NBER WORKING PAPER SERIES

THE IMPACT OF NO CHILD LEFT BEHIND ON STUDENT ACHIEVEMENT

Thomas Dee
Brian Jacob

Working Paper 15531
http://www.nber.org/papers/w15531

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
November 2009

We would like to thank Rob Garlick, Elias Walsh, Nathaniel Schwartz and Erica Johnson for their research assistance. We would also like to thank Kerwin Charles, Robert Kaestner, Ioana Marinescu and seminar participants at the Harris School of Public Policy and at the NCLB: Emerging Findings Research Conference for helpful comments. An earlier version of this work was also presented by Jacob as the David N. Kershaw Lecture at the Annual meeting of the Association of Public Policy and Management (November 2008). All errors are our own. The views expressed herein are those of the author(s) and do not necessarily reflect the views of the National Bureau of Economic Research.

NBER working papers are circulated for discussion and comment purposes. They have not been peer-reviewed or been subject to the review by the NBER Board of Directors that accompanies official NBER publications.

© 2009 by Thomas Dee and Brian Jacob. All rights reserved. Short sections of text, not to exceed two paragraphs, may be quoted without explicit permission provided that full credit, including © notice, is given to the source.

The Impact of No Child Left Behind on Student Achievement
Thomas Dee and Brian Jacob
NBER Working Paper No. 15531
November 2009
JEL No. H52,I20,I21,I28,J01,J08,J18

ABSTRACT

The No Child Left Behind (NCLB) Act compelled states to design school-accountability systems based on annual student assessments. The effect of this Federal legislation on the distribution of student achievement is a highly controversial but centrally important question. This study presents evidence on whether NCLB has influenced student achievement based on an analysis of state-level panel data on student test scores from the National Assessment of Educational Progress (NAEP). The impact of NCLB is identified using a comparative interrupted time series analysis that relies on comparisons of the test-score changes across states that already had school-accountability policies in place prior to NCLB and those that did not. Our results indicate that NCLB generated statistically significant increases in the average math performance of 4th graders (effect size = 0.22 by 2007) as well as improvements at the lower and top percentiles. There is also evidence of improvements in 8th grade math achievement, particularly among traditionally low-achieving groups and at the lower percentiles. However, we find no evidence that NCLB increased reading achievement in either 4th or 8th grade.

Thomas Dee
Department of Economics
Swarthmore College
Swarthmore, PA 19081
and NBER
dee@swarthmore.edu

Brian Jacob
Gerald R. Ford School of Public Policy
University of Michigan
735 South State Street
Ann Arbor, MI 48109
and NBER
bajacob@umich.edu

1. Introduction

The No Child Left Behind (NCLB) Act is arguably the most far-reaching education-policy initiative in the United States over the last four decades. This legislation, which was signed by President Bush in January of 2002, dramatically expanded Federal influence over the nation's more than 90,000 public schools. The hallmark features of this legislation compelled states to conduct annual student assessments linked to state standards, to identify schools that are failing to make "adequate yearly progress" (AYP) towards the stated goal of having all students achieve proficiency in reading and math by 2013-14 and to institute sanctions and rewards based on each school's AYP status. A fundamental motivation for this reform is the notion that publicizing detailed information on school-specific performance and linking that "high-stakes" test performance to the possibility of meaningful sanctions can improve the focus and productivity of public schools.

On the other hand, critics charge that test-based school accountability has several unintended, negative consequences for the broad cognitive development of children (e.g., Nichols and Berliner 2007). They argue that NCLB and other test-based accountability policies cause educators to shift resources away from important but non-tested subjects (e.g., social studies, art, music) and to focus instruction in math and reading on the relatively narrow set of topics that are most heavily represented on the high-stakes tests (Rothstein et al. 2008, Koretz 2008). In the extreme, some suggest that high-stakes testing may lead school personnel to intentionally manipulate student test scores (Jacob and Levitt 2003).

Though the reauthorization of NCLB is currently under consideration, the empirical evidence on the impact of NCLB on student achievement is, to date, extremely limited. There have been a number of studies of NCLB that analyze national achievement trends. Interestingly,

however, different studies in this tradition come to starkly different conclusions (see, for example, Fuller, Wright, Gesicki, and Kang 2007; Center on Education Policy 2008). A likely explanation for these divergent results is that time series studies of NCLB lack a credible control group that allows them to distinguish the effects of the Federal reforms from the myriad of other factors taking place over the past 8 years. On the other hand, studies of school-level performance during the post-NCLB era often focus on what one might consider the "partial effects" of NCLB (e.g., comparing achievement gains across schools that make or miss AYP) and frequently rely on "high-stakes" state assessment scores that may be susceptible to "teaching to the test."

In this paper, we present new evidence on whether NCLB influenced student achievement using state-level panel data on student test scores from the National Assessment of Educational Progress (NAEP). This study identifies the impact of NCLB using a comparative interrupted time series design that relies on comparisons of test-score changes across states that already had school-accountability policies similar to NCLB in place prior to the implementation of NCLB and those that did not. We consider not only average effects, but look at effects separately by race, gender and free-lunch eligibility and at effects at various points on the achievement distribution.

This study builds on the existing literature in at least three critical ways. First, by using state-year NAEP data instead of state or city-specific data, this study relies on consistent measures of student achievement that are more nationally representative and that span the periods both before and well after the implementation of NCLB. Second, by relying on the "low-stakes" NAEP data rather than the "high-stakes" data from state assessments, the results we present should be comparatively immune to concerns about whether policy-driven changes in

achievement merely reflect "teaching to the test" rather than broader gains in cognitive performance. Third, the panel-based research design we use provides a credible way to distinguish the impact of NCLB from other social, economic and educational changes that were taking place over the same time period.

We find that NCLB generated large and statistically significant increases in the math achievement of 4th graders (effect size = 0.22 by 2007) and that these gains were concentrated among white and Hispanic students, among students who were eligible for subsidized lunch, and among students at all levels of performance. We find more moderate positive effects in 8th grade math achievement. These effects are concentrated at lower achievement levels and among students who were eligible for subsidized lunch. In contrast, our results suggest that NCLB had no impact on reading achievement among either 4th or 8th graders.

