Working Paper:

The Impact of Intensifying State Accountability Pressures on Student Achievement under No Child Left Behind

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

Accountability reform has been at the forefront of the domestic policy agenda. In 2015, the House and Senate passed bills to replace No Child Left Behind (NCLB). In both versions of the bill, the centerpiece of NCLB—identifying failing schools through the standardized testing—remained intact. What remain in contention, however, are the policy levers that states use to identify and address failing schools. As states move to implement their revised accountability plans under the Every Student Succeeds Act (ESSA), we present new evidence on how increased state accountability pressures affect student achievement. We address prior methodological challenges for evaluating NCLB by introducing a new implementation measure of states' accountability policies from 2003 to 2011 (pre-waiver period). Importantly, the measure describes variation in states' accountability standards but is independent of school and student population characteristics within each state. Using our implementation measure, we estimate the causal effects of increasing state accountability pressures on student reading and math performance. Overall, we find that increased implementation stringency resulted in small improvements in eighth-grade math achievement but had no effects on fourth-grade math or reading outcomes. The study concludes that ratcheting accountability pressures alone is not enough to sustain improvements in student achievement.

[1]University of Virginia [2]Indiana University

Updated July 2018

EdPolicyWorks
University of Virginia, PO Box 400879
Charlottesville, VA 22904

Acknowledgements: The authors wish to thank Jim Wyckoff, Daphna Bassok, Luke Miller, Dan Player, Tom Dee, and all the members of Ed Policy Works for their valuable contributions and feedback. All errors are our own. This research was supported by a grant from the American Educational Research Association which receives funds for its "AERA Grants Program" from the National Science Foundation under NSF Grant #DRL-0941014. Opinions reflect those of the author(s) and do not necessarily reflect those of the granting agencies.

EdPolicyWorks Working Paper Series No. 63. July 2018.
Available at http://curry.virginia.edu/edpolicyworks/wp
Curry School of Education | Frank Batten School of Leadership and Public Policy | University of Virginia

1. INTRODUCTION

The No Child Left Behind Act (NCLB) of 2002 gave the federal government the authority to hold schools accountable to uniform standards. The headline goal of NCLB was to make every student "proficient" in math and reading by 2014. Early impact evaluations found improvements in math but not in reading (Dee & Jacob, 2011; M. Wong, Steiner, & Cook, 2015). But even in math, improvements fell far short of the goal of 100 percent proficiency. Today, NCLB is synonymous with the overuse of standardized testing and the limitations of top-down, one-size approaches to education policy. In 2015, Congress replaced NCLB with the Every Student Succeeds Act (ESSA). ESSA maintains NCLB's focus on annual testing in reading and math but devolves many responsibilities associated with school accountability to state and local actors.

Some observers argue that ESSA marks the return of state governments to American education policy (Burnette, 2016). It is not clear, however, that NCLB was as centralized as its critics imply. State governments had substantial discretion in implementing NCLB standards (Davidson, Reback, Rockoff, & Schwartz, 2015; Taylor, Stecher, O'Day, Naftel, & Le Floch, 2010). They selected their own standardized tests, defined their own standards for determining proficiency, and set their own trajectories for reaching complete proficiency by 2014. Moreover, many states applied for and were granted exemptions that allowed them to declare some schools "proficient" even when they did not meet Annual Measurable Objectives (AMO). Implementation discretion introduced considerable variation in NCLB accountability stringency across states and over time. The same students, teachers, principals, and schools deemed "proficient" in one state would have been candidates for remediation or even school closure in another state (Wong, Wing, Martin, & Krishnamachari, 2018).

In this study, we examine how variations in NCLB implementation stringency affected student performance on the National Assessment of Educational Progress (NAEP). The NAEP is sometimes referred to as the "nation's report card of what students know and can do in different subject areas" (NCES, 2018). To measure NCLB implementation stringency, we built a database of state accountability rules from 2003 to 2011 and developed an AYP Calculator, which determines whether a particular school would have been evaluated as "proficient" under each state's rules for the year. We used the calculator to estimate the fraction of a *fixed sample* of schools that would have failed to meet Adequate Yearly Progress (AYP) standards in every state and year. Because they measure how each state would "treat" the same population of schools, these simulated failure rates provide a concrete measure of the stringency of each state's NCLB implementation. Simulated stringency rates vary across states and years because of differences in state implementation details. Importantly, the simulated rates do not depend on the characteristics and realized educational outcomes of the students living in the state. We estimate the effects of NCLB stringency on reading and math outcomes on the NAEP using a generalized differences-in-differences strategy (Wing, Simon, & Bello-Gomez, 2018). Our results suggest that higher NLCB accountability stringency led to small but positive effects on students' eighth-grade math scores. These effects were largest for the lowest-performing (10th percentile) students and for students who were at risk of low performance (Hispanic, English-language learners, and students with disabilities). NCLB stringency may have had small effects on eighth-grade reading scores, but the estimates are sensitive to model specification. We found no evidence that stringency affected fourth-grade math or reading scores.

2. BACKGROUND

Under NCLB, states used multiple indicators to determine whether schools met Adequate Yearly Progress (AYP) standards. A school "passed" AYP if all of its student subgroups met test participation and proficiency thresholds, and the school as a whole met Other Academic Indicator (OAI) standards for attendance, graduation, and other measures. A school "failed" if even one subgroup missed the participation or proficiency threshold, or if the school did not meet OAI standards. Schools may have up to eight student subgroups held accountable for performance. They include: five racial/ethnic groups, students from low-income families, students with limited English proficiency, and students with disabilities. Schools that failed to make AYP over the course of several consecutive years faced corrective action from the state, such as increased oversight, reconstitution, or closure.

Previous quasi-experimental evaluations of NCLB took advantage of the fact that many states had adopted their own accountability reforms before the federal NCLB passed in 2002. The idea is that the federal mandate was (essentially) non-binding in states with existing standards and consequential in states where accountability standards were genuinely new. Using this approach, Dee and Jacob (2011) found that by 2007, average fourth-grade math achievement improved by approximately 8.21 points on the NAEP. Effects were largest for at-risk students and students in the lowest-performing percentiles. There was also evidence of improvement in average eighth-grade math performance.

Despite evidence of NCLB's initial impact, there was tremendous variation in how the federal mandate was implemented by states. In many states, students performed better on their state's NCLB assessments than they would on the NAEP, which suggests that some states may have adopted easier standards than others (McLaughlin, Bandeira de Mello, Blankenship, Chaney, Esra, Hikawa, Rojas, William & Wolman, 2008). Carey (2007) found that some states pursued steady trajectories for reaching the 2014 performance target, while others adopted a more backloaded trajectory. And states applied for and were granted "exemptions" that effectively adjusted proficiency requirements for schools and subgroups. For example, some states allowed schools to construct a "confidence interval" around their AMO target so that they only needed to achieve a target score that was equivalent to the lower bound of the confidence interval. Under Safe Harbor rules, states examined school subgroup performance from the prior year and exempted failing schools as long as some specified fraction (e.g. 10 percent) of previously non-proficient students became proficient the following year. Multiyear Averaging rules allowed schools to adjust AMOs by averaging subgroup percent proficiency from a given year with proficiency performance from previous years. Some exemption policies may have promoted reliable and valid AYP designations, but there was controversy over the legitimacy of the adjustments (Rogasa, 2003).

More recent evaluations have linked state implementation choices with school and student outcomes. Davidson, Reback, Rockoff, and Schwartz (2015) found that school failure rates correlated with whether states had confidence interval and minimum subgroup size rules. Wei (2012) developed a predictive model that linked state population characteristics with the adoption of more stringent AYP rules and studied the correlation between these predictions and students' NAEP scores. She found that state NCLB stringency was associated with negative achievement outcomes for White and Hispanic, but not Black, students. Finally, M. Wong et al. (2015) constructed a measure of difficulty in the assessment chosen by each state and found that math test scores improved more in states that adopted rigorous exams than in states with easier tests.

These studies provide some evidence that implementation stringency may matter. But they generally do not attempt to summarize the overall accountability pressure created by the package of implementation details adopted in each state. A core question is whether schools under more intense accountability regimes produce greater gains in student achievement than schools in states with more lenient standards. This question remains relevant as states revise their accountability plans under ESSA.

3. METHODS

The overarching goal of our study is to measure the effects of NCLB implementation stringency on math and reading outcomes. The idea is easy to understand in terms of a simple (naïve) regression model:

$$[1] \qquad Achievement_{st} = \beta_1 Stringency_{st} + \epsilon_{st} \;.$$

In the model, s indexes states and t indexes years. $Achievement_{st}$ is the average math or reading score in the state-year. And $Stringency_{st}$ is the difficulty of NCLB accountability standards in the state-year. β_1 represents the incremental effect of a small increase in state accountability stringency on student achievement.

The model poses two methodological challenges. First, there is no commonly accepted implementation measure for describing accountability stringency. State NCLB implementations involve multiple rules that relate in complicated ways. Previous research has focused on a single dimension of accountability policy (M. Wong et al., 2015) or examined multiple indicators of stringency in small subsets of states and years (Hamilton, Stecher, Marsh, & McCombs, 2007; Srikantaiah, 2009; Taylor et al., 2010). Second, implementation stringency is likely correlated with other state characteristics that determine student achievement, so estimates of the effect of stringency on outcomes based on the model would likely suffer from omitted variable bias.

In this paper, we make progress on both of these challenges. We make it feasible to study the link between stringency and outcomes by computing simulated AYP failure rates in a fixed sample of schools. We use the simulated failure rates to summarize implementation stringency in each state over time. Because the simulated failure rates are determined for a fixed population of schools, they are not correlated with the characteristics of schools in the state or with changes in school performance in a state from year to year. To further control for possible confounding from unmeasured variables, we estimate the effects of NCLB stringency in a state using models that control for a full set of state and year fixed effects.

