WHAT REALLY HAPPENED IN THE NCLB ERA?

A State-Level Analysis of Achievement as Influenced by

Critical School and Non-school Factors

Jeffrey Michael Fouquet
B.A. Fort Hays State University, 2002
M.A. University of Kansas, 2005

Submitted to the graduate degree program in Educational Leadership and Policy Studies and the Graduate Faculty of the University of Kansas in partial fulfillment of the requirement for the degree of Doctor of Education.

_____

Chair: Dr. Argun Saatcioglu

_____

Dr. Lisa Wolf-Wendel

_____

Dr. Thomas M. Skrtic

_____

Dr. Susan Twombly

_____

Dr. Joe Novak

Date Defended: 23 August 2019

The Dissertation Committee for Jeffrey Michael Fouquet certifies that this is the approved version of the following dissertation:

WHAT REALLY HAPPENED IN THE NCLB ERA?

A State-Level Analysis of Achievement as Influenced by

Critical School and Non-school Factors

Chair: Dr. Argun Saatcioglu

Date Approved: 4 September 2019

ABSTRACT

This study analyzes the results of 4th grade reading and mathematics as well as critical school and non-school factors to determine what happened to student outcomes during the peak years of NCLB, 2003-2009. Converting each state's reported percent proficient to a NAEP scale equivalent (NSE) using equipercentile linking has been done in numerous federally commissioned reports. Such a conversion enables comparisons of student performance across states and over time utilizing a standardized and stable metric. This study adds to the existing literature by examining how each NSE correlates to policy-relevant and reform-relevant school factors such as location (urban/rural), class size, and per pupil revenue in addition to critical non-school factors such as race, income, family structure, and parents' level of education. Furthermore, using each NSE conversion's relative error, or RE, these correlations are weighted down or up using the inverse of RE (1/RE) to determine the validity of the patterns being observed. In fourth grade mathematics, the NSE was negatively affected by percent African American, percent of poverty, and percent of urban and rural schools. In fourth grade reading, the NSE was negatively affected by percent African American, percent urban/rural schools, and class size. This study suggests that policies aimed at improving student outcomes have a deleterious impact on already disadvantaged students, and that policy decisions based on state-reported percent proficient have the unfortunate consequence of masking lower standards of learning for these same students.

Keywords: assessment, NAEP, NSE, NCLB, achievement, reform, education policy

ACKNOWLEDGMENTS

This dissertation would not be possible without the influence of several people who deserve my gratitude.

First, I'd like to thank my advisor, Argun Saatcioglu, for making this not only possible, but enjoyable. Even when the process was frustrating, being able to have real talk about the problems facing education always reminded me why I hope to have more influence over the experiences of my students. Thanks for your candor, intellect, and insights.

Next I'd like to thank my maternal grandmother, Doneta Parsons. She didn't know anything about college or what it would require, but she dragged me there, essentially against my will, because she knew I was capable of living a better life than the one I'd reluctantly accepted. She's gone now, but none of this would have been possible without her unconditional, unwavering – and at times, unfounded – belief in my potential.

To my grandfather, Calvin Parsons, who taught me the value of hard work and honoring my commitments: my decisions may not always be popular, but I always stay true to my own values because of your no-nonsense, fearless approach to living – what a gift that has been.

And finally, to my dearest Shanon. Your support through this has meant everything to me, and I hope the growth that follows fulfills me in ways that spill over into our shared life together; that's always been, and will always be, my greatest commitment.

Thanks, Nabs.

TABLE OF CONTENTS

Chapter 1: Introduction

1.1 Purpose of this Study

The central challenge in evaluating No Child Left Behind (NCLB) has been associated with the absence of a uniformly common set of outcomes across the 50 individual states, each of which was charged with developing its own assessment system (tests, reporting conventions, evaluation practices, etc.) aligned with its own statewide curricular standards. This has prohibited an overall national assessment as well as between-state comparisons (Rothstein, 2004). Moreover, since assessment systems and curricula can and do change over time, even longitudinal evaluations for the same states have been problematic (Cronin et al., 2007, 2009).

This dissertation takes a critical step back to address what happened to student outcomes in NCLB's peak years, from 2003 to 2009: Was the No Child Left Behind Act (NCLB) successful in accomplishing its basic goal ("to leave no child behind")? Did states improve student performance? How did they vary in this regard and why? What factors explain differences across the states and over time?

Although the federal government has an established history of using funding strategies to influence the priorities and curricula of local schools, the 2002 reauthorization of The Elementary and Secondary Act – more commonly known as No Child Left Behind – ushered in an era of high-stakes accountability that included a cascading menu of punitive sanctions for schools that failed to achieve performance targets. Ostensibly, this reform was designed to provide historically disadvantaged students with a better education and close the achievement gap between minority students and their white classmates (Lowell, 2004; Hursh, 2007; Loveless, 2007). But decades of education reform research dating to the landmark Coleman Report (1966) provide compelling evidence that the most significant variables influencing student learning lay outside the educational sphere of control (see also Coleman, 1988; Jeffrey, 1978; Davis-Kean,

2005; Darling-Hammond, 2007; Sharkey and Elwert, 2011).

Social scientist Donald Campbell established Campbell's Law, a theoretical premise asserting that "the more any quantitative social indicator is used for social decision making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it was intended to monitor" (Nichols and Berliner, 2007, pp. 43). Similarly, Holmstrom and Milgrom (1991) show that incentives based on quantifiable measures will lead organizations to prioritize the most easily observable aspects of a complex task. As a result, any system of educational accountability that fails to account racial, economic, or regional disparities – differences that manifest themselves in critical school and non-school characteristics – incentivizes an organizational response that is favorable to the self-preservation and good standing of the organization.

In this dissertation, I will evaluate the effectiveness of NCLB by using both the state-reported proficiency information and the results of each state's performance on the National Assessment of Educational Progress (NAEP), the gold standard in educational assessment. I utilize a novel measure, the NAEP scale equivalency (NSE) utilized in numerous federally commissioned studies to map the results of fourth grade mathematics and reading onto the more stable and standardized scale of the NAEP.

However, this study goes one step further by examining how each NSE correlates to policy-relevant and reform-relevant school factors such as location (urban/rural), class size, and per pupil revenue in addition to critical non-school factors such as race, income, family structure, and parents' level of education. Furthermore, using each NSE conversion's relative error, or RE, these correlations are weighted down or up using the inverse of RE (1/RE) to determine the validity of the patterns being observed. This provides a critical degree of accuracy in estimated effects.

1.2 Research Hypotheses

Because decades of research have shown that SES-related disparities are a major factor in performance gaps between have and have-nots in the U.S., I hypothesize that:

> ***Hypothesis 1:*** *Inequities associated with key non-school factors (racial/ethnic and SES backgrounds) are negatively related to learning outcomes under NCLB.*

Additionally, because achieving the desired levels of academic proficiency can be impacted by relevant school factors such as location, teacher expectations, and per pupil expenditures, I hypothesize that:

> ***Hypothesis 2:*** *Inequities associated with key school factors (location, class size, and funding) are negatively related to learning outcomes under NCLB.*

1.3 Significance of the Study

High stakes accountability remains a central factor in the logic of education reform. States continue to test their students in selected subjects, publicize results, and rely on the results to guide policy and administrative actions in pressuring schools to improve student learning. The federal No Child Left Behind Act of 2002 (NCLB) – the pinnacle of the standards and accountability movement – was reauthorized in 2016 as the Every Student Succeeds Act (ESSA) despite significant questions about improvement in student achievement. Because studies often focus solely on *whether* high stakes accountability improves achievement, the question of *why* outcomes may vary across states and over time tends to be neglected. Addressing reasons for variance in policy outcomes is integral to a more complete understanding of how to genuinely improve student learning. Non-school factors related to race

and economics, as well as critical school factors such as location, class size, and revenue all influence the learning outcomes states are able to achieve.

An analysis of these factors will show that achieving equity in education has more to do with addressing the social and economic disparities in our society rather than continuing to blame the inadequacies of an educational system that has been subject to continuous reform efforts for at least 150 years.

Chapter 2: Review of Literature

This dissertation seeks to examine whether No Child Left Behind (NCLB) achieved its stated purpose "to leave no child behind." By using a novel measure, the NAEP scale equivalency (NSE), I analyze achievement data for fourth grade mathematics and reading from the key years of NCLB, 2003-2009, to determine if this reform resulted in higher levels of learning, as well as to uncover to what extent relevant school and non-school factors were related to student achievement.

This study seeks to examine two key hypotheses:

> ***Hypothesis 1:*** *Inequities associated with key non-school factors (racial/ethnic and SES backgrounds) are negatively related to learning outcomes under NCLB.*
>
> ***Hypothesis 2:*** *Inequities associated with key school factors (location, class size, and funding) are negatively related to learning outcomes under NCLB.*

In order to examine these hypotheses and the extent to which NCLB was able to achieve its intended outcomes, this study relies on several key literatures. An overview of the history of the standards and accountability movement, including promises of increased federal funding will show how previous attempts at accountability had struggled to have a measurable impact. That relative ineffectiveness can then be sharply contrasted with the punitive approach to accountability characteristic of the peak years of NCLB. A review of organizational sociology will show that even under these impossible proficiency mandates, the states may have been compelled to demonstrate – in sometimes symbolic or ceremonial ways that had detrimental impacts on student learning – their compliance with federal law. Finally, an analysis of the non-

school and school inequities that are difficult for any system of education to overcome will suggest that some states have characteristics that are more likely to exert a negative influence on performance outcomes.

2.1 History of Standards and Accountability

The 2002 reauthorization of the Elementary and Secondary Education Act (ESEA), commonly referred to as No Child Left Behind (NCLB) continues to be a controversial moment in the history of educational reform (Darling-Hammond, 2007; Hombo, 2003). Prior to this law, the federal government had used funds to influence education (a state responsibility) for decades. Earlier educational reforms such as The Smith-Hughes Act of 1917 and the National Defense Education Act of 1958 had also introduced new standards to public education that were designed to support federal initiatives (Hillson, 1995; Davies, 2007). These earlier initiatives were accompanied with additional money to support implementation of the new programs and all were passed with relatively little pushback from local stakeholders (Tyack & Cuban, 1995).

In 1965, the Johnson administration's passage of the Elementary and Secondary Education Act (ESEA) was a controversial foray into state affairs by federal powers at that time (McLaughlin, 1975). Because this law attempted to build upon the promise of the Civil Rights Act of 1964 and expand Johnson's "War on Poverty," the ESEA experienced some difficult days before becoming law. When it passed, those federal dollars were used to expand equal access to quality education, specifically for historically disadvantaged students (McLaughlin, 1975; Jeffrey, 1978), but accountability remained difficult to define.

Subsequent presidential administrations began to disparage the inadequacy of public education and pushed through incremental advancements without overtly threatening state and local control. This decades-long discourse began with the release of the Reagan Administration's

A Nation at Risk (Gardner, 1983). Published in the context of an ongoing Cold War with the Soviet Union, this report warned of the "rising tide of mediocrity" in America's schools and tried to emphasize the connection between strong schools and national security. In this politically conservative context, the evaluation of schools had a decidedly business-like model, including ideas such as outputs and quality control (Sahlberg, 2011). This explosion in purpose statements defining what students should be learning lead to the standards movement so prevalent today; the corresponding ideas as to how quantifiable outcomes and achievement data could measure a school's success would eventually solidify into the accountability movement (Fuhrman, 2001).

The first Bush Administration proposed its own version of standards and accountability in "America 2000," which advocated national standards, school choice through the development of 535 New American Schools, and alternative licensure for teachers (America 2000); however, in the process of gaining bipartisan support to pass this bill into law, the core provisions of accountability had been stripped, and schools no longer had to participate in assessments in order to receive the federal support ("George H. W. Bush").

Similarly, the Clinton Administration tried to capitalize on the national appetite for "systemic reform" when it authorized the Goals 2000 legislation in 1994. This level of accountability made some uneasy, and both Republicans and Democrats expressed concerns over federal overreach. As a concession, Clinton abandoned his hopes of developing a National Education Standards and Improvement Council, the governing body that would have vetted the standards and assessments developed in each state. Consequently, "the act could not assure that state standards were of uniformly high quality" ("Clinton Years"). This equated to federal funding that required only "good faith" compliance and failed to substantively improve on the previous administration's policies, still falling far short of *A Nation at Risk*'s demand for greater accountability in education.

Nichols and Berliner (2008) argue that many of the facts and figures presented in *A Nation at Risk* were at best, flawed and at worst, misleading. However, because the administration had signaled that federal funds were contingent upon compliance, state-level policymakers responded by raising standards for teachers and students (Fuhrman, 2001). In response, governors commissioned studies to assess how well schools were preparing students (Kress et al., 2011). Within ten years, Texas, Massachusetts, and North Carolina would pass school accountability laws (Kress et al., 2011). By 1996, 13 states had accountability systems; by 2000, 39 had adopted ways of measuring student growth and school effectiveness (Hanushek & Raymond, 2003). Despite these symbolic gestures, four presidential administrations had been unable to introduce lasting reform that could address the failures each had decried while on the campaign trail.

In the presidential campaign leading up to his election in 2001, then-candidate and former governor, George W. Bush promised to bring many of his Texas-based practices with him to Washington (Hayes, 2008). The failings of previous administrations were well known by the time he entered the Oval Office, and he found strong bipartisan will to pass the first federal mandate for school improvement. His closest allies, including politicians and private corporations who had a stake in the success of NCLB, argued that without some form of accountability and sanctions – including privatization of services and competitive pressures – there would be no incentives to address what he referred to as the "soft bigotry of low expectations" that was plaguing public education (Metcalf, 2002).

The strategy to gain support for these market-based reforms was simple. When political advisors were outlining an educational platform for then-candidate George W. Bush, the key findings of Reagan's *A Nation at Risk* were still relevant and timely, and there was increased concern for national security and global competitiveness brought on by free-trade agreements

and a decade of outsourcing (Hayes, 2008;). By requiring curriculum standards and assessments and threatening underperforming schools with the harsh, cascading public sanctions outlined in the bill, proponents argued that unmotivated teachers and administrators would find the impetus they needed to impart their subjects on students, redirect resources to support these efforts, and somehow become more efficient in these straight-forward technical processes of skill transmission and student retention (Hanushek & Raymond, 2003).

As has been shown, this focus on standards and accountability had a long history with previous presidents, so when it passed in 2002, the reauthorization of the Elementary and Secondary Education Act of 2001 – commonly referred to as No Child Left Behind (NCLB) – felt a like a bipartisan political victory decades in the making. On a more practical level, it introduced a new era of accountability wherein failure to meet proficiency targets could result in punitive sanctions, both economic and organizational in nature. Under this regulatory system, all students were to take standardized tests, including students with special needs and second-language learners, and a school's future rested on its ability to demonstrate progress toward the goal of 100% proficiency by 2014, regardless of the challenges it faced.

2.2 Accountability Under NCLB

Public polls from 1980 indicate that "education was rated twenty-third out of forty-one issues. By 2000, the final year of George W. Bush's presidential campaign, education was named the most important problem facing the nation" (as cited in Hayes, 2008, pp. 10). Sponsors of NCLB unified popular and political consensus by capitalizing on the perception that the academic needs of minority students, second-language learners, and children with special needs were not being met (Darling-Hammond, 2007; Hayes, 2008; Hursh, 2007; Dee & Jacob, 2011). Proponents argued NCLB was a long-overdue response to middling academic performance and

an attempt at satisfying the public expectation for educational excellence and equity. Building off the purported success of the controversial Texas system of standards, assessments, and outcomes, NCLB was signed into law, introducing *federal* provisions for testing and assessment, academic accountability, and the threat of corrective measures for persistently struggling schools (Tollefson, 2008).

For the first time in history, federal funding of public education required that schools administer approved tests in English and math[1] in grades 3-8 and once in high school. As a reauthorization of the ESEA of 1965 – the act that created Title I funds meant to improve the academic performance of low-income and minority students – NCLB promised to hold states accountable for closing the achievement gaps of minorities and other historically underperforming subgroups (Hursh, 2007). Consequently, the law insisted that student performance be disaggregated and analyzed by racial subgroups, economic indicators, second-language learners, and students with special needs (Hayes, 2008). To mitigate complaints of federal overreach, the Bush administration made it clear that it had no role in setting the curriculum, nor would it provide a precise, uniform definition of proficiency; states were left to determine their benchmarks, proficiency levels and assessment instruments. Beginning in 2003, schools needed to show Adequate Yearly Progress (AYP) toward the lofty goal of achieving 100% student proficiency by 2014 to remain in compliance with NCLB.

Once students had taken their respective state assessments, the results of each school's performance were made publicly available on an "Annual Report Card." Parents, community members, students, and teachers would be able to see if their schools were making AYP toward the long-term proficiency goals. The achievement data was also reported for each of the

[1] Although the law also requires tests in science and social studies, controversies surrounding funding made the administration of these assessments inconsistent; as such, they are not considered defining aspects of the legislation.

subgroups mentioned previously, so the public could apply pressure on schools to close the achievement gaps or improve the effectiveness of special programs. This public shaming of failing schools was based on market ideology, which endorses competition and the threat of consumer choice as appropriate and effective motivating factors that can turn an organization around – or justify its demise. Because provisions in the law allowed students of failing schools to "shop the market" for a higher performing school, experts argued that the fear of failure would motivate public educators to do what they had not yet done (Lowell, 2004; Metz, 2008; Hombo, 2003).