The mixed results presented here pose difficult but important questions for policymakers questioning whether to "end" or "mend" NCLB. The evidence of substantial and almost universal gains in math is undoubtedly good news for advocates of NCLB and school accountability. On the other hand, the lack of any effect in reading, and the fact that NCLB appears to have generated only modestly larger impacts among disadvantaged subgroups in math (and thus only made minimal headway in closing achievement gaps), suggests that, to date, the impact of NCLB has fallen short of its ambitious "moon-shot rhetoric" (Hess and Petrilli 2009).

The remainder of the paper proceeds as follows. Section 2 briefly reviews the literature on prior school-accountability policies and NCLB and situates the contributions of this study within that literature. Sections 3 and 4 discuss the methods and data used in this study. Section 5 summarizes the key results and robustness checks. Section 6 concludes with suggestions for further research.

2. NCLB and School Accountability

The NCLB legislation was actually a reauthorization of the historic Elementary and Secondary Education Act (ESEA), the central Federal legislation relevant to K-12 schooling. The ESEA, which was first enacted in 1965 along with other "Great Society" initiatives and previously reauthorized in 1994, introduced Title I, the Federal government's signature program for targeting financial assistance to schools and districts serving high concentrations of economically disadvantaged students. NCLB dramatically expanded the scope and scale of this Federal legislation by requiring that states introduce school-accountability systems that applied to *all* public schools and students in the state. In particular, NCLB requires annual testing of public-school students in reading and mathematics in grades 3 through 8 (and at least once in grades 10-12) and that states rate schools, both as a whole and for key subgroups, with regard to whether they are making "adequate yearly progress" (AYP) towards their state's proficiency goals.

NCLB requires that states introduce "sanctions and rewards" relevant to every school and based on their AYP status. However, NCLB also mandates explicit and increasingly severe sanctions for persistently low-performing schools that receive Title I aid (e.g., public school choice, staff replacement, and school restructuring). According to data from the Schools and Staffing Survey, 54.4 percent of public schools participated in Title I services during the 2003-04 school year. However, it should be noted that some states applied these explicit sanctions to schools not receiving Title I assistance as well. For example, 24 states introduced accountability systems that threatened all low-performing schools with reconstitution, regardless of whether they received Title I assistance (Olson 2004).

2.1 Theoretical Underpinnings of School Accountability

A basic perception that has motivated the widespread adoption of school-accountability policies like NCLB is that the system of public elementary and secondary schooling in the United States is "fragmented and incoherent" (e.g., Ladd 2007). In particular, proponents of school-accountability reforms argue that too many schools, particularly those serving the most at-risk students, have been insufficiently focused on their core performance objectives and that this organizational slack reflected the weak incentives and lack of accountability that existed among teachers and school administrators. For example, Hanushek and Raymond (2001) write that accountability policies are "premised on an assumption that a focus on student outcomes will lead to behavioral changes by students, teachers, and schools to align with the performance goals of the system" and that "explicit incentives… will lead to innovation, efficiency, and fixes to any observed performance problems."

The theoretical framework implicitly suggested by this characterization of public schools is a principal-agent model. The interests of parents and voters are viewed as imperfectly aligned with those of teachers and school administrators. Furthermore, parents and voters cannot easily monitor or evaluate the input decisions made by these agents. The performance-based sanctions and rewards that characterize accountability policies are effectively output-based incentives that can be understood as a potential policy response to this agency problem. Similarly, some of the provisions in NCLB with regard to teacher qualifications can be construed as an "agent selection" approach to a principal-agent problem.

The principal-agent lens is also useful for understanding criticisms of accountability-based reforms. The assumption that teachers and school administrators have misaligned self-

interest implies that they may respond to accountability policies in unintentionally narrow or even counterproductive ways. For example, in the presence of a high-stakes performance threshold, schools may reallocate instructional effort away from high and low-performing students and towards the "bubble kids" who are most likely, with additional attention, to meet the proficiency standard (e.g., Neal and Schanzenbach, in press). Similarly, concerns about "teaching to the test" reflect the view that schools will refocus their instructional effort on the potentially narrow cognitive skills targeted by their high-stakes state assessment at the expense of broader and more genuine improvements in cognitive achievement. Schools may also reallocate instructional effort away from academic subjects that are not tested or even attempt to shape the test-taking population in advantageous ways.

Our empirical analysis of NCLB provides evidence on whether this reform generated its conjectured benefits as well as the hypothesized deleterious consequences. In particular, by focusing on low-stakes NAEP scores rather than high-stakes state assessments, our analysis can circumvent concerns about the construct validity of state assessment scores. Furthermore, in addition to focusing on average test scores in subjects targeted by NCLB (i.e., mathematics and reading), we also focus on NCLB impacts at different points in the test-score distributions, on whether it influenced the exclusion of students from testing, and on its effects on performance in other subjects.

2.2 Research on Pre-NCLB Accountability Reforms Adopted by States

School-accountability reforms similar to those brought about by NCLB were adopted in a number of states during the 1990s. Several research studies have evaluated the achievement consequences of these reforms. Because of the similarities between the NCLB and aspects of

these pre-NCLB accountability systems, this body of research provides a useful backdrop against which to consider the potential achievement impacts of NCLB. In a recent review of this diverse evaluation literature, Figlio and Ladd (2008) suggest that three studies (Carnoy and Loeb 2002, Jacob 2005, and Hanushek and Raymond 2005) are the "most methodologically sound" (Ladd 2007).

The study by Carnoy and Loeb (2002), which was based on state-level achievement data from the National Assessment of Educational Progress (NAEP), found that the within-state growth in math performance between 1996 and 2000 was larger in states with higher values on an accountability index, particularly for Black and Hispanic students in 8th grade.[1] Similarly, Jacob (2005) found that, following the introduction of an accountability policy, math and reading achievement increased in Chicago Public Schools, relative both to the prior trends and relative to the contemporaneous changes in other large urban districts in the region. However, Jacob (2005) also found that, for younger students, there were not similar gains on a state-administered, low-stakes exam and that teachers responded strategically to accountability pressures (e.g., increasing special-education placements).