This section of the paper provides a short summary of the AYP Calculator and the simulated failure rates. The Methodology Appendix in Wong et al. (2018) describes the AYP Calculator and simulated failure rates in detail, reports our analysis of the sensitivity of the simulated failure rates to alternative sample specifications, and presents empirical validation checks of the measure.

The AYP Calculator

We compiled publicly available information about state accountability plans to create a database of all AYP rules for each state and year in the NCLB pre-waiver period from 2002-2003 to 2010-2011.[1] Our database accounts for all state AYP requirements for 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

[1] All state AYP rules were collected and coded from archived files of "State Accountability Plans" and "State Decision Letters" on the Department of Education Website. See the "Accountability Rule Database" section of the Methodology Report in Wong et al. (2018).

information about Confidence Intervals, Safe Harbor, Confidence Intervals Around Safe Harbor, Multi-Year Averages, Performance and Proficiency Indices, and modified versions of these rules.

We developed an algorithm that takes school characteristics and outcomes (e.g. subgroup sizes and proficiency levels), applies state accountability rules, and determines whether the school failed to meet AYP standards in the state and year. Specifically, the algorithm classifies a school as passing AYP if—after applying all relevant exemptions and criteria—the school met the OAI school standard and all student subgroups in the school met the test participation and proficiency requirements.

To get a sense of how this works, let α_{gmcst} and β_{gmcst} be the minimum thresholds for the participation and proficiency requirements that subgroup *g* must achieve in subject area *m* in school *c* in state *s* in year *t*. Let γ_{cst} be the OAI requirement for school *c* in state *s* and year *t*. These are the "final" thresholds that prevail after the state's exemption rules (e.g. Confidence Interval, Safe Harbor) have been applied.[2] Next, let X_{1gmc}, X_{2gmc}, and X_{3mc} represent school-specific characteristics, where X_{1gmc} and X_{2gmc} are the participation and proficiency rates for each subgroup g in subject area *m* in school *c*, and X_{3c} is school *c*'s performance on the OAI measure. We focus on a fixed sample of schools, which is why there are no state or time subscripts associated with the school input variables. A school's AYP status (*P*) in state *s* and year *t* is:

$$[2] \qquad A_{cst}^{P} = 1\left[X_{1gmc} \geq \alpha_{gmcst}\right] \times 1\left[X_{2gmc} \geq \beta_{gmcst}\right] \times 1[X_{3c} \geq \gamma_{cst}]$$

[2] The Methodology Appendix (Wong et al., 2018) describes how each accountability rule was operationalized in the Calculator.

In the expression, 1[.] is an indicator function that takes a value of 1 if the statement in the brackets is true. Thus, $A_{cst}^{P} = 1$ if school *c* met all three batches of AYP requirements in state *s* and year *t*. A_{cst}^{P}= 0 if the school failed *any* of the requirements. By evaluating the expression using the parameters from each state and year, we are able to evaluate how the same vector of school attributes would be classified in each state and year in our study.

State Stringency Scores

In the second step, we "fed" a sample of school input characteristics through the AYP Calculator. The result is a school-level dataset that received 51 new variables for each year of the simulation. For example, $A_{c,NY,2005}$ indicates whether school *c* would have met AYP standards in New York state in 2005. $\overline{A_{st}} = \frac{1}{C}\sum_{c=1}^{C} A_{cst}$ is the proportion of schools in the fixed sample that would have met AYP standards in state *s* in year *t*. For ease of interpretation, we recoded $\overline{A_{st}}$ to represent the percent of schools in the fixed sample that was simulated to have failed AYP for the year, such that $\overline{F_{st}} = (1 - \overline{A_{st}}) \times 100$. States with higher simulated AYP failure rates had more stringent accountability standards, and states with lower simulated failure rates had more lenient policies.

Adjusting for State Test Difficulty

One concern with the implementation measure is that it fails to capture differences in state test difficulty. This is problematic because prior research on NCLB implementation suggests that some states used easier tests for assessing student proficiency (Taylor et al., 2010). To incorporate test difficulty in our measure, we used NCES Reports that map state proficiency thresholds onto NAEP equivalent scales for fourth- and eighth-grade students. We obtained NAEP equivalent cutoff scores for 2003, 2005, 2007, 2009, and 2011.[3]

We used the NAEP equivalent scores to adjust the AYP Calculator so that it allowed a NAEP fixed sample of students and determined the proficiency status for each student by comparing his or her NAEP achievement score to the NAEP equivalent proficiency cutoff. For example, the NAEP equivalent cutoff scores for fourth-grade reading proficiency in 2007 were 163 in Mississippi and 199 in Indiana. In our Calculator, a fourth-grade student with a reading score of 170 would be proficient in Mississippi but not proficient in Indiana. The procedure incorporates test difficulty into the stringency measure. States with more difficult test assessments had higher NAEP equivalent cutoffs, and states with easier tests had lower cutoffs. If test difficulty changed over time, changes would be reflected in the NAEP equivalence cutoffs and incorporated into our stringency measure.

NAEP Fixed Sample

The logic of our simulated failure rate approach does not depend on the details of the specific fixed sample we use. We could use a sample from a single state, or a completely hypothetical sample. The key is to understand how each state's AYP policies would evaluate the same collection of students and schools. Nevertheless, we used a nationally representative sample of students and schools from the 2009 NAEP survey.[4]

[3] See NCES NAEP State Mapping Project at: https://nces.ed.gov/nationsreportcard/studies/statemapping. In years where NCES-NAEP scores were unavailable, we interpolated using the state's equivalent cutoffs from previous and subsequent years. Our final analysis includes NAEP equivalent cutoffs for every state, subject, and year.

[4] We used participants from the 2009 NAEP because we were able to link to prior year information from earlier samples of the 2005 and 2007 NAEP. Prior year information was important for implementing Safe Harbor and Multiyear Averaging rules in the Calculator. We were unable to link to prior year information from earlier NAEP samples (pre-2003).

In practice, the NAEP does not provide a single reading and math achievement score. Instead, it reports five "plausible" subject scores for each student in the sample. Thus, our fixed sample compared the average of five subject-specific plausible value scores on NAEP achievement tests to the scale equivalent cutoffs for each state and year to determine students' proficiency performance.[5] Because the NAEP provides information about students' subgroup membership, we could calculate school subgroup performance by averaging the proficiency status of students identified as part of each school subgroup.

However, determination of schools' AYP status required additional information that was not included in the NAEP fixed sample. Because the NAEP includes random samples of students within schools, it does not have school-level information about subgroup sizes, test participation rates, or attendance and graduation rates. Because the Calculator requires only that we use fixed input characteristics, we could have in theory simulated these values. However, we used actual (or approximated) school information to inject realism into our simulations. Specifically, we used NCES School IDs to link NAEP schools with information from the Common Core Data (CCD) and the Barnard Columbia NCLB Data Project. The CCD provides information about school subgroup sizes, attendance rates, and graduation rates. The Barnard NCLB dataset provides information about test participation rates for most schools in the U.S.[6]

[5] We used input information for the fixed sample from the 2009 NAEP because we were able to link 24 percent of schools (N=3,831) to prior performance information for the school. This was needed to incorporate AYP rules such as Multiyear Averaging and Safe Harbor.

[6] The Barnard NCLB data provided test participation rates from 2002 to 2003, while our fixed sample includes schools from the 2008-2009 survey. Although the years from the two surveys do not line up, we believe they provide a reasonable approximation of schools' actual test participation rates. We examined school test participation rates from California and Florida for 2003-2004 and 2008-2009 and found that the year-to-year correlation was around 0.61 (for reading and math combined). Although there is some year-to-year fluctuation, the annual test participation rates are positively and moderately correlated over time.

AYP rules such as Safe Harbor, Multi-Year Averaging, and Confidence Interval around Safe Harbor also required one to two years of prior information about school subgroup performance. To obtain prior information for our fixed sample of schools, we looked at data from additional NAEP surveys and restricted the fixed sample to include only schools with performance information in earlier NAEP samples. Although this drastically reduced the number of NAEP schools in the fixed sample (24 percent of the 2009 NAEP sample), the consistent sampling design across years ensured that the remaining students and schools were demographically similar to NAEP students overall, and to students in the U.S. population overall (see Table A1 in Appendix A of Wong et al. (2018)).

For states that included writing and science proficiency standards as other academic indicators, we used subgroup proficiency performance on the 2009 NAEP science assessment and the NAEP 2007 writing assessment. However, because there are no NAEP equivalent scale cutoffs for alternative subject areas, the Calculator does not account for test difficulty in subject areas other than reading and math.

The final fixed sample includes 3,831 schools, with 94,615 students who took the NAEP assessments in reading and math. About half the sample of students were White (51 percent), 11 percent were Black, 23 percent were Hispanic, 43 percent were economically disadvantaged, and 10 percent had an Individualized Education Program (IEP). One limitation of the NAEP fixed sample is that it includes only fourth and eighth graders, so the measure reflects accountability stringency only for elementary and middle school students. Sensitivity analyses indicate that our stringency ratings are robust to alternative specifications of the fixed sample (Wong et al., 2018).

