Working Paper:

Did States Use Implementation Discretion to Reduce the Stringency of NCLB? Evidence from a Database of State Regulations

Vivian C. Wong[1], Coady Wing[2], & David Martin[1]

When No Child Left Behind (NCLB) became law in 2002, it was viewed as an effort to create uniform standards for students and schools across the country. More than a decade later, we know surprisingly little about how states actually implemented NCLB and the extent to which state implementation decisions managed to undo the centralizing objectives of the law. This paper introduces a state level measure of NCLB stringency that helps shed light on these issues. The measure is available for 43 states and covers most years under NCLB (2003-2011). Importantly, the measure does not depend on population characteristics of the state. It varies only because of state level decisions about rule exemptions, standards, and proficiency trajectories. Results show that despite national trends in states' implementation of accountability stringency, schools' and students' experiences of NCLB varied greatly by region and state characteristics.

[1]University of Virginia
[2]Indiana University

Updated June 2016

EdPolicyWorks
University of Virginia
PO Box 400879
Charlottesville, VA 22904

Acknowledgements: The authors wish to thank participants of the University of Virginia's Center on Education Policy and Workforce Competitiveness, as well as Daniel Player (UVA) and Christina LiCalsi (AIR) for their thoughtful comments and feedback. All errors are our own.

DID STATES USE IMPLEMENTATION DISCRETION TO REDUCE THE STRINGENCY OF NCLB? EVIDENCE FROM A DATABASE OF STATE REGULATIONS

Vivian C. Wong, Coady Wing, & David Martin

Introduction

On January 8, 2002, President George W. Bush signed the No Child Left Behind (NCLB) Act into law. The law provided the federal government with authority to hold schools accountable to uniform standards. One of the headline goals of NCLB was to ensure that 100% of students were "proficient" in math and reading by 2014. Early impact evaluations of NCLB found improvements in math but not for reading (Dee & Jacob, 2011; M. Wong, Steiner, & Cook, 2015), but it is clear that NCLB failed to achieve its goal of 100% proficiency. Today, NCLB is synonomous not with achievement but with the overuse of standardized testing and the problems of a Washington oriented one-size-fits-all approach to education policy. In a rare showing of bipartisan support, Congress replaced NCLB with the Every Student Succeeds Act (ESSA) in 2015. ESSA maintains some provisions from NCLB, including annual testing for 3rd to 8th grade students in reading and math. But it devolves many responsibilities of school accountability to state and local levels.

Observers of education reform in the United States argue that ESSA marks a return of state governments in American education policy (Burnette, 2016). But was NCLB really such a centralized effort? Several researchers have already noted that state governments had substantial discretion in implementing NCLB standards (Davidson, Reback, Rockoff, & Schwartz, 2015; Taylor, Stecher, O'Day, Naftel, & Le Floch, 2010). States had authority to select their own assessment measures for determining whether students were proficient, as well as their own trajectories for reaching the 100% proficiency target in 2014. Many states applied for and were granted exemption rules, such as confidence intervals for small subgroup sizes or multi-year averaging of test performance that allowed for schools to be considered "proficient" even when they failed to meet the state Annual Measureable Objective (AMO). Combined, the ratcheting up of proficiency requirements as well as the inclusion of exemption rules introduced variation in accountability stringency across states. The result was that the same students, teachers, principals and schools deemed "proficient" in one state, could have been candidates for remediation – or even school closure – in a different state with more stringent accountability standards. Thus, under NCLB, schools' and students' experience with accountability policies depended not just on their performance, but on the state in which they resided in, and the implementation stringency of their states' accountability policies.

EdPolicyWorks Working Paper Series No. 51. June 2016.
Available at http://curry.virginia.edu/edpolicyworks/wp
Curry School of Education | Frank Batten School of Leadership and Public Policy | University of Virginia

In this study, we construct a stringency measure of state level implementations of NCLB that accounts for the complicated array of ways that the national policy differed across states. To develop this measure, we created a database of state accountability rules from 2003 to 2011 (NCLB pre-waiver period). We used the database to develop a proficiency calculator that would determine how a particular school would be evaluated under the rules in each state and year. With the calculator in hand, we tallied up the percentage of a *fixed sample* of schools that would have failed to meet the standards for Adequate Yearly Progress (AYP) in each state and year. Simulated failure rates in the fixed sample provide a concrete measure of the stringency of each state's NCLB implementation. It takes account of most AYP decisions made by the state, but is independent of school attributes and student characterstics in the state. This is important because it helps us separate the issue of implementation stringency from actual school performance and student outcomes, which may be determined by other non-NCLB factors. We use the implementation measure to describe state accountability stringency under NCLB, and to document the ways that accountability stringency has changed over time from 2003 to 2011. We also look at variation in states' accountability plans, and whether NCLB succeeded in creating more uniform proficiency standards across states over time. Finally, we examine state-level characteristics that were predictive of accountability stringency under NCLB.

Our study shows that the introduction of exemption rules decreased accountability stringency during the early years of NCLB (2004-2007). However, it also shows that most states increased accountability standards over time. Moreover, accountability standards across the country became less discrepant over time because states with the lowest standards increased their accountability requirements to catch up with the rest of the country. Despite some convergence over time, our stringency measure suggests that, even under NCLB, accountability standards vary substantially with regional and state characteristics. Northeastern and Southern states, states with more educated populations, and higher percentages of white students were related to more stringent accountability plans. However, Western states, states with larger proportions of black students, and higher baseline reading achievement scores in 8th grade were negatively associated with stringent accountability standards.

Background

Since the introduction of NCLB, researchers have noted substantial variation in the way states implemented accountability policies (Taylor et al., 2010). A fair bit of attention has been

EdPolicyWorks Working Paper Series No. 51. June 2016.
Available at http://curry.virginia.edu/edpolicyworks/wp
Curry School of Education | Frank Batten School of Leadership and Public Policy | University of Virginia

devoted to the sometimes wide discrepancy in student achievement scores between state assessments and the National Assessment of Educational Progress (NAEP), which is considered a more general measure of students' knowledge and skill (McLaughlin et al. 2008). Other researchers noted variation in state trajectories for reaching the 2014 performance target (Carey, 2007), with some states raising AMOs in equal increments each year and other states ratcheting up proficiency requirements only in the final four years of the mandate. In addition, states applied for and were granted a number of exemptions that allowed certain schools to be considered "proficient" even when they failed to meet the state AMO. Although the stated purpose of the exemption policies was to promote reliable and valid AYP designations, there was controversy about the legitimacy of the adjustments (Rogasa, 2003) and the rules varied by state.

Researchers have made efforts to link some specific components of states' accountability rules with school outcomes. Davidson, Reback, Rockoff, and Schwartz (2015) examined how specific NCLB rules were related to AYP failure rates in the earliest years of NCLB (2003 to 2005). They observed that AYP failure rates were associated with the implementation of state accountability rules, including confidence interval and Safe Harbor rules, minimum subgroup sizes for determining which students were held accountable, and alternative assessments for testing students with disabilities. Taylor et al. (2010) used data collected in 2004-05 and 2006-07 to examine the way states implemented NCLB. They also observed that school failure rates increased when AMO targets were raised.