For schools unable to make AYP or wiggle through a loophole, public, political, and market pressures could be applied to "compel" them to strive harder to show academic gains. After too many years of failing to make annual yearly progress (AYP), schools could be shut down or restructured under charter management or other for-profit organizations. As part of the NCLB legislation, schools were to reserve 20% of Title I funds to offer students the option to transfer to higher performing schools within the district and to pay for supplemental educational services (SES) from private providers free of charge (Henig, 2007; NCLB Action Brief, 2003). If students chose to exercise the NCLB provision for school choice, schools could use up to 15% of Title I funds on transportation for these students; at least 5% of title funds were designated for approved SES. Figure 1 summarizes the full list of cascading corrective actions for those schools failing to meet AYP (Manna, 2007).

A notable element absent in the framework shown in Figure 1 are incentives for meeting AYP. Tyack and Cuban (1995) note the disproportionate use of "the stick" (threat of punitive action) versus "carrots" (incentives for achievement) in educational reform efforts. Such trends betray an undercurrent of distrust that has come to characterize the discourse surrounding public education, painting the trouble with schools as a lack of professional will rather than inequitable

Figure 1. Sanctions for schools not meeting AYP

Not meeting AYP	Recommended Sanctions under NCLB
First year	1. Noted in "Annual Report Card" 2. Warning of further sanctions if AYP is not met
Second year	1. Designated *In Need of Improvement* in "Annual Report Card" 2. Parents are notified and children are offered school choice 3. District is mandated to create PD plan to address issues 4. 10% of Title I funds must be used on research-based PD
Third year	*Penalties 1-4 carryover from previous year, plus* 5. 20% of Title I funds used for supplemental educational services
Fourth year	*Penalties 1-5 carryover from previous year, plus* 6. "Corrective Action" taken: administration, schedule or curriculum
Fifth year	*Penalties 1-5 carryover from previous year, plus* 7. "Restructuring Plan" must be developed and submitted to the state
Sixth year	"Restructuring Plan" is enacted, turning the school over to charter management organization, replacement of current personnel, or other forms of external management or state intervention

cultural capital and advantage (Lowell, 2004). As stated earlier, since the Reagan administration's endorsement of *A Nation at Risk* in 1983, inflammatory, cautionary rhetoric has been used to portray the need for urgent, drastic measures and tighter controls of an inefficient bureaucratic institution. Sunderman, Kim and Orfield (2005) agree that effects of the policy – of its high-stakes, test-driven accountability – are to exert control. These punitive sanctions were the long and looming shadows introduced by accountability under NCLB, and these harsh conditions compelled schools already fighting to close the achievement gap to find new ways to meet these tough new federal mandates (Mathis & Trujillo, 2016).

2.3 Strategic Institutional Responses

Organizations are primarily interested in survival (Meyer & Rowan, 1977). In order to continue functioning, secure a portion of the available resources, and protect the livelihood of those in the organization, every organization must "justify to a peer or superordinate system its

right to exist" (Dowling & Pfeffer, 1975, pp. 123). This unspoken justification is called *legitimacy*, and the most common way to establish this currency is to conform to established values or expectations. However, decades of educational and sociological research suggest that genuine efforts to improve the results of public education face significant obstacles outside the school's sphere of influence (Coleman, 1966).

Symbolic compliance with organizational expectations and preserving legitimacy are important for organizations that are resource dependent (Pfeffer & Salancik, 1978). Any time a new educational reform is introduced, this is another opportunity for schools to demonstrate their legitimacy (Tyack & Cuban, 1995). From the perspective of institutional survival, implementing reform not only satisfies legal requirements, it signals efforts to satisfy the public's desire for effective and equitable schooling (Labaree, 2006). But in largely hierarchical and loosely coupled systems like public education, systemic change is difficult to monitor and enforce (Tyack, 1990). And any system of accountability that penalizes schools for poor performance without also acknowledging those factors that lay outside the educational sphere of control incentivizes a strategic response that takes precedence over bestowing the broad range of academic skills the standards and accountability movement is designed to promote (Bohte & Meier, 2000; Dobbin & Sutton, 1998).

Schools are state-governed institutions, so NCLB had no real authority to impact curriculum and instruction, standardize state assessments, or define proficiency in ways that could have impacted the quality and level of student learning (Metz, 2008). The law was written without a clear and objective definition of proficiency, let alone a prescriptive set of universally adopted methods or instruments for measuring whether or not students were actually meeting these targets (Loveless, 2007). This level of ambiguity in the legal framework paved the ways for states to define proficiency in largely symbolic and favorable ways that made it difficult to

ascertain whether proficiency scores were actually measuring academic improvement (Metz 2008). Because implementation was left to each state, reform-driven changes in curriculum, instruction, and assessments led to differing definitions of proficiency and inconsistency from year to year meant to satisfy the legal requirements of the legislation (Cronin et al., 2007, 2009; Edelman et al., 1999).

The locally subjective interpretation of assessment and proficiency had proven problematic even before NCLB in many of the states said to be pioneers in the proficiency-based accountability movement. In his analysis of math assessments from four different states, Porter (2002) found that the mean assessment-to-standard alignment was 0.40 on range of zero to one. In a similar study, Resnick, Rothman, Slattery and Vranek (2003) found that assessments in both math and reading significantly restricted the standards students were expected to master – some as low as 27 percent (see also, Wixson, Fisk, Dutro, & McDaniel, 2002). In all of these cases, the assessments exhibited wide disparities not only in the enacted curriculum (what gets taught), but also in the definition of proficient (the standard for student achievement) – and these realities limit students' opportunities to achieve higher levels of learning (Berliner, 2009).

Even before NCLB, high-stakes accountability had negatively impacted the students that should have been learning more under the spirit of this reform. Koretz and Barron (1998) found that teachers in Kentucky lowered standards for learning by reducing instructional time in math and science to dedicate more time to the tested subjects. Jacob (2002) discovered that teachers responded strategically to the accountability system adopted in Chicago in the 1990's through exclusionary practices such as increasing special education placement and the rates of retention – practices which were unjust for those students whose grade-level learning was affected. Similar patterns of exclusion were examined in Texas, and studies there found this phenomenon disproportionately in the grades that test as part of the accountability system (Figlio & Getzler,

2006; Deere & Strayer, 2001a, 2001b; Cullen & Reback, 2006).

Perhaps it is no surprise that there is ample evidence of a wide variety of similar strategic (and often misleading) institutional responses during NCLB (Linn, 2003; Hornback, 2013; Darling-Hammond, 2007; Peterson & Hess, 2006). Although the law was intended to close the achievement gaps between certain subgroups, instruction was often focused on "bubble kids" – students who were close to demonstrating the desired proficiency levels – at the expense of other students' learning (Booher-Jennings, 2005; Jennings & Sohn, 2014). Other researchers have documented manipulation of the testing pool either through "push out" practices that encourage low performers to drop out of school before the testing dates (Hursh, 2007), increased rates of suspension for low performing students (Figlio, 2006), or disproportionate increases in the rate of students identified as "disabled" who can either be exempted from taking the test or given an easier, modified assessment (Figlio & Getzler, 2006; Cullen & Reback, 2006). Jacobs and Levitt (2003) and Copeland (2013) reported the most extreme instances of teachers trying to preserve the legitimacy of their organizations by distributing test answers prior to the assessment or changing selected responses to the correct answers after students have concluded their tests.

Multiple studies suggest that another way states complied with these institutional pressures for legitimacy was by inflating measures of student proficiency (Linn, Graue, & Sanders, 1990; Shepard, 1990; Koretz, Dunbar, & Shepard, 1991; Stecher & Barron 1998; Klein et. al., 2000). By arbitrarily adjusting the standards for proficiency and manipulating the reporting conventions, states could symbolically comply in order to maintain their standing and remain eligible for the external resources that would ensure their organizational survival. Such patterns of strategic response are often used to explain why student achievement has been slow to improve on low-stakes national assessments compared to the impressive proficiency gains reported on high-stakes state assessments (Barton & Coley, 2010; Dee & Jacob, 2011).

Unfortunately, over-reporting percent proficient also creates the unintended consequence of restricting the range of skills students are being taught or are expected to master.

In the high-stakes context of NCLB, with the threat of punitive sanctions looming overhead, many states found ways to meet the requirements of this federal mandate by exploiting the ambiguities in the law (Edelman, Uggen, & Erlanger 1999; Dobbin & Sutton, 1998). These schools did not deliberately choose to impact students and student learning in negative ways, but accountability pressures, compounded with the difficulty of overcoming racial and economic variables, incentivized strategic and symbolic responses that allowed these systems and states to demonstrate their compliance with the mandates of NCLB.

2.4 Understanding School and Non-school Factors

Proficiency-based accountability lacks a thorough consideration of whether public schools alone have the capacity to ensure excellence and equity for all children (Mehta, 2013). This elusive capacity is impacted by numerous non-school and school factors that can have negative consequences on the ability of any actor – be it the teacher, the school, or the state – to improve student outcomes. Campbell's Law states that “the more any quantitative social indicator is used for social decision making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it was intended to monitor” (Nichols & Berliner, 2007). In highly complex and non-standardized institutions like public schools, any system of accountability that fails to consider the racial and economic disparities that impact a student’s physical and cognitive development (Brooks-Gunn, Klebanov, & Liaw, 1995; Guo &Harris, 2000; Votruba-Drzal, 2003), social capital (Ream & Palardy, 2008; Coleman, 1988), and critical attitudes (Sampson, Morenoff, & Gannon-Rowley, 2002; Davis-Kean, 2005) in ways that make achieving high levels of academic success difficult is going to be

subject to corruption.

Under NCLB, states were pressured to demonstrate "proficiency" on state assessment regardless of the state-level school and non-school factors they were already struggling to overcome. At the same time, they were being told that failure to show improvement would result in severe penalties, up to and including government takeovers of struggling schools. Even in states facing challenges such as racial diversity, low SES, or systemic problems like large class sizes or low per pupil funding, this kind of threat incentivized a variety of institutional responses (Holmstrom & Milgram, 1991). In order to preserve resources and legitimacy, some actors may have found ceremonial and symbolic compliance – satisfying the letter of the law – easier than achieving the spirit of the reform. These kinds of institutional responses can be beneficial in terms of legal compliance, but they also hurt the very students the reform was designed to benefit (Berliner, 2009; Dee & Jacob, 2011; Edelman et al., 1999). This is especially true in states affected by key non-school and school factors that are historically difficult to overcome.

Public schools have limited control over several non-school factors, which undermines their ability to guarantee high achievement for all learners. Because public policies should evaluate agents – in this case, states – on results these agents can meaningfully influence, NCLB's plan to sanction states for failure was an ill-advised approach (Harris, 2011). In many cases, such undue pressure to improve student learning without regard for established and historical inequities prompted strategic responses that satisfied the letter of the law. In other words, finding ways to signal compliance with the mandates became more important than actually improving student learning (Bohte & Meyer, 2000; Edelman et al., 1999).

Non-school inequities can be considered broadly in terms of those related to race/ethnicity and socioeconomic status (SES). The share of underserved racial/ethnic groups in the state population, such as African Americans and Hispanics, is critical in this regard, because

outcomes for children in these groups are disproportionately affected by problems outside the school, such as residential isolation (Sharkey & Elwert, 2011), community social disorganization (Wilson, 1996), poor health (Williams & Collins, 2001), low quality social services (Iglehart & Becerra, 2011), disenchantment with mainstream pathways to success (Mickelson, 1990), and discrimination (Blum, 2002). These are key obstacles to closing racial/ethnic achievement gaps (Barton & Coley, 2010). While low SES often intersects with racial/ethnic disadvantages, family economics can also be problematic in its own right (Owens, Reardon, & Jencks, 2016), particularly in terms of the effects of low parental income, education, and efficacy.

Low income constrains access to materials (Mayer, 1997), extracurricular activities (Lee & Burkham, 2002), quality healthcare and residential facilities (Rothstein, 2004). It can also hinder basic endowments such as birth weight, undermining cognitive growth (Brooks-Gunn, Klebanov, & Liaw, 1995). These problems can be coupled with parental stress, with dire outcomes for child development (Guo & Harris, 2000). Moreover, low income is associated with exposure to neighborhoods with limited social control and negative peer groups (Sampson, Morenoff, & Gannon-Rowley, 2002). As for low parental education, it can limit cognitive stimulation at home (Votruba-Drzal, 2003) and support for activities such as homework (Posner & Vandell, 1994). It can also undermine parental expectations for children's performance as well as guidance for future success (Davis-Kean, 2005). Finally, low parental efficacy impedes school involvement.

While families can support students at home, effective parenting also can bridge home and school, such as interacting with teachers and advocating for children's success (Lareau and Horvat, 1999). This "social capital" potentially compensates for other SES disadvantages (Ream & Palardy, 2008), but it can be undermined by inequities in time and effort parents can allocate to school involvement (Rothstein, 2004), limits on parental resiliency (Amatea, Smith-Adcock,

& Villares, 2006), and shortages in social support for parents (Hashima & Amato, 1994). In all, SES-related disparities are a major factor in performance gaps between haves and have-nots in the U.S., in some ways above and beyond racial/ethnic gaps (Reardon, 2011).

Ultimately, the greater the proportion of minority and/or low SES students in a state, the more difficult it can be for the state as a whole to improve percent proficient. This may motivate officials to compromise state content standards and/or narrow state assessments to facilitate proficiency gains (Bohte & Meyer, 2000; Edelman et al., 1999). In response, districts and schools serving largely minority and/or low SES students may adopt curricula narrower than those in other settings. Also, teachers in such settings may teach to narrow assessments more consistently. Given racial/ethnic and SES segregation, both within districts (Orfield & Eaton, 1996) and increasingly between them (Clotfelter, 2004; Owens, Reardon, & Jencks, 2016), these curricular and instructional patterns may not limit opportunities for many white or middle-class students. Nonetheless, proficiency increases resulting from restriction of learning opportunities for minority and/or low SES students' may help states comply with legal mandates at the same time these "gains" are hurting students.

Similar dynamics can result from disparities in school quality as well. Tendencies for adopting narrower curricula and/or teaching to narrower assessments can be driven by proficiency struggles in schools lacking adequate resources and other internal features necessary to help raise achievement (Elmore, 2003). The first factor to consider at the state level is the share of students in urban and rural schools. These schools experience problems in raising achievement partly because they serve predominantly racial/ethnic minority and/or low SES students affected by greater non-school problems, a key reason why non-school disparities and disparities in school quality are linked. But urban and rural schools can aggravate low achievement in their own right, due to lack of adequate funding (Baker, Sciarra, & Farrie, 2010),

lower teacher quality and expectations (Lankford, Loeb, & Wyckoff, 2002), disruptive climate (Bryk & Schneider, 2003), and Inequities associated with key peer effects (Hanushek, Kain, & Rivkin, 2009).

Another key school factor that can impact student learning is teacher-to-pupil ratio. The larger the size of the average classroom, the more difficult it can be for a teacher to cover a larger set of skills due to potentially greater variety of student needs. This increases the odds of teacher stress and turnover (Russell, Altmeier, & Van Velzen, 1987) and can thus impede student performance (Nye, Hedges, & Konstantopoulos, 2000) and complicate compliance with proficiency mandates.

Finally, school quality can vary across states as a function of per pupil revenue from state and local (district) sources. This is a fiscal expression of state-specific capability and/or willingness to fund public schools (Gramlich & Rubinfeld, 1982) and is affected by the state's income and wealth, as well as its emphasis on education relative to other services such as law enforcement, healthcare, and highways. States with lower per pupil state and local revenue spend less on hiring high quality teachers and provide students with lower quality programs and facilities (Card & Payne, 2002). They also allocate less to students whose needs cost more (Duncombe & Yinger, 2008). To the extent that these factors undermine achievement, they can stimulate skill range restriction to facilitate proficiency gains. Such patterns can be aggravated by failures to couple proficiency mandates with supporting funds. NCLB's failure to provide sufficient funding for after-school tutoring and other supplemental services to support achievement has been viewed by many state and district officials as a critical obstacle to achieving genuine proficiency gains (Center on Education Policy, 2006, 2007).

In states where large disparities in school quality might motivate officials to compromise the rigor of assessments and/or narrow the scope of curriculum to facilitate proficiency gains,

these shifts do not have the same impact on students attending better-resourced schools. Since most urban and rural schools function jurisdictionally apart from their suburban counterparts, and since class size and per pupil revenue vary within a state, districts and schools where markers of low school quality are less prevalent may not limit their curricula to basic statewide content standards, and they may not limit instruction to skills covered by state assessments. Rather, to the extent that proficiency gains for the state are the result of strategic responses, these patterns may reflect harms endured largely by students in lower quality schools.

2.5 Examining the Impact of NCLB

What really happened to achievement in the NCLB era? Did states improve in student performance? How did they vary in this regard and why? What factors explain differences across the states and over time? These questions remain open for two reasons, both of which this dissertation seeks to address.