Hanushek and Raymond (2005) evaluated the impact of school-accountability policies on state-level NAEP math and reading achievement measured by the difference between the performance of a state's 8th graders and that of 4th graders in the same state four years earlier. This gain-score approach applied to the NAEP data implied that there were two cohorts of state-level observations in both math (1992-1996 and 1996-2000) and reading (1994-1998 and 1998-2002). Hanushek and Raymond (2005) classified state accountability policies as either "report-card accountability" or "consequential accountability." Report-card states provided a public

[1] The accountability index constructed by Carnoy and Loeb (2002) ranged from 1 to 5 and combined information on whether a state required student testing and performance reporting to the state, whether the state imposed sanctions or rewards and whether the state required students to pass an exit exam to graduate from high school.

report of school-level test performance. States with consequential accountability both publicized school-level performance and could attach consequences to that performance. The types of potential consequences states could implement were diverse. However, virtually all of the accountability systems in consequential-accountability states included key elements of the school-accountability provisions in NCLB (e.g., identifying failing schools, replacing a principal, allowing students to enroll elsewhere, and the takeover, closure, or reconstitution of a school). Hanushek and Raymond (2005) note that "all states are now effectively consequential accountability states (at least as soon as they phase in NCLB)."

Hanushek and Raymond (2005) find that the introduction of consequential accountability within a state was associated with statistically significant increases in the gain-score measures. The achievement gains implied by consequential accountability were particularly large for Hispanic students and, to a lesser extent, White students. However, the estimated effects of consequential accountability for the gains scores of Black students were statistically insignificant as were the estimated effects of report-card accountability. The authors argue that these achievement results provide support for the controversial school-accountability provisions in NCLB because those provisions were so similar to the consequential-accountability policies that had been adopted in some states.

2.3 Analyses of National Achievement Trends

The broad interest in understanding whether NCLB has influenced the distribution of student achievement, both overall and for key subgroups, has motivated careful scrutiny of the most recent trend data. For example, in a report commissioned by the U.S. Department of Education's Institute of Education Sciences (IES), Stullich, Eisner, McCrary and Roney (2006)

note that achievement trends on both state assessments and the NAEP are "positive overall and for key subgroups" through 2005. Similarly, using more recent data, a report by the Center on Education Policy (2008) concludes reading and math achievement measures based on state assessments have increased in most states since 2002 and that there have been smaller but similar patterns in NAEP scores. Both reports were careful to stress that these national gains are not necessarily attributable to the effects of NCLB. However, a press release from the U.S. Department of Education (2006) pointed to the improved NAEP scores, particularly for the earlier grades where NCLB was targeted, as evidence that NCLB is "working."

Other studies have taken a less sanguine view of these achievement gains. For example, Fuller, Wright, Gesicki, and Kang (2007) are sharply critical of relying on trends in state assessments, arguing that they are misleading because states adjust their assessment systems over time. They also document a growing disparity between student performance on state assessments and the NAEP since the introduction of NCLB and conclude that "it is important to focus on the historical patterns informed by the NAEP." Using NAEP data on fourth graders, they conclude that the *growth* in student achievement has actually become flatter since the introduction of NCLB. Similarly, an analysis of NAEP trends by Lee (2006) concludes that reading achievement is flat over the NCLB period while the gains in math performance simply tracked the trends that existed prior to NCLB.

2.4 Research on NCLB in Specific States and/or Districts

Several more recent studies have directly assessed the achievement consequences of NCLB through analyses of cross-sectional and panel data. Most of these studies have focused on the distributional consequences of NCLB within particular cities and states and using data that

are exclusively from the post-NCLB period. For example, Neal and Schanzenbach (in press) present evidence that, following the introduction of NCLB in Illinois, the performance of Chicago school students near the proficiency threshold (i.e., those in the middle of the distribution) improved while the performance of those at the bottom of the distribution of was the same or lower. Similarly, using data from the state of Washington, Krieg (2008) finds that the performance of students in the tails of the distribution is lower when their school faces the possibility of NCLB sanctions. However, in a study based on data from seven states over four years, Ballou and Springer (2008) conclude that NCLB generally increased performance on a low-stakes test, particularly for lower-performing students. Their research design leveraged the fact that the phased implementation of NCLB meant that some grade-year combinations mattered for calculating AYP while others did not.

An earlier study by Lee (2006) evaluated the achievement effects using state-year panel data and a research design similar to that used in this study. Specifically, the study by Lee (2006) relied partly on comparing the pre/post changes in states that had "strong" accountability to the contemporaneous changes in states that did not. Lee (2006) concluded that NCLB did not have any achievement effects. However, these inferences might be underpowered both because the study could only use the NAEP data through 2005 and because it failed to exploit the precision gains associated with conditioning on state fixed effects.[2] Furthermore, the definition of "strong" accountability used by Lee (2006) was based on a study by Lee and Wong (2004) and seems overly narrow in this context because it fails to identify multiple states that actually had NCLB-like consequential-accountability polices (e.g., column (5) of Table 2). Furthermore, this

[2] In fact, like our study, Lee (2006, Table C-7) finds evidence for a positive NCLB effect on math scores among 4th graders. Lee (2006, page 44) dismisses these results because they become statistically insignificant after conditioning on additional covariates. However, the estimated NCLB effect actually increases by roughly 20 percent after conditioning on these controls so the insignificance of this estimate reflects a substantial loss of precision in the saturated specification.

taxonomy may also be subject to measurement error because it relies on aspects of accountability (e.g., student-focused accountability) that are not actually a part of NCLB.

3. Research Design

This section describes the empirical strategy we pursue to assess the impact of NCLB on student performance. The national time trends in student achievement are a natural point of departure for considering the impact of NCLB. Figure 1 presents national trends on the Main NAEP from 1990 to 2007. The dashed horizontal line in 2002 visually identifies the point at which NCLB was implemented. The trends shown in these figures suggest that NCLB may have had some positive effects on 4th grade math achievement but, with a few exceptions, provide little suggestion of impacts in the other three grade-subject combinations.[3]

Given the myriad of other social, economic and educational factors occurring over this time period, however, it is not clear that one should draw strong causal inferences from these data. For example, the nation was suffering from a recession around the time NCLB was implemented, which may have been expected to reduce student achievement in the absence of other forces. Conversely, there were a number of national education policies or programs that may have influenced student achievement at this time. For example, the National Council of Teachers of Mathematics (NCTM) adopted new standards in 2000, which likely shifted the content of math instruction in many elementary classrooms over this period (NCTM website). Similarly, the Reading Excellence Act of 1999 (the precursor to the Reading First program within NCLB) provided more than $750 million to states and LEAs to adopt scientifically-based instructional practices and professional development activities (Moss 2006).