The implementation literature on state accountability policies is limited in key ways. Most studies focus on one to three years of states' accountability policies (Davidson et al., 2015; Taylor et al., 2010), making it challenging for researchers and policy-makers to understand and link state responses over the span of NCLB. Implementation studies have often included small, purposive sample of states (Mann, 2010; Srikantaiah, 2009; Hamilton et al., 2007) that may not be representative of the United States as a whole. And most measures accountability stringency have been based on one-dimensional indices of academic proficiency standards, such as test difficulty (Taylor et al., 2010; M. Wong et al., 2015). One dimensional measures may not capture the complex set of ways that state governments may weaken or strengthen the NCLB accountability standards. For example, one state may have a difficult test assessment, but low AMO requirements and exemption rules that help schools make AYP. Another state may have an easier test assessment for labeling students proficient, but high percent proficiency targets and more stringent exemption rules. In these cases, it is hard to determine which state actually had the more stringent accountability rules.

EdPolicyWorks Working Paper Series No. 51. June 2016.
Available at http://curry.virginia.edu/edpolicyworks/wp
Curry School of Education | Frank Batten School of Leadership and Public Policy | University of Virginia

Finally, implementation studies of state accountability systems often are challenged by the possible correlation between implementation stringency and the population characteristics of schools and students within states. For example, one possible measure of states' implementation stringency is the percentage of schools that fail to meet AYP requirements. In fact, Davidson et al. (2015) report substantial state variation in the percentage of schools that failed AYP in the first two years of NCLB, ranging from less than 1 percent in Iowa to more than 80 percent in Florida. Taylor et al. found similar state-by-state differences in AYP failure rates in 2005-2006 and 2006-2007 data. But how should one interpret such variation in school AYP failure rates? High failure rates may be due to stringent standards in state AYP rules, or it may be because schools failed to perform to meet these standards. By itself, school AYP failure rates cannot provide much information about the nature of how states' implemented their accountability policies.

Alternative Measure for Examining States' Responses to NCLB: Simulated AYP School Failure Rates

Our proposed approach provides a quantitative summary of state accountability plans from 2003 to 2011. The intuition of the approach is to estimate the fraction of a fixed basket of schools would fail to meet AYP accountability standards under each state's accountability plan. A strength of our stringency measure is that it reflects the multiple implementation decisions states made under NCLB. By focusing on how each state would score a fixed basket of schools, our measure is independent of population characteristics of the state, which may be useful for later efforts to understand the causal effects of NCLB on school and student outcomes.

To illustrate our approach, we start with the population of Pennsylvania schools in 2007-2008. During this school year, 3,105 public schools were held to AYP accountability standards. Compared to the national average of schools, Pennsylvania schools had similar average percentages of Black (16%) students, students with IEPs (16%), as well as students who were economically disadvantaged (37%). However, the state's schools had higher percentages of White students (74%), and lower percentages of Hispanic (7%) and English Language Learner (2%) students. Under Pennsylvania's NCLB rules in 2007-08, 28% of public elementary, middle and high schools failed to make AYP.

Next, we consider the percentage of the 3,105 Pennsylvania schools that would have failed to make AYP if these *same schools* were located in different states. We do this by first examining "input characteristics" of Pennsylvania schools, including their enrollment sizes and percent

EdPolicyWorks Working Paper Series No. 51. June 2016.
Available at http://curry.virginia.edu/edpolicyworks/wp
Curry School of Education | Frank Batten School of Leadership and Public Policy | University of Virginia

proficiencies for each subgroup, grade, and subject area, as well as attendance, test participation, and graduation rates, and then by determining which schools that would have failed to make AYP in a different state based on that state's accountability policies. For example, a school with 46 Asian students and a 4th grade ELA proficiency rate of 65% would have met accountability standards in Pennsylvania but not in Tennessee because the AMO cut-off was 63% in Pennsylvania and 89% in Tennessee. In Texas, the subgroup would not have been held accountable at all because the minimum subgroup size was 50 students.

Table 1 summarizes AYP rules for four states, Pennsylvania, Alaska, Tennessee and Texas. The table demonstrates considerable state variation in percent proficiency thresholds and minimum subgroup sizes, as well as in exemption rules. The second to last row shows the percentage of schools that actually failed AYP in the state during 2006-07; the last row shows the percentage of Pennsylvania schools that would have failed AYP under each state's accountability rules. We see that for Pennsylvania, the actual AYP failure rate in 2006-07 was 22%, but the simulated failure rate (for the population of Pennsylvania schools in 2007-08) is 19%. The second column shows that 47% of Pennsylvania schools would have failed to meet Alaska's AYP 2006-07 requirements. Compared to Pennsylvania, Alaska had higher AMO standards and no multi-year averaging, but had a wider confidence interval adjustment, and lower requirements for attendance and graduation rates. Tennessee had the most stringent AMO requirements, did not allow for multi-year averaging or confidence interval adjustments around the Safe Harbor target, and it had the highest attendance and graduation rate requirements. Here, we see that Tennessee's accountability rating reflects the apparent rule stringency, with 62% of Pennsylvania schools failing to meet the state's AYP requirements. Finally, although Texas had the lowest AMO and graduation requirements, it did not allow schools to make AMO thresholds through confidence interval adjustments. Under these accountability rules, 32% of Pennsylvania schools would have failed. Table 1 demonstrates how simulated failure rates may be used to compare state accountability plans from least stringent (Pennsylvania) to most stringent (Tennessee) without relying on population characteristics of schools and students within the state. It also makes it clear how we arrive at a one-number summary of stringency even though states may use a diverse set of tools to affect stringency. The basic approach may be applied for each state and year, as long as rules for determining school AYP are observable.

EdPolicyWorks Working Paper Series No. 51. June 2016.
Available at http://curry.virginia.edu/edpolicyworks/wp
Curry School of Education | Frank Batten School of Leadership and Public Policy | University of Virginia

Implementing the Method

To implement our approach, we used publicly available information on state accountability plans to create a database of most AYP rules for each state and year from 2002-2003 to 2010-2011 (NCLB pre-waiver period). Our database accounts for all state AYP requirements about minimum school and subgroup participation rates, AMO thresholds, and other academic indicators (e.g. attendance, graduation, and writing and science proficiency performance). It also includes information about minimum subgroup sizes, confidence intervals, Safe Harbor, confidence intervals around Safe Harbor, and multi-year averages for computing proficiency. However, the calculator does not account for rules about growth models, performance indexes, and alternative/modified tests for students with disabilities and limited English proficiencies. Because some states (e.g. California and New York) have AYP processes that diverge significantly from other states, and/or rely on growth models and performance indices, we omit seven states (California, Colorado, Idaho, New York, Oklahoma, Vermont, and West Virginia) and the District of Columbia from our analysis sample.

Using AYP rule data, we developed an "AYP calculator," which takes the percentage of proficient students, cell sizes and other performance metrics of subgroups in schools, and returns a variable indicating whether a given school would make AYP according to each state's rules for each year. We then constructed a fixed basket of schools and "fed" these schools – with their input characteristics – through the calculator to determine the percentage of schools in the sample that would make AYP for the state and year. The result was a state-by-year level dataset showing the simulated AYP failure rates (our measure of implementation stringency) for each state and year. Importantly, because the fixed basket of schools did not change across states and time periods, the variation in simulated pass rates arose purely from differences in rules used to determine AYP, and not on changes in the population of schools.