As noted earlier, studies have lacked a proper metric for examining achievement under the high stakes accountability regime since assessments and curricular guidelines, as well as the related proficiency criteria vary across states and over time (Rothstein, 2004). The lack of a common and stable metric has been an obstacle in evaluating high-stakes accountability. Determining the true degree of progress denoted by state-reported outcomes has been a challenge in the absence of a common, stable frame of reference for achievement. Under the NCLB (and now ESSA) framework, states are required to *develop their own* achievement standards and related assessments and proficiency criteria (Loveless, 2007). But frequent changes in these elements complicate efforts to track progress for any given state over time (Nichols and Berliner, 2007; Cronin, Dahlin, Xiang, & McCahon, 2009). Additionally, between-state variance prohibits comparative and nationwide analyses. As Mehta (2013) stresses, lack of uniformity in curricular

guidelines, tests, and reporting practices is the most ironic feature of high-stakes accountability as a "test-based" national reform policy.

A few notable studies (e.g., Dee & Jacob, 2011; Lee & Reeves, 2012) have relied on results of the National Assessment of Educational Progress (NAEP), which are comparable across states and over time, but this approach is open to criticisms of conceptual misalignment with the high stakes accountability paradigm because NAEP is not a test designed for accountability purposes. Studies utilizing NAEP provide compelling evidence on achievement shortfalls under NCLB, which likely will recur under ESSA (Mathis & Trujillo, 2016). But these studies remain open to the criticism that NAEP is inappropriate for evaluating NCLB or any other accountability policy because it is not *designed* for accountability purposes and that the states do not pressure schools to improve on NAEP in the same way they pressure schools to improve on state tests (e.g., Popham, 2005; Phelps, 2005; Stoneberg, 2007). A state assessment ideally is used to detect struggling students in the state and guide remedial strategies while NAEP is less sensitive to these efforts as it is developed independently of state-specific problems and is not designed to guide remedies (Cizek, 2005). From this standpoint, discrepancies between NAEP and state assessment outcomes may reflect the modest success of this accountability policy, not its failure.

For those who ascribe to this view, the very features of NAEP that are attractive in evaluating NCLB – its low-stakes, rigorous, and independent nature – represent a conceptual misalignment with the high-stakes accountability paradigm. Some go as far as to suggest that expecting significant changes in NAEP outcomes in response to a policy like NCLB is "antithetical" to the thinking behind NAEP because NAEP's designers ensure that scores are unaffected by local curricular and instructional changes over short periods such as three, five, or even ten years (Popham, 2005). NAEP, it is argued, is constructed for long-term trend analysis,

not for short-term policy evaluation. This argument implies that, despite the broad consensus on NCLB's ineffectiveness, there may be a hidden virtue behind most curricular, assessment, and instructional decisions in the states that lead to progress in learning which NAEP cannot fully or properly detect. Accountability advocates argue that narrower (or more "focused") curricula and assessments, and teaching to the test are essentially unproblematic if they help schools concentrate energies on well-defined goals for each grade level, which in the aggregate would benefit students (Firestone, Schorr, & Monfils, 2004).

In this study, I utilize a novel measure that overcomes the challenges: the NAEP scale equivalent, or NSE, for the state-reported proficiency level. Commissioned by the National Center for Educational Statistics (NCES) and developed by members of the American Institutes for Research, this measure was created as a way to convert state-reported proficiency levels onto a more standardized scale. Different than a state's raw NAEP score, the NSE is the translation of the state's percent proficient (on its own assessment) to the NAEP scale based on the performance of students in the state's NAEP sample, which includes students who take both the state assessment *and* NAEP. NSE enables examining not just between-state and longitudinal variance in achievement, but also overall (nationwide) patterns over time. Moreover, analyses using NSE can be adjusted for problems in converting the state's percent proficient to the NAEP scale (e.g., differences in skills assessed by the state and NAEP, testing conditions, etc.).

According to NCES, such analyses are useful for the following reasons:

> First, they allow each state to compare the stringency of its criteria for proficiency with that of other states. Second, mapping analyses inform states whether the rigor of their proficiency standards as represented by NAEP scale equivalents changed [over time]. Significant differences in NAEP scale equivalents might reflect changes in state assessments and standards and/or other changes such as

> changes in policies or practices that occurred between the years. Finally, when key aspects of a state's assessment or standards remained the same, these mapping analyses allow NAEP to corroborate state-reported changes in student achievement and provide states with an indicator of the construct validity and generalizability of their test results. (Bandeira de Mello, Blankenship, & McLaughlin, 2009, pp. 9)

In the pages that follow, I will outline this theoretical framework, describe the data and methods utilized during the course of this dissertation, and discuss my analysis and findings. I will finish with a discussion of implications for future research and public policy.

Chapter 3: Data and Methods

3.1 Goals of this Dissertation

This dissertation takes a critical step back to address what happened to student outcomes in NCLB's peak years, from 2003 to 2009. Was the No Child Left Behind Act (NCLB) successful in accomplishing its basic goal ("to leave no child behind")? Did states improve student performance? How did they vary in this regard and why? What factors explain differences across the states and over time?

The central challenge in evaluating NCLB has been associated with the absence of a uniformly common set of outcomes across the 50 individual states, each of which was charged with developing its own assessment system (tests, reporting conventions, evaluation practices, etc.) aligned with its own statewide curricular standards. This has prohibited an overall national assessment as well as between-state comparisons (Rothstein, 2004). Moreover, since assessment systems and curricula can and do change over time, even longitudinal evaluations for the same states have been problematic (Cronin et al., 2007, 2009).

In this dissertation, I will evaluate the effectiveness of NCLB by using both the state-reported proficiency information and the results of each state's performance on the National Assessment of Educational Progress (NAEP), the gold standard in educational assessment. I utilize a novel measure, the NAEP scale equivalency (NSE) for state-reported proficiency, before examining the relationship of state NSE with fundamental non-school and school factors that characteristically underlie achievement patterns. This approach constitutes a direct test of how learning outcomes were related to non-school and associated school factors, a test that no study to date has provided.

3.2 NAEP Scale Equivalence (NSE)

In this study, I use NAEP in a novel way that transcends criticisms of earlier work relying on NAEP to evaluate NCLB. My approach compliments this earlier body of work. I utilize the NAEP scale equivalent (NSE) for percent proficient reported by the state given its own achievement standards (its own assessment, curricular guidelines, and proficiency criteria). NSE, reported in numerous federally commissioned reports (Bandeira de Mello, 2011; Bandeira de Mello, Blankenship, & McLaughlin, 2009; McLaughlin et al., 2008a, 2008b) is a *translation* of the state's percent proficient to a NAEP score based on the performance of students in the state's NAEP sample (the students who take both the state assessment and NAEP). Rather than denoting achievement on NAEP itself, NSE expresses the state-reported percent proficient as a NAEP score (see Table 4 in Appendix). Thus, using the NSE to examine achievement does not amount to using actual NAEP outcomes. It involves using the measure most pertinent to NCLB – that is, percent proficient – by way of converting that measure to a scale common across states and stable over time (the NAEP scale).

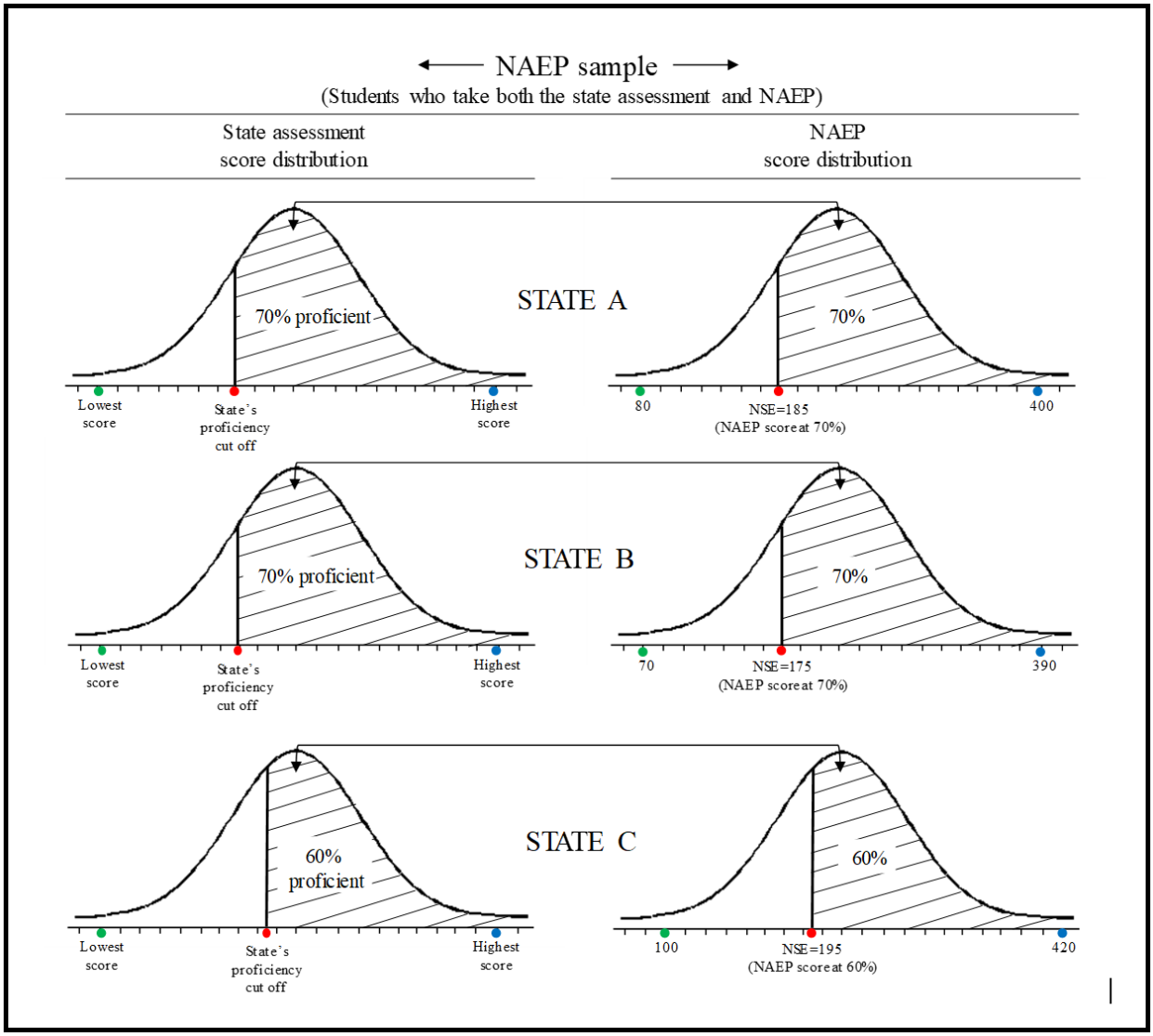
It is important to understand the method of converting each state-reported percent proficient onto the NAEP in these federally commissioned reports; the process is called equipercentile linking. Any time the two same subjects – whether these are the same people, teams, states, or students – are being assessed on different instruments, their scores can be translated onto the other assessments, even if the two assessments do not have any substantive or topical relationships (Holland, 2007; Dorans & Holland, 2007). For example, imagine that ten students take an assessment over American History, and the *same* ten students measure their shoe sizes and chart the distribution for the results of each assessment. Although the results of American History exam and distribution of shoe sizes have *nothing* to do with one another, because the same ten students took both, the results of one assessment can be translated to the

other assessment using these score distributions. If 70% of the students tested pass the American History assessment, that distribution, along with the cut score establishing the 70% pass rate in this example, can be placed onto the results of the distribution of responses for the student's shoe sizes. Although this equipercentile linking can be done, the translation will produce a relative error, which is an estimate of this conversion problem. In this example, the conversion of scores will have a relative error of 1, showing that a comparison of these results is not valid because the substantive connections of what is being measured are not related in any way (Thissen & Wainer, 2001).

NSE is constructed by mapping – through equipercentile linking – percent proficient on the state assessment onto the NAEP score distribution in the same sample (Feuer, Holland, Green, Bertenthal, & Hemphill, 1999; Braun & Qian, 2007). For each tested grade level, a state's NAEP sample includes nearly 2,500 students from about 100 schools that are representative of the state.[i] Because these assessments are dealing with the same population – namely, the 2500 students who took both the NAEP and the state assessment – and because the NAEP sample is rigorously constructed to be representative of state-level characteristics, equipercentile linking is applied here to translate percent proficient to the NAEP score to produce the NSE, as has been done in multiple federal studies (Bandeira de Mello, 2011; Bandeira de Mello, Blankenship, & McLaughlin, 2009; McLaughlin et al., 2008a, 2008b). Because the populations are the same, and because percent proficient and the NAEP are both established measures of academic skills such as reading and mathematics, this methodology is not only practically accomplishable, it is substantively meaningful.

The procedure involves broadly two steps, which can be addressed in reference to State A in Figure 2. First, the percentage of students in the state's NAEP sample who are proficient on

Figure 2: A stylized depiction of mapping percent proficient onto the NAEP



the state assessment is determined based on State A's proficiency criteria, resulting in 70 percent as shown by the curve on the left. This percentage mirrors proficiency in the entire state given the representativeness of the NAEP sample. Next, a cut point on the NAEP score distribution in State A's NAEP sample is identified that provides a proportion equal to percent proficient on the state assessment, resulting in an NSE, shown as 185 under the curve on the right for State A (for more detail, see Braun & Qian, 2007; Bandeira de Mello, Blankenship, & McLaughlin, 2009).

Comparing State A and State B illustrates how two states with the same percent proficient may have different NSEs (175 for State B), given differences in respective NAEP score distributions. Also, as in the case of State C, a lower percent proficient (60 percent) may map onto a higher NSE (195). And a higher percent proficient may map onto a lower NSE (not shown in figure). Finally, the same state's NSE may change over time depending on changes in the state's achievement standards and/or its NAEP score distribution. While state assessments, curricular guidelines, and proficiency criteria vary across states and over time, the stable, nationwide, and rigorous nature of NAEP offers a robust frame of reference for converting state-reported proficiency outcomes to a cross-sectionally and longitudinally common metric. NAEP remained essentially unchanged from 2002 to 2010, NCLB's peak years, on which I focus (see Dee & Jacob, 2011; Jacob, 2007; Kolstad, 2002).

The unique advantage of the NSE is that expresses percent proficient on a common metric, regardless of differences in assessment systems across states or over time. This is because (1) NAEP content and practices remain stable over time (Hombo, 2003; Yeager, 2007), and (2) the percent proficient is taken as "state-reported" regardless of state assessment content, cut score, and curricular alignments. *That means NSE is a robust conversion of whatever a state itself says its proficiency outcome is for a given year and grade level under the NCLB regime*, not an outcome that is pedagogically or operationally unrelated (or loosely related) to NCLB.

The procedure to construct NSE also produces a "relative error" in mapping percent proficient onto the NAEP scale. This error is critical with regard to the "substantive" validity of the mapping. Any two distributions can be equipercentally linked when they are common to the same students, but distributions may differ in terms of the substantive content area they represent. Recall the hypothetical example using equipercentile linking to equate the results of the American History exam with the distribution of the same students' shoe sizes. The results of

one assessment can be translated to the other assessment using their respective distributions, even though the assessments have nothing to do with one another. Relative error varies from 0 to 1, where 1 indicates total invalidity for the mapping. In this hypothetical example of American History and shoe size, the relative error would be 1; the conversion is accomplishable, but it is completely meaningless.

Linking is fully appropriate when the assessed skills are the same but the scales are different (e.g., inches versus centimeters). Since NAEP measures an amalgamation of skills students are expected to learn in all states, the skills it addresses are never exactly the same as the skills evaluated by any given state assessment, which focuses on a narrower set of skills in light of local curricular standards and preferences. Thus, the conversion of percent proficient to the NAEP scale is never fully valid. The relative error (RE) in mapping is an estimate of this conversion problem. It is based on how well the NSE can reproduce percent proficient in individual NAEP schools (for details, see Bandeira de Mello, Blankenship, & McLaughlin, 2009). Every one of these conversions will produce a relative error, but it will never be 1, because this conversion is substantively meaningful because both assessments are of academic skills. Similarly, the RE will never be 0 because the skills assessed on the NAEP are too broad for any state assessment – let alone curriculum – to cover all that the NAEP measures. The RE for these conversions will always be between 0 and 1.

In the broad testing context of high stakes assessments and the NAEP, factors such as differences in student motivation regarding the two tests, in time of test administration (the month), and in-test format (e.g., open-ended versus multiple-choice questions) can also contribute to the conversion problem in limited degrees (Feuer et al., 1999), each producing a relative error. As described below, I use the inverse of the RE (i.e., 1/RE as the basis for observation weights in my estimation, such that observations involving NSEs with greater

mapping errors are weighted down while those with smaller mapping errors are weighted up. This provides a critical degree of accuracy in estimated effects.

3.3. Data Measures

My analysis addresses effects on NSE for fourth grade proficiency in math and reading, between 2003 and 2009. I rely on a state-level panel dataset with two-year intervals starting in 2003, resulting in four time points (n=50x4=200). As noted above, NSE information for states is available in federally commissioned reports (Bandeira de Mello, 2011; Bandeira de Mello, Blankenship, & McLaughlin, 2009; McLaughlin et al., 2008a, 2008b). I obtained corresponding yearly non-school data for each state from American Community Survey (ACS), which reports various state means for adults, ages 30 to 50.[ii] It is critical to note that I deliberately limited my predictors to an essential set of state features even though more variables are available. Because basic characteristics at the state level tend to be strongly correlated and operationally redundant, my study focuses on those school and non-school factors the existing literature suggests would have a strong impact on student achievement.

