Lee, J. (2008). Is test-driven external accountability effective?: Synthesizing the evidence from cross-state causal-comparative correlational studies, *Review of Educational Research*, *78*(3), 608-644.

Linn, R. L. (1993). Linking results of distinct assessments. *Applied Measurement in Education*, 6*(*1) 83-102.

Linn, R. L., & Kiplinger, V. L. (1995). Linking statewide tests to the National Assessment of Educational Progress: Stability of results. *Applied Measurement in Education*, *8*(2), 135-155.

Manno, B. (2004). Chartering and the idea of accountability consequences: Adding performance value to schooling. *Journal of Education*, *185*(3), 27-40.

McLester, S. (2006). Stepping up to AYP: More than a quarter of American schools have been labeled failing under the provisions of NCLB: What interventions are working? *Technology & Learning*, *27*(2), 20.

Mintrop, H., & Trujillo, T. (2005). Corrective action in low performing schools: Lessons for NCLB implementation from first-generation accountability systems. *Education Policy Analysis Archives*, *13*(48), 1-30.

Mislevy, R., Educational Testing Service. (1992). Linking educational assessments: Concepts, issues, methods, and prospects. Retrieved from ERIC database.

National Center for Education Statistics. (2001). *The NAEP 1998 technical report.* Retrieved from http://nces.ed.gov/nationsreportcard/pubs/main1998/2001509.asp

National Center for Education Statistics. (2008). *NAEP technical documentation.* Retrieved from http://nces.ed.gov/nationsreportcard/tdw/

National Center for Education Statistics. (2009a). *A closer look at exclusion and accommodation results as related to assessment results.* Retrieved from http://nces.ed.gov/nationsreportcard/about/effect_exclusion.asp

National Center for Education Statistics. (2009b). *More about NAEP reading.* Retrieved from http://nces.ed.gov/nationsreportcard/reading/moreabout.asp

National Center for Education Statistics. (2009c). *NAEP overview.* Retrieved from http://nces.ed.gov/nationsreportcard/about/

National Center for Education Statistics. (2009d). *NAEP state mapping.* Retrieved from http://nces.ed.gov/nationsreportcard/studies/statemapping/statemapping_faq.asp#quest1

National Center for Education Statistics. (2010a). *The nation's report card: Frequently asked questions.* Retrieved from http://nces.ed.gov/nationsreportcard/faq.asp

National Center for Education Statistics. (2010b). *NAEP data explorer* [Data file]. Retrieved from http://nces.ed.gov/nationsreportcard/naepdata/dataset.aspx

National Center for Education Statistics. (2010c). *NAEP questions tool.* Retrieved from http://nces.ed.gov/nationsreportcard/itmrlsx/detail.aspx?subject=reading

National Center for Education Statistics. (2010d). *The NAEP reading achievement levels by grade.* Retrieved from http://nces.ed.gov/nationsreportcard/reading/achieveall.asp#2009_grade8

National Center for Education Statistics. (2010e). *The NAEP reading scale.* Retrieved from http://nces.ed.gov/nationsreportcard/reading/scale.asp

National Center for Education Statistics. (2010f). *NAEP reporting groups.* Retrieved from http://nces.ed.gov/nationsreportcard/reading/interpret-results.asp#repgroups

New York State Department of Education. (2009a). *New York State testing program 2009: English language arts, grades 3-8*. Monterey, CA: CTB McGraw Hill. Retrieved from http://www.p12.nysed.gov/osa/reports/2009/ela-techrep-09.pdf

New York State Department of Education. (2009b). *New York State testing program: English language arts test book one grade eight.* Monterey, CA: CTB McGraw Hill. Retrieved from http://www.nysedregents.org/Grade8/EnglishLanguageArts/20090120book1.pdf

New York State Department of Education (2009c). *New York State testing program: English language arts test book two grade eight.* Monterey, CA: CTB McGraw Hill. Retrieved from http://www.nysedregents.org/Grade8/EnglishLanguageArts/20090120book2.pdf

Olson, L. (2002). Accountability studies find mixed impact on achievement. *Education Week*, *21*(41), 13.

Peterson, P.E., & Hess, F.M. (2005). Johnny can read… in some states: Assessing the rigor of state assessment systems. *Education Next*, *5*(3), 52-53.

Pommerich, M., Hanson, B. A., Harris, D. J., & Sconing, J. A. (2004). Issues in conducting linkages between distinct tests. *Applied Psychological Measurement*, *28*(4), 247-273.

Porter, A., Linn, R., & Trimble, C. (2005). The effects of state decisions about NCLB adequate yearly progress targets. *Educational Measurement: Issues & Practice*, *24*(4), 32-39.

Prowker, A., & Camilli, G. (2007). Looking beyond the overall scores of NAEP assessments: Applications of generalized linear mixed modeling for exploring value-added item difficulty effects. *Journey of Educational Measurement*, *44*(1), 69-87.

Public Education Network and National Coalition for Parent Involvement in Education. (2010). State accountability system and adequate yearly progress. Retrieved from http://www.ncpie.org/nclbaction/ayp.html

Rustique-Forrester, E. (2005). Accountability and the pressures to exclude: A cautionary tale from England. *Education Policy Analysis Archives*, *13*(26), 1-39.

Springer, M. (2008). The influence of an NCLB accountability plan on the distribution of student test score gains. *Economics of Education Review*, *27*(5), 556-563.

Taylor, R., & Gordon, W. R. (In press). National Assessment of Educational Progress and state assessments: What do the data really mean? *ERS Spectrum.*

Texas Education Agency: Student Assessment Division, *Texas Assessment of Knowledge and Skills Performance Level Descriptors (*2004). Retrieved from http://www.tea.state.tx.us/index3.aspx?id=3222&menu_id=793

Texas Education Agency: Student Assessment Division, *TAKS 2009 Mean P-Values and Internal Consistency Values by Objective and Subjet Area* (2009). Retrieved from http://www.tea.state.tx.us/index3.aspx?id=3654&menu_id=793

U.S. Department of Education (2002). No Child Left Behind Act of 2001. Public Law 107-110. Retrieved from http://www.ed.gov/policy/elsec/leg/esea02/index.html

U.S. Department of Education (2003). *No Child Left Behind: A parents guide.* Retrieved from http://www2.ed.gov/parents/academic/involve/nclbguide/parentsguide.pdf

U.S. Department of Education (2009a). *Adequate yearly progress.* Retrieved from http://answers.ed.gov

U.S. Department of Education (2009b). *No Child Left Behind*. Retrieved from http://answers.ed.gov

U.S. Department of Education (2010). *SY 2008-2009 consolidated state performance reports: Part I.* Retrieved from http://www2.ed.gov/admins/lead/account/consolidated/sy08-09part1/index.html

U.S. National Commission on Excellence in Education. (1983). *A nation at risk: The imperative for educational reform : A report to the nation and the Secretary of Education, United States Department of Education*. Washington, DC

Vinovskis, M. A. (1998). Overseeing the nation's report card: The creation and evolution of the National Assessment Governing Board. Retrieved from http://www.nagb.org/publications/95222.pdf

Waltman, K. (1997). Using performance standards to link statewide achievement results to NAEP. *Journal of Educational Measurement*, *34*(2), 101-121.