One concern with the stringency measure described above is that it fails to capture state differences in test difficulty, or changes in test assessments. This is may be problematic because prior research on NCLB implementation has noted tremendous variation in test difficulty, especially compared to a national benchmark such as the NAEP (Taylor et al., 2010). To address this concern, we created an alternative AYP calculator that begins with a NAEP fixed sample of students, and compares NAEP scores to NAEP equivalent cutoffs for proficiency standards in each state and year. We obtain NAEP equivalent score cutoffs from a series of NCES reports that map state proficiency standards onto NAEP scale scores for 4th and 8th grade students in 2003, 2005, 2007, 2009, and 2011

(see NCES NAEP State Mapping Project at: https://nces.ed.gov/nationsreportcard/studies/statemapping). Once we adjusted the AYP calculator to reflect NAEP equivalent cutoffs, we constructed a fixed sample of NAEP 4th and 8th grade students (using their NAEP reading and math test scores as well as accompanying school information from the CCD as input values for the AYP calculator). From this fixed sample of NAEP students, we created the state-by-year stringency measure by calculating the percentage of schools from the NAEP fixed sample that would have failed to make AYP based on the state's proficiency standards, exemption rules, and NAEP equivalent thresholds. This procedure incorporates test difficulty in the stringency measure, where states with more difficult test assessments had higher NAEP equivalent cutoffs and states with easier tests had lower cutoff values. If test difficulty changed over time, the change is reflected in the NAEP equivalence cutoffs, and incorporated in our stringency measure.

Description of Fixed Samples

For our AYP calculator to work, the specific details of the fixed sample characteristics are not important. In principle, one could use a sample from a single state or a completely hypothetical sample. The key is to understand how each state's policies would evaluate the same set of students or schools. However, we used two fixed samples to assess the sensitivity of our results to sample characteristics. Our main results use a national dataset of students (NAEP), which we hope injects some realism into our fixed sample, and which ensures that there is sufficient variation to reflect state policies that are targeted at specific subgroups. The NAEP fixed sample consists of 33,285 students who took the 2009 assessment, where approximately 57.7% students were white, 16.4% were African American, 18.2% were Hispanic, 27.6% were Economically Disadvantaged, and 11.7% have an IEP. A limitation of the NAEP fixed sample is that it includes only 4th and 8th graders in the sample, so our stringency measure is based only on standards that pertain to 4th and 8th grade students. NAEP equivalent cutoffs were not available for every state and year so our annual stringency measure uses interpolated NAEP equivalent cutoff scores for years in which information was not available.[1]

As a a robustness check, we also examine results from a second fixed sample based on the population of Pennsylvania schools in 2007-2008. This sample has the advantage of including school inputs from all elementary and high school grades. The key limitation is that the results do not

[1] Results presented in this paper were not sensitive in alternative sample specifications in which we included only states and years where NAEP equivalent proficiency scores were observed.

account for differences in test difficulty. Appendix A1 plots simulated failure rates using the NAEP (green line) and Pennsylvania (purple line) fixed samples, and NAEP equivalent cutoff scores are indicated by the gray dots. The plot shows that generally, trends in simulated failure rates between the two fixed samples mirror each other and diverge in cases where state test assessments become more or less difficult (as compared to the NAEP benchmark). These results provide additional reassurance that our stringency rates are not sensitive to characteristics of the fixed sample, and reflect differences in test difficulty.

Validation of the Implementation Measure

It is critical that our implementation measure correctly describes the AYP process for each state and year. We validated the implementation measure by comparing our simulated AYP failure rates using the population of schools in the state with the actual AYP failure rates of schools in the same state and year. If our calculator correctly accounted for the AYP decision process, then our simulated AYP failure rates using the actual population of schools should replicate the state's reported AYP failure rates. Overall, our validation checks demonstrate that our calculator performed well in reproducing AYP failure rates that match states' actual AYP failure rates. In Pennsylvania, our predicted failure rates were 14% in 2004 and 27% in 2008, and the actual failure rates were 14% in 2004 and 28% in 2008. For Texas, our predicted failure rates were 15.2% in 2004 and 36.8% in 2011, and the actual AYP failure rates were 16.6% and 33.9%.

States' Implementation of NCLB Accountability Rules

National Trends in State Accountability Policies

Using the simulated failure rates obtained from our AYP calculator, we examine national trends in how states responded to the federal NCLB mandate, whether these trends varied by geographic region, as well as state demographic and legislative factors that predict states' adoption of more (or less) stringent policies. Figure 1 depicts national trends in state accountability policies under NCLB. Panel 1 shows stringency (simulated failure) rates at the 10^{th}, 25^{th}, 50^{th}, 75^{th}, and 90^{th} percentiles from 2003 to 2011. On the whole, accountability stringency rose from 2003 to 2011, where the median stringency rate increased from 32% in 2003 to 56% in 2011. State accountability policies also became less disparate over time. Panel 2 shows the $90/10^{th}$ ratio of state stringency rates from 2003 to 2011. In 2003, the most stringent states had simulated failure rates that were five times larger than simulated failure rates for the least stringent states (12% at the 10^{th} percentile versus 63% at the 90^{th} percentile); by 2011, the most and least stringent states differed by a factor of

EdPolicyWorks Working Paper Series No. 51. June 2016.
Available at http://curry.virginia.edu/edpolicyworks/wp
Curry School of Education | Frank Batten School of Leadership and Public Policy | University of Virginia

two (36% at 10th percentile versus 79% at the 90th percentile). These trends suggest that states responded to federal requirements by increasing stringency in their own accountability policies. At the same time, the gap in accountability standards between the most and least stringent states became smaller over time, as states with weak accountability rules ratcheted up their proficiency standards under NCLB.

Variation in State Accountability Responses

Despite national trends in state accountability policies, there was variation in the intensity of states' accountability policies over time. Figure 2 summarizes simulated AYP failure rates across the United States in 2003 and 2011. States with lower simulated failure rates are shaded in light gray, while states with higher simulated failure rates are shaded in dark gray (states shaded white are not included in the sample because their AYP criteria depended heavily on student growth models). The figure shows that in the first year of NCLB implementation, most states had relatively low simulated failure rates. Mississippi (4%), Texas (5%), Louisiana (8%), Georgia (10%), and Arizona (12%) had the lowest stringency scores, while New Hampshire (77%), Minnesota (76%), and Massachusetts (74%%) had the highest. However, by the end of the NCLB pre-waiver period, almost all states had ratcheted up stringency in their accountability policies. In 2011, no state had a simulated failure rate less than 22%. Mississippi (31%) and Arkansas (36%) continued to have the lowest simulated failure rates, along with Alabama (24%) and Arizona (35%), while Kentucky (89%), Minnesota (88%), New Hampshire (86%), and North Dakota (81%) had the most stringent rules. In looking across time, Texas, Wisconsin, Florida, Kentucky, and North Carolina had the largest increases in accountability stringency from 2003 to 2011, while Arizona, South Carolina, New Mexico, Wyoming, and Nebraska had the smallest changes in accountability stringency from 2003 to 2011 (see Table A1 in Appendix A).

Figure 3 shows state simulated failure rates in 2003, 2007 and 2011 by Census regions. On average, Northeastern states had the highest simulated failure rates, while Southern and Western states had the lowest. However, some Southern states (Kentucky, North Carolina, and Florida) ratcheted up their accountability policies so intensely during NCLB that they had the most stringent accountability policies in the nation by the end of the pre-waiver period. On the other hand, Western states such as New Mexico and Wyoming began with relatively tough accountability standards in 2003 (70% and 53%, respectively), but scaled back their standards via changes in confidence interval rule in 2004 (47% for NM) or test difficulty in 2007 (18% for WY). By 2011, both states had stringency rates comparable or lower than levels observed in 2003 (66% for NM and

41% for WY). Finally, the plot shows tremendous variation in accountability stringency among Midwestern states. However, here too, we saw increases in accountability standards under NCLB. The only exceptions were Ohio and Nebraska, which decreased accountability stringency over time.