Racial demographics came from the American Community Survey (ACS) data, and were separated into the two largest minority groups: percent African American (RACE) and percent Hispanic (HISPANIC). As mentioned in the literature review, academic outcomes for these populations are disproportionately affected by a variety of non-school factors (see Sharkey and Elwert, 2011; Wilson, 1996; Williams & Collins, 2001; Iglehart & Becerra, 2011; Mickelson, 1990; Blum, 2002) that represent significant obstacles to closing racial/ethnic achievement gaps (Barton & Coley, 2010).

Data regarding percent poor in the state (POVERTY), was also taken from the American Community Survey (ACS). Socio-economic disadvantages often overlap racial and ethnic

divides, but lower SES can present problems regardless of race and ethnicity (Owens, Reardon, & Jencks, 2016). Low income not only limits access to opportunities (Mayer, 1997; Lee & Burkham, 2002; Rothstein, 2004), but can also have negative effects on healthy physical and cognitive development (Brooks-Gunn, Klebanov, & Liaw, 1995; Guo & Harris, 2000; Votruba-Drzal, 2003). Children who grow up in lower SES are also negatively impacted by the attitudes within these environments (Sampson, Morenoff, & Gannon-Rowley, 2002; Davis-Kean, 2005). Obviously, these economic and environmental factors can contribute to circumstances that negatively impact student achievement.

Similarly, the percent of single parent homes (MARST>1 and NCHILD>0) in each state was also obtained from the American Community Survey. Invested parents can be considered to possess a kind of "social capital" that can help mitigate other SES disadvantages (Ream & Palardy, 2008), but such potential advantages can be undermined by inequities in time and effort parents can allocate to school involvement (Rothstein, 2004). This limited capacity for involvement is a characteristic of lower SES neighborhoods, and these economic disadvantages are critical non-school factors that should be understood (Reardon, 2011).

Conversely, having at least one parent with a college degree can not only improve the student's chances of academic success, but that experience endows its own "social capital" that can improve the parent's and student's abilities to advocate in ways that support academic success. For these reasons, this study also looks at percent of adults (ages 30-50) with college degree (EDUC).

Data on school factors were obtained from the federal Common Core of Data (CCD). The first measure taken from the CCD was percent of students in urban and rural districts (ULOCAL) Not only do these schools serve more racially and economically disadvantaged students than suburban schools do, they often have a host of systemic problems that present challenges to

achieving at high levels (see Baker, Sciarra, & Farrie, 2010; Lankford, Loeb, & Wyckoff, 2002; Bryk & Schneider, 2003; Hanushek, Kain, & Rivkin, 2009).

Also taken from CCD, average class size is another school factor that can exert a significant impact on one's education. Studies show that large class sizes contribute to teacher stress and turnover (Russell, Altmeier, & Van Velzen, 1987) which can have negative consequences on academic outcomes (Nye, Hedges, & Konstantopoulos, 2000). In this study of fourth grade math and reading, the mean pupil/teacher ratio in elementary school is used as a proxy measure for class size so that the effects of this environmental condition can be analyzed.

Because school quality is somewhat dependent on the availability of financial resources, I also obtained per pupil state and local revenue ((TSREV+TLOCREV)/MEMBER) from the CCD. Lower state and local revenue has implications for recruitment, hiring, and retention of teachers, quality of programs and facilities (Card & Payne, 2002), and on the availability and effectiveness of student supports (Duncombe &Yinger, 2008). Table 1 offers a full explanation of the measures used in this dissertation.

3.4 Analytical approach

Although each state (50) is being observed four times (2003, 2005, 2007, and 2009), there are constants within each state that remain unchanged. In longitudinal datasets that observe the same unit over time, these unit-specific constants "color" the other variables in ways that keep each of the four state-level observations from being truly independent. In order to limit the effects of this, I used hierarchical linear regression. This approach limits biases related to non-independence of repeated observations of the same unit.

Table 1: Measures used in the analysis

Measure	Source	Variable Used	Description	Mean					Standard deviation				
				2003	2005	2007	2009	2003-09	2003	2005	2007	2009	2003-09
NAEP scale equivalency (NSE) for percent proficient	NCES		NAEP score conversion of state percent proficient										
Fourth grade math				222.462	223.382	223.648	222.166	222.917	7.739	13.354	12.146	11.878	11.422
Fourth grade reading				194.589	195.861	198.827	199.223	197.138	6.839	14.946	14.398	12.540	12.676
Percent proficient	State dept. of educ. websites		Percent of students proficient on the stateassessment										
Fourth grade math				0.648	0.691	0.720	0.719	0.694	0.167	0.157	0.136	0.138	0.152
Fourth grade reading				0.688	0.719	0.745	0.731	0.721	0.133	0.137	0.113	0.112	0.124
Percent African American (ages 30-50)	ACS	RACE	African American	0.096	0.097	0.101	0.103	0.099	0.092	0.093	0.095	0.096	0.094
Percent Hispanic (ages 30-50)	ACS	HISPANIC	Mexican, Cuban, Puerto Rican, and other Hispanic	0.081	0.090	0.098	0.106	0.094	0.092	0.095	0.099	0.101	0.096
Percent poor (ages 30-50)	ACS	POVERTY	Proportion at or below the federal poverty line	0.118	0.122	0.188	0.125	0.121	0.031	0.029	0.031	0.029	0.030
Percent with college degree (ages 30-50)	ACS	EDUC	Proportion with a B.S. degreee or higher	0.187	0.187	0.192	0.193	0.189	0.031	0.031	0.032	0.031	0.031
Percent single parent (ages 30-50)	ACS	MARST>1 and NCHILD>0	Family structure	0.106	0.109	0.112	0.115	0.111	0.016	0.016	0.014	0.017	0.016
Percent urban and rural	CCD	ULOCAL	Proportion of students in large cities and rural districts	0.379	0.381	0.385	0.394	0.385	0.132	0.131	0.129	0.129	0.129
Pupil/teacher ratio in elementary grades	CCD		Mean number of students per teacher in grades K-5	15.768	16.048	15.762	15.265	15.710	2.627	3.956	3.866	3.540	3.301
Per pupil state and local revenue	CCD	(TSTREV+ TLOCREV)/ MEMBER	Total CWI-adjusted revenue from state and district sources per student	7,732	9,075	10,980	12,015	9,950	1,620	2,095	2,758	2,963	2,920

Note: CWI (comparable wage index) is a measure of cost of living, reflecting regional variations in salaries of college graduates who are not educators. It enables robust comparisons of income and other fiscal measures across different geographic locations. Measures from the American Community Survey (ACS) are weighted averages constructed using the person weight variable (PERWT).

I fit the following models using the xtmixed procedure in STATA 14.2 separately for fourth grade math and reading:

$$NSE_{jt} = \beta_{0j} + \beta_{1j}t + \beta_{2j}t^2 + \varepsilon_{jt} \quad (1a)$$

$$\beta_{0j} = \lambda_{00} + \Sigma_{i=1}(\psi_{0i}S_{ijt}) + \upsilon_{0j} \quad (1b)$$

$$\beta_{1j} = \lambda_{10} + \Sigma_{i=1}(\psi_{0i}S_{ijt}) + \upsilon_{1j} \quad (1b)$$

NSE_{jt} is the NAEP scale equivalent for percent proficient for state j in year t, where t=0 for 2003, t=1 for 2005, t=2 for 2007, and t=3 for 2009. This coding facilitates interpretation of time interactions of predictors (see "Results" chapter for a further explanation). S_{ijt} is the ith state characteristic where the series i runs from 1 to 8 (see figure 3) for state j in year t. Equation 1b represents main effects of state characteristics, while equation 1b specifies the interactions of those characteristics with time (since this equation is a level-2 specification for β_{1j}, the t coefficient at level-1). Notice that I also account for nonlinear time effects at level-1 ($\beta_{2j}t^2$).

Figure 3: Measures – i series state characteristics, 1-8

Number	Measure	Code	Source
1	Percent African American	RACE	ACS
2	Percent Hispanic	HISPANIC	ACS
3	Percent poor	POVERTY	ACS
4	Percent college graduate	EDUC	ACS
5	Percent single parent	(MARST>1 and NCHILD>0)	ACS
6	Percent urban and rural	ULOCAL	CCD
7	Pupil/teacher ratio in elementary grades		CCD
8	Per pupil state and local revenue	(TSREV+TLOCREV)/MEMBER	CCD

Most importantly, as noted earlier, the relative error (RE) in generating the NSE is used as the basis for observational weights in estimation. For each state-by-year observation, the weight is specified as 1/RE, the inverse of the relative error, consistent with insights from the

conventional literature on inverse variance weighting. The RE is essentially a variance measure underlying the NSE (see Bandeira de Mello, Blankenship, & McLaughlin, 2009).

Chapter 4: Results

This study set out to examine whether No Child Left Behind (NCLB) achieved its stated purpose "to leave no child behind." By using a novel measure, the NAEP scale equivalency (NSE), I analyzed achievement data for fourth grade mathematics and reading from the key years of NCLB, 2003-2009, to determine if this reform resulted in higher levels of learning and to examine the effects of relevant school and non-school factors historically correlated to student achievement. In this chapter I will discuss the results of the two hypotheses posed in this dissertation:

Hypothesis 1: *Inequities associated with key non-school factors (racial/ethnic and SES backgrounds) are negatively related to learning outcomes under NCLB.*

Hypothesis 2: *Inequities associated with key school factors (location, class size, and funding) are negatively related to learning outcomes under NCLB.*

4.1 Analysis: Fourth Grade Mathematics

Estimated effects on NSE are shown in Table 2. In Model 1, for mathematics, I entered predictors without time interactions. Thus, each effect shown in this model reflects the aggregate effect of that particular (school or non-school) factor's relationship with NSE across all four time points (2003, 2005, 2007, and 2009). In this model, percent African American is the only state feature significantly related to NSE: a standard deviation increase in percent African American is related to -0.325 ($p \leq 0.050$) standard deviation units of decrease in that state's NSE. This model is helpful because when the effects of time are taken out, only one non-school factor – percent African American – is statistically significant, at -0.325 ($p \leq 0.050$). Model 1 is represented below.

Model 1: Fourth Grade Mathematics (aggregate - no time interactions)
NSE = -0.325 RACE (African American)

In Model 2, I introduce time interactions to test whether effects vary across the key years of NCLB: 2003, 2005, 2007, and 2009. When measured this way, the main effect of race loses the strength of its significance (-0.041, $p \geq 0.100$), but this effect is only for 2003 (t=0). In other words, the negative effects were less negative, less statistically significant, and less perceptible when observed over this single unit of time as compared to Model 1. However, the interaction of percent African American (RACE) with time is strongly significant (-0.192, $p \leq 0.050$). This means that for each consecutive time point after 2003, the effect of percent African American on fourth grade math proficiency became increasingly negative. Recall that t=1 for 2005, t=2 for 2007, and t=3 for 2009. These calculations for each consecutive time period reveal more trends that should inform any analysis of NCLB and its attempts at accountability.

Model 2: Fourth Grade Mathematics (with time interactions)
2003: NSE = -0.041 RACE – 0.168 POV + 0.000 ULOCAL +0.530 (TSTREV+TLOCAL)/MEMBER
2005: NSE = -0.233 RACE – 0.168 POV – 0.120 ULOCAL +0.367 (TSTREV+TLOCAL)/MEMBER
2007: NSE = -0.425 RACE – 0.168 POV – 0.240 ULOCAL +0.204 (TSTREV+TLOCAL)/MEMBER
2009: NSE = -0.617 RACE – 0.168 POV – 0.360 ULOCAL +0.041 (TSTREV+TLOCAL)/MEMBER

In Model 2, the estimated effect of percent African Americans in 2003 is -0.233 (i.e., 0.041+(1*-0.192)); for 2005: -0.425 (i.e., -0.041+(2*-0.192)); and for 2007: -0.617 (i.e., -0.041+(3*-0.192)). This demonstrates that although the main effect of race in Model 2 for the year 2003 (-0.041, $p \geq 0.100$) is statistically weaker than the effect measured in Model 1 (-0.325, $p \leq 0.050$), by the end of the decade a one unit (STD) increase in African Americans results in a decrease in NSE (-0.617, $p \leq 0.050$), more than one-half of one standard deviation. To clarify,

having more African Americans in *any* state between the years 2003-2009 resulted in a NAEP scale equivalency that continued to decline over time. Regardless of these state-level trends in self-reported percent proficient (which will be discussed below), the NAEP scale equivalence of these states with a high percent of African American students demonstrate that the standards for meeting state-defined proficiency were *less* rigorous at the end of the reform than they were at the beginning. Since NCLB was designed to ensure that all students were achieving at high levels, regardless of their location or background, this empirical evidence of decline is fundamentally antithetical to the spirit of the reform – and a direct contradiction of George W. Bush's stated goal to "end the soft bigotry of low expectations."

Model 2 also shows that the effect of location (percent urban/rural = ULOCAL) is time dependent. In 2003 (the main effect), location had no relationship with NSE (0.000, $p \geq 0.100$), but for each consecutive time point, the effect of percent urban/rural students became increasingly negative (-0.120, $p \leq 0.050$). Similar to the analysis of percent African American, this highlights the value of both models: although in Model 1 the effects of percent urban/rural were not significant, when observed over time this important school-related factor is correlated with a decrease in the NSE. As explained above, this means that in states with more students attending urban and rural schools, the *actual* benchmark to meet proficiency targets was lower at the end of the reform than it was at the beginning, leaving students who attended schools in often-already disadvantaged and under-resourced communities to a lower standard than they were before this accountability-based reform was initiated.

Finally, the effect of funding (per pupil state/local revenue = (TSTREV+TLOCAL) /MEMBER)) became increasingly *less positive* over time. In 2003 (the main effect), per pupil revenue had a positive interaction with time (0.530, $p \leq 0.010$), which meant that for this year, the

Table 2. Weighted HLM estimates of effects on NAEP scale equivalent (NSE) for percent proficient fourth grade math and reading, 2003-2009

	Mathematics		Reading	
	Model 1	Model 2	Model 3	Model 4
Time	0.096	-0.165	0.129	-0.294
	(0.233)	(0.262)	(0.241)	(0.259)
$Time^2$	-0.061	0.002	-0.055	0.059
	(0.060)	(0.068)	(0.075)	(0.080)
Percent African American	-0.325 **	-0.041	-0.219	0.053
	(0.143)	(0.190)	(0.140)	(0.185)
Percent Hispanic	-0.079	0.045	0.084	0.177
	(0.101)	(0.124)	(0.098)	(0.120)
Percent poor	-0.103	-0.168 *	0.030	0.145
	(0.091)	(0.097)	(0.122)	(0.112)
Percent single parent	0.193	-0.036	0.167	-0.115
	(0.125)	(0.181)	(0.136)	(0.197)
Percent college graduate	0.069	0.009	0.072	0.041
	(0.135)	(0.109)	(0.112)	(0.088)
Percent urban and rural	-0.148	0.000	-0.273 ***	-0.147 *
	(0.112)	(0.108)	(0.087)	(0.085)
Pupil/teacher ratio in elementary grades	-0.086	0.067	-0.233 ***	0.096
	(0.135)	(0.142)	(0.086)	(0.104)
Per pupil state and local revenue	0.177	0.530 ***	0.332 ***	0.883 ***
	(0.136)	(0.209)	(0.093)	(0.175)
Time x Percent African American		-0.192 **		-0.161 **
		(0.083)		(0.079)
Time x Percent Hispanic		-0.077		-0.050
		(0.048)		(0.049)
Time x Percent poor		0.075		-0.063
		(0.053)		(0.065)
Time x Percent single parent		0.134		0.146
		(0.091)		(0.092)
Time x Percent college graduate		0.017		0.003
		(0.069)		(0.061)
Time x Percent urban and rural		-0.120 **		-0.080
		(0.060)		(0.054)
Time x Pupil/teacher ratio in elementary grades		-0.079		-0.185 ***
		(0.075)		(0.059)
Time x Per pupil state and local revenue		-0.163 **		-0.277 ***
		(0.073)		(0.076)
Constant	0.099	0.363 *	0.037	0.431
	(0.149)	(0.202)	(0.104)	(0.150)

Table 2. Continued

	Mathematics		Reading	
	Model 1	Model 2	Model 3	Model 4
Variance components				
Between-state				
Constant	0.113 *	0.165 **	0.044	0.064
Time	0.020	0.016	0.009	0.007
Cov(Time, Constant)	0.048 **	0.044 **	0.019	0.021
Within-state	0.496 ***	0.458 ***	0.616 ***	0.564 ***
Wald chi-square	34.662 ***	50.553 ***	90.381 ***	149.912 ***

Note: Standard errors are in parentheses. Estimates are based on standardized scores. *** p≤0.010, ** p≤0.050, * p≤0.100.

more money a state spent per pupil, that was strongly correlated to a high NSE. Despite this positive impact measured in 2003, the effect of funding decreased by -0.120 (p≤0.050) in each subsequent time point.