Validation Analysis

A key assumption in our study is that the AYP Calculator is able to recreate the procedures the states used to identify failing schools. As explained above, we were able to incorporate most of the discretionary implementation requirements into the Calculator directly, including school, subgroup, and subject requirements for test participation, test proficiency, attendance rates, graduation rates, and various exemption rules (e.g. Confidence Interval, Safe Harbor, Confidence Interval around Safe Harbor, and Multiyear Performance Averages). Nevertheless, a few states adopted rules that could not be easily implemented in the Calculator. For example, we could not include growth model requirements because we lacked longitudinal student data in the NAEP. And we worked with approximations of some of the more complicated state procedures related to performance indices, proficiency indices, modified proficiency and exemption rules, and/or additional academic indicator requirements for additional subject areas, such as science or writing.

Table 1 categorizes states by our subjective assessment of the Calculator's fidelity to the actual rules. In "high-fidelity" states, we were able to describe all of the rules for making school-level AYP decisions and were able to apply these rules to a fixed sample of schools with suitably defined input characteristics. In "partial fidelity" states, we applied procedures that approximated states' performance and proficiency indices, their modified and grade-span rules, and requirements for other subject areas. Eight of the partial fidelity states (Alaska, Arizona, Arkansas, Delaware, Florida, Iowa, North Carolina, and Tennessee) had accountability plans with growth model rules, which we could not include in the Calculator. However, state accountability reports show that growth model rules were inconsequential from 2003 to 2011 in every state except Ohio. Thus, we list Ohio as a state where the Calculator has low coding fidelity.

To empirically validate our qualitative assessment of state coding fidelity, we obtained data from the population of schools for a specific state and year. We then used the Calculator to determine the percent of schools predicted to have failed AYP. The idea here is that if our Calculator performs well, the predicted AYP failure rates should be identical to the state's actual failure rates for the year. Because the validation process is time consuming and data intensive, we were not able to implement it for all states and years. Instead, we obtained data for seven states – California, Pennsylvania, Texas, Maryland, New York, North Carolina, and Ohio – and evaluated the Calculator performance for two time periods in each state (14 validation efforts in total).

The scatterplot in Figure 1 recreates the simulated and reported failure rates for validation efforts originally published in Wong et al. (2018) (See Table B1 in the Methodology Appendix of Wong et al. (2018) for the exact failure rates for each state). Most observations lie close to the 45-degree line, indicating close correspondence between the AYP Calculator and the actual failure rates in the state. Overall, when we were able to obtain complete information about the population of schools, our Calculator performs well in replicating reported AYP failure rates, even in the partial fidelity states where we had to rely on approximations for some rules (California, New York, and North Carolina). When there were substantial differences between our predicted and actual AYP failure rates, it was because we lacked key input information about schools in the state, such as attendance (North Carolina) or grade-specific data (Maryland). In California, we lacked information about small schools in the state. Finally, in Ohio, our predicted failure rate was 35 percent for 2006-2007, compared to the actual AYP failure rate of 38 percent.

However, after the consequential growth model rule was introduced in 2007-08, the predicted versus actual AYP failure rates in Ohio in 2010-11were 55 and 40 percent, respectively. As a result, we omit Ohio from most of our analysis.

Effects of NCLB Implementation Stringency

We studied the effects of NCLB implementation stringency on NAEP math and reading outcomes at the state-year level. The NAEP was administered to a representative random sample of schools and students in each state in 2005, 2007, 2009, and 2011 NAEP assessments. We estimated the effects using the following two-way fixed effects regression model:

$$[3] \qquad Achievement_{st} = \beta_1 \ln\left(\bar{F}_{st-1}\right) + X_{st-1}\beta + \theta_s + \delta_t + \varepsilon_{st}.$$

In the model, $Achievement_{st}$ is the average reading or math performance on NAEP in the state-year. $\ln(\bar{F}_{st-1})$ is the log of the simulated AYP failure rate from the prior academic year. θ_s and δ_t represent a full set of state and year fixed effects, and X_{st-1} is a vector of lagged time varying covariates that were determined prior to the NAEP outcome scores.[7] β_1 represents the effect of state accountability stringency on student achievement. In the empirical analysis, we report estimates of $\beta_1 \times \ln(1.01)$, which is approximately equal to the effect of a 1% change in NCLB implementation stringency on average NAEP scores. In practice, most states changed their stringency by close to 38 percent over the study period. To understand the magnitude of the stringency effects given the change in stringency in the typical state, we also report estimates of

[7] The covariates include log population size, unemployment rate, poverty rate, percent over age 25 with a bachelor's degree, governor's party affiliation, student population characteristics (population of students enrolled in public school; percent of students who are White, Black, or Hispanic; percentage of students who receive free or reduced-price lunch; NAEP exclusion restrictions for students), and education policy factors (student-teacher ratio, expenditures per student).

$\beta_1 \times \ln(1.38)$, which represents the effects of the *average percent change* in state accountability stringency from 2003 to 2011. All standard errors are clustered at the state level.

The research design in our study assumes that—after accounting for state and year fixed effects—changes in state-level NCLB implementation stringency are independent of the entire history of unmeasured determinants of student outcomes in the state. The assumption is not directly testable, but we assess the sensitivity of our results to model specifications that include specific time trends in student achievement, as well as models that include a vector of lagged time-varying covariates.

3. RESULTS

Figure 2 shows national trends in NCLB stringency, AYP failure rates, and NAEP math and reading scores from 2003 to 2011. The top panel shows that from 2003 to 2011, the average simulated failure rate increased by 38 percent from 40 percent to 55 percent[8]. The figure also shows the evolution of the cross-state average of actual AYP failure rates. These rates are lower but follow the same general trend of the simulated failure rates. The bottom panel shows changes in state-average NAEP scale scores. Between 2003 and 2011, fourth and eighth graders scored 4 to 5 points higher in math and 2 to 3 points higher in reading.

Table 2 summarizes implementation details during the pre-waiver period, during which proficiency thresholds increased. In 2003, to meet AYP standards in the average state, 37 and 48 percent of students in each subgroup had to exceed minimum proficiency standards on the eighth-grade reading and math assessments, respectively. By 2011, the average standard had

[8] The cross-state averages in Figure 2 mask heterogeneity across the states. Typically, the standard deviation of simulated failure rates across states in a given year was between 16 and 18 percentage points. See Appendix Table A1.

risen to 73 percent proficiency in math and 78 percent proficiency in reading. When we adjust for test difficulty using the NAEP equivalent cutoffs, we find that the thresholds increased by only 3 to 4 scale points in fourth-grade reading and math and about 1 scale point in eighth-grade reading and math. Most states set the minimum subgroup size for holding schools accountable at 40 students. And most chose attendance as the OAI for elementary schools and graduation as the OAI for secondary schools.

State Implementation Stringency

In prior analyses, we used linear regressions to form a low dimensional summary of the connection between specific implementation features and simulated failure rates. Table 3 presents regression results originally presented in Wong et al. (2018). The results imply that higher AMO and NAEP equivalent cutoffs, the use of growth models, and smaller subgroup sizes are associated with higher simulated failure rates. Performance Indices, Confidence Intervals, Safe Harbor, and Confidence Intervals for Safe Harbor are associated with lower simulated failure rates. A one standard deviation change in proficiency standards (16 scale points on the NAEP, see descriptive table in appendix A1) accounted for a 9.44 percentage point increase in simulated failure rates. Adopting a performance index was associated with a decrease in simulated failure rates of 8 percentage points. The Confidence Interval, Safe Harbor, and Confidence Interval for Safe Harbor all had negative coefficients but were not statistically significant, perhaps because there is not much variation in the adoption of some of these rules. These analyses show that states' implementation decisions did affect the stringency of the NCLB within their state.

Effects of NCLB Implementation Stringency

Table 4 presents the results of our main impact analyses, with math results in panel A and reading results in panel B. Each column in both panels gives estimates of the coefficient on log simulated failure rates from a sequence of models. The base model includes state and year fixed effects. The subsequent models add covariates, state-specific linear trends, and models in which states are weighted by public school enrollment.

The results suggest that more stringent NCLB implementation led to small and statistically significant improvements in eighth-grade math scores. Using the results from the model in column 6, a one percent change in state accountability stringency increased state-level average NAEP math scores by about 0.013 scale points. In the pre-waiver period, the average state increased NCLB stringency by about 38 percent. The average increase in stringency implies an improvement in eighth-grade math of about a half point (0.43 scale points; effect size = 0.05 standard deviations[9]). We found no effects of stringency on fourth-grade math achievement.

The estimates in panel B suggest that more stringent NCLB implementation led to small and statistically significant improvements in eighth-grade reading achievement. A one percent change in accountability stringency improved eighth-grade reading achievement by between 0.009 and 0.026 scale points. During the pre-waiver period, the average percent change in state accountability stringency (38 percent) resulted in reading scale score improvements of between 0.28 (effect size = 0.04 standard deviations) and 0.83 points (effect size = 0.12 standard deviations). These gains were not evident for fourth-grade reading achievement.

[9] Effect sizes were calculated by dividing by the standard deviations of the state-level 2003 NAEP achievement scores (Appendix Table 3).

Sensitivity Tests

One threat to the internal validity of our analysis arises if states adopt weaker or stronger NCLB implementations in response to other unmeasured determinants of student outcomes. For example, we might be worried that NCLB implementation strategies responded to changes in family residential mobility patterns generated by the 2007-2009 housing crisis and recession, which coincided with implementation of NCLB. States may have lowered their accountability stringency if families of low socio-economic status moved into the state or families of high socio-economic status moved out.