Predictors of Stringency in States' Accountability Policies

To assess whether state characteristics explain trends in accountability stringency over time, as well as regional differences between states, we ran a series of regressions in which state *i*'s accountability stringency at time *t* is a function of fixed effects for NCLB year (YEAR) and census regions (REGION), as well as lagged state demographic (DEMO) characteristics, student characteristics in the state (STU DEMO), education policy decisions (ED POLICY), 1998/2000 student achievement performance (ACH), and a random error term:

$$Stringency_{it} = \alpha_0 + \beta_1 \sum YEAR_t + \beta_2 REGION_i + \beta_3 DEMO_{it-1} + \beta_4 STU\ DEMO_{it-1} + \beta_5 ED\ POLICY_i + \beta_6 ACH_i + \epsilon_{it}$$

Here, YEAR includes period fixed effects from 2003 through 2011, and REGION includes indicators for whether a state is in the Midwest, Northeast, South, or West. DEMO includes lagged covariates for state population size (logged), unemployment rate, and whether the governor was Democratic, Republican, or Independent (U.S. Census). STU DEMO includes lagged covariates for the percentage of student population in the state who were Black, White, or Hispanic (NCES), as well as the percentage who graduated from college (U.S. Census). ED POLICY includes an indicator for whether the state had a consequential accountability policy before NCLB was implemented (Dee & Jacob, 2011; Hanushek & Raymond, 2005). Following Hanushek and Raymond, we define "consequential accountability" to include states in which schools were awarded sanctions or rewards based on student test performance.[2] Finally, our last set of covariates include 1998/2000 NAEP performance in reading and math for 4th and 8th grade students, as well as flags for whether baseline NAEP scores were available (NCES). To aid in interpretation of the intercept, all quantitative variables were centered at their 2003 means.

Table 2 summarizes results from our regressions of states' simulated AYP failure rates on lagged characteristics. Table 2 provides coefficient estimates for six models in which a new set of covariates was systematically included in the model. Overall, we observe that the U-shaped trend in accountability stringency holds, even after controlling for state characteristics. Notably,

[2] We also ran models in which lagged total expenditures per students, and the percentage of funds from federal and state sources were included as covariates. These predictors were not related to accountability stringency and we did not include these variables in the final models due to endogeneity concerns.

accountability stringency dipped in 2004 by approximately five percentage points (p-value<.01) when exemption rules were introduced, but ratcheted up again over time. By 2011, accountability stringency was about 17 percentage points higher than 2003 levels (p-value<.01). Moreover, regional differences (from Western states) remained robust across all six models, although only Southern and Northeastern states had statistically significant differences in Model 6. Overall, Northeastern states had stringency rates that were 21 percentage points (p-value<.01) higher than Western States. However, the difference between Southern and Western states became apparent only after controlling for student population characteristics (Models 4-6), where Southern states had stringency rates that were approximately 18 percentage points (p-value<.05) higher than Western states. Higher unemployment rates (p-value<.10) and percentage of state populations with Bachelors degrees (p-value<.01) were positively related to accountability stringency. Moreover, compared to Democratic Governors, having an Independent Governor was related to higher stringent accountability standards (p-value<.05). However, this difference was driven entirely by Governors Jesse Ventura in Minnesota and Angus King in Maine, whose terms ended just as NCLB was implemented in January 2003. Interestingly, having a consequential accountability system pre-NCLB was associated with a three percentage point decrease in accountability stringency, but the result was not statistically significant.

Finally, the composition of students in the state appear related to adoption of more stringent accountability standards. Larger percentages of Black students were related to lower stringency rates (p-value<.05), while higher percentages of White students were related to increased stringency (p-value<.05). However, the magnitude of these relationships were small – a one percentage point increase of black students in the state is associated with .63 percentage point (p-value<.01) decrease in accountability stringency, while a one percentage point increase in white students was associated with a .42 percentage point increase (p-value<.05). In terms of student performance pre-NCLB, 8th grade reading performance on the NAEP appears negatively related to accountability stringency – that is, a one-point increase on the 8th grade NAEP reading assessment is associated with nearly a three percentage point decrease in accountability stringency (p-value<.05). These results are interesting in light of the fact that an early NCLB goal was to improve reading achievement among elementary school students. These results suggest that states with higher performing students in reading at baseline tended to implement less stringent accountability rules under NCLB.

Discussion

EdPolicyWorks Working Paper Series No. 51. June 2016.
Available at http://curry.virginia.edu/edpolicyworks/wp
Curry School of Education | Frank Batten School of Leadership and Public Policy | University of Virginia

Over the last 10 years, considerable attention has been devoted to the prescriptive nature of No Child Left Behind. In reforming NCLB, Senator Alexander (R-Tennessee) – a co-author of ESSA – urged governors to "return control to states and local school districts," and "push back against any attempt by the federal government to shape education policy in the coming years" (Burnette, 2016). This paper demonstrates that even under NCLB, states had considerable latitude in implementing accountability policies, and their choices were related to population characteristics of those living within their states.

Using our stringency measure, we observed broad national trends in states' accountability policies under NCLB, but also variation by region and state characteristics. Northeastern states had the most stringent accountability standards, followed by Southern states; Midwestern and Western states had the least stringent accountability standards. States with more highly educated populations, and larger percentages of white students were positively correlated with more stringent accountability standards, while states with higher percentages of black students and higher reading achievement scores were negatively related to stringent accountability standards.

As a nation, we will continue to grapple with the design and implementation of accountability policies for improving student learning and achievement. In this study, we have focused on describing states' responses to the federal accountability mandate, but it is worth pointing out that the measure is well-suited for also uncovering causal linkages between states' adoption of accountability policies and state and student outcomes. Policy-makers and education researchers need empirically-based resources for describing and understanding the role that states have in determining schools' and students' experiences with accountability reform.

EdPolicyWorks Working Paper Series No. 51. June 2016.
Available at http://curry.virginia.edu/edpolicyworks/wp
Curry School of Education | Frank Batten School of Leadership and Public Policy | University of Virginia

References

Burnette, D. (2016). Senator Lamar Alexander Tells Governors to Hold Their Ground on ESSA. *Education Week,* Retrieved from: http://blogs.edweek.org/edweek/state_edwatch/2016/02/lamar_alexander.html

Bush, G.W. (2002). President Signs Landmark No Child Left Behind Education Bill, The White House, President George W. Bush, Retrieved from: https://georgewbush-whitehouse.archives.gov/news/releases/2002/01/20020108-1.html

Carey, K. (2007). *The Pangloss index: How states game the No Child Left Behind Act*. Washington, DC: Education Sector.

Common Core of Data, https://nces.ed.gov/ccd/

Davidson, E., Reback, R., Rockoff, J.E., Schwartz, H.L. (2013). Fifty ways to leave a child behind: Idiosyncrasies and discrepancies in states' implementations of NCLB. NBER Working Paper 18988.

Dee, T., & Jacob, B. (2011). The impact of No Child Left Behind on student achievement. *Journal of Policy Analysis and Management, 30*(3), 418-446.

Hamilton, L. S., Stecher, B. M., Marsh, J. A., McCombs, J. S., Robyn, A., Russell, J. L., et al. (2007). *Standards-based accountability under No Child Left Behind: Experiences of teachers and administrators in three states*. Santa Monica, CA: RAND Corporation.

Hanushek, E. A., & Raymond, M. E. (2005). Does school accountability lead to improved student performance? *Journal of Policy Analysis and Management*, 24, 297–327

McLaughlin, D.H., Bandeira de Mello, V., Blankenship, C., Chaney, K., Esra, P., Hikawa, H., Rojas, D., William, P., and Wolman, M. (2008). Comparison Between NAEP and State Mathematics Assessment Results: 2003 (NCES 2008-475). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.

McLaughlin, D.H., Bandeira de Mello, V., Blankenship, C., Chaney, K., Esra, P., Hikawa, H., Rojas, D., William, P., and Wolman, M. (2008). Comparison Between NAEP and State Reading Assessment Results: 2003 (NCES 2008-474). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.