[3] One exception is a noticeable improvement in 8th grade math scores among African-Americans. Data from the Long-Term Trend NAEP tell a similar story for 9- and 13-year olds in math and reading.

3.1 Comparative Interrupted Time Series

To circumvent these concerns, we rely on a comparative interrupted time series (CITS) approach (also known as an interrupted time series with a non-equivalent comparison group). Specifically, we compare the deviation from prior achievement trends among a "treatment group" that was subject to NCLB with the analogous deviation for a "comparison group" that was arguably less affected by NCLB, if at all. The intuition is that the deviation from trend in the comparison group will reflect other hard-to-observe factors (e.g., the economy, other education reforms) that may have influenced student achievement in the absence of NCLB. This strategy has a long tradition in education research (see, for example, the discussion in Bloom 1999 and Shadish et al. 2002), and has been used recently to evaluate reforms as diverse as Accelerated Schools (Bloom et al. 2001) and pre-NCLB accountability policies (Jacob 2005).

One interesting and potentially important question is exactly which academic year we should consider as the first one in which NCLB may have influenced school performance. NCLB secured final Congressional approval (December 18, 2001) and was signed by President Bush (January 8, 2002) in the middle of the 2001-02 academic year. Our preferred approach is to view NCLB as first in effect during the next academic year (i.e., AY 2002-03). NCLB is often characterized as having been implemented during this year, in part because states were required to use testing outcomes from the prior 2001-02 year as the starting point for determining whether a school was making adequate yearly progress (AYP) and to submit draft "workbooks" that described how school AYP status would be determined (Palmer and Coleman 2003, Olson 2002). Furthermore, state data collected during the 2002-03 year also suggest that states had moved quickly to adapt to NCLB's new testing requirements and to introduce school-level

performance reporting (Olson 2002). Interestingly, we also find evidence that this conventional definition of NCLB's start date is supported by how changes in measures of student achievement during this period corresponded with prior trends.

However, one could reasonably conjecture that the discussion and anticipation surrounding the adoption of NCLB would have influenced school performance during the 2001-02 school year. In particular, both major presidential candidates in the 2000 election had signaled support for school-based accountability and President Bush sent a 26-page legislative blueprint titled "No Child Left Behind" to Capitol Hill within days of taking office in January of 2001 (Hess and Petrilli 2006). Alternatively, it could also be argued that NCLB should not be viewed as in effect until the 2003-04 academic year when new state accountability systems were more fully implemented as well as more informed by guidance from and negotiations with the U.S. Department of Education (Olson 2002, 2003). The flexible functional form of the CITS specification we describe below actually allows for the kinds of dynamically heterogeneous effects that this sort of phased implementation might imply. Nonetheless, as a check on the robustness of our impact estimates, we also report the results of specifications where NCLB is considered first in effect during the 2001-02 school year as well as during the 2003-04 school year.

As discussed in more detail below, there are several other important threats to causal inference in a CITS design. One such example involves the endogenous student mobility, as might occur if NCLB caused families to leave or return to the public schools. If this NCLB-induced mobility were random with respect to characteristics influencing achievement, it would not be a concern. On the other hand, if the most motivated parents pulled their children from public schools at the onset of NCLB, the resulting compositional change may have decreased

student achievement in the absence of any changes to the schools themselves. A similar concern arises if NCLB induced states to selectively change the composition of students tested for the NAEP (e.g., increasing exclusion rates). In the analysis that follows, we take particular care to examine a variety of such potential concerns, and find no evidence that our findings are biased.

However, it is worth noting that all NCLB-induced changes do not necessarily invalidate our research design. For example, states may have responded to NCLB by increasing funding for schools, or instituting kindergarten testing for early identification of at-risk students. In this case, one could still interpret the estimates presented below as the causal, reduced-form effect of NCLB, where funding and early identification are viewed as mediating mechanisms through which the policy operated. We are exploring a variety of channels through which NLCB may have influenced student achievement in ongoing work.

The central challenge for any CITS design is to identify a plausible comparison group that was unaffected by the intervention under study. In the case of NCLB, this is particularly difficult. As noted earlier, the policy was signed into law on January 8, 2002 and implemented nationwide during the 2002-03 school year. It simultaneously applied to all public schools in the United States but with particularly explicit sanctions for schools receiving Federal Title I funds.

One seemingly compelling comparison group is the set of Catholic schools in the U.S. (Jacob 2008). Though students in Catholic schools are eligible to participate in a number of major programs under the Elementary and Secondary Education Act (ESEA), the NCLB reauthorization of ESEA left these prior provisions "largely intact" (U.S. Department of Education 2007), implying that the NCLB reforms were comparatively irrelevant for Catholic schools. However, just two days prior to President Bush's signing of the NCLB legislation, the *Boston Globe* (Carroll et al. 2002) began publishing the results of investigative reporting, which

indicated that the leadership of the Boston Archdiocese had continually reassigned priests known to have sexually abused children and allowed those priests to continue working with children. This scandal received extensive, nationwide coverage and ultimately led to similar revelations in Catholic dioceses throughout the United States. In Appendix A, we present evidence that the press coverage of this broad scandal coincided with large drop in Catholic-school enrollments. Because the onset of NCLB was closely aligned with this large attrition from Catholic schools (and its other unobservable effects on the quality of Catholic schools), we conclude that Catholic schools do not constitute a convincing control group for evaluating the achievement consequences of NCLB.

3.2 Consequential Accountability Prior to NCLB

Instead, the research design emphasized in this study relies on comparing trends in student achievement across states that had varying degrees of experience, prior to NCLB, with state school-accountability policies similar to those brought about by NCLB. The intuition behind this approach is that NCLB represented less of a "treatment" in states that had already adopted NCLB-like school accountability policies prior to 2002. To the extent that NCLB-like accountability had either positive or negative effects on measured student achievement, we would expect to observe those effects most distinctly in states that had not previously introduced similar policies.

Here we are relying on the assertion that pre-NCLB school accountability policies were comparable to NCLB – that is, the two types of accountability regimes are similar in the most relevant respects. The fact that some state officials forcefully criticized NCLB, arguing that it "needlessly duplicates" their prior accountability systems (Dobbs 2005), suggests the functional

equivalence of earlier state consequential-accountability policies and state policies under NCLB. To ensure that this is the case, we categorize states according to whether the features of their pre-NCLB accountability policies closely resemble the key aspects of NCLB.