We examined covariate balance (in DID form) to assess whether changes in NCLB implementation stringency were associated with changes in state composition after accounting for state and year fixed effects. Specifically, we fit two-way fixed effects regressions of time-varying state population characteristics on log simulated failure rates. The dependent variables in the balance regressions are measures of state population size; unemployment rate; poverty rate; governor's party affiliation; percent of adults over 25 with bachelor's degrees; public school enrollment; percent of Hispanic, Black, or White students; and percentage of students receiving free or reduced price lunch. Because the housing crisis may have affected other education policies that occurred concurrently with NCLB, we also looked at whether implementation stringency was associated with K-12 student-teacher ratios and K-12 expenditures per student. Finally, because critics of accountability reform have suggested that NCLB incentivized the exclusion of more students from testing, we examined whether implementation stringency resulted in higher NAEP exclusion rates by grade and subject areas. Each row of table 5 presents the results from a separate regression with the same specification but a different dependent variable. Each model includes state fixed effects, year fixed effects, and the log simulated failure rate. The table also includes the mean and standard deviation of each variable in 2002, which can be used to interpret the coefficient on implementation stringency in standard deviation units. Overall, the balancing regressions provide little evidence that within-state changes in stringency were associated with changes in the composition of the sample.

We also fit event-study models that included lagged and leading values stringency scores. These models helped us assess whether the downstream implementation decisions were associated with current period student outcomes and whether stringency scores took some time to affect outcomes. Figure 3 plots the coefficients on the leading and lagged stringency scores. The top panel of Figure 3 shows that there is no evidence of anticipatory or persistent effects of stringency on fourth- or eight-grade math scores. However, a one percent change in accountability stringency had immediate effects[10] of 0.03 and 0.05 points in fourth- and eighth-grade math achievement, respectively. Both immediate effects are statistically significant.

The pattern changes for reading outcomes. For fourth-grade reading achievement, there appears to be an immediate negative effect of 0.04. For eighth-grade reading achievement, scores improved by a statistically significant 0.03 points the year *before* accountability stringency increased. One interpretation of the eighth-grade reading effect is that it is statistically significant by chance: over the four outcomes, there were 24 statistical tests and no other evidence of anticipatory effects. Another interpretation is that schools were able to anticipate changes in NCLB requirements and make adjustments in instructional practice before state standards were actually implemented. In that case, one might consider both the stringency of AYP rules and anticipation of these policies as the total implementation effect. Finally, it is possible that the

[10] Here, “immediate effect” is defined as the NAEP achievement score observed in the winter or spring after the prior academic year.

anticipatory effects may signal other omitted confounders that explain changes in AYP rules and eighth-grade reading achievement. As a result, we interpret the implementation effects on eighth-grade reading achievement with caution.

Implementation Effects by Student Subgroup Status

Table 6 shows estimates of the effects of NCLB implementation stringency on math and reading outcomes for key sub-populations of students. The table shows coefficient estimates from models with state fixed effects, year fixed effects, and time varying covariates. The results were similar in the other specifications, which are available on request. These models include state specific linear trends, and weights determined by public school enrollment.

The subgroup effects mirror the main effects in Table 4. That is, even among different subgroups, more stringent NCLB implementation led to small improvements in eighth-grade math and possibly in eighth-grade reading achievement, but no effects on fourth-grade math or reading achievement. The effects on eighth-grade math achievement for Students with Disabilities (SWDs) and White and Free or Reduced Price Lunch (FRPL) students are statistically significant. The coefficients are less precisely estimated in the other sub-populations, but the direction of the effects is the same and the magnitude of effects is larger for Hispanics, English Language Learners (ELLs), and SWDs. The average increase in state implementation stringency during the pre-waiver period improved average eighth-grade math scores by 1.49 scale points for ELL students (effect size = 0.17 standard deviations), 0.91 scale points for SWDs (effect size = 0.10 standard deviations), and 0.80 scale points for Hispanic students (effect size = 0.09 standard deviations). In comparison, the average implementation effect was 0.33 points for White students (effect size = 0.04 standard deviations) and 0.44 points for FRPL students (effect size = 0.05 standard deviations).

Implementation Effects by Percentile Groups

Table 7 examines the impact of implementation stringency at different points on the achievement distribution. Because schools were held accountable based on student proficiency status, one concern was that schools would focus their efforts on the lowest performing students. Table 7 shows that there were small but statistically significant effects for eighth-grade math achievement for all percentile groups. However, students at the lowest end of the distribution improved most under stringent NCLB implementation. Among students in the 10th percentile, the average percent increase in implementation stringency led to a 0.78 scale score improvement on eighth-grade math achievement (effect size = 0.09 standard deviations); among students in the 25th percentile, implementation stringency increased math achievement by 0.50 scale points (effect size = 0.06 standard deviations). For students in the top half of the achievement distribution (50th percentile and higher), accountability stringency raised math achievement between 0.31 and 0.35 points (effect size = 0.04 standard deviations).

For eighth-grade reading, the largest (and statistically significant) effect was again observed at the bottom of the distribution. For students in the 10th percentile, an average percent change in state implementation stringency increased eighth-grade reading scores by 0.55 points (effect size = 0.07 standard deviations). At the 25th percentile, implementation stringency raised eighth-grade reading achievement by 0.31 points (effect size = 0.05 standard deviations). However, for the top half of the achievement distribution, eighth-grade reading improvements were small and not statistically significant, between 0.15 and 0.22 scale points (effect sizes = 0.02 – 0.03 standard deviations). Again, we see no evidence that implementation stringency improved fourth-grade reading or math achievement, regardless of location along the achievement distribution.

Implementation Effects by Math and Reading Subscales

Table 8 examines whether eighth-grade achievement effects were driven by gains in specific math and reading skills. The results imply that higher levels of implementation stringency led to significant improvements in algebra, measurement, math operations, and statistics, but not in geometry. The average percent change in implementation stringency resulted in approximately a half-point improvement in all math subscale areas except geometry (effect size = 0.06 standard deviations). The implementation effect was approximately 0.31 scale points and statistically significant at the 0.10 level for both eighth-grade informational and literary reading (effect size = 0.05 standard deviations). Table 8 also shows that there were generally no implementation effects on subscale skills for fourth-grade reading or math. The only exception was fourth-grade geometry, which had a negative and statistically significant coefficient.

Implementation Effects by Pre-NCLB Accountability Status

Finally, we examined whether the impact of implementation stringency varied by states with and without pre-NCLB accountability plans. Prior evaluations of NCLB have relied on differential responses from states depending on whether they had accountability plans before 2002 (Dee & Jacob, 2011). Table 9 summarizes the results of separate regressions for each grade and subject area. In addition to state and year fixed effects and time-varying covariates, these models include interaction terms that allow for differential implementation effects for states with and without pre-NCLB accountability policies.

Overall, we find that implementation effects for fourth- and eighth-grade math achievement were larger for states without pre-NCLB policies than for states with existing accountability rules. For these states, a one percent increase in implementation stringency resulted in a 0.02-point increase in eighth-grade math achievement. This means that during the full pre-waiver period, the 38% increase in stringency lead to an increase of 0.62 points on the average eighth-grade math achievement score (effect size = 0.07 standard deviations). However, the implementation effect was smaller for states with pre-NCLB accountability policies. A one percent increase in accountability stringency in these states resulted in only a 0.01 point (0.01=0.02-0.01) increase in math or a 0.23 point (0.23=0.62-0.38) increase for the full pre-waiver period (effect size = 0.03 standard deviations). This difference in effect, however, is not statistically significant.

In looking at fourth-grade math achievement, while there is not much evidence of an implementation effect for states without pre-NCLB accountability policies, there is evidence of a *difference* in implementation effects for states with and without pre-NCLB accountability policies. That is, the implementation effect was 0.02 points lower in states with pre-NCLB policies than in states without accountability rules ($p < 0.10$). Thus, for states with pre-NCLB accountability stringency, the average percent change in implementation stringency reduced fourth-grade achievement scores by 0.34 scale points (0.34=0.22-0.56) (effect size = 0.05 standard deviations).

The pattern reverses itself for fourth- and eighth-grade reading achievement. There is little evidence that increased implementation stringency improved fourth- and eighth-grade reading achievement scores in states without pre-NCLB accountability policies. However, the difference in implementation stringency between states with and without pre-NCLB accountability policies is large enough (and statistically significant) to explain positive implementation effects observed for reading achievement. For example, for states with pre-NCLB accountability policies, the average percent change in accountability stringency resulted in a 0.5-point gain in eighth-grade reading achievement (effect size = 0.08 standard deviations).

4. DISCUSSION

In reforming NCLB, Senator Lamar 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). Although ESSA continues to require that states hold schools accountable for student performance, the law allows states discretion over how schools are held accountable. ESSA expressly forbids the Education Secretary from pressuring states to adopt uniform performance standards, such as the Common Core. This means that while accountability systems remain the law of the land, the standards schools must meet will vary across states and school districts.

This study examines whether intensifying accountability pressures on schools improves student achievement. To address this question, we leveraged state variation in implementation stringency under NCLB to uncover the causal connection between increasing accountability standards and student performance. We found that during the NCLB pre-waiver period, increased accountability stringency resulted in small but statistically significant gains in eighth-grade math achievement. This was equivalent to an almost half-scale point improvement on the NAEP (effect size = 0.05 standard deviations). Effects were largest for ELLs, SWDs, and Hispanic students, as well as for students at the lowest ends of the achievement distribution. In addition, the improvements were evident in all math subject areas, except for geometry.

The results for eighth-grade reading outcomes were similar. Increased implementation stringency raised NAEP reading scores by 0.29 scale points (effect size = 0.04 standard deviations). Here too, the largest effects were for ELLs, SWDs, and Hispanic students, and for students at the 10th percentile in the achievement distribution. NCLB stringency did not seem to affect fourth-grade reading or math achievement.