Rogasa, D. (2003). The "99% confidence" scam: Utah-style calculations. Stanford University.

Srikantaiah, D. (2009). *How state and federal accountability policies have influenced curriculum and instruction in three states: Common findings from Rhode Island, Illinois, and Washington.* Washington, DC: Center on Education Policy, 2.

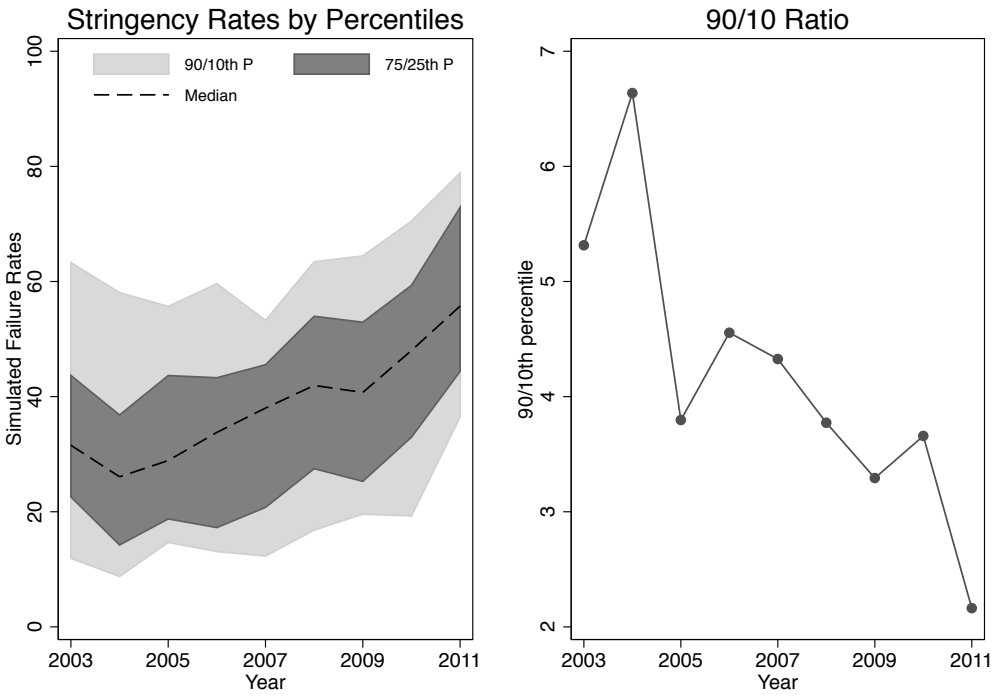
Taylor, Stecher, O'Day, et al. (2010). State and Local Impelmentation of the No Child Left Behind Act. Volume IX – Accountability Under NCLB: Final Report. Washington, DC: U.S. Department of Education.

U.S. Census Bureau; American Community Survey, http://factfinder2.census.gov.

U.S. Department of Education. Institute of Education Sciences, National Center for Education Statistics.

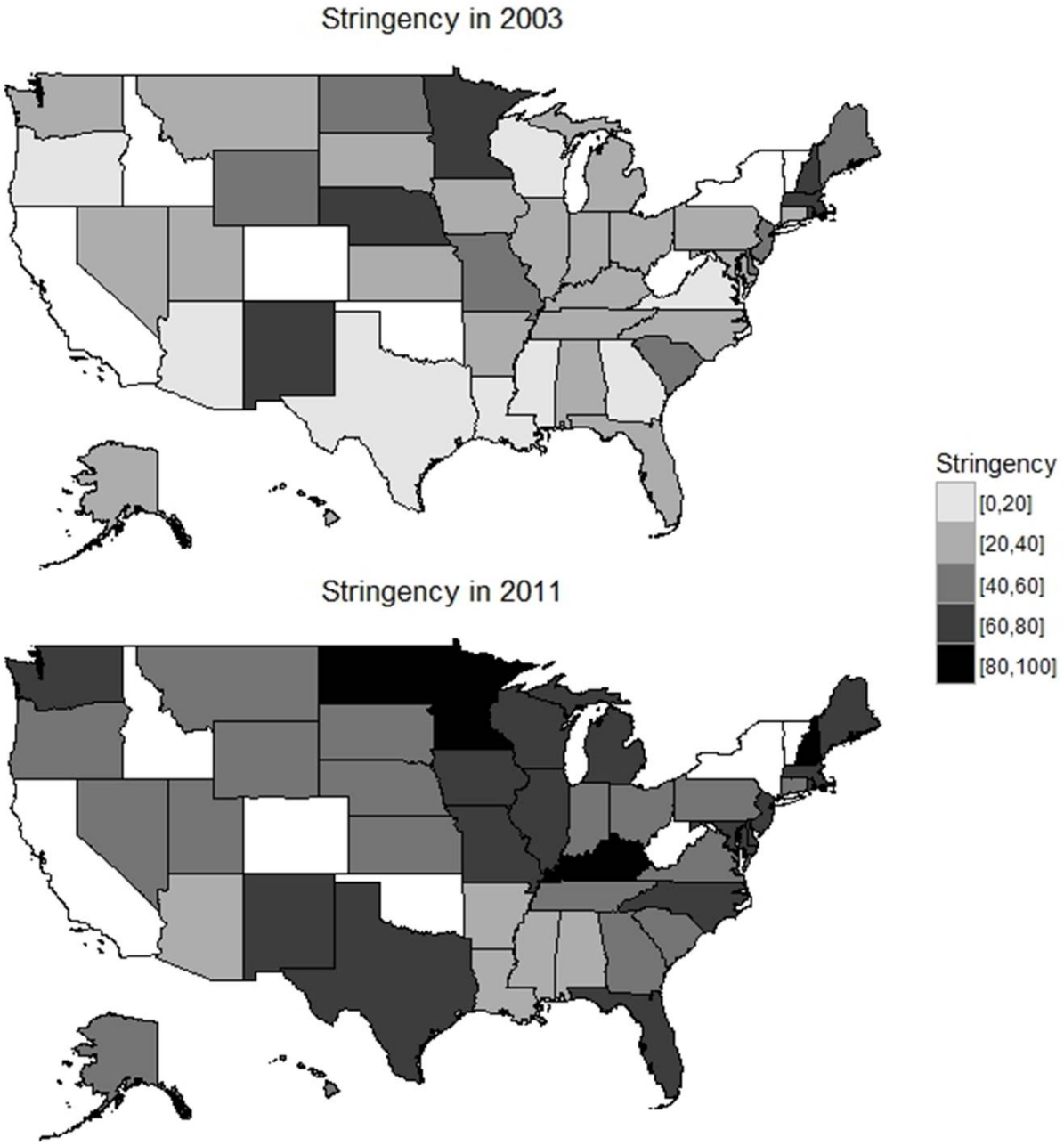
Wong, M., Cook, T. D., & Steiner, P. M. (2014). No Child Left Behind: An interim evaluation of its effects on learning using two interrupted time series each with its own non-equivalent comparison series. *Journal of Research on Educational Effectiveness.*