While we relied on a number of different sources to categorize pre-NCLB accountability policies across states (including studies of such policies by Carnoy and Loeb 2002, Lee and Wong 2004, and Hanushek and Raymond 2005), the taxonomy developed by Hanushek and Raymond (2005) is particularly salient in this context because it most closely tracked the key school-accountability features of NCLB. The authors identified 25 states that implemented "consequential accountability" prior to NCLB by coupling the public reporting of data on school performance to the possibility of meaningful sanctions based on that performance.[4] We reviewed their coding with information from a variety of sources including the Quality Counts series put out by Education Week (1999), the state-specific "Accountability and Assessment Profiles" assembled by the Consortium for Policy Research in Education (Goertz and Duffy 2001), annual surveys on state assessment programs fielded by the Council of Chief State School Officers (CCSSO), information from state Department of Education web sites, Lexis-Nexis searches of state and local newspapers, and conversations with academics and state officials in several states.

Our review generally confirmed their coding for the existence and timing of these state "consequential accountability" policies.[5] Furthermore, our review indicated that these pre-NCLB school-accountability systems closely resembled the state policies subsequently shaped by NCLB in that they both reported school performance and attached the possibility of sanctions

[4] States that publicize information on school performance without attaching sanctions to that performance are categorized as having "report card" accountability (Hanushek and Raymond 2005).

[5] However, there are also a few notable distinctions between our classification of consequential-accountability states (Table 2) and the coding reported by Hanushek and Raymond (2005). These discrepancies are discussed more fully in Appendix C.

to school performance (e.g., ratings, takeover, closure, reconstitution, replacing the principal and/or allowing student mobility). The strong similarities between the pre-NCLB consequential-accountability policies and post-NCLB state policies suggest that states with prior school accountability policies may be a good comparison group.

We also assessed, in a more data-driven manner, whether NCLB was effectively irrelevant in consequential-accountability states. We relied on data from several recent studies (Braun et al. 2008, NCES 2007, Bandeira de Mello et al. 2009) that have converted the test-based proficiency thresholds in state assessment systems to a common metric benchmarked to the National Assessment of Educational Progress (NAEP). These studies calculate the NAEP scale score that corresponds to the proficiency standard in a particular state, year, grade and subject. In 2007, for example, the state proficiency standard in 4th grade mathematics in Alabama corresponded to a NAEP scale score of 205 whereas the identical proficiency standard in Arkansas was equivalent to a NAEP scale score of 229. The difference of 24 NAEP scale score points in the proficiency levels across these states suggests that Arkansas had substantially more rigorous standards at the time.

Using these NAEP equivalence measures to compare the rigor of proficiency standards across states, we found that the proficiency thresholds during the NCLB era were, on average, similar across states with and without prior consequential accountability. Given the recent evidence that some states lowered proficiency standards since the introduction of NCLB (Bandeira de Mello et al. 2009), we also sought to examine *changes* in proficiency standards for states with and without prior accountability. Unfortunately, it is difficult to do this in a comprehensive manner because NAEP equivalence measures prior to NCLB are available for only a limited number of states. In fact, these equivalence measures are available for only

slightly more than half of states in 2003 and 2005.[6] Focusing on the states in our analysis sample for which NAEP equivalence measures were available post-NCLB (roughly half of our sample), we find that states that had pre-NCLB accountability policies did *not* lower proficiency standards from 2003 to 2007. In contrast, we find that states with no prior school accountability policies prior to NCLB did lower proficiency standards from 2003 to 2007.[7]

Overall, these results suggest that the state-level policies catalyzed by NCLB were quite similar to the first generation of state-level consequential-accountability policies. Furthermore, the evidence that pre-NCLB school-accountability policies closely resembled NCLB and that the states that adopted these earlier reforms did not change their proficiency standards after NCLB was implemented, suggests that these states can serve as a plausible comparison group for identifying the impact of NCLB. In the following section, we outline the specific models we use to generate our impact estimates. Before doing so, however, it is worth considering exactly how one should interpret the resulting estimates. First, it is important to realize that our estimates will capture the impact of the accountability provisions of NCLB, but will not reflect the impact of other NCLB provisions such as Reading First or the highly qualified teacher provision. Second, our estimates will identify the impact of NCLB-induced school accountability provisions on states without prior accountability policies. To the extent that one believes that states that expected to gain the most from accountability policies adopted them prior to NCLB, one might

[6] Only about 20 states have NAEP equivalence measures prior to NCLB because of a combination of reasons, including (i) many states did not administer state-representative NAEP prior to 2003, (ii) many states did not report proficiency levels as part of their state testing regime, (iii) many states did not test at all the two NAEP grades (i.e., grades 4 and 8) and (iv) the authors of the report only calculated equivalence measures for a subset of states with available data prior to 2003 because these early years were viewed as a "trial run" for developing the equating procedures. Even in the years 2003, 2005 and 2007, NAEP equivalence measures are only available for a limited set of states because (a) not all states tested 4th and 8th graders, and (b) there are were a handful of states that did not have sufficient NAEP data in a given grade x year x subjects to justify the equivalence exercise.

[7] For example, our point estimates suggest that states without prior accountability lowered proficiency standards in 4th grade math by 8 to 10 NAEP scale points between 2003 and 2007, although these estimates are very imprecise and not statistically different than zero.

view the results we present as an underestimate of the average treatment effect of school accountability. Finally, based on the evidence presented above, it appears that our estimates will indeed capture the full impact of NCLB-induced accountability provisions. In fact, to the extent that states induced to implement school accountability by NCLB were lowering proficiency standards over this period, one might consider the results presented below as an *underestimate* of the true causal impact of school accountability.