Policy Implications

An important concern from critics of accountability reform is that as accountability pressure increases, schools focus on ensuring proficiency among a subset of at-risk students who score slightly below the proficiency standard. This effort might come at the expense of supporting average or higher achieving students. We found that while increased implementation stringency did help those students who were at greatest risk for low achievement in eighth-grade reading and math, there was not much evidence that that schools withdrew resources from higher performing students. In particular, there were no negative effects on eighth-grade reading and math achievement for students at the top half of the achievement distribution.

It is striking that there were consistent effects for eighth-grade outcomes but not for fourth-grade outcomes. This suggests that implementation effects may be evident only when students have had sufficient time in school to be "exposed" to increased accountability stringency. Follow-up research should examine the impacts of implementation stringency on later student outcomes, such as graduation rates, high school dropout, and post-secondary attendance.

Finally, there appears to be variation in effects by state policy contexts. Here we note that implementation effects on math achievement in particular were beneficial in states without previous accountability requirements. This result is consistent with prior evidence (Ottmar et al., 2014) suggesting that while student literacy is determined by both in-school and out-of-school factors, math achievement best represents the quality of instruction that students experience in school. Thus, schools in states with no prior accountability plans may have had more opportunities (or incentives) to adopt new instructional practices once accountability requirements increased. Similar gains were not observed in states with pre-existing accountability plans, perhaps because schools in these states had already adopted these instructional changes. Since we do not have an implementation measure for the pre-NCLB period, it remains unclear whether similar gains in math achievement were observed for states that adopted earlier accountability plans.

Interpreting the Magnitude of the Implementation Effects

NCLB was both costly to implement and politically unpopular. States spent more than $1.7 billion a year on standardized tests alone (Chingos, 2012), and as Wong et al. (2018) observed, just as accountability stringency was intensifying in most states (2007), public opinion of the education sector also became sharply more negative. Given that there were improvements in eighth-grade reading and math achievement, but not in fourth-grade achievement, how should we think about the magnitude of these effects in assessing their policy relevance?

In terms of the absolute scale score of the NAEP, the relative size of these effects is small. Out of a 500-point scale, implementation effects over the pre-waiver period ranged from 0.30 to 0.60 points. This is equivalent to 0.04 to 0.10 standard deviations in effect size units. Relative to introducing a new accountability plan, the benefits of increasing implementation stringency also appear minor. Dee and Jacob (2011) estimated effects of 8.21 scale points (effect size = 1.21 standard deviations) for fourth-grade math and 5.25 scale points (effect size = 0.59 standard deviations) for eighth-grade math.[11] This suggests that the impact of introducing a new

[11] Dee and Jacob reported effect sizes of 0.26 standard deviations for fourth-grade math and 0.14 standard deviations for eighth-grade math. They standardized by the standard deviation of student outcomes prior to NCLB. Since these are state-level estimates, we divided by the standard deviation of NAEP achievement outcomes in 2003 (before implementation effects

accountability system was about 9 times larger for improving eighth-grade math achievement than for increasing accountability stringency over the pre-waiver period.

Perhaps it is not surprising that relative to the overall NAEP scale itself, and the introduction of a new accountability system, implementation effects under NCLB were small. Does this mean that once an accountability system is in place, state policy-makers have no leverage to raise performance standards to increase student achievement? These results suggest that the ratcheting of test-based accountability pressures alone is not enough to sustain improvements in student achievement. Schools and teachers also need additional resources to improve instructional practice. It remains unclear how states will implement ESSA, but the federal law will likely not succeed if performance requirements are not accompanied by additional support for educators (Ladd, 2017).

5. CONCLUSION

An important contribution of this study is that it introduces simulated failure rates as a measure of accountability stringency. We are not the first to use simulated instruments for studying policy implementations.[12] However, this method has been underutilized in education evaluation settings and is well-suited for uncovering the connections between states' adoption of accountability policies and state and student outcomes. This paper is a first step toward understanding how states, schools, and students responded to accountability pressures under NCLB. Further work is needed to provide policy-makers and education researchers with tools for

would have taken place). We divided the Dee and Jacob effects by the standard deviations in our analysis, so the effects are comparable.

[12] See Currie and Gruber's (1996) examination of Medicaid expansion and Gruber and Saez's (2002) evaluation of the elasticity of taxable income.

describing and understanding the role that states have in determining schools' and students' experiences with accountability reform.

References

Burnette, D. (2016). ESSA poses capacity challenges for state education agencies. *Education Week*, *35*(18), 1.

Carey, K. (2007). The Pangloss Index: How States Game the No Child Left Behind Act. *Education Sector*.

Chingos, M. (2012). Strength in numbers: State spending on K-12 assessment systems. Washington, DC: Brookings Institute.

Davidson, E., Reback, R., Rockoff, J., & Schwartz, H. L. (2015). Fifty ways to leave a child behind: Idiosyncrasies and discrepancies in states' implementation of NCLB. *Educational Researcher*, *44*(6), 347-358.

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

Gruber, J., & Simon, K. (2008). Crowd-out 10 years later: Have recent public insurance expansions crowded out private health insurance? *Journal of health economics*, *27*(2), 201-217.

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

Jennings, J. L., & Bearak, J. M. (2014). "Teaching to the test" in the NCLB era: How test predictability affects our understanding of student performance. *Educational Researcher*, *43*(8), 381-389.

Ladd, H. F. (2017). No Child Left Behind: A Deeply Flawed Federal Policy. *Journal of Policy Analysis and Management*. https://doi.org/10.1002/pam.21978

McLaughlin, D., Bandeira de Mello, V., Blankenship, C., Chaney, K., Esra, P., Hikawa, H., ... & Wolman, M. (2008). Comparison between NAEP and State Mathematics Assessment Results: 2003. Volume 2. Research and Development Report. NCES 2008-475. *National Center for Education Statistics*.

Neal, D., & Schanzenbach, D. W. (2010). Left behind by design: Proficiency counts and test-based accountability. *The Review of Economics and Statistics*, *92*(2), 263-283.

Ottmar, E. R., Decker, L. E., Cameron, C. E., Curby, T. W., & Rimm-Kaufman, S. E. (2014). Classroom instructional quality, exposure to mathematics instruction and mathematics achievement in fifth grade. *Learning Environments Research*, *17*(2), 243-262.

Rogosa, D. R. (2003). The NCLB “99% confidence” scam: Utah-style calculations. *CRESST deliverable. Retrieved December*, *5*, 2004.

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. *Center on Education Policy*.

Stiefel, L., Schwartz, A. E., & Chellman, C. C. (2007). So many children left behind: Segregation and the impact of subgroup reporting in No Child Left Behind on the racial test score gap. *Educational Policy*, *21*(3), 527-550.

Taylor, J., Stecher, B., O'Day, J., Naftel, S., & Le Floch, K. C. (2010). State and Local Implementation of the" No Child Left Behind Act". Volume IX--Accountability under" NCLB". *US Department of Education*.

Wei, X. (2012). Are more stringent NCLB state accountability systems associated with better student outcomes? An analysis of NAEP results across states. *Educational Policy*, *26*(2), 268-308.

Wing, C., Simon, K., & Bello-Gomez, R. A. (2018). Designing Difference in Difference Studies: Best Practices for Public Health Policy Research. *Annual review of public health*.

Wong, M., Cook, T. D., & Steiner, P. M. (2015). Adding design elements to improve time series designs: No Child Left Behind as an example of causal pattern-matching. *Journal of Research on Educational Effectiveness*, *8*(2), 245-279.

Wong, V. C., Wing, C., Martin, D., & Krishnamachari, A. (2018). Did States Use Implementation Discretion to Reduce the Stringency of NCLB? Evidence from a Database of State Regulations. *Educational Researcher*, 0013189X17743230.

Table 1: Fidelity Coding of States AYP Rules

High Coding Fidelity	Partial Coding Fidelity	Low Coding fidelity
	Adjusted AYP Rules	Consequential growth models
District of Columbia, Georgia, Idaho, Illinois, Indiana, Kansas, Louisiana, Maine, Maryland, Minnesota, Missouri, Montana, Nevada, New Jersey, New Mexico, North Dakota, Oregon, Pennsylvania, South Carolina, South Dakota, Texas, Utah, Virginia, West Virginia, Wisconsin	Alabama, Alaska*, Arizona*, Arkansas*, California, Colorado, Connecticut, Delaware*, Florida*, Hawaii, Iowa*, Kentucky, Massachusetts, Michigan, Mississippi, Nebraska, New Hampshire, New York, North Carolina*, Oklahoma, Rhode Island, Tennessee*, Vermont, Washington, Wyoming	Ohio
25	25	1