Figure 1: National Trends in States' Implementation of Accountability Policies



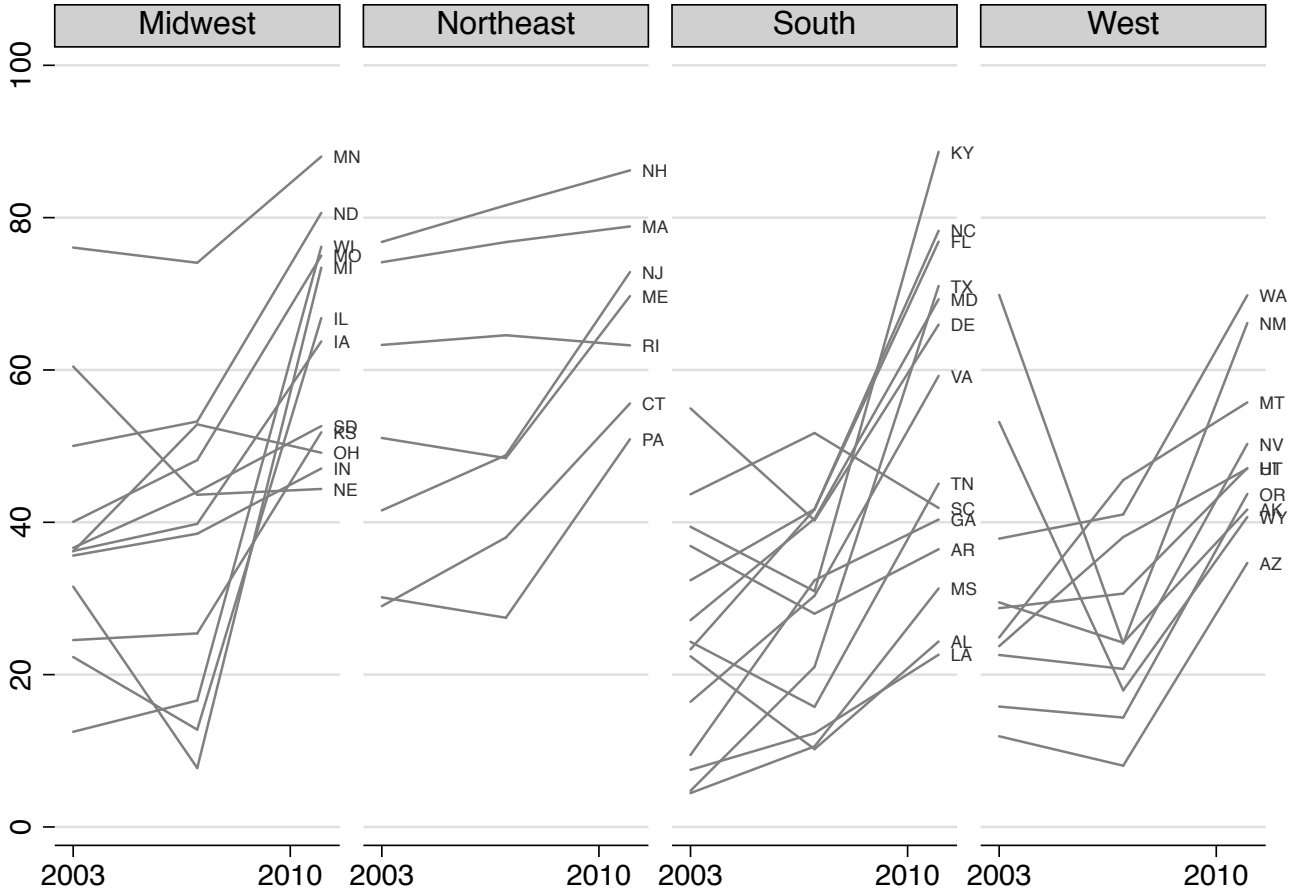
EdPolicyWorks Working Paper Series No. 51. June 2016.
Available at http://curry.virginia.edu/edpolicyworks/wp
Curry School of Education | Frank Batten School of Leadership and Public Policy | University of Virginia

Figure 2: Simulated Failure Rates Across the US in 2003 and 2011



EdPolicyWorks Working Paper Series No. 51. June 2016.
Available at http://curry.virginia.edu/edpolicyworks/wp
Curry School of Education | Frank Batten School of Leadership and Public Policy | University of Virginia

Figure 3: Changes in Simulated AYP Stringency by Census Region



EdPolicyWorks Working Paper Series No. 51. June 2016.
Available at http://curry.virginia.edu/edpolicyworks/wp
Curry School of Education | Frank Batten School of Leadership and Public Policy | University of Virginia

Table 1. Summary of AYP Rules for four states in 2006-2007

	Pennsylvania	Alaska	Tennessee	Texas
Participation requirement	95%	95%	95%	95%
Minimum subgroup size	40	20	45	50
State AMO				
Elem Math	56	66.1	86	50
Middle Math	56	66.1	86	50
High Math	56	66.1	83	50
Elem ELA	63	77.2	89	60
Middle ELA	63	77.2	89	60
High ELA	63	77.2	93	60
Confidence Interval Rule	95%	99%	95%	No
Safe Harbor Rule	Yes	Yes	Yes	Yes
Confidence Interval around Safe Harbor	75%	75%	No	No
Multi-year averages	2 Years	No	No	No
Attendance rate	90%	85%	93%	90%
Graduate rate	80%	55.6%	90%	70%
Actual AYP school failure rate	.22	.34	.13	.09
Simulated AYP school failure rate for Pennsylvania Schools	.19	.47	.62	.32

EdPolicyWorks Working Paper Series No. 51. June 2016.
Available at http://curry.virginia.edu/edpolicyworks/wp
Curry School of Education | Frank Batten School of Leadership and Public Policy | University of Virginia

Table 2: Predictors of State Accountability Stringency

	(1)	(2)	(3)	(4)	(5)	(6)
Trend						
2003	--	--	--	--	--	--
	--	--	--	--	--	--
2004	-5.31**	-5.31**	-4.46**	-4.72**	-4.83**	-4.96**
	(0.95)	(0.95)	(1.40)	(1.36)	(1.35)	(1.33)
2005	-1.17	-1.17	-0.97	-0.77	-0.81	-0.71
	(1.73)	(1.74)	(1.88)	(1.93)	(1.92)	(1.99)
2006	-0.13	-0.27	0.18	0.69	0.67	1.30
	(2.01)	(1.99)	(2.18)	(2.25)	(2.24)	(2.34)
2007	0.49	0.49	1.27	2.47	2.61	3.59
	(2.11)	(2.12)	(2.48)	(2.63)	(2.67)	(2.72)
2008	6.15**	6.15**	6.81*	8.24**	8.37**	9.20**
	(2.14)	(2.15)	(2.56)	(2.72)	(2.76)	(2.77)
2009	6.93**	6.93**	7.35**	7.76**	7.61**	7.76**
	(2.17)	(2.18)	(2.32)	(2.61)	(2.61)	(2.78)
2010	11.8**	11.8**	12.3*	9.26+	8.25	6.24
	(2.80)	(2.81)	(4.63)	(4.68)	(4.91)	(4.93)
2011	23.3**	23.3**	23.3**	19.9**	18.8**	16.6**
	(2.89)	(2.90)	(5.05)	(5.06)	(5.20)	(5.22)
Census Regions						
West		--	--	--	--	--
		--	--	--	--	--
Midwest		10.1+	13.5*	14.4*	14.7*	13.1
		(5.38)	(5.11)	(6.00)	(6.11)	(8.18)
Northeast		23.8**	19.8**	19.5**	20.1**	20.7**
		(7.41)	(6.30)	(6.26)	(6.04)	(6.23)
South		-2.46	4.47	18.5*	20.9*	17.3*
		(4.61)	(4.92)	(7.14)	(8.82)	(8.50)
State Population Characteristics						
Ln(Population)			-4.97*	-3.80+	-3.96+	-2.86
			(1.90)	(2.13)	(2.16)	(1.95)
Unemployment Rate			-0.082	1.00	1.26	1.93+
			(0.95)	(0.94)	(1.02)	(0.98)
Democrat			--	--	--	--
			--	--	--	--
Independent			20.1+	17.0	16.5	17.5*