3.3 Estimation

Following the intuition of the CITS research design we have outlined, we estimate the following regression model:

$$\begin{aligned} Y_{st} = {} & \beta_0 + \beta_1 YEAR_t + \beta_2 NCLB_t + \beta_3 \left(YR\_SINCE\_NCLB_t\right) + \\ & \beta_4 \left(T_s \times YEAR_t\right) + \beta_5 \left(T_s \times NCLB_t\right) + \beta_6 \left(T_s \times YR\_SINCE\_NCLB_t\right) + \\ & \beta_7 X_{st} + \mu_s + \varepsilon_{st} \end{aligned} \tag{1}$$

where Y_{st} is NAEP-based measure of student achievement for state s in year t, $YEAR_t$ is a trend variable (defined as $YEAR_t$ – 1989 so that it starts with a value of 1 in 1990), and $NCLB_t$ is a dummy variable equal to one for observations from the NCLB era. For the majority of our analysis, we assume the NCLB era begins in the academic year 2002-03, which was effectively the first year of full implementation since the legislation was signed into law in January 2002. In sensitivity analyses, we confirm that assuming NCLB began in spring 2002 or even spring 2001 yields comparable results. $YR\_SINCE\_NCLB_t$ is defined as $YEAR_t$ – 2002, so that this variable takes on a value of 1 for the 2002-03 year, which corresponds to the 2003 NAEP testing. X_{st} represents covariates varying within states over time (e.g., per pupil expenditures, NAEP test exclusion rates, etc.). The variables, μ_s and ε_{st} represent state fixed effects and a mean-zero random error respectively.

T_s is a time-invariant variable that measures the treatment imposed by NCLB. For example, in our most basic application, T_s is a dummy variable that identifies whether a given state had *not* instituted consequential accountability prior to NCLB. This regression specification then allows for an NCLB effect that can be reflected in both a level shift in the outcome variable (i.e., β_5) as well as a shift in the achievement trend (i.e., β_6). Thus, the total estimated NCLB effect as of 2007 would be $\hat{\beta}_5 + 5 \times \hat{\beta}_6$.

While this simple case highlights the intuition behind our approach, there are ways in which it is probably more accurate to view the "treatment" provided by the introduction of NCLB in the framework of a dosage model. In particular, slightly more than half of the states that introduced consequential school accountability prior to NCLB did so just four years or fewer prior to NCLB's implementation. Given the number of states that implemented consequential accountability shortly before the implementation of NCLB, the simple binary definition of T_s defined above could lead to attenuated estimates of the NCLB effect. That is, the "control" group includes some states for which the effects of prior state policies and NCLB are closely intertwined. To address this concern, we report the results from some specifications that simply omit data from states that adopted state accountability within several years of NCLB. However, this approach has two important disadvantages: (1) it reduces our statistical power and (2) it requires one to make largely arbitrary decisions about which states to omit from the analysis.

As a preferred alternative, we also define T_s as the number of years during our panel period that a state did *not* have school accountability. Specifically, we define the treatment as the number of years *without* prior school accountability between the 1991-92 academic year and the onset of NCLB. Hence, states with no prior accountability have a value of 11. Illinois, which adopted its policy in the 1992-03 school year, would have a value of 2. Texas would have a value

of 4 since its policy started in 1994-95, and Vermont would have a value of 9 since its program started in 1999-2000. Our identification strategy implies that the larger the value of this treatment variable, the greater potential impact of NCLB.

An alternative sense in which NCLB's impacts may have been heterogeneous is that it may have represented a more substantial treatment in states that had adopted relatively weaker accountability provisions during the 1990s. To address this possibility, we also report as a robustness check the results of specifications where T_s is defined in a manner that reflects the weakness of a state's prior accountability system (and, consequently the strength of the treatment implied by NCLB). More specifically, we define T_s as the difference in the percent of students attaining proficiency on the *state* test and the percentage attaining proficiency on the NAEP tests in 2000 for math and 2002 for reading. In 2000, for example, 62 percent of Georgia 4th graders attained proficiency on the Georgia state math exam while only 17 percent attained proficiency on the NAEP math exam, yielding a difference of 45 percentage points in the proficiency rates. In Arkansas, 14 percent of 4th graders were proficient in math according to NAEP compared with 37 percent according to the Arkansas state exam, yielding a difference of 23 percentage points. Therefore, by this metric, Arkansas would have a substantially "tougher" pre-NCLB accountability policy than Georgia. If a state did not have a consequential accountability policy prior to NCLB, we assign the state a value of 100 percent on this measure. As in the other strategies, larger values of T_s correspond with *weaker* pre-NCLB accountability and thus a *greater* potential impact of NCLB.

Unfortunately, this approach may be underpowered relative to our preferred definition of T_s (i.e., years without prior school accountability) for at least two reasons. One is that, with this definition of T_s, correcting for the downward bias implied by the late-adopting consequential

accountability states implies omitting data from some states. Second, because not all states collected proficiency information on students prior to NCLB (e.g., a number of states used norm-referenced exams that simply provided national percentile ranks), we are not able to calculate this measure of treatment intensity for a subset of our analysis sample.[8] Nonetheless, we report the key results from specifications where T_s is based on these proficiency measures as a check on our main results.

A more fundamental concern with the inferences from our CITS approach involves the reliability of the identifying assumptions it uses to estimate the impact of NCLB. In particular, our approach assumes that the deviations from prior achievement trends within the "control" states (i.e., those with lower values of T_s) provide a valid counterfactual for what would have happened in "treatment states" if NCLB had not been implemented. The internal validity of this identification strategy would be violated if there were unobserved determinants of student achievement that varied both contemporaneously with the onset of NCLB and uniquely in either the treatment or control states. For example, if the socioeconomic status of families deteriorated during our study period but did so particularly in the states with prior consequential accountability as well as during the implementation of NCLB, our CITS approach would overstate the achievement gains associated with NCLB. While it is not possible to assess these sorts of concerns definitively, we provide indirect evidence on this important question by reporting the results of auxiliary regressions like equation (1) but where the dependent variables are state-year measures of observed traits that may influence student achievement (e.g., parental

[8] Another limitation of this measure is the fact that it utilizes state proficiency results from the end of the pre-NCLB period. Hence, a state that initially implemented a very stringent proficiency cutoff and realized substantial student improvement would appear to have a very weak policy under this measure (insofar as NAEP scores did not rise as quickly as state exam scores). More generally, state proficiency cutoffs are endogenous insofar as policymakers determine them with an eye toward potential student performance and various other social, economic and political factors.

education, poverty rate, and median household income). The estimated "effect" of NCLB on these measures provides evidence on whether achievement-relevant determinants appear to vary along with the adoption of NCLB in a manner that could confound our key CITS inferences. In addition to this evidence and the previously discussed robustness checks, we also assess the sensitivity of our CITS results to changes in the set of regression controls (e.g., introducing year fixed effects and state-specific trend variables) and to alternative estimation procedures (e.g., weighted least squares).