Although much was made of NCLB being an "unfunded mandate" (Lowell, 2004; Mathis, 2005), there are a few possible explanations for why per pupil expenditures had a less positive impact over time. Because in an accountability-based system, outcomes – in this case learning, skills, or concepts – are standardized, the individual actors within these systems may have become more proficient at communicating, transmitting, and measuring these desired outcomes. As the state actors became more skilled in this new educational context, that increased capacity may have resulted in funding became less consequential.

In a similar vein, because learning outcomes are standardized under this kind of reform, states with lower per pupil expenditures may have improved how they satisfied the requirements of this reform over time. The longer these lower revenue states operated under and became familiar with the system, the more they increased their capacity to achieve the letter of

the law. Because Model 2 shows interactions over time, these states might have, in essence, been "gaining" on states that had historically higher per pupil revenue, effectively diminishing the effect of funding on each state's ability to demonstrate their compliance with the mandate.

4.2 NCLB's Impact: Fourth Grade Mathematics

With regard to dissertation hypothesis 1, *Inequities associated with key non-school factors (racial/ethnic and SES backgrounds) are negatively related to learning outcomes under NCLB,* the results of the Model 1 and Model 2 analysis clearly show that the effects of race on fourth grade math proficiency were not only significant in the aggregate (-0.325, $p \leq 0.050$), but that this factor became increasingly deleterious over time. Because NCLB was designed to result in higher levels of learning for all students by raising the expectations and outcomes for historically disadvantaged students, this pattern is fundamentally antithetical to NCLB's objectives and to purposes of proficiency-based accountability reform more generally.

My first hypothesis also suggested that another critical non-school factor, percent of poverty in the state, would also have a negative impact on NSE. Model 1 does not pick up any effect of poverty on a state's NSE in the aggregate, and although it is detectable in Model 2, it does not vary in strength as it interacts with time, remaining constant at -0.168 ($p \geq 0.100$). Although this is counter to the assumptions I made in my research hypothesis 1, the effects of race on the NSE clearly demonstrate that, in total, NCLB's effect on fourth grade mathematics may have *hurt* the very students it was officially designed to help.

Turning to dissertation hypothesis 2, *Inequities associated with key school factors (location, class size, and funding) are negatively related to learning outcomes under NCLB*, Models 1 and 2 make clear that the key school factors indicated in the existing literature such as percent urban/rural and per pupil state and local revenue do have implications for a state's NSE,

although not as predicted. In terms of fourth grade mathematics, percent urban/rural had an increasingly negative effect over time, while the effect of per pupil funding had a (2003) main effect (0.530, $p \leq 0.010$) that as increasingly *less positive* over time. Although hypothesis 2 also predicted class size would have a negative impact on learning outcomes under NCLB, as measured by a negative impact on NSE over time, this effect did not show up in Model 1 or Model 2. However, because NCLB was intended to level the playing field by ensuring that students in disadvantaged and under-resourced schools were learning at the same high levels as many of their better-funded suburban counterparts, this results of this study show that in terms of overcoming well-documented school factors that have historically presented obstacles to learning, NCLB did not achieve its stated purpose.

4.3 Analysis: Fourth Grade Reading

Models 3 and 4 address effects on NSE for fourth grade reading. They replicate the same specification procedure as in Models 1 and 2. Model 3 is represented below.

Model 3: Fourth Grade Reading (aggregate - no time interactions)
NSE = -0.273 ULOCAL – 0.233 PT/RATIO + 0.332 (TSTREV+TLOCAL)/MEMBER

As seen in the calculation, percent urban/rural (-0.273, $p \leq 0.010$) and pupil/teacher ratio (a proxy for class size) (-0.233, $p \leq 0.010$) undermined reading proficiency, while per pupil state/local revenue improved it. Unlike the results for my analysis of fourth grade mathematics in either Model 1 or 2, percent African American does not present as a significant factor in NSE for fourth grade reading when studied in the aggregate, 2003-2009. But these effects are longitudinal aggregates, and much like the earlier analysis showed, that is the value of Models 1 and 3; with the effects of time removed, the results can be quite different than the effects of critical school

and non-school factors as measured when they interact with time; these interactions are broken down in Model 4.

Model 4: Fourth Grade Reading (with time interactions)
2003: NSE = 0.053 RACE – 0.147 ULOCAL + 0.096 PT/RATIO + 0.883 (TSTREV+TLOCAL)/MEMBER
2005: NSE = -0.108 RACE – 0.147 ULOCAL - 0.089 PT/RATIO + 0.606 (TSTREV+TLOCAL)/MEMBER
2007: NSE = -0.269 RACE – 0.147 ULOCAL - 0.274 PT/RATIO + 0.329 (TSTREV+TLOCAL)/MEMBER
2009: NSE = -0.430 RACE – 0.147 ULOCAL – 0.459 PT/RATIO + 0.052 (TSTREV+TLOCAL)/MEMBER

Unlike in Model 3 where the effect was not at all significant, Model 4 shows that the percent of African American students in a state had an increasingly negative effect on the NSE (-0.161, $p \leq 0.50$) of fourth grade reading over time. This is a critical point because although the effects were almost undetectable in the aggregate, by the end of the decade in 2009, a one-unit increase (STD) in percent African Americans would have a -0.430 ($p \leq 0.50$) decrease in NAEP scale equivalency.

While Model 3 suggests that percent urban/rural negatively impacted NSE, Model 4 shows that this factor loses significance in terms of both its (2003) main effect (-0.147, $p \leq 0.100$) and its interaction with time (-0.080, $p \geq 0.100$). This simply suggests that while the location effect is strongly negative in the aggregate, its size in any given year is not large enough to be statistically significant. In contrast, the effect of pupil/teacher ratio on fourth grade reading outcomes became increasingly negative over time. While its (2003) main effect is non-significant, its interaction with time is strongly negative (-0.185, $p \leq 0.010$). There are a few possible explanations as to why under the exact same proficiency-based reform, this school factor would impact reading more than it did mathematics.

Math skills are segmented and sequentially ordered (Reys et al., 2014). Because of this inherent structure, students can be good in some grade-appropriate mathematics without

performing well in other ones. For instance, in eighth grade, a student may have exceptional geometry skills while struggling to understand basic statistics. Because the skills in math can be taught in isolation or in ways that can be logically scaffolded for students, math instruction is easier to standardize in large group settings. When students require remediation, the objective nature and standardized processes characteristic of math make it easier for a teacher to diagnose and address a variety of learner needs within one class.

Reading proficiency, on the other hand, is dependent on a greater number of nuanced skills that are far more subjective in nature and that occur simultaneously in the act of reading fluently. To read proficiently, students need to understand orthography and phonology (Alexander & Fox 2013; Dresser 2012), have the ability to decode (Smith, 1985), apply contextual interpretation of meanings (Adams, 2103), as well as actively practice comprehension and inference-making (Vacca & Vacca, 2005). Because of the "nested" and interdependent skills required of proficient readers, it can be much more difficult to diagnose and address skill deficiencies that students exhibit. These complications may explain why both Models 3 and 4 reveal that a greater pupil/teacher ratio (a proxy for class size) has a larger impact on fourth grade reading instruction than it does on fourth grade mathematics.

And finally, while the effect of per pupil state/local revenue was strongly positive in 2003 (0.883, $p \leq 0.010$), it became increasingly *less positive* over time (-0.277, $p \leq 0.010$). Potential explanations for this decreasing impact of funding were explained above (see section 4.1)

4.4 NCLB's Effect: Fourth Grade Reading

My first research hypothesis states that, *Inequities associated with key non-school factors (racial/ethnic and SES backgrounds) are negatively related to learning outcomes under NCLB*. Although not statistically significant in Model 3, the analysis in Model 4 demonstrate that the

critical non-school factor of racial diversity – specifically, percent African American in the state – has a positive (2003) main effect (0.053, $p \leq 0.050$), but that it is affected by an increasingly negative effect in subsequent years (-0.161, $p \leq 0.050$). Unlike I hypothesized, states with high percentages of low-income families did not impact the NSE over time. However, the analysis of both fourth grade reading and fourth grade mathematics show that increases in percent African American students has a deleterious consequence on these state's respective NAEP scale equivalence. Because it did not live up to its promise of closing the achievement gap for these students of color, instead lowering the real benchmark for states to demonstrate their proficiency (as measured by the decrease in NSE over time), NCLB clearly negatively impacted learning outcomes for African American students.

The second hypothesis I forwarded in this dissertation suggests that, *Inequities associated with key school factors (location, class size, and funding) are negatively related to learning outcomes under NCLB*. While the school-related variable of location (defined as percent urban/rural) had a negative effect on fourth grade reading proficiency in aggregate in the 2003-2009 period (-0.273, $p \leq 0.010$), the effect size when interacting with time did not show up as statistically significant in Model 4; again, this highlights the value of running two models of analysis, one in the aggregate (2003-2009) and one that breaks down these effects as they interact with time. From the peak years of NCLB, 2003-2009, the impact class size had on NSE became increasingly more negative during every period (-0.185, $p \leq 0.010$). When observed by the end of Model 4, a one unit (STD) increase in pupil/teacher ratio results in nearly one-half of one standard deviation decline in NSE for fourth grade reading. That effect makes clear that larger class sizes made teaching reading a more difficult endeavor, and the NSE over time suggests that *actual* student outcomes suffered under these conditions.

Finally, money became less beneficial over time with regard to reading proficiency, as was the case with math. As explained previously, this probably had to do with the standardization of expectations and individual actors' increasing familiarity with and capacity to achieve these outcomes. Taken as a whole, it is clear that school-level factors such as class size and location had negative consequences on the levels of learning being measured on those state's respective assessments as measured when these are mapped onto the NAEP. For this reason, it is clear that No Child Left Behind did not improve the educational achievement of all students, but rather resulted in symbolic compliance that satisfied the letter of the law without benefitting the very students it was intended to help.

As an auxiliary analysis, I examined the relationship of all these predictors with state-reported percent proficient (*unconverted* to the NAEP scale), relying on the same modeling approach as above. My findings are shown in Table 3 below.

This is important for two reasons. First, it demonstrates how conclusions based on raw proficiency information will be different than those based on the NSE for proficiency. And second, based on the nature and magnitude of differences, one can infer how distorted state-reported proficiency information can be under the proficiency-based accountability regime, which ultimately undermines the entire reform.

As seen across all four models in Table 3, no state feature – school or non-school - statistically predicts percent proficient as reported by the state. Time is the only significant predictor in virtually all models. This suggest that states reported better proficiency regardless of differences in key non-school and school factors that are well known to influence performance. The average state-reported fourth grade reading proficiency steadily increased over time (as all

Table 3. HLM estimates of effects on percent proficient in fourth grade math and reading, 2003-2009

	Mathematics		Reading	
	Model 1	Model 2	Model 3	Model 4
Time	0.300 ***	0.517 ***	0.420 ***	0.600 ***
	(0.110)	(0.180)	(0.122)	(0.204)
$Time^2$	-0.081 ***	-0.127 ***	-0.096 ***	-0.135 ***
	(0.026)	(0.043)	(0.032)	(0.050)
Percent African American	0.108	0.222	0.075	0.041
	(0.137)	(0.189)	(0.134)	(0.203)
Percent Hispanic	-0.126	-0.121	-0.184	-0.283 *
	(0.122)	(0.166)	(0.114)	(0.185)
Percent poor	0.112	0.084	0.104	0.101
	(0.079)	(0.105)	(0.088)	(0.122)
Percent single parent	0.034	0.026	-0.072	0.075
	(0.098)	(0.142)	(0.111)	(0.165)
Percent college graduate	0.025	0.144	0.058	0.161
	(0.100)	(0.115)	(0.105)	(0.130)
Percent urban and rural	0.003	0.031	0.084	0.013
	(0.130)	(0.152)	(0.122)	(0.158)
Pupil/teacher ratio in elementary grades	-0.064	-0.043	0.003	-0.088
	(0.043)	(0.084)	(0.051)	(0.097)
Per pupil state and local revenue	0.185 *	-0.113	0.012	-0.258
	(0.109)	(0.208)	(0.108)	(0.229)
Time x Percent African American		-0.030		0.050
		(0.064)		(0.071)
Time x Percent Hispanic		0.013		0.062
		(0.048)		(0.054)
Time x Percent poor		0.012		0.001
		(0.052)		(0.058)
Time x Percent single parent		-0.033		-0.115
		(0.070)		(0.077)
Time x Percent college graduate		-0.107		-0.080
		(0.069)		(0.053)
Time x Percent urban and rural		-0.024		0.035
		(0.044)		(0.048)
Time x Pupil/teacher ratio in elementary grades		-0.020		0.053
		(0.049)		(0.055)
Time x Per pupil state and local revenue		0.110		0.114
		(0.068)		(0.077)
Constant	-0.167	-0.391 *	-0.294 *	-0.483 **
	(0.182)	(0.223)	(0.177)	(0.238)

Table 3. Continued

	Mathematics		Reading	
	Model 1	Model 2	Model 3	Model 4
Variance components				
Between-state				
Constant	1.111 ***	0.964 ***	0.981 ***	0.966 ***
Time	0.064 ***	0.045 ***	0.054 ***	0.048 ***
Cov(Time, Constant)	-0.162 ***	-0.112 ***	-0.157 ***	-0.152 ***
Within-state	0.122 ***	0.126 **	0.183 ***	0.182 ***
Wald chi-square	33.872 ***	48.384 ***	22.992 **	30.583 **

Note: Standard errors are in parentheses. Estimates are based on standardized scores. *** $p \leq 0.010$, ** $p \leq 0.050$, *$p \leq 0.100$.

rate of increase (as all second-order time effects [$Time^2$] are modestly negative and significant at $p \leq 0.010$).

This pattern is consistent with insights from classic studies of compliance with legal mandates (Tolbert & Zucker, 1983). When a reform is legally mandated, all actors subject to the mandate are coerced to comply. Individual traits are inconsequential in time and degree of compliance in terms of official deliverables, whether these involve symbolic responses or genuine ones. For example, all drivers, regardless of personal traits, are expected to comply with speed limits. Coercive compliance in this regard is different from normative and mimetic processes where characteristics and needs of complying actors play a greater role in time and degree of compliance with the law (Tolbert & Zucker, 1983).

Under the proficiency-based accountability regime, all states are *obligated* to improve proficiency, irrespective of their characteristics. Hence, no state characteristic predicts proficiency outcomes. But the NSE analysis in Table 2 reveals that effects on patterns of learning

were very different than those estimated based on state-reported percent proficient information as represented in Table 3.

Chapter 5: Discussion

This dissertation takes a critical step back to address what happened to student outcomes in NCLB's peak years, from 2003 to 2009: Was the No Child Left Behind Act (NCLB) successful in accomplishing its basic goal ("to leave no child behind")? Did states improve student performance? How did they vary in this regard and why? What factors explain differences across the states and over time?

In this dissertation, I set out to evaluate the effectiveness of NCLB by using both the state-reported proficiency information and the results of each state's performance on the National Assessment of Educational Progress (NAEP) using equipercentile linking (see Figure 2). By using the NAEP scale equivalency (NSE) for state-reported proficiency (see Table 4 in Appendix), this study looks at fourth grade mathematics and reading in order to analyze the relationship of state NSE with fundamental non-school and school factors that characteristically underlie achievement patterns.

Because decades of research have shown that SES-related disparities are a major factor in performance gaps between have and have-nots in the U.S., I hypothesized that:

> ***Hypothesis 1:*** *Inequities associated with key non-school factors (racial/ethnic and SES backgrounds) are negatively related to learning outcomes under NCLB.*

Additionally, because achieving the desired levels of academic proficiency can be impacted by relevant school factors such as location, teacher expectations, and per pupil expenditures, I hypothesized that:

> ***Hypothesis 2:*** *Inequities associated with key school factors (location, class size, and funding) are negatively related to learning outcomes under NCLB.*

5.1 Key Findings

Campbell's Law is Strong

Campbell's Law states that "the more any quantitative social indicator is used for social decision making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it was intended to monitor" (Nichols and Berliner, 2007). Under the structure of NCLB, states were not only forced to find a way to 100% proficiency by the end of 2014, but they were threatened with a series of cascading punitive sanctions if they failed to show improvement, without regard or mercy for any significant school-related or non-school factors that have been proven to negatively impact academic achievement.

According to the National Center for Educational Statistics, one of the purposes of analyses utilizing the NSE is to "inform states whether the rigor of their proficiency standards as represented by NAEP scale equivalents changed [over time]. Significant differences in NAEP scale equivalents might reflect changes in state assessments and standards and/or other changes such as changes in policies or practices that occurred between the years" (Bandeira de Mello, Blankenship, & McLaughlin, 2009, pp. 9).

NCLB was a federal mandate with a cascading menu of punitive sanctions for states that failed to comply. Although Table 3 shows that state-reported proficiency rates are impressive and improve steadily over time, Table 2 reveals that these state-reported gains were actually distorted in ways that had negative consequences for the very students the reform was intended to benefit. In states facing challenges such as racial diversity (defined as a high percent of African American residents), or that have systemic problems like a large number of communities located in an historically under-resourced urban or rural locations, or large class sizes, the standard for proficiency did *not* increase as NCLB had intended. Additionally, because the legal

mandates of this reform required compliance from all actors regardless of school or non-school factors, even per pupil revenue did not improve a state's ability to improve student achievement over time.