EdPolicyWorks Working Paper Series No. 51. June 2016.
Available at http://curry.virginia.edu/edpolicyworks/wp
Curry School of Education | Frank Batten School of Leadership and Public Policy | University of Virginia

			(11.9)	(11.8)	(11.7)	(8.15)
Republican			3.03	3.90	3.87	1.53
			(3.01)	(3.01)	(3.01)	(2.65)
% 25 and Older with Bachelors			1.30*	1.47**	1.52**	2.02**
			(0.50)	(0.47)	(0.48)	(0.46)
Student Population Characteristics						
% of Black Students				-0.54**	-0.55**	-0.63**
				(0.19)	(0.20)	(0.19)
% of White Students				0.14	0.14	0.42*
				(0.092)	(0.094)	(0.18)
% of Hispanic Students				-0.011	0.026	0.16
				(0.18)	(0.19)	(0.20)
Education Policy						
Has Consequential Accountability Pre-NCLB					-2.62	-2.99
					(4.77)	(5.27)
Student Achievement Pre-NCLB						
1998 NAEP Grade 4 Reading						0.51
						(0.73)
1998 NAEP Grade 8 Reading						-2.84**
						(0.83)
2000 NAEP Grade 4 Math						0.28
						(0.62)
2000 NAEP Grade 8 Math						0.58
						(0.54)
Constant	34.7**	28.9**	23.7**	18.6**	21.6**	27.7**
	(2.94)	(3.60)	(4.40)	(4.73)	(6.96)	(7.01)
Observations	386	386	386	386	386	386
Adjusted R-squared	0.147	0.362	0.459	0.516	0.517	0.573

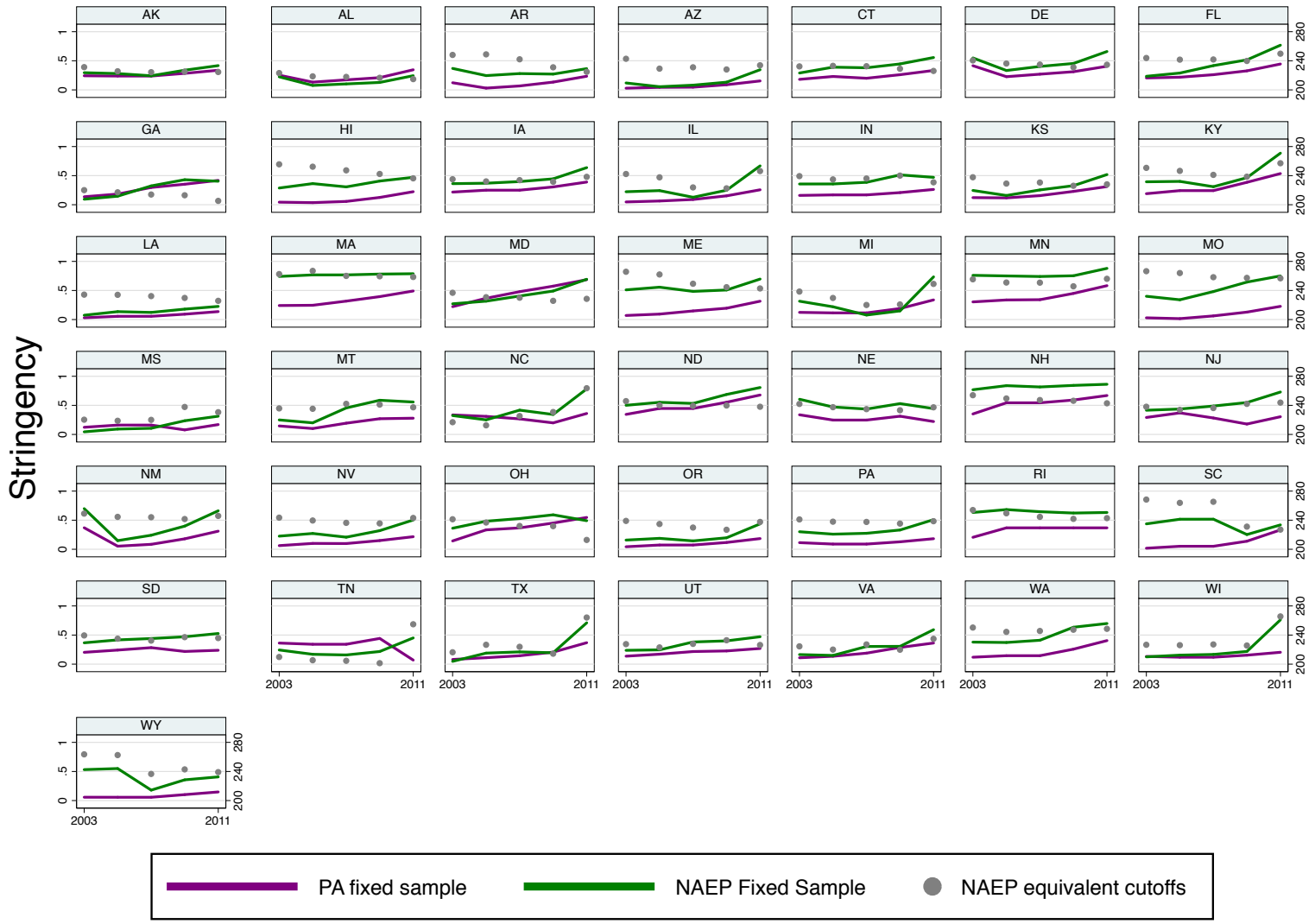
Standard error in parentheses; p-values + < 0.10; * < 0.05; ** < 0.10

EdPolicyWorks Working Paper Series No. 51. June 2016.
Available at http://curry.virginia.edu/edpolicyworks/wp
Curry School of Education | Frank Batten School of Leadership and Public Policy | University of Virginia

Appendix A

Figure A1: Comparison of Stringency Rates for NAEP versus Pennsylvania Fixed Samples with NAEP Equivalent Test Benchmarks

This figure demonstrates simulated AYP failure rates using the Pennsylvania (purple line) and NAEP (green line) fixed samples. The grey dots represent the NCES NAEP equivalent cutoff scores (McLauglin et al., 2008). Although there are differences in population characteristics between the two fixed samples, trends in the simulated failure rates are similar for most states. Divergence in simulated failure rates for the NAEP and Pennsylvania fixed samples often occur when test difficulty changes (as indicated by changes in the grey dots). Increases in NAEP equivalent cutoff scores result in more stringent accountability policies, decreases in NAEP equivalent scores result in less stringent accountability policies. All results presented in this paper are robust for both the NAEP and Pennsylvania fixed samples (see Tables A2 and A3).