4. The National Assessment of Educational Progress (NAEP)

This analysis uses data on math and reading achievement from the state-representative NAEP. There are several advantages to utilizing NAEP data for our analysis. First, it is a low-stakes exam that is not directly tied to a state's own standards or assessments. Instead, the NAEP aims to assess a broad range of skills and knowledge within each subject area. Second, it is viewed as a well-designed assessment from a psychometric perspective and is scaled to allow for comparisons across time and between states. For these reasons, the NAEP data should be relatively immune to concerns about accountability-driven test score inflation (Jacob 2005, Fuller et al. 2007, Koretz 2009).

One important factor to consider when using the NAEP data is that the rules regarding the permissibility of test accommodations changed shortly before the introduction of NCLB. Prior to the 2000 math administration (1998 for the reading administration), schools were not allowed to offer test accommodations to students with special needs. In all subsequent years, schools were permitted to do so. Test accommodations might influence aggregate achievement levels in at least two different ways: (i) they may encourage schools to test students with special

needs who had previously been completed excluded from testing, which may lower scores on average; (ii) they may allow students with special needs who had previously been tested without accommodations to perform better on the exam, thus raising scores on average. In the year of the switch (2000 for math and 1998 for reading), there were two different administrations of the NAEP – one with and one without accommodations permitted. In our baseline specifications, we use data from the 2000 math and 1998 reading administrations with accommodations permitted. We later show that our results are not sensitive to using data from the alternative administration, or to using data from all administrations.

Because our identification strategy depends on measuring achievement trends prior to NCLB, we limit our sample to states that administered the state NAEP at least two times prior to the implementation of NCLB.[9] Because so few states administered the 8th grade math exam in 1990, when looking at math we focus on the pre-NCLB years of 1992, 1996 and 2000. For reading, we focus on 1994, 1998 and 2002. We chose to include 2002 as a pre-NCLB data point in our analysis because, given the timing of the passage and implementation of the law, it seems unlikely that Spring 2002 scores could have been substantially influenced by NCLB. All states administered NAEP in 2003, 2005 and 2007.

Our final sample includes 39 states (227 state x years) for 4th grade math, 38 states (220 state x years) for 8th grade math, 37 states (249 state x years) for 4th grade reading and 34 states (170 state x years) for 8th grade reading. A complete list of states in our sample can be found in Appendix Table B1. Since our estimates will rely on achievement changes across these states over time, it is worth exploring how representative these states are with respect to the nation.

[9] In order to ensure that we are accurately capturing the pre-NCLB trends, in addition to requiring that a state have at least two NAEP scores prior to 2003, we also require that states in our math sample participated in the 2000 NAEP and states in our reading sample participated in both the 1998 and 2002 NAEP. However, as shown in Table 5, our results are not sensitive to this sample restriction.

Table 1 presents some descriptive statistics that compare traits of our analysis sample to nationally representative NAEP data. With a few exceptions, our analysis sample closely resembles the nation in terms of student demographics (e.g., percent black and percent Hispanic), observed socioeconomic traits (e.g., the poverty rate) and measures of the levels and pre-NCLB trends in NAEP test scores.

5. Results

5.1 Achievement Trends by Pre-NCLB Accountability Status

Before presenting formal estimates from equation (1), we show the trends in NAEP scores by pre-NCLB accountability (Figures 2-5). These figures are meant to illustrate the intuition underlying our research design, and to provide tentative evidence with regard to the achievement effects of NCLB. In each case, we present trends for two groups: (i) states that adopted school accountability between 1994 and 1998 and (ii) states that did not adopt school accountability prior to NCLB. The dots reflect the simple mean for each group x year, and the connecting lines show the predicted trends from the model described above.

Consider first Figure 2a, which shows trends in 4th grade math achievement. We see that in 1992, states that never adopted accountability scored roughly 5 scale points (0.18 standard deviations) higher on average than states that adopted school accountability policies by 1998. While all states made modest gains between 1992 and 2000, the states that adopted accountability policies prior to 1998 experienced more rapid improvement during this period. Indeed, this is the type of evidence underlying the conclusions in Carnoy and Loeb (2002) and Hanushek and Raymond (2005). Mean achievement in both groups jumped noticeably in 2003, although relative to prior trends, this shift was largest among the "no prior accountability" group.

Interestingly, there was little noticeable change in the growth *rate* across period for the prior accountability states. That is, the slope of the achievement trend before and after 2002 is roughly equivalent for this group. In contrast, states with no prior accountability grew at a faster rate from 2003 to 2007 than from 1992 through 2000, such that the growth rates after 2002 were roughly equivalent across both groups of states. The trends for percent of students meeting the basic standard, shown in Figure 13b, are similar. These figures suggest that NCLB had a positive impact on 4th grade math achievement.

The trends for 8th grade math (Figure 3) are similar to those for 4th grade math, but somewhat less clear in showing a positive achievement effect. The level shift following the introduction of NCLB was smaller than for 4th grade math scores, and neither group of states experienced a substantial improvement in growth rates after 2002.

The pattern for 4th grade reading in Figure 4 is much less clear. The pre-NCLB reading trends for both groups are much noisier than the math trends. In particular, both groups experienced a decline in achievement in 1994, little change in 1998 (relative to 1992) and then very large gains in 2002.[10] The prior accountability group experienced a drop in achievement from 2002 to 2003, both in absolute terms and relative to trend. The other group experienced very little increase following NCLB. Perhaps most importantly, however, a visual inspection of the data in these plots indicates that the prior achievement trend was not linear, which is a central assumption of the CITS model specified in equation (1). In the analysis that follows, we explore alternative modeling strategies in an effort to identify the impact of NCLB on reading achievement, none of which provides any indication of a positive impact. For example, if one compares the four-year trend from 1998 to 2002 to the four-year trend from 2003 to 2007, there

[10] Note that the graph is scaled to accentuate what are really quite small absolute changes from year to year.

is no evidence of any NCLB effect. Similarly, Figure 5 provides no evidence of an NCLB effect on 8th grade reading achievement.