Instead, as Campbell's Law predicts, the pressure to show continued gains year after year lead to a corruption of the intended outcome (Nichols and Berliner, 2007; Holmstrom and Milgram, 1991). Although the states reported ever-increasing levels of proficiency, students in these states achieved less, year after year. When viewed through the lens of established principles of organizational sociology such as resource dependence (Pfeffer & Salancik, 1978), organizational legitimacy (Tyack & Cuban, 1995), and goal displacement (Bohte & Meier, 2000), one could conclude that states already facing disadvantages due to established non-school and school inequities responded in largely symbolic ways.

Because the unit of analysis in this student is the state, the results do not allow conclusions to be drawn on any units smaller than that. Without further investigation, it is difficult to conclude whether these changes in the NSE were due to strategic gaming (Linn et al., 1990; Shepard, 1990; Koretz et al., 1991; Stecher & Barron 1998; Klein et. al. 2000) or were motivated by organizational survival (Dowling & Pfeffer, 1975; Pfeffer & Salancik, 1978). Either way, it is clear that the real standard for student proficiency in these states diminished over time, and diminished at statistically significant rates in states with large percentages of African American students and students attending urban and rural schools.

NCLB Failed

Although that section heading might seem harsh and absolute, my analysis of the effects of school and non-school factors on the NSE for fourth grade mathematics and reading very

clearly show that this policy did not, as President George W. Bush suggested it would, "end the soft bigotry of low expectations." In fact, the *only* significant predictors of lower NSE for both of these academic subjects was percent African American within the state. This means that over time, although states with large populations of African American students continued to show improvements in state-reported proficiency (see Table 3) to signal their compliance with the federal mandate, when those same state-reported proficiency rates were mapped onto the NAEP for the years 2003, 2005, 2007, and 2009, the *real* benchmark to achieve proficiency continued to decline over time.

Again, because the state is the smallest unit of analysis, it is impossible to say whether the deleterious effects of each respective coefficient (African American or urban/rural schools) affected only these students, or whether or not the disproportionate percentages of these factors within these states had the effect of holding down the expectations for all students, especially since state assessments and cut scores are determined by state departments of education. What can be seen, whether intentional or not, is that the declining rigor of assessments in these states suggests that already disadvantaged students were even more "left behind" as a result of the policy.

Although critics could argue that the states could have been more successful had NCLB not been an "unfunded mandate," per pupil expenditures had less of positive impact over the life of the reform rather than exerting a greater effect. Another way to interpret this is that the positive effect of more per pupil funding was not a significant enough variable to cancel out the well-documented obstacles to academic success found in urban and rural schools or those serving high numbers of African American students. In fourth grade reading, larger class size also had a negative effect on NSE. Under-resourced and over-stressed reading teachers were not

able to improve outcomes for their students, which likely had negative implications for these students in other academic subjects.

In total, this analysis of trends using the NSE demonstrates NCLB not only negatively impacted the academic achievement in states with large percentages of African American students and students attending urban and rural schools, but in both cases the policy had an increasingly negative effect on these states over time. Furthermore, not only did the policy fail the very students it was designed to benefit, but it may be possible that in states with decreasing NSEs, all students were negatively impacted by the decreasing expectations of what one must know to demonstrate "proficiency" under No Child Left Behind.

5.2 Limitations of this Study

Studies that utilize NAEP data provide compelling evidence on achievement trends under NCLB, but using NAEP for NCLB evaluation is inherently problematic. Critics note that NAEP outcomes can be insensitive to NCLB since, unlike state assessments, NAEP, as a *national* test, is not designed to reveal and help remedy state-specific weaknesses in achievement (e.g., specific groups or regions that underperform or specific subject areas that are weak across the state) (e.g., Popham, 2005). NAEP is not a high-stakes test for which schools prepare their students. It tests an array of skills that is far wider than the average statewide curriculum standards, and is therefore a policy-independent assessment (Feuer et al., 1999; National Assessment Governing Board, 2002).

In this regard, lack of or limited progress in NAEP for a given state, may be an indication that NCLB worked in that state, particularly if such NAEP outcomes are coupled with growth in proficiency gains in the state assessment itself. This argument implies that, despite the broad

consensus on NCLB's ineffectiveness, there may be a hidden virtue behind most curricular, assessment, and instructional decisions in the states that lead to progress in learning which NAEP cannot fully or properly detect. Accountability advocates argue that narrower (or more "focused") curricula and assessments, and teaching to the test are essentially unproblematic if they help schools concentrate energies on well-defined goals for each grade level, which in the aggregate would benefit students (e.g., Phelps, 2003; Edelman, 2015).

This dissertation attempts to address this limitation by taking advantage of both the state-reported proficiency information and the state-level NAEP outcomes, using the NAEP scale equivalent (NSE) for percent proficient in the state.

As mentioned previously, constructing the NSE also produces a "relative error" in mapping percent proficient onto the NAEP scale that has affected the validity of previous attempts to use the NSE. Factors such as differences in student motivation regarding the two tests, the time of year tests are administered, and the format of the test can also contribute to the conversion problem in limited degrees (Feuer et al., 1999). The relative error (RE) is based on how well the NSE can reproduce percent proficient in individual NAEP schools – in previous studies, the RE has proven to be a problem that plagues attempts at conversion and diminishes the value of such attempts at analysis. To address this limitation, I used the RE as the basis for observation weights in my estimation. In other words, observations involving NSEs with greater mapping errors are weighted less than those with smaller mapping errors. This adjustment not only improves on previous attempts at mapping, but it also provides an unprecedented degree of accuracy.

Although there could be many state-level characteristics that impact the NSE, this study deliberately limited its analysis to an essential set of school and non-school state features

available through the Common Core of Data and the American Communities Survey. A more robust analysis might have also accounted for the effects of state-level politics on the NSE, as such an analysis might have "washed out" or overridden the effects of these other variables. Because NCLB mandated federally enforced state sanctions against struggling schools, news coverage and the public controversy over NCLB made it clear that there were heavily politicized attempts to push back against this federal intrusion in the name of local control (Heck, 2004). Such dynamics are more likely in states that are politically more conservative. Elazar (1984) attributes this to political orientations underlying a preference for "small government." By 2005, various bills critical of NCLB were introduced in 20 states. The severest ones came from conservative states. Key examples include Utah, which enacted a bill stating that Utah law superseded federal law, and Colorado, which allowed districts to decline to participate in NCLB without incurring state penalties (Center on Education Policy, 2006). In some states, actions by districts sparked broader resistance to evaluation under NCLB, some districts going as far as rejecting funds to avoid compliance (Loveless, 2007). With all of this as background, it would have been interesting to see if the analysis shows more conservative states having a lower NSE over time, which would reduce the risk of federal involvement. Although this could also be a direction for future study, because the state's political orientation might have diminished, overridden, or "washed out" the effects of other school or non-school factors, I view this as a potential limitation.

5.3 Implications for Future Research

This study measures learning outcomes for fourth grade mathematics and fourth grade reading during the peak years of NCLB, 2003-2009. It looks at the effects of key non-school and

school factors known to influence performance by mapping state-reported proficiency onto the NAEP using a novel measure, the NSE, while measuring the respective effects of critical variables both in the aggregate and broken down in time interactions for each of the four periods.

Future studies can look at additional grade levels and subjects to see how the effects of these predictor variables change over time. If certain effects persist across multiple grades, subjects, and years, the implications for developing policy to address these systemic trends could be beneficial.

Because most of the punitive aspects of NCLB were dislodged by the adoption of Common Core State Standards (CCSS) and the Every Student Succeeds Act (ESSA), it would also be of value to study the trends in NSE after 2009. Because institutional legitimacy and continued access to resources are not in as great a jeopardy as they were under NCLB, what are the patterns revealed by studying the NSE now? In the absence of punitive sanctions after 2009, did academic expectations improve (as measured by the NSE) for the student subgroups that suffered under No Child Left Behind? If so, this strengthens the validity of Campbell's Law, which more policy makers should attempt to understand before supporting educational reforms that prioritize quantitative growth (as measured by percent proficient) on a largely qualitative endeavor (such as public education).

5.4 Implications for Future Policy

State-Reported Proficiency is Unreliable

Multiple studies suggest that states complied with the federal requirements of NCLB by inflating measures of student proficiency (Linn et al., 1990; Shepard, 1990; Koretz et al., 1991; Stecher & Barron 1998; Klein et. al. 2000). By arbitrarily adjusting the standards for proficiency

and manipulating the reporting conventions, states could symbolically comply in order to maintain their standing and remain eligible for the external resources that would ensure their organizational survival. These states may not have consciously chosen to impact students and student learning in negative ways, but accountability pressures, compounded with the difficulty of overcoming racial and economic variables, may have incentivized finding creative ways to demonstrate their effectiveness as educational institutions. If this tendency is real, the federally enforced punitive sanctions introduced under NCLB may have only broadened and accelerated these strategic responses.

Such patterns of strategic response are often used to explain why student achievement has been slow to improve on low-stakes national assessments compared to the impressive proficiency gains reported on high-stakes state assessments (Barton & Coley, 2010; Dee & Jacob, 2011). As the results of Tables 2 and 3 demonstrate, evaluations of educational policy based on percent proficient can have the unfortunate consequence of masking the lower standards for academic success for already disadvantaged students.

A more rigorous analysis of academic outcomes from the peak years of this reform is critical as the nation has recently adopted NCLB's successor, ESSA (Every Student Succeeds Act), which is predicated on the same assumption that proficiency-based accountability is an effective approach to achieving equity in public education. Efforts to evaluate NCLB by relying on state-reported proficiency outcomes are problematic in that such outcomes are based on assessment systems and parameters that vary over time as well as across states. Although NSE is one way to evaluate the impacts of this policy, future policy decisions should be based on analytical approaches that have either standardized (or circumvented) percent proficient as reported by the state.

Addressing School and Non-School Factors

Sociology is uniquely positioned to address social, economic, and political factors that affect learning. My analysis suggests that high-stakes accountability incentivizes strategic or symbolic responses because the law requires compliance without regard for key school and non-school factors that have long-proven to have adverse effects on academic achievement. Correcting this may require policy makers to give education the same degree of concern various other policies receive, such as national defense and sound energy policy. Although it would be difficult to do, accountability can be refocused to address non-school and school disparities that would – as a natural outcome of more equitable social policies – improve educational outcomes for historically struggling subgroups.

Future educational policy may be more productive if it avoids punishing schools (Center on Education Policy, 2006). A more supportive approach for struggling schools, which, for instance, brings new resources and attempts to proactively address social and economic challenges in the surrounding communities would not only be more effective, but may also be more popular within the communities it aims to improve. By negating the emphasis on quantitative measure to evaluate educational systems, there may be less incentive for struggling schools and states to comply in strategic and symbolic ways. By focusing on public policies that attempt to mitigate the effects of critical school and non-school factors indicated in this dissertation, stakeholders can focus on achieving the spirit of high-minded reforms rather than scrambling to avoid the threats of non-compliance introduced under policies like NCLB.

Anticipating Strategic and Symbolic Responses

Regardless of its incarnation, accountability in public education continues to present challenges. Policy makers should also do more to understand the conditions that provoke largely strategic and symbolic responses so they can design policy to reduce the likelihood of an institutional response that does not achieve the intended outcomes. An abbreviated history of the Common Core State Standards (CCSS) illustrates this point.

For states, the adoption of these new standards was one of the ways they could escape the demands of the previous accountability mandates under NCLB. Although adoption of CCSS was said to be voluntary, there were clear and early institutional incentives for states to embrace this initiative. In July 2009, Obama's Education Department announced Race to the Top (RTTT), a $3.4 billion grant program for states willing to embrace significant reform. The RTTT application awarded points to states that had agreed to adopt CCSS (McNeil & Klein, 2011), and in the post-recession economy at the time, that alone might have been a great incentive.

By the end of 2010, 38 states, the District of Columbia and the territories of America Samoa and the Virgin Islands had adopted the CCSS (Center for Educational Policy, 2015). 2011 saw six more states added the list of those supporting the initiative in addition to the adoption of CCSS by the Northern Marina Islands (2015). The endorsements by Wyoming in 2012 and Arizona in 2013 signaled the high-water mark for CCSS, with 46 states having officially embraced this new common core curriculum as the next step in educational improvement.

In August of 2011, President Obama revealed the details of a long-awaited waiver plan to help states escape the punitive sanctions that were increasing under NCLB as AYP targets became harder to meet. To be eligible for these NCLB waivers, states had to have adopted college and career readiness standards (McNeil & Klein, 2011). As institutional expectations

shifted – as the definition of legitimacy and the requirements ensure continued access to resourced changed – the number of waivers states were granted to escape various aspects of accountability increased dramatically. Two months after the announcement of program, 42 states had expressed their intent to apply for these waivers (Center for Educational Policy, 2015). By January of 2015, 42 of the 50 states had received waivers of some kind, essentially wiping out all of the emphasis on accountability that had been the hallmark of NCLB (Center for Educational Policy, 2015).

Because early allegiance to CCSS was a necessary condition for getting waivers from the federal mandates of NCLB and satisfied the requirements for Race to the Top funds under the Obama Administration, making these decisions may have been motivated by a desire to preserve the state's legitimacy. Even though states had done all they could to satisfy the requirements of NCLB, when the public conversation about educational reform started to change, there were significant incentives for symbolic compliance and ceremonial adoption of CCSS. As the reform has materialized, been challenged, and become politicized, states have backed out by the dozens. In most cases, they have kept the substance of the common core (which means they are still eligible for all of the federal benefits), but renamed their curricular standards to signal their desire to be seen as an autonomous state, independent from federal oversight – a position that both ensures continued access to resources and imbues them with institutional legitimacy.

This short history of CCSS demonstrates the power both sociological concepts – organizational legitimacy and resource dependence – have on schools and states. Any future policy that fails to understand the conditions that provoke strategic or symbolic, rather than substantive, institutional responses may motivate individual actors to distort or corrupt their processes in ways that benefit the organization.

References

Adams, M. J. (2013). *Theoretical models and processes of reading*. Newark, DE: International Reading Association.

Alexander, P. A. & Fox. E. (2013). *A Historical Perspective on Reading Research and Practice.* Newark, DE: International Reading Association.

Amatea, E. S., Smith-Adcock, S., & Villares, E. (2006). From family deficit to family strength: Viewing family's contributions to children's learning from a family resilience perspective. *Professional School Counseling, 9,* 277-313.

America 2000: An Education Strategy Sourcebook. (1991). Department of Education, Washington, DC.

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

Bandeira de Mello, V. (2011). *Mapping State Proficiency Standards onto the NAEP Scales: Variation and Change in State Standards for Reading and Mathematics, 2005–2009 (NCES 2011-458)*. Washington, D.C.: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education.

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

Barton, P. E., & Coley, R. J. (2010). *The black-white achievement gap: When progress stopped.* Princeton, NJ: Educational Testing Service.

Berliner, D. C. (2009). MCLB (Much Curriculum Left Behind): A U.S. calamity in the making. *The Educational Forum, 73,* 284-296.

Blum, L. (2002). *"I'm not a racist, but . . ." The moral quandary of race.* Ithaca, NY: Cornell University Press.

Bohte, J. & Meier, K.M. (2000). Goal Displacement: Assessing the Motivation for Organizational Cheating. *Public Administration Review* 60: 173-182.

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

Braun, H. I. and Qian, J. (2007). An enhanced method for mapping state standards onto the NAEP scale. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), Linking and aligning scores and scales (pp. 313-338). New York: Springer.

Brooks-Gunn, J., Klebanov, P. K., & Liaw, F. R. (1995). The learning, physical, and emotional environment of the home in the context of poverty: The Infant Health and Development Program. *Children and Youth Services Review, 17,* 251-276.

Bryk, A. S., & Schneider B. (2003). Trust in schools: A core resource for school reform. *Educational Leadership, 60, 6,* 40-45.

Card, D., & Payne, A. A. (2002). School finance reform, the distribution of school spending, and the distribution of student test scores. *Journal of Public Economics, 83,* 49-82.

Center on Education Policy. (2006). *From the capital to the classroom: Year 4 of the No Child Left Behind Act*. Washington, DC.

Center on Education Policy. (2007). *Choices, changes, and challenges: Curriculum and instruction in the NCLB era. A report in the series 'From the capital to the classroom: Year 5 of the No Child Left Behind Act.'* Washington, DC.

Center for Educational Policy. (2015). NCLB/ESEA Waiver Watch. Online.

"Clinton Years: State Initiatives for Systemic Reform." (n.d). US Department of Education, New York State Archives. Retrieved from http://www.archives.nysed.gov/

Clotfelter, C. T. (2011). After *Brown*: The rise and retreat of school desegregation. Princeton, NJ: Princeton University Press.

Coleman, J. S. (1966). Equality of Educational Opportunity Study (EEOS). Ann Arbor, MI: Inter-university Consortium for Political and Social Research.

Coleman, J. S. (1988). Social capital in the creation of human capital. *American Journal of Sociology, 94,* S95-S120.

Copeland, L. (2013, April 14). School cheating scandal shakes up Atlanta. *USA Today*.