Table 2: States' Implementation of AYP Rules During NCLB Pre-Waiver Period

	2003	2005	2007	2009	2011
Actual School AYP Failure Rate	33.66	24.58	27.40	32.84	45.28
	(17.53)	(16.56)	(16.16)	(17.83)	(21.62)
AMO Thresholds					
Math Grade 4	43.13	51.03	56.40	65.24	74.27
	(19.08)	(17.33)	(15.07)	(12.68)	(12.56)
Math Grade 8	37.43	46.14	52.49	62.49	73.04
	(18.38)	(16.84)	(15.64)	(13.00)	(12.20)
ELA Grade 4	52.43	59.88	63.16	69.56	78.11
	(18.07)	(16.07)	(14.55)	(12.17)	(10.33)
ELA Grade 8	47.49	55.68	60.89	68.56	77.52
	(16.68)	(15.55)	(14.72)	(12.09)	(10.19)
NAEP Equivalent Thresholds					
Math Grade 4	219.93	226.22	224.18	222.64	222.64
	(14.28)	(13.20)	(12.31)	(12.03)	(12.03)
Math Grade 8	266.78	273.41	272.05	268.17	268.17
	(17.30)	(16.32)	(13.72)	(13.84)	(13.84)
Reading Grade 4	194.98	198.36	199.45	199.29	199.29
	(15.27)	(16.11)	(14.42)	(12.94)	(12.94)
Reading Grade 8	241.57	250.14	245.70	243.25	243.25
	(18.06)	(16.04)	(13.80)	(13.45)	(13.45)
Used Growth Models	0.00	0.00	0.12	0.16	0.16
	(0.00)	(0.00)	(0.33)	(0.37)	(0.37)
Min subgroup participation < 40	0.26	0.22	0.22	0.24	0.24
	(0.44)	(0.42)	(0.42)	(0.43)	(0.43)
Min subgroup participation = 40	0.42	0.52	0.56	0.58	0.58
	(0.50)	(0.50)	(0.50)	(0.50)	(0.50)
Min subgroup participation > 40	0.06	0.14	0.14	0.14	0.14
	(0.24)	(0.35)	(0.35)	(0.35)	(0.35)
Attendance as OAI	0.72	0.72	0.72	0.72	0.72
	(0.45)	(0.45)	(0.45)	(0.45)	(0.45)
Graduation as OAI	0.90	0.92	0.98	0.98	0.98
	(0.30)	(0.27)	(0.14)	(0.14)	(0.14)

Exemption Rules					
Used Performance Indices	0.14	0.20	0.30	0.30	0.32
	(0.35)	(0.40)	(0.46)	(0.46)	(0.47)
Used MultiYear Average	0.18	0.32	0.38	0.38	0.38
	(0.39)	(0.47)	(0.49)	(0.49)	(0.49)
Used Confidence Intervals	0.66	0.94	0.94	0.94	0.94
	(0.48)	(0.24)	(0.24)	(0.24)	(0.24)
Used Safe Harbor	0.88	1.00	1.00	1.00	1.00
	(0.33)	(0.00)	(0.00)	(0.00)	(0.00)
Used Confidence Intervals for Safe Harbor	0.04	0.38	0.42	0.42	0.42
	(0.20)	(0.49)	(0.50)	(0.50)	(0.50)
States and District of Columbia	50	50	50	50	50

Notes: Average annual state-level means and standard deviations (in parentheses) presented above. Ohio is omitted from analysis sample.
Recreated from Table 3 in Wong et al. (2018).

Table 3: Regressions of AYP Rules on States' Implementation Stringency

	Stringency Score
AYP Requirements	
AMO Threshold (avg)	0.48**
	(0.071)
NAEP Equivalent Cutoff (avg)	0.59**
	(0.14)
Used Growth Models	8.07+
	(4.14)
Min Subgroup Size (N = 40)	--
	--
Min Subgroup size (N < 40)	8.05*
	(3.05)
Min Subgroup Size (N > 40)	0.48
	(4.76)
Attendance as OAI	8.15*
	(3.26)
Exemption Rules	
Used Performance Indices	-8.58*
	(3.22)
Used MultiYear Averages	0.93
	(2.74)
Used Confidence Intervals	-5.21
	(3.53)
Used Safe Harbor	-1.26
	(8.99)
Used Confidence Intervals for Safe Harbor	-4.36
	(2.90)

+ < 0.10, * p <0.05, ** p<0.01

Notes: Robust standard errors are presented in parentheses.
Recreated from Table 4 in Wong et al. (2018)

Table 4: Impact of Accountability Stringency on Math and Reading NAEP Achievement Scores

Panel A	4th Grade Math				8th Grade Math			
	1	2	3	4	5	6	7	8
Simulated Failure Rate (logged)	-0.0035	-0.0056	-0.0093+	-0.0092	0.0116*	0.0131**	0.0116+	0.0191+
	(0.0048)	(0.0054)	(0.0054)	(0.0063)	(0.0049)	(0.0038)	(0.0067)	(0.0077)
Implementation Effect During Pre-Waiver Period	*-0.1127*	*-0.1804*	*-0.2995*	*-0.2963*	*0.3768*	*0.4252*	*0.3768*	*0.6184*
Panel B	4th Grade Reading				8th Grade Reading			
	9	10	11	12	13	14	15	16
Simulated Failure Rate (logged)	-0.0046	-0.0054	-0.0073	-0.0044	0.0086	0.0090+	0.0158*	0.0256**
	(0.0071)	(0.0056)	(0.0045)	(0.0039)	(0.0064)	(0.0052)	(0.0050)	(0.0055)
Implementation Effect During Pre-Waiver Period	*-0.1482*	*-0.1739*	*-0.2351*	*-0.1417*	*0.277*	*0.2899*	*0.5121*	*0.8278*
Model specifications for each column								
State FE	x	x	x	x	x	x	x	x
Year FE	x	x	x	x	x	x	x	x
Covariates		x	x	x		x	x	x
State*Year			x				x	
Enrollment Weights				x			x	x
N Years	4	4	4		4	4	4	
N States	50	50	50		50	50	50	
N Observations	200	200	200		200	200	200	

+ p<0.10, * p<0.05, ** p < 0.01

Notes: Treatment effects are estimated for 2005, 2007, 2009, and 2011 NAEP achievement outcomes.
49 states and the District of Columbia are included in the analysis sample. Ohio is omitted from the analysis.
Depending on the model specification, the Simulated Failure Rate coefficient is estimated by separate regressions of NAEP achievement scores on simulated failure rates, state and year fixed effects, time-varying covariates, and interaction terms of state and year fixed effects.

Coefficient values for Models 4, 8, 12, and 16 are estimated using weighted least squares with public school student enrollment weights. These models control for state and year fixed effects, and time-varying covariates.

Coefficient values for Actual Failure Rate, Simulated Failure Rates, and standard errors have been multiplied by $\ln(1.01)$ to reflect the change in NAEP scale score related to a 1% change in the failure rate.

The "implementation effect during pre-waiver period" is the difference in NAEP scale scores that is the result of the average percent change in simulated failure rates from 2003 to 2011 (38%). This was calculated by $\hat{\beta} \times \ln(1.38)$.

Standard errors are clustered at the state level and shown in parentheses.

Table 5: Estimated Effects of Simulated Failure Rates on Alternative State Outcomes

Alternative State Outcomes	Mean 2002	Standard Deviation 2002	Simulated Failure Rate (Implementation Period)	Standard Error	Effect size difference (Cohen's D)
State Population Characteristics					
State population	5,587,026	6,438,277	-8,059	20,725	0.00
Unemployment Rate	5.64	1.06	0.01	0.06	0.01
Poverty Rate	11.79	3.23	0.02	0.06	0.00
Governor's Affiliation (Republican)	0.58	0.50	0.01	0.01	0.01
% with Bachelor's Degree	26.53	5.54	0.16*	0.06	0.03
Student Characteristics					
Total Population of Students	916,679	1,102,915	-968.99	2,996	0.00
% Hispanic Students	10.53	11.91	-0.10	0.06	-0.01
% Black Students	15.79	16.55	-0.03	0.06	0.00
% White Students	66.83	19.83	0.06	0.10	0.00
% of Students on FRPL	29.17	11.35	-0.03	0.19	0.00
State Education Policies					
Pupil Teacher Ratio	15.52	2.34	-0.02	0.03	-0.01
Expenditure Per Student	7,736	1,698	37.72	69.89	0.02
NAEP Exclusion Restrictions					
4th Grade Math	3.34	1.36	-0.06	0.06	-0.05
8th Grade Math	3.34	1.72	-0.06	0.13	-0.04
4th Grade Reading	5.84	2.21	0.06	0.13	0.03
8th Grade Reading	4.94	2.21	-0.19+	0.13	-0.09

+ p<0.10, * p<0.05, ** p < 0.01

Notes: Each row was estimated using separate regressions of the covariate on state simulated AYP failure rates, and state and year fixed effects. Coefficients have been multiplied by ln(1.38) to reflect the metric of the dependent variable for the implementation period.

Cohen's D was calculated by dividing the Simulated Failure Rate for the implementation period by the standard deviation in 2002.

Standard errors are clustered at the state level and shown in parentheses.

Table 6: Impact of Accountability Stringency on Different Subgroups

Panel A			4th Grade Math			
	White	Black	Hispanic	ELL	SWD	FRPL
Simulated Failure Rate (logged)	-0.005	-0.020	-0.012	-0.019	-0.016+	-0.003
	(0.006)	(0.010)	(0.012)	(0.020)	(0.008)	(0.006)
Implementation Effect During Pre-waiver Period	*-0.174*	*-0.631*	*-0.393*	*-0.606*	*-0.506*	*-0.090*

Panel B			8th Grade Math			
	White	Black	Hispanic	ELL	SWD	FRPL
Simulated Failure Rate (logged)	0.010*	0.014	0.025+	0.046	0.028*	0.014*
	(0.004)	(0.010)	(0.014)	(0.033)	(0.014)	(0.006)
Implementation Effect During Pre-waiver Period	*0.329*	*0.438*	*0.802*	*1.488*	*0.908*	*0.438*

Panel C			4th Grade Reading			
	White	Black	Hispanic	ELL	SWD	FRPL
Simulated Failure Rate (logged)	-0.003	-0.012	0.002	0.002	-0.006	-0.001
	(0.005)	(0.017)	(0.014)	(0.022)	(0.013)	(0.010)
Implementation Effect During Pre-waiver Period	*-0.084*	*-0.396*	*0.081*	*0.074*	*-0.187*	*-0.039*

Panel D			8th Grade Reading			
	White	Black	Hispanic	ELL	SWD	FRPL
Simulated Failure Rate (logged)	0.005	-0.003	0.021	0.033	0.028*	0.011
	(0.005)	(0.017)	(0.017)	(0.033)	(0.011)	(0.010)
Implementation Effect During Pre-waiver Period	*0.177*	*-0.084*	*0.696*	*1.056*	*0.905*	*0.367*

+ p<0.10, * p<0.05, ** p < 0.01

Notes: Treatment effects are estimated for 2005, 2007, 2009, and 2011 NAEP achievement outcomes.
49 states and the District of Columbia are included in the analysis sample. Ohio is omitted from the analysis.
The Simulated Failure Rate coefficient was estimated by separate regressions of the NAEP achievement scores (for each subgroup) on simulated failure rates, state and year fixed effects, and a vector of time-varying covariates.
Coefficient values for Simulated Failure Rates (logged) and standard errors have been multiplied by $\ln(1.01)$ to reflect the change in NAEP scale score related to a 1% change in the failure rate.
The "implementation effect during pre-waiver period" is the difference in NAEP scale scores that is the result of the overall percent change in simulated failure rates from 2003 to 2011 (38%). This was calculated by $\hat{\beta} \times \ln(1.38)$.
Standard errors are clustered at the state level and shown in parentheses.