EdPolicyWorks Working Paper Series No. 51. June 2016.
Available at http://curry.virginia.edu/edpolicyworks/wp
Curry School of Education | Frank Batten School of Leadership and Public Policy | University of Virginia

Table A1 (NAEP Fixed Sample):
Simulated AYP Failure Rates (%) in 2003, 2005, 2007, 2009, and 2011

State	2003	2005	2007	2009	2011
AK	29	28	24	34	42
AL	22	8	10	13	24
AR	37	25	28	27	36
AZ	12	5	8	13	35
CT	29	39	38	45	56
DE	55	33	40	45	66
FL	23	29	42	52	77
GA	9	15	32	43	40
HI	29	36	31	41	47
IA	36	37	40	45	64
IL	22	24	13	25	67
IN	36	36	39	51	47
KS	25	16	25	33	52
KY	39	40	31	47	89
LA	7	14	12	18	23
MA	74	77	77	78	79
MD	27	32	40	49	69
ME	51	56	48	51	70
MI	32	22	8	15	73
MN	76	75	74	75	88
MO	40	34	48	64	75
MS	4	9	11	24	31
MT	25	20	46	59	56
NC	32	25	42	34	78
ND	50	55	53	69	81
NE	60	47	44	53	44
NH	77	84	82	84	86
NJ	42	44	49	55	73
NM	70	15	24	40	66
NV	23	27	21	32	50
OH	36	48	53	59	49
OR	16	19	14	20	44
PA	30	26	27	33	51
RI	63	68	65	62	63
SC	44	52	52	25	42
SD	37	42	44	47	53
TN	24	17	16	22	45
TX	5	19	21	20	71
UT	24	25	38	40	47
VA	16	15	30	31	59

EdPolicyWorks Working Paper Series No. 51. June 2016.
Available at http://curry.virginia.edu/edpolicyworks/wp
Curry School of Education | Frank Batten School of Leadership and Public Policy | University of Virginia

WA	38	37	41	63	70
WI	12	15	17	22	76
WY	53	55	18	36	41

EdPolicyWorks Working Paper Series No. 51. June 2016.
Available at http://curry.virginia.edu/edpolicyworks/wp
Curry School of Education | Frank Batten School of Leadership and Public Policy | University of Virginia

Table A2 (Pennsylvania Fixed Sample):
Simulated AYP Failure Rates (%) in 2003, 2005, 2007, 2009, and 2011

State	2003	2005	2007	2009	2011
AK	46	47	47	54	60
AL	36	41	44	49	61
AR	17	5	15	28	47
AZ	7	9	13	16	33
CT	42	49	49	57	64
DE	50	33	50	55	65
FL	50	53	58	67	76
GA	27	26	49	55	62
HI	9	11	15	27	47
IA	31	31	45	53	61
IL	13	16	21	35	51
IN	33	35	35	43	51
KS	64	39	32	47	57
KY	74	65	39	54	68
LA	6	9	14	21	31
MA	36	45	60	67	74
MD	39	65	74	82	88
ME	14	16	31	40	58
MI	18	17	26	37	61
MN	26	54	58	67	77
MO	52	27	16	28	46
MS	10	16	29	29	42
MT	20	15	40	52	54
NC	56	56	52	42	61
ND	33	44	64	74	83
NE	55	55	55	62	51
NH	42	57	79	82	86
NJ	43	49	54	41	57
NM	44	16	23	36	56
NV	16	19	27	35	45
OH	29	57	58	66	74
OR	7	10	16	25	38
PA	19	16	19	27	37
RI	23	62	62	62	62
SC	7	8	16	35	58
SD	26	28	50	41	45
TN	53	51	62	70	27
TX	19	25	33	44	63
UT	21	24	45	47	53
VA	15	18	39	51	59

EdPolicyWorks Working Paper Series No. 51. June 2016.
Available at http://curry.virginia.edu/edpolicyworks/wp
Curry School of Education | Frank Batten School of Leadership and Public Policy | University of Virginia

WA	18	21	30	48	69
WI	17	16	25	33	41
WY	12	10	14	24	33

Table A3 (Pennsylvania Fixed Sample):
Predictors of State Accountability Stringency

	(1)	(2)	(3)	(4)	(5)	(6)
Trend						
2003	--	--	--	--	--	--
	--	--	--	--	--	--
2004	-2.69+	-2.69+	-3.56**	-3.82**	-3.89**	-3.64**
	(1.35)	(1.35)	(1.28)	(1.30)	(1.30)	(1.33)
2005	2.16	2.16	0.54	0.72	0.69	0.92
	(1.98)	(1.99)	(2.08)	(2.16)	(2.17)	(2.20)
2006	7.20*	6.95*	5.83*	6.34*	6.33*	6.69*
	(2.79)	(2.77)	(2.69)	(2.78)	(2.78)	(2.88)
2007	9.57**	9.57**	9.01**	10.2**	10.3**	10.6**
	(2.66)	(2.67)	(2.78)	(2.84)	(2.87)	(2.97)
2008	15.2**	15.2**	14.4**	15.8**	15.9**	16.1**
	(2.69)	(2.70)	(3.00)	(3.08)	(3.12)	(3.17)
2009	17.1**	17.1**	15.5**	15.9**	15.8**	15.9**
	(2.56)	(2.57)	(2.63)	(2.85)	(2.85)	(2.94)
2010	18.4**	18.4**	15.4**	12.3*	11.6*	11.6*
	(2.89)	(2.90)	(5.09)	(4.98)	(5.19)	(5.13)
2011	27.0**	27.0**	23.2**	19.7**	19.0**	19.0**
	(2.78)	(2.79)	(5.39)	(5.25)	(5.48)	(5.47)
Census Regions						
West		--	--	--	--	--
		--	--	--	--	--
Midwest		12.7*	16.1**	16.9**	17.1**	8.95
		(4.92)	(4.13)	(4.28)	(4.46)	(5.80)
Northeast		18.4**	11.8*	11.1*	11.5*	6.48
		(6.81)	(5.10)	(5.32)	(5.62)	(5.99)
South		11.9*	18.7**	32.4**	34.0**	29.7**
		(5.54)	(4.98)	(7.62)	(8.89)	(8.22)
State Population Characteristics						
Ln(Population)			-4.08+	-3.16	-3.26	-2.14
			(2.06)	(1.98)	(1.99)	(2.03)
Unemployme nt Rate			0.33	1.41	1.58	1.62
			(1.22)	(1.13)	(1.19)	(1.14)
Democrat			--	--	--	--
			--	--	--	--

EdPolicyWorks Working Paper Series No. 51. June 2016.
Available at http://curry.virginia.edu/edpolicyworks/wp
Curry School of Education | Frank Batten School of Leadership and Public Policy | University of Virginia

Independent			-12.8**	-15.7**	-16.0**	-9.48+
			(3.50)	(3.70)	(3.77)	(5.55)
Republican			3.27	4.22+	4.20+	3.38
			(2.71)	(2.44)	(2.45)	(2.72)
% 25 and Older with Bachelors			1.87**	2.06**	2.09**	2.22**
			(0.46)	(0.47)	(0.48)	(0.46)
Student Population Characteristics						
% of Black Students				-0.49*	-0.50*	-0.48*
				(0.24)	(0.25)	(0.22)
% of White Students				0.18	0.18	0.27
				(0.13)	(0.12)	(0.18)
% of Hispanic Students				0.059	0.084	0.18
				(0.18)	(0.19)	(0.18)
Education Policy						
Has Consequential Accountability Pre-NCLB					-1.73	1.42
					(3.80)	(4.13)
Student Achievement Pre-NCLB						
1998 NAEP Grade 4 Reading						0.99
						(0.83)
1998 NAEP Grade 8 Reading						-1.97
						(1.17)
2000 NAEP Grade 4 Math						-0.70
						(0.68)
2000 NAEP Grade 8 Math						0.95
						(0.62)
Constant	29.6**	19.2**	16.0**	11.1*	13.0*	11.7
	(2.68)	(4.01)	(3.89)	(4.35)	(5.09)	(7.06)
Observations	386	386	386	386	386	386

Adjusted R-squared	0.212	0.310	0.455	0.518	0.518	0.567

Standard error in parentheses; p-values + < 0.10; * < 0.05; ** < 0.10

EdPolicyWorks Working Paper Series No. 51. June 2016.
Available at http://curry.virginia.edu/edpolicyworks/wp
Curry School of Education | Frank Batten School of Leadership and Public Policy | University of Virginia