Cronin, J., Dahlin, M., Adkins, D., & Kingsbury, G. G. (2007). *The proficiency illusion*. Washington, D.C.: Thomas B. Fordham Institute.

Cronin, J., Dahlin, M., Adkins, D., & Kingsbury, G. G. (2009). *The Accountability Illusion*. Washington, D.C.: Thomas B. Fordham Institute.

Cullen, J., & Reback, R. (2006). Tinkering toward accolades: School gaming under a performance accountability system. Improving School Accountability: Check-ups or choice. *Advances in Applied Economics*, *14*, 1-34.

Darling-Hammond, L. (2007). Race, inequality and educational accountability: The irony of "No Child Left Behind." *Race Ethnicity and Education, 10,* 245-260.

Davies, G. (2007). *See Government Grow: Education Politics from Johnson to Reagan*. Lawrence, KS: University of Kansas.

Davis-Kean, P. E. (2005). The influence of parent education and family income on child achievement: The indirect role of parental expectations and the home environment. *Journal of Family Psychology, 19,* 294-304.

Dee, T.S. & Jacob. B.A. (2011). The Impact of No Child Left Behind on Student Achievement. *Journal of Policy Analysis and Management, 30:* 418-64.

Deere, D., & Strayer, W. (2001a). Closing the gap: School incentives and minority test scores in Texas. Department of Economics, Texas A&M University.

Deere, D., & Strayer, W. (2001b). Putting schools to the test: School accountability, incentives, and behavior. Working Paper 113, Private Enterprise Research Center, Texas A&M University (March 2001).

Dobbin, F., & Sutton, J.R. (1998). The Strength of a Weak State: The Rights Revolution and the Rise of Human Resources Management Divisions. *American Journal of Sociology, 104:* 441-76.

Dorans, N.J., Pommerich, M., & Holland, P.W. (2007). Linking and Aligning Scores and Scales. New York: Springer

Dowling, J., & Pfeffer, J. (1975). Organizational Legitimacy. *Pacific Sociological Review* 18: 122-136.

Dresser, R. (2012). The impact of scripted literacy instruction on teachers and students. *Issues in Teacher Education, 21,* 71-87.

Driscoll, D., Halcoussis, D., & Svorny, S. (2004). School District Size and Student Performance. *Economics of Education Review, 22:* 193-201.

Duncombe, W., & Yinger, J. (2008). Measurement of cost differentials." In H. Ladd & E. B. Fiske (Eds.), *Handbook of research in education finance and policy* (pp. 203-221). New York, NY: Routledge.

Edelman, J. (2015). Parents Should Welcome the New Common Core Tests. Online.

Edelman, L.B., Uggen, C., & Erlanger, H.S. (1999). The Endogeneity of Legal Regulation: Grievance Procedures as Rational Myth. *American Journal of Sociology, 105*: 406-54.

Elazar, D. J. (1984). *American federalism. Third edition*. New York, NY: Harper & Row.

Elmore, R. (2003). Accountability and capacity. In M. Carnoy, R. Elmore, & L. S. Siskin (Eds.), *The new accountability: High schools and high-stakes testing* (pp. 188-207). New York, NY: RoutledgeFalmer.

Feuer, M. J., Holland. P.W., Green, B. F., Bertenthal, M. W. and Hemphill, F. C. (1999). *Uncommon Measures: Equivalence and Linkage among Educational Tests*. Washington, DC: National Academies Press.

Figlio, D. N. (2006). Testing crime and punishment. *Journal of Public Economics*, *90*, 837-851.

Figlio, D. N., & Getzler, L. S. (2006). Accountability, ability and disability: Gaming the system. Improving School Accountability: Check-ups or choice. *Advances in Applied Economics*, *14*, 35-49.

Firestone, W.A., Schorr, R.Y. & Monfils, L.F. (Eds.). (2004). *The Ambiguity of Teaching to the Test: Standards, Assessment, and Educational Reform.* Mahwah, NJ: Lawrence Erlbaum Associates.

Fuhrman, S. (Ed.). (2001). *From the capitol to the classroom: Standards-based reform in the states* (Vol. 100). National Society for the Study of Education.

Gardner, D. P. (1983). *A Nation at Risk.* Washington, DC: The National Commission on Excellence in Education, US Department of Education.

"George H.W. Bush Years: Systemic Reform." (n.d). US Department of Education, New York State Archives. Retrieved from http://www.archives.nysed.gov/

Gramlich, E. M., & Rubinfeld, D. L. (1982). Micro estimates of public spending demand functions and tests of the Tiebout and median voter hypotheses. *The Journal of Political Economy, 90,* 536-560.

Guo, G., & Harris, K. M. (2000). The mechanisms mediating the effects of poverty on children's intellectual development. *Demography, 37,* 431-447.

Hanushek, E. A., & Raymond, M. E. (2003). Lessons about the Design of State Accountability Systems. In P. E. Peterson and M.R. West (Eds.), *No Child Left Behind? The Politics and Practice of Accountability*, (pp. 126-151). Washington, DC: Brookings.

Hanushek, E. A., Kain, J. F., & Rivkin, S. G. (2009). New evidence about *Brown v. Board of Education*: The complex effects of school racial composition on achievement. *Journal of Labor Economics, 27,* 349-383.

Harris, D. N. (2011). *Value-added measures in education: What every educator needs to know.* Cambridge, MA: Harvard Education Press.

Hashima, P. Y., & Amato, P. R. (1994). Poverty, social support, and parental behavior. *Child Development, 65,* 394-403.

Hayes, W. (2008). *No Child Left Behind: past, present and future*. Lanham, MD: Rowman & Littlefield.

Heck, R. H. (2004). *Studying educational and social policy: Theoretical concepts and research methods*. New York, NY: Routledge.

Henig, J. R. (2007). The political economy of supplemental educational services. In F. M. Hess & C. E. Finn (Eds.), *No remedy left behind: lessons from a half decade of NCLB,* (pp. 66-95). Washington, DC: American Enterprise Institute.

Hillson, J. (1995). The coalition that supported the Smith-Hughes Act. *Journal of Vocational and Technical Education 11*(2). 4-11. Retrieved from http://eric.ed.gov/?id=EJ504569

Holland, P. W. (2007). *A framework and history for score linking,* in Dorans, N.J., Pommerich, M., & Holland, P.W. (Editors). Linking and Aligning Scores and Scales. New York: Springer.

Holmstrom, B., & Milgrom, P. (1991). Multitask principal-agent analyses: incentive contracts, asset ownership and job design. *Journal of Law, Economics and Organization 7*, 24–52.

Hombo, C. (2003). NAEP and No Child Left Behind: Technical Challenges and Practical Solutions. *Theory into Practice, 42*: 59-65.

Hornback, J.E. (2013). State Level Cheating on High Stakes Assessments. Unpublished Doctoral Dissertation. Department of Educational Leadership and Policy Studies. Lawrence, KS: University of Kansas, Lawrence.

Hursh, D. (2007). Exacerbating inequality: The failed promise of the No Child Left Behind Act. *Race Ethnicity and Education*, *10*, 295-308.

Iglehart, A. P., & Becerra, R. M. (2011). *Social services and the ethnic community: History and analysis. Second edition*. Long Grove, IL: Waveland Press.

Jacob, B.A. (2002). "Making the grade: The impact of test-based accountability in schools." Kennedy School of Government, Harvard University (mimeo) (April).

Jacob, B. A. (2007). Test-Based Accountability and Student Achievement: An Investigation of Differential Performance on NAEP and State Assessments. *NBER Working Paper No. 12817*. Cambridge, MA: National Bureau of Economic Research.

Jacobs, B. A., & Levitt, S. D. (2003). Rotten apples: An investigation of the prevalence and predictors of teacher cheating. *Quarterly Journal of Economics, 118*, 843-877.

Jeffrey, J. (1978). Education for children of the poor: A study of the origins and implementation of the Elementary and Secondary Education Act of 1965. Columbus: Ohio State University Press.

Jennings, J., & Sohn, H. (2014). Measure for Measure How Proficiency-based Accountability Systems Affect Inequality in Academic Achievement. *Sociology of Education*, *87*(2), 125-141.

Klein, S. P., Hamilton, L., McCaffrey, D. F., Stecher, B. (2000). *What Do Test Scores in Texas Tell Us?* Santa Monica, CA: RAND Corporation.

Koretz, D., Linn, R. L., Dunbar, S. B., & Shepard, L. A. (1991). *The Effects of High-Stakes Testing: Preliminary Evidence About Generalization Across Tests.* American Educational Research Association, Chicago.

Koretz, D., & Barron, S. (1998). *The Validity of Gains in Scores on the Kentucky Instructional Results Information System (KIRIS*). Santa Monica, CA: RAND Corporation.

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

Labaree, D. (2006). The Trouble with Ed Schools. New Heaven, CT: Yale University Press.

Lankford, H., Loeb, S., & Wyckoff, J. (2002). Teacher sorting and the plight of urban schools: A descriptive analysis. *Educational Evaluation and Policy Analysis, 24,* 37-62.

Lareau, A., & Horvat, E. M. (1999). Moments of social inclusion and exclusion: Race, class, and cultural capital in family-school relationships. *Sociology of Education, 72,* 37-53.

Lee, J., & Reeves, T. (2012). Revisiting the Impact of NCLB High-Stakes School Accountability, Capacity, and Resources: State NAEP 1990–2009 Reading and Math Achievement Gaps and Trends. *Educational Evaluation and Policy Analysis, 34,* 209-23.

Lee, V. E., & Burkam, D. T. (2002). *Inequality at the starting gate: Social background differences in achievement as children begin school.* Washington, DC: Economic Policy Institute.

Lieske, J. (2009). The Changing Regional Subcultures of the American States and the Utility of a New Cultural Measure. *Political Research Quarterly, 63:* 538-52.

Linn, R. (2003). Accountability: Responsibility and reasonable expectations. *Educational Researcher, 32,* 3-13.

Linn, R. L., Graue, M. E., & Sanders, N. M. (1990). Comparing state and district results to national norms: the validity of the claim that 'Everyone is above average'. *Educational Measurement, Issues and Practice 9(3),* 5–14.

Lowell, C.R. (2004). No Child Left Behind": Failed Schools—Or Failed Law? *Educational Horizons 82:2,* 121-130.

Loveless, Tom. 2007. "The Peculiar Politics of No Child Left Behind." Pp. 253-85 in Standards-Based Reform and the Poverty Gap: Lessons for No Child Left Behind, edited by A. Gamoran. Washington, DC: Brookings Institution.

Manna, P. (2007). NCLB in the states: fragmented government, uneven implementation. In F. M. Hess & C. E. Finn (Eds.), *No remedy left behind: lessons from a half decade of NCLB* (pp. 17-42). Washington, DC: American Enterprise Institute.

Mathis, W.J. (2005). The Cost of Implementing the Federal No Child Left Behind Act: Different Assumptions, Different Answers. *Peabody Journal of Education, 80*:2, 90-119.

Mathis, W.J. & Trujillo, T.M. (2016). Lessons from NCLB for the Every Student Succeeds Act. Boulder, CO: National Education Policy Center.

Mayer, S. (1997). *What money can't buy: Family income and children's life chances.* Cambridge, MA: Harvard University Press.

McLaughlin, D., V. et al. (2008a). *Comparison between NAEP and State Mathematics Assessment Results: 2003 (NCES 2008-474)*. Washington, DC: National Center for Educational Statistics, Institute of Education Sciences, U.S. Department of Education.

McLaughlin, D., V. et al. (2008b). *Comparison between NAEP and State Reading Assessment Results: 2003 (NCES 2008-474)*. Washington, DC: National Center for Educational Statistics, Institute of Education Sciences, U.S. Department of Education.

McLaughlin, M. (1975). Evaluation and reform: The Elementary and Secondary Education Act of 1965, Title I. Cambridge, Massachusetts: Ballinger Publishing Company.

McNeil, M., & Klein, A. (2011). Obama offers waivers from key provisions of NCLB. *Education Week*, *31*(5), 20-21.

Mehta, J. (2013). How paradigms create politics: The transformation of American educational policy, 1980–2001. *American Educational Research Journal, 50, 285-324.*

Metcalf, S. (2002, January 10). Reading between the lines. *The Nation*. Retrieved from http://www.thenation.com/article/reading-between-lines

Metz, M. H. (2008). Symbolic Uses of NCLB: Reaffirmation of Equality of Educational Opportunity or Delegitimization of Public Schools. In A. Sadovnik, J.A. O'Day, G.W. Bohrnstedt, & K.M. Borman (Eds.), *No Child Left Behind and the Reduction of the Achievement Gap: Sociological Perspectives on Federal Educational Policy* (pp. 343-58). New York, NY: Routledge.

Meyer, J. W., & Rowan, B. (1977). Institutionalized Organizations: Formal Structure as Myth and Ceremony. *American Journal of Sociology* 83:340-63.

Mickelson, R. (1990). The attitude-achievement paradox among black adolescents. *Sociology of Education, 63,* 44-61.

National Assessment Governing Board. (2002). *Using the National Assessment of Educational Progress to confirm state test results*. Washington, DC: National Assessment Governing Board.

NCLB Action Brief. (2003, August 22). Supplemental Educational Services. Washington, D.C.: Center on Education Policy.

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

Nichols, S.L., & Berliner, D.C. (2007). *Collateral Damage: How High-Stakes Testing Corrupts America's Schools*. Cambridge, MA: Harvard Education Press.

Nye, B., Hedges, L. V., & Konstantopoulos, S. (2000). The effects of small classes on academic achievement: The results of the Tennessee class size experiment. *American Educational Research Journal, 37,* 123-151.

Orfield, G., & Eaton, S. E. (1996). *Dismantling desegregation: The quiet reversal of* Brown v. Board of Education*.* New York, NY: The New Press.

Owens, A., Reardon, S. F., & Jencks, C. (2016). Income segregation between schools and school districts. *American Educational Research Journal, XX,* 1-39.

Peterson, P. E., & Hess. F. M. (2006). Keeping an eye on state standards. A race to the bottom? *Education Next, 6,* 28-29.

Pfeffer, J. & Salancik, G.R., (1978). *The External Control of Organizations.* New York, NY: Harper and Row.

Popham, J. W. (2005). All about accountability/NAEP: Gold standard or fools' gold? *Educational Leadership, 62,* 79-81.

Porter, A. C. (2002). Measuring the Content of Instruction: Uses in Research and Practice. *Educational Researcher* 31:3-14.

Posner, J. K., & Vandell, D. L. (1994). Low-income children's after-school care: Are there beneficial effects of after-school programs? *Child Development, 65,* 440-456.

Ream, R. K., & Palardy, G. J. (2008). Reexamining social class differences in the availability and the educational utility of parental social capital. *American Educational Research Journal, 45,* 238-273.

Reardon, S.F. (2011). The Widening Academic Achievement Gap between the Rich and the Poor: New Evidence and Possible Explanations. In G.J. Duncan & R.J. Murnane (Eds.), *Withier Opportunity? Rising Inequality, Schools, and Children's Life Chances* (pp. 91-115). New York, NY: Russell Sage Foundation.

Resnick, L.B., Rothman, R., Slattery, J. B., & Vranek, J. L. (2003). Benchmark and Alignment of Standards and Testing. *Educational Assessment* 9:1-27.

Reys, R., Lindquist, M., Lambdin, D., & Smith, N. L. (2014). *Helping children learn mathematics. 11th edition*. New York, NY: Wiley.

Rothstein, R. (2004). *Class and schools: Using social, economic, and educational reform to close the achievement gap*. New York, NY: Teachers College Press.

Russell, D. W., Altmaier, E., & Van Velzen, D. (1987). Job-related stress, social support, and burnout among classroom teachers. *Journal of Applied Psychology, 72,* 269-274.

Sahlberg, P. (2011). *Finnish lessons*. New York: Teachers College Press.

Sampson, R. J., Morenoff, J. D., & Gannon-Rowley, T. (2002). Assessing 'neighborhood effects': Social processes and new directions in research. *Annual Review of Sociology, 28,* 443-478.

Sharkey, P., & Elwert, F. (2011). The legacy of disadvantage: Multigenerational neighborhood effects on cognitive ability. *American Journal of Sociology, 116,* 1934-1981.

Shepard, L. (1990). Inflated test score gains: is the problem old norms or teaching the test? *Educational Measurement, Issues and Practice 9(3),* 15–22.

Smith, F. (1985). A metaphor for literacy: Creating worlds or shunting information? In R. Shaw & J. Bransford (Eds.), *Perceiving, acting, and knowing: Toward an ecological psychology* (pp. 1-39). Hillsdale, NJ: Erlbaum.

Stecher, B., & Barron, S. (1999). Quadrennial Milepost Accountability Testing in Kentucky. Center for the Study of Evaluation, UCLA.

Sunderman, G. L., Kim, J. S., & Orfield, G. (2005). *NCLB meets school realities: lessons from the field.* Thousand Oaks, CA: Corwin.

Thissen, D. & Wainer, H. (2001). *Test Scoring.* Hillsdale, NJ: Lawrence Erlbaum.

Tolbert, P. S., & Zucker, L. G. (1983). Institutional sources of change in the formal structure of organizations: The diffusion of civil service reform, 1880-1935. *Administrative Science Quarterly, 28,* 22-39.