Table 7: Impact of Accountability Stringency by Percentile Groups

Panel A	4th Grade Math				
	10th	25th	50th	75th	90th
Simulated Failure Rate (logged)	-0.008	-0.006	-0.003	-0.004	-0.004
	(0.006)	(0.006)	(0.006)	(0.006)	(0.007)
Implementation Effect During Pre-waiver Period	*-0.271*	*-0.196*	*-0.100*	*-0.132*	*-0.145*

Panel B	8th Grade Math				
	10th	25th	50th	75th	90th
Simulated Failure Rate (logged)	0.024***	0.015**	0.011*	0.011*	0.010+
	(0.007)	(0.005)	(0.004)	(0.004)	(0.005)
Implementation Effect During Pre-waiver Period	*0.779*	*0.496*	*0.345*	*0.351*	*0.312*

Panel C	4th Grade Reading				
	10th	25th	50th	75th	90th
Simulated Failure Rate (logged)	-0.006	-0.009	-0.005	-0.006	-0.009
	(0.009)	(0.007)	(0.006)	(0.005)	(0.006)
Implementation Effect During Pre-waiver Period	*-0.196*	*-0.277*	*-0.167*	*-0.190*	*-0.277*

Panel D	8th Grade Reading				
	10th	25th	50th	75th	90th
Simulated Failure Rate (logged)	0.017*	0.010+	0.007	0.004	0.006
	(0.007)	(0.006)	(0.005)	(0.006)	(0.006)
Implementation Effect During Pre-waiver Period	*0.554*	*0.312*	*0.222*	*0.145*	*0.180*

+ $p<0.10$, * $p<0.05$, ** $p < 0.01$

Notes: Treatment effects are estimated for 2005, 2007, 2009, and 2011 NAEP achievement outcomes.
49 states and the District of Columbia are included in the analysis sample. Ohio is omitted from the analysis.
The Simulated Failure Rate coefficient was estimated by separate regressions of the NAEP achievement scores (for each percentile group) on simulated failure rates, state and year fixed effects, and a vector of time-varying covariates.
Coefficient values for Simulated Failure Rates (logged) and standard errors have been multiplied by $\ln(1.01)$ to reflect the change in NAEP scale score related to a 1% change in the failure rate.
The "implementation effect during pre-waiver period" is the difference in NAEP scale scores that is the result of the overall percent change in simulated failure rates from 2003 to 2011 (38%). This was calculated by $\hat{\beta} \times \ln(1.38)$.
Standard errors are clustered at the state level and shown in parentheses.

Table 8: Impact of Accountability Stringency on Math and Reading Subscales

Panel A	4th Grade Math				
	Algebra	Geometry	Measurement	Math Operations	Statistics
Simulated Failure Rate (logged)	-0.006	-0.013**	-0.004	-0.002	-0.002
	(0.005)	(0.005)	(0.005)	(0.007)	(0.007)
Implementation Effect During Pre-waiver Period	*-0.190*	*-0.432*	*-0.135*	*-0.064*	*-0.071*

Panel B	8th Grade Math				
	Algebra	Geometry	Measurement	Math Operations	Statistics
Simulated Failure Rate (logged)	0.016**	0.005	0.016*	0.015***	0.018***
	(0.005)	(0.005)	(0.007)	(0.004)	(0.005)
Implementation Effect During Pre-waiver Period	*0.506*	*0.164*	*0.525*	*0.470*	*0.580*

Panel C	4th Grade Reading	
	Reading Information	Reading Literacy
Simulated Failure Rate (logged)	-0.005	-0.007
	(0.006)	(0.007)
Implementation Effect During Pre-waiver Period	*-0.177*	*-0.213*

Panel D	8th Grade Reading	
	Reading Information	Reading Literacy
Simulated Failure Rate (logged)	0.010+	0.010+
	(0.005)	(0.005)
Implementation Effect During Pre-waiver Period	*0.309*	*0.312*

+ p<0.10, * p<0.05, ** p < 0.01, *** p < 0.001

Notes: Treatment effects are estimated for 2005, 2007, 2009, and 2011 NAEP achievement outcomes.
49 states and the District of Columbia are included in the analysis sample. Ohio is omitted from the analysis.
The Simulated Failure Rate coefficient was estimated by separate regressions of the NAEP achievement scores (for each subscale) on simulated failure rates, state and year fixed effects, and a vector of time-varying covariates.

Coefficient values for Simulated Failure Rates (logged) and standard errors have been multiplied by $\ln(1.01)$ to reflect the change in NAEP scale score related to a 1% change in the failure rate.
The “implementation effect during pre-waiver period” is the difference in NAEP scale scores that is the result of the overall percent change in simulated failure rates from 2003 to 2011 (38%). This was calculated by $\hat{\beta} \times \ln(1.38)$.
Standard errors are clustered at the state level and shown in parentheses.

Table 9: Accountability Stringency by Pre-NCLB Accountability Plans

	Math		Reading	
	4th Grade	8th Grade	4th Grade	8th Grade
Simulated Failure Rate	0.007	0.019**	-0.010	0.000
	(0.007)	(0.007)	(0.007)	(0.007)
Implementation Effect During Pre-waiver Period	*0.222*	*0.615*	*-0.338*	*0.005*
Pre-NCLB accountability x Simulated Failure Rate	-0.017+	-0.012	0.017+	0.016+
	(0.009)	(0.010)	(0.009)	(0.009)
Implementation Effect During Pre-waiver Period	*-0.560*	*-0.383*	*0.557*	*0.502*

+ p<0.10, * p<0.05, ** p < 0.01, *** p < 0.001

Notes: Treatment effects are estimated for 2005, 2007, 2009, and 2011 NAEP achievement outcomes.
49 states and the District of Columbia are included in the analysis sample. Ohio is omitted from the analysis.
The Simulated Failure Rate coefficient was estimated by separate regressions of the NAEP achievement scores (for each subgroup) on simulated failure rates, an interaction term for simulated failure rates and pre-NCLB accountability status, state and year fixed effects, and a vector of time-varying covariates.
Coefficient values for Simulated Failure Rates (logged) and standard errors have been multiplied by $\ln(1.01)$ to reflect the change in NAEP scale score related to a 1% change in the failure rate.
The “implementation effect during pre-waiver period” is the difference in NAEP scale scores that is the result of the overall percent change in simulated failure rates from 2003 to 2011 (38%). This was calculated by $\hat{\beta} \times \ln(1.38)$.
Standard errors are clustered at the state level and shown in parentheses.

Figure 1: Predicted AYP Failure Rates vs Actual AYP Failure Rates

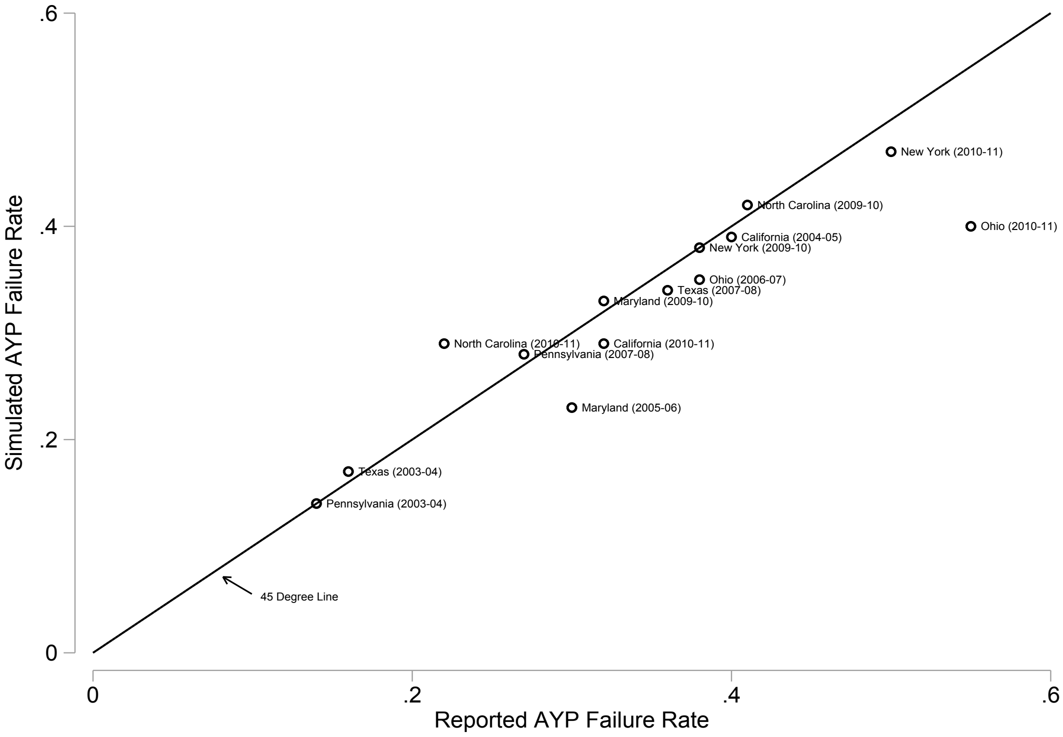


Figure 2: Accountability Stringency, AYP Failure Rates, and NAEP Achievement Trends (2003-2011)