Tollefson, K. (2008). *Volatile Knowing*. Lanham, MD: Lexington.

Tyack, D. & Cuban, L. (1995). T*inkering Toward Utopia.* Cambridge, MA: Harvard University.

Vacca, R. T., & Vacca, J. A. L. (2005). *Content area reading: Literacy and learning across the curriculum. 8th edition.* Boston, MA: Allyn & Bacon.

Votruba-Drzal, E. (2003). Income changes and cognitive stimulation in young children's home learning environments. *Journal of Marriage and Family, 65,* 341-355.

Williams, D. R., & Collins, C. (2001). Racial residential segregation: A fundamental cause of racial disparities in health. *Public Health Reports, 116,* 404-416.

Wilson, W. J. (1996). *When work disappears: The world of the urban poor.* Chicago, IL: University of Chicago Press.

Wixson, K. K., Fisk, M. C., Dutro, E., & McDaniel, J. (2002). *The alignment of state standards and assessments in elementary reading.* Ann Arbor, MI: Center for the Improvement of Early Reading Achievement.

Table 4: State-level percent proficient and NAEP scale equivalent (NSE)

(Alaska through Hawaii)

stateaby	year	RE READ4	RE MATH4	PP READ4	PP MATH4	NSE READ4	NSE MATH4
AK	2003	0.301	0.362	0.730	0.710	191.245	227.488
AK	2005	0.201	0.258	0.780	0.750	182.123	222.019
AK	2007	0.114	0.286	0.790	0.780	183.190	216.251
AK	2009	0.292	0.296	0.780	0.740	182.672	217.555
AL	2003	0.054	0.105	0.530	0.460	195.393	215.134
AL	2005	0.246	0.446	0.830	0.730	172.143	206.843
AL	2007	0.353	0.792	0.840	0.780	178.956	205.208
AL	2009	0.090	0.490	0.870	0.790	179.160	206.526
AR	2003	0.450	0.426	0.610	0.610	210.144	220.368
AR	2005	0.283	0.361	0.520	0.590	216.963	235.601
AR	2007	0.241	0.200	0.600	0.750	212.859	228.858
AR	2009	0.208	0.339	0.700	0.810	199.727	216.018
AZ	2003	0.376	0.503	0.570	0.490	179.607	204.265
AZ	2005	0.735	0.909	0.670	0.680	184.295	207.499
AZ	2007	0.092	0.114	0.700	0.690	197.631	213.425
AZ	2009	0.255	0.153	0.730	0.720	192.924	212.199
CA	2003	0.305	0.396	0.710	0.610	204.041	230.754
CA	2005	0.053	0.260	0.750	0.680	210.072	230.658
CA	2007	0.084	0.414	0.770	0.700	210.379	225.518
CA	2009	0.092	0.378	0.830	0.630	202.250	220.399
CO	2003	0.266	0.420	0.870	0.870	200.122	230.724
CO	2005	0.107	0.184	0.690	0.620	186.174	200.838
CO	2007	0.111	0.242	0.680	0.640	186.688	200.745
CO	2009	0.150	0.039	0.690	0.630	183.032	201.566
CT	2003	0.248	0.347	0.680	0.800	199.134	227.527
CT	2005	0.031	0.233	0.680	0.790	212.338	221.120
CT	2007	0.074	0.109	0.710	0.810	213.133	220.276
CT	2009	0.065	0.156	0.700	0.810	207.898	213.847
DE	2003	0.842	0.929	0.790	0.730	182.665	212.625
DE	2005	0.300	0.579	0.840	0.790	186.547	214.441
DE	2007	0.330	0.214	0.810	0.780	202.391	225.110
DE	2009	0.396	0.165	0.810	0.780	198.588	220.244
FL	2003	0.296	0.320	0.630	0.630	194.015	213.435
FL	2005	0.227	0.305	0.670	0.680	202.241	230.403
FL	2007	0.107	0.244	0.690	0.740	209.043	229.718
FL	2009	0.247	0.430	0.710	0.780	206.288	224.830
GA	2003	0.508	0.400	0.900	0.890	197.644	225.740
GA	2005	0.533	0.630	0.920	0.890	174.137	214.822
GA	2007	0.546	0.903	0.850	0.900	185.123	213.381
GA	2009	0.600	0.794	0.930	0.780	178.072	217.898
HI	2003	0.518	0.436	0.420	0.240	198.255	235.886
HI	2005	0.088	0.130	0.510	0.280	204.747	246.661
HI	2007	0.226	0.237	0.610	0.490	211.913	237.729
HI	2009	0.127	0.218	0.620	0.480	203.135	238.824

Table 4: (Continued) State-level percent proficient and NAEP scale equivalent (NSE)

(Iowa through Michigan)

stateaby	year	RE READ4	RE MATH4	PP READ4	PP MATH4	NSE READ4	NSE MATH4
IA	2003	0.438	0.419	0.700	0.750	187.814	222.404
IA	2005	0.472	0.372	0.790	0.800	196.773	219.146
IA	2007	0.429	0.344	0.800	0.810	198.558	219.535
IA	2009	0.469	0.321	0.810	0.810	194.012	220.883
ID	2003	0.603	0.554	0.870	0.860	177.720	218.642
ID	2005	0.506	0.344	0.840	0.820	184.697	207.039
ID	2007	0.416	0.507	0.810	0.860	196.642	217.231
ID	2009	0.242	0.468	0.870	0.860	185.929	212.540
IL	2003	0.351	0.393	0.630	0.770	199.656	220.916
IL	2005	0.599	0.860	0.680	0.800	196.758	220.833
IL	2007	0.306	0.253	0.740	0.880	199.679	207.942
IL	2009	0.116	0.273	0.730	0.860	198.181	207.143
IN	2003	0.869	0.909	0.730	0.730	198.132	227.509
IN	2005	0.229	0.412	0.750	0.750	198.980	225.259
IN	2007	0.094	0.571	0.740	0.740	199.427	227.787
IN	2009	0.222	0.266	0.750	0.750	203.439	229.375
KS	2003	0.691	0.652	0.680	0.730	197.772	219.958
KS	2005	0.747	0.413	0.760	0.830	185.277	217.860
KS	2007	0.349	0.504	0.800	0.820	191.685	218.647
KS	2009	0.545	0.260	0.830	0.860	185.904	217.010
KY	2003	0.692	0.964	0.620	0.380	201.172	210.152
KY	2005	0.281	0.717	0.680	0.450	206.219	214.631
KY	2007	0.272	0.437	0.720	0.590	204.661	229.290
KY	2009	0.122	0.266	0.740	0.640	204.934	223.470
LA	2003	0.369	0.438	0.580	0.570	193.140	208.251
LA	2005	0.264	0.256	0.520	0.540	197.570	222.671
LA	2007	0.514	0.290	0.580	0.550	193.096	222.904
LA	2009	0.230	0.183	0.710	0.640	192.249	220.932
MA	2003	0.439	0.401	0.560	0.400	197.341	232.979
MA	2005	0.154	0.167	0.500	0.410	234.292	255.371
MA	2007	0.196	0.263	0.550	0.480	232.038	254.484
MA	2009	0.195	0.242	0.530	0.480	233.899	255.058
MD	2003	0.381	0.435	0.750	0.690	193.136	225.641
MD	2005	0.108	0.278	0.810	0.820	186.573	214.779
MD	2007	0.272	0.490	0.850	0.840	185.591	206.439
MD	2009	0.617	0.918	0.860	0.890	186.994	207.787
ME	2003	0.481	0.629	0.490	0.280	190.977	219.488
ME	2005	0.362	0.196	0.530	0.390	223.930	248.672
ME	2007	0.212	0.184	0.670	0.610	213.835	235.838
ME	2009	0.417	0.289	0.710	0.660	207.386	234.330
MI	2003	0.589	0.569	0.750	0.650	196.242	229.124
MI	2005	0.909	0.932	0.830	0.810	182.343	222.064
MI	2007	0.393	0.604	0.840	0.850	178.292	203.756
MI	2009	0.212	0.729	0.840	0.920	194.276	199.886

Table 4: (Continued) State-level percent proficient and NAEP scale equivalent (NSE)

(Minnesota through Nevada)

stateaby	year	RE READ4	RE MATH4	PP READ4	PP MATH4	NSE READ4	NSE MATH4
MN	2003	0.432	0.479	0.710	0.700	201.334	235.967
MN	2005	0.896	0.899	0.710	0.700	198.797	244.128
MN	2007	0.250	0.237	0.710	0.700	214.679	237.467
MN	2009	0.081	0.271	0.740	0.740	203.588	232.519
MO	2003	0.613	0.633	0.340	0.370	189.014	219.386
MO	2005	0.735	0.619	0.350	0.430	195.503	242.267
MO	2007	0.344	0.441	0.460	0.450	226.721	245.061
MO	2009	0.214	0.107	0.470	0.440	228.635	245.684
MS	2003	0.409	0.413	0.880	0.790	188.673	214.695
MS	2005	0.335	0.214	0.880	0.780	160.678	206.014
MS	2007	0.263	0.475	0.950	0.940	163.173	204.406
MS	2009	0.294	0.401	0.520	0.580	209.662	222.941
MT	2003	0.412	0.471	0.760	0.730	198.038	228.104
MT	2005	0.078	0.162	0.750	0.570	197.302	220.402
MT	2007	0.430	0.304	0.800	0.680	202.884	233.762
MT	2009	0.275	0.243	0.810	0.670	197.889	234.562
NC	2003	0.342	0.579	0.810	0.920	185.907	220.865
NC	2005	0.309	0.276	0.820	0.910	183.367	202.712
NC	2007	0.251	0.251	0.850	0.670	182.552	231.149
NC	2009	0.151	0.175	0.690	0.810	204.387	220.255
ND	2003	0.451	0.522	0.850	0.730	189.842	230.705
ND	2005	0.251	0.524	0.870	0.820	203.822	223.829
ND	2007	0.448	0.447	0.890	0.780	201.362	225.660
ND	2009	0.399	0.694	0.770	0.780	202.694	225.095
NE	2003	0.909	0.285	0.780	0.870	188.701	227.553
NE	2005	0.825	0.952	0.840	0.870	196.112	224.444
NE	2007	0.543	0.235	0.890	0.910	197.348	231.633
NE	2009	0.091	0.367	0.940	0.950	201.977	225.566
NH	2003	0.661	0.994	0.750	0.690	201.911	233.124
NH	2005	0.805	0.991	0.750	0.690	199.109	218.369
NH	2007	0.431	0.623	0.750	0.690	210.018	238.995
NH	2009	0.343	0.375	0.780	0.720	210.591	236.596
NJ	2003	0.326	0.405	0.770	0.670	201.475	217.466
NJ	2005	0.356	0.355	0.810	0.800	190.812	220.668
NJ	2007	0.206	0.386	0.800	0.840	201.410	219.722
NJ	2009	0.229	0.435	0.630	0.730	221.327	230.866
NM	2003	0.355	0.449	0.510	0.390	192.550	219.697
NM	2005	0.256	0.475	0.510	0.390	208.398	232.499
NM	2007	0.334	0.350	0.540	0.450	210.000	232.901
NM	2009	0.321	0.325	0.510	0.420	207.343	236.475
NV	2003	0.303	0.410	0.430	0.490	192.773	231.184
NV	2005	0.910	0.921	0.420	0.500	196.418	221.370
NV	2007	0.192	0.257	0.510	0.480	206.738	223.710
NV	2009	0.152	0.410	0.500	0.520	202.099	225.285

Table 4: (Continued) State-level percent proficient and NAEP scale equivalent (NSE)

(New York through Utah)

stateaby	year	RE_READ4	RE_MATH4	PP_READ4	PP_MATH4	NSE_READ4	NSE_MATH4
NY	2003	0.345	0.325	0.640	0.790	199.498	220.318
NY	2005	0.276	0.284	0.700	0.850	206.515	207.275
NY	2007	0.098	0.235	0.680	0.800	209.106	219.238
NY	2009	0.221	0.388	0.770	0.870	199.955	207.333
OH	2003	0.473	0.424	0.660	0.580	204.673	227.777
OH	2005	0.323	0.759	0.760	0.780	198.838	232.693
OH	2007	0.353	0.478	0.800	0.750	198.338	224.615
OH	2009	0.146	0.129	0.820	0.780	192.182	219.458
OK	2003	0.701	0.723	0.630	0.590	190.308	215.014
OK	2005	0.469	0.197	0.910	0.820	182.169	218.428
OK	2007	0.354	0.794	0.900	0.820	171.901	213.331
OK	2009	0.295	0.389	0.630	0.670	211.465	228.039
OR	2003	0.795	0.988	0.800	0.760	186.126	227.363
OR	2005	0.773	0.640	0.820	0.840	193.473	211.160
OR	2007	0.428	0.347	0.710	0.720	185.596	220.198
OR	2009	0.335	0.149	0.760	0.760	177.348	213.839
PA	2003	0.388	0.399	0.580	0.560	193.072	211.686
PA	2005	0.559	0.694	0.630	0.690	200.767	218.821
PA	2007	0.098	0.170	0.700	0.780	211.145	222.974
PA	2009	0.184	0.238	0.730	0.840	205.602	217.761
RI	2003	0.320	0.513	0.610	0.410	201.362	232.791
RI	2005	0.773	0.421	0.630	0.540	192.589	235.376
RI	2007	0.180	0.113	0.630	0.540	210.217	235.840
RI	2009	0.063	0.246	0.680	0.630	208.873	231.466
SC	2003	0.444	0.538	0.760	0.810	191.303	223.956
SC	2005	0.070	0.133	0.790	0.780	228.460	246.406
SC	2007	0.163	0.175	0.820	0.780	222.525	245.028
SC	2009	0.175	0.256	0.750	0.760	193.636	215.220
SD	2003	0.495	0.414	0.850	0.730	188.268	222.248
SD	2005	0.595	0.710	0.870	0.820	187.943	226.801
SD	2007	0.361	0.215	0.880	0.780	185.444	224.323
SD	2009	0.320	0.110	0.770	0.770	198.868	224.074
TN	2003	0.063	0.455	0.840	0.800	181.021	215.976
TN	2005	0.239	0.244	0.870	0.860	169.701	199.519
TN	2007	0.359	0.413	0.870	0.890	174.587	198.086
TN	2009	0.112	0.644	0.900	0.900	170.272	195.398
TX	2003	0.953	0.982	0.810	0.800	194.272	213.825
TX	2005	0.265	0.430	0.790	0.810	190.352	219.389
TX	2007	0.591	0.534	0.840	0.860	187.601	216.922
TX	2009	0.558	0.705	0.840	0.860	188.474	214.362
UT	2003	0.014	0.113	0.760	0.740	185.404	209.517
UT	2005	0.907	0.895	0.760	0.740	181.093	215.104
UT	2007	0.362	0.339	0.770	0.730	196.793	223.491
UT	2009	0.181	0.349	0.770	0.720	195.630	225.161

Table 4: (Continued) State-level percent proficient and NAEP scale equivalent (NSE)

(Virginia through Wyoming)

stateaby	year	RE_READ4	RE_MATH4	PP_READ4	PP_MATH4	NSE_READ4	NSE_MATH4
VA	2003	0.644	0.959	0.630	0.560	197.384	218.194
VA	2005	0.484	0.956	0.560	0.550	195.311	223.470
VA	2007	0.533	0.592	0.510	0.380	190.751	219.301
VA	2009	0.776	0.945	0.540	0.330	185.673	212.917
VT	2003	0.664	0.910	0.690	0.640	204.129	235.580
VT	2005	0.719	0.464	0.690	0.640	210.620	233.123
VT	2007	0.501	0.277	0.680	0.610	213.790	239.120
VT	2009	0.612	0.333	0.690	0.680	213.675	235.749
WA	2003	0.517	0.626	0.660	0.550	204.179	222.024
WA	2005	0.340	0.180	0.790	0.600	196.747	235.530
WA	2007	0.385	0.174	0.760	0.580	202.633	239.868
WA	2009	0.189	0.165	0.730	0.520	205.216	243.480
WI	2003	0.555	0.427	0.820	0.740	194.400	221.514
WI	2005	0.369	0.319	0.820	0.720	188.708	224.496
WI	2007	0.277	0.318	0.810	0.760	192.813	222.389
WI	2009	0.174	0.102	0.810	0.800	188.893	218.685
WV	2003	0.908	0.059	0.810	0.750	190.869	216.950
WV	2005	0.530	0.206	0.810	0.750	186.099	214.606
WV	2007	0.281	0.372	0.830	0.790	182.011	216.834
WV	2009	0.679	0.379	0.640	0.640	206.241	225.425
WY	2003	0.669	0.601	0.440	0.370	195.693	220.749
WY	2005	0.402	0.641	0.470	0.390	228.133	250.987
WY	2007	0.486	0.508	0.750	0.770	204.379	216.311
WY	2009	0.611	0.391	0.630	0.720	208.010	225.760

[i] For details on NAEP's sample design, see https://nces.ed.gov/nationsreportcard/assessment_process/samplesfaq.aspx.

[ii] Data source: http://usa.ipums.org/usa. State means used in my analysis are weighted averages constructed using the person weight variable "PERWT" in ACS. Analysis results using unweighted versions of the measures are similar to those reported here.