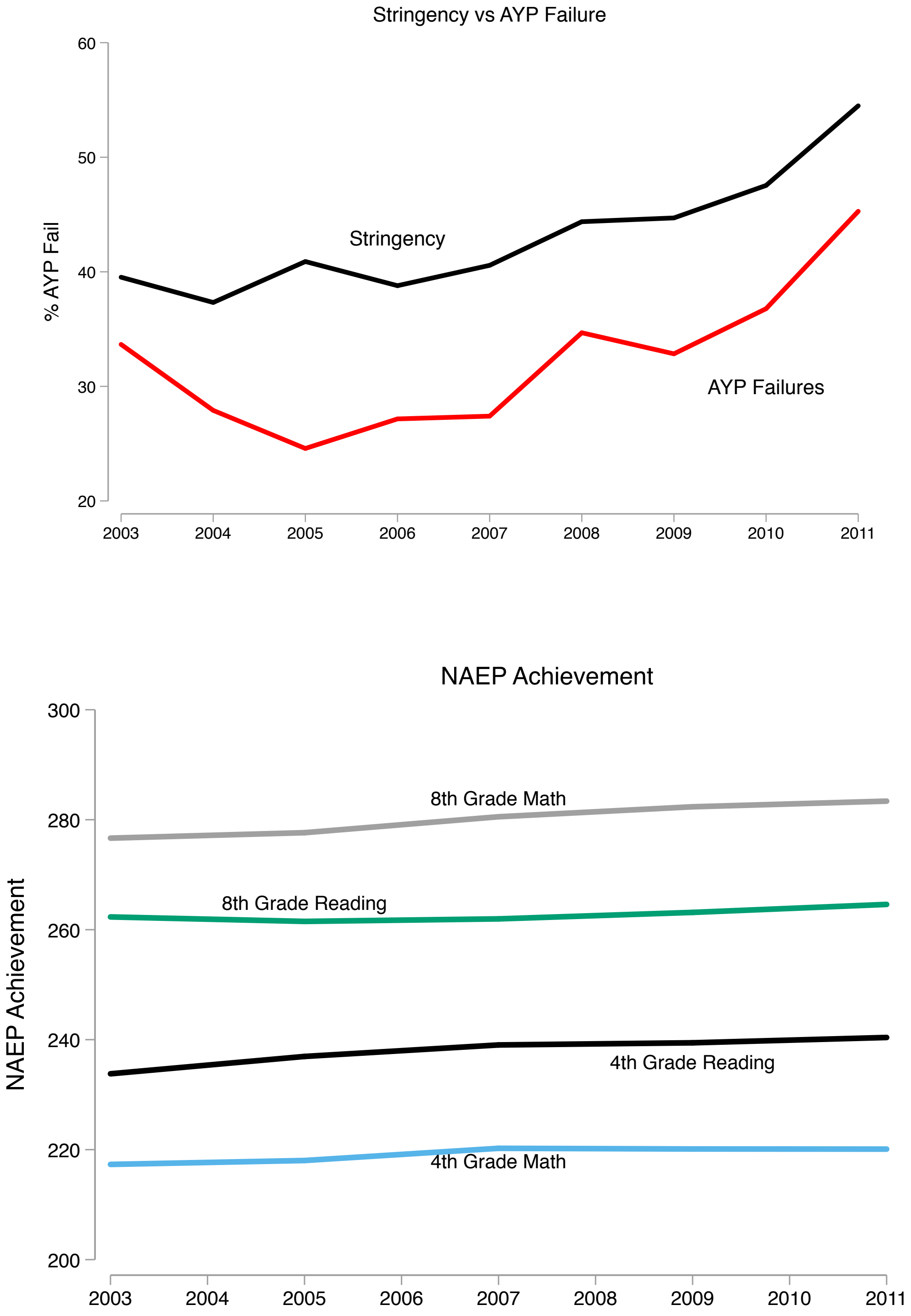
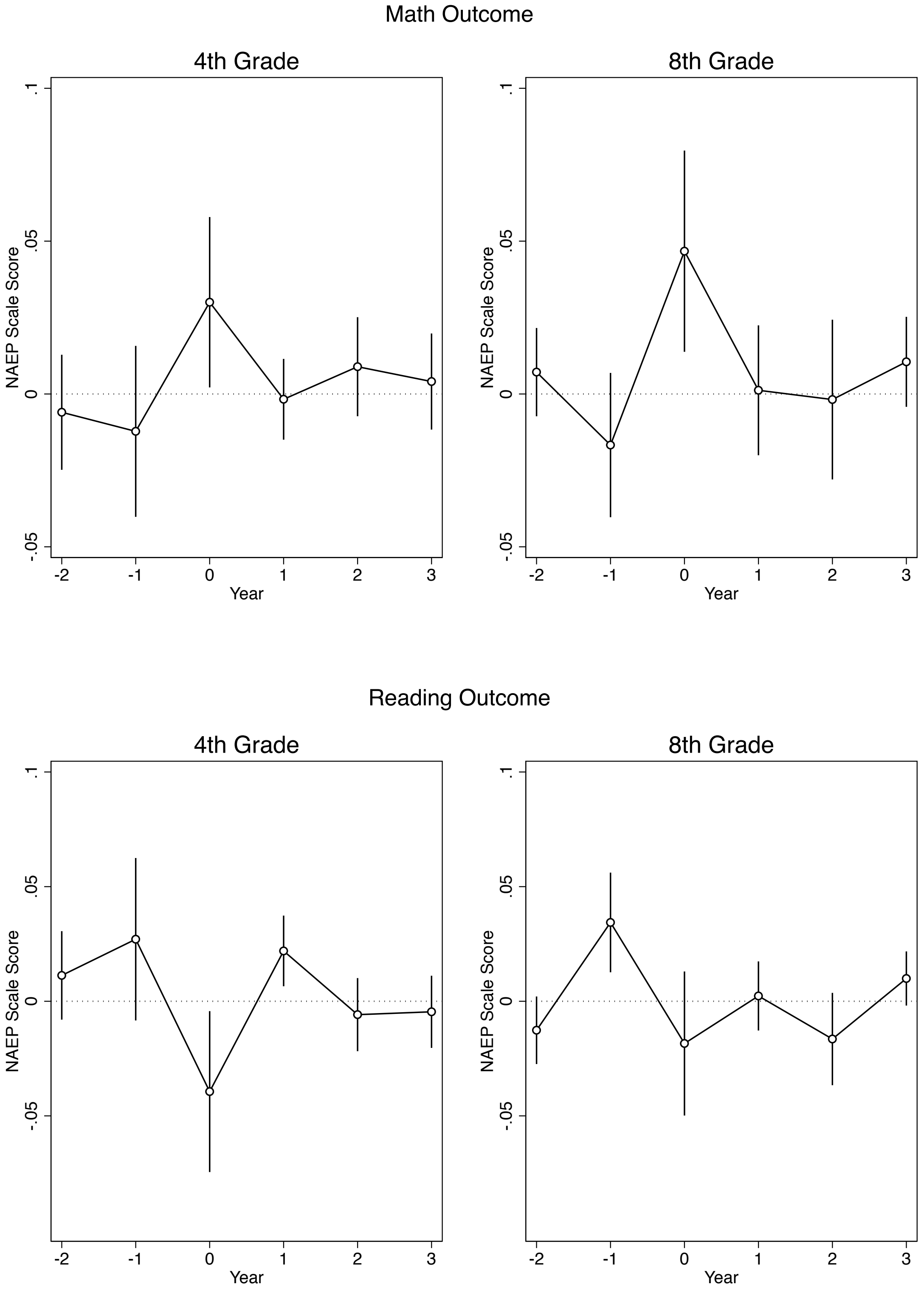


Figure 3: Differential Treatment Effects by Time Period



Appendix A

Table A1: States' Simulated Failure Rates and Average NAEP Achievement Scores

	2003	2005	2007	2009	2011
Implementation Stringency					
Simulated Failure Rate	39.53	41.61	40.66	44.87	54.72
	(18.46)	(16.35)	(16.40)	(16.57)	(17.15)
Average NAEP Performance					
Grade 4 Math	233.80	236.95	239.04	239.43	240.40
	(6.81)	(6.73)	(6.80)	(6.44)	(5.87)
Grade 8 Math	276.65	277.64	280.53	282.35	283.38
	(8.83)	(8.62)	(8.78)	(8.62)	(7.74)
Grade 4 Reading	217.32	218.03	220.25	220.12	220.10
	(7.68)	(7.53)	(7.10)	(6.71)	(6.80)
Grade 8 Reading	262.32	261.51	261.96	263.14	264.60
	(6.77)	(7.11)	(6.91)	(6.63)	(6.52)
States	50	50	50	50	50

Notes: The table presents average annual state level means and standard deviations (in parentheses). Ohio is omitted from analysis sample.