5.2 Main Estimation Results

Table 3 shows our baseline estimates of equation (1). The outcome measure in all cases is the mean NAEP scale score for all students in a particular state x year. All models include linear and quadratic terms for the state-year exclusion rate as well as state fixed effects. Standard errors clustered at the state level are shown in parentheses. In Panel A, we define our treatment group to include only states that did not adopt school accountability prior to NCLB. We find that NCLB increased 4th grade math achievement by roughly 4.7 points by 2007 in states with no prior accountability relative to other states. Given a standard deviation of 31, this reflects an effect size of 0.15. We find no effect for 8th grade math or reading and a small but statistically significant effect for 4th grade reading.

As discussed earlier, the inclusion of these "late adopter" states may understate any positive effects of NCLB. Hence, in panel B we estimate the same models but exclude states that adopted school accountability policies between 1999-2001. Hence, the results shown in Panel B correspond to the trends shown in Figures 2-5. In this specification, the impact of NCLB on 4th grade math achievement is roughly 8.2 scale points (0.26 standard deviations) and the effect on 8th grade math is 5.2 scale points (0.14 standard deviation). While this specification does avoid confounding the impact of NCLB with the impact of a state's own accountability policy, it also reduces the precision of our estimates and relies on a somewhat arbitrary decision of which states to exclude.

Panel C presents results in which the treatment is defined as the number of years *without* prior school accountability. The total effect we report is the impact of NCLB in 2007 for states

with no prior accountability relative to states that adopted school accountability in 1997 (the mean adoption year among states that adopt prior to NCLB). The results suggest moderate positive effects for 4th grade math (7.2 scale points or 0.23 standard deviations) and smaller effects for 8th grade math that are not statistically different than zero at conventional levels (a 0.10 standard deviation effect with a p-value of 0.12). The effects for 4th and 8th grade reading are small in magnitude and statistically indistinguishable from zero.

In an effort to test our identifying assumption, Table 4 examines the impact of NCLB on a variety of variables besides student achievement. The specification we use is identical to the one shown in panel C from Table 3. Column 1 shows the "effect" of NCLB on NAEP exclusion rates. A common concern with test-based accountability is that it provides school personnel an incentive to exclude low-performing children from testing. In theory, this should not be a major concern in our context because neither schools nor states are held accountable on the basis of their NAEP scores. And, indeed, we find no significant association between the NCLB and test exclusion. Columns 2-4 show the relationship between NCLB and state-year poverty rates and median household income as measured by the Current Population Survey and Census and the state-year employment-to-population ratio.[11] The results of our auxiliary regressions indicate that there are no statistically significant relationships between NCLB and any of these state-year observables.[12]

Column 5 reports the key results of an additional robustness check based on the fraction of students in the state-year enrolled in public schools. These measures are based on grade-

[11] For poverty, median household income and employment rates, we use the state x year rates with the year prior to the NAEP exam. The reason for this is that the NAEP exam is given by March of a calendar year, making the prior calendar year's value more predictive of the achievement outcome.

[12] We did find that NCLB appeared to have a statistically significant, negative effect on state-year unemployment rates. However, this result appears to be driven by several small states. In particular, regressions that weight by student enrollment in the state x year show very small and statistically significant point estimates. Importantly, similar weighted regressions in our main specifications yield achievement effects comparable to our baseline results (see Table 5). In addition, our main results are robust to conditioning on state-year unemployment rates.

specific enrollment data from the Common Core and the Private School Universe survey. We find an extremely small, marginally significant, effect in 4th grade math that suggests that NCLB may have reduced the fraction of students attending public school by roughly one percent. Columns 6 and 7 show the fraction of public school students in the state x year who were Black and Hispanic respectively, as measured by the same NAEP data on which the outcomes are measured. We see some evidence that NCLB is associated with an increase in the fraction of students who are Black. However, column 8 indicates that there was no effect on the fraction of students eligible for free lunch. Moreover, student-reported parental education data from the NAEP (available only for 8th graders) indicates that NCLB was associated with an *increase* in parental education (p-value 0.12) (not reported here; results available upon request). Hence, there is some evidence of small changes in student composition associated with NCLB, although the predicted effect is ambiguous. And we will see in Table 5, the inclusion of these variables as time-varying covariates does not change our results.

Finally, columns 9 and 10 show that NCLB was not associated with the fraction of students in a cohort that attended preschool or full-day kindergarten. These results allay concerns that states that did not adopt school accountability in the mid-1990s were instead focusing on early childhood policies. If this were the case, one might be concerned that the impacts we document are really driven by lagged policy changes at the state level – namely, students who started attending preschool and/or full-day (as opposed to half-day) kindergarten in the late-1990s were entering elementary school better prepared which is reflected on state NAEP tests.

Table 5 presents a series of sensitivity analyses. Perhaps most importantly, column 2 presents results from a model that includes as time-varying covariates all of the variables that

appear in Table 4 as well as measures of the pupil-teacher ratio and log per-pupil expenditures in 2007 dollars (see the table notes for the full list of variables included in this specification). The estimates are virtually identical to our baseline results shown in column 1. We find comparable results with weighted least squares (WLS) based on public-school enrollments (column 3).

Columns 4 and 5 present specifications that utilize alternative coding for consequential accountability. In column 4, we code four states that may be viewed as marginal cases of consequential accountability (VA, WI, IN and KS; see appendix for a more detailed discussion of these states) as *not* having consequential accountability. In column 5, we report results from the specification described above that uses the difference between state proficiency rates on state vs. NAEP exams as a measure of treatment intensity. In both cases, we find that our baseline results are robust.

The remainder of the table reports the estimated NCLB effect for the following specifications: the inclusion of a full set of year fixed effects (column 6), the inclusion of state-specific time trends (column 7), the omission of state fixed effects (column 8), using a sample that includes all states with at least one year (instead of at least two years) of pre-NCLB data (column 9), using a sample that includes only states with at least three years of pre-NCLB data (column 10), using data from the administration that did not permit accommodations in 2000 for math and 1998 for reading (column 11), alternative years for the first impact of NCLB (columns 12 and 13), and including the recently released math data from the 2009 NAEP.[13] The results of these various sensitivity analyses suggest that the baseline results presented in Table 3 are quite robust.

[13] In results not reported here but available upon request, we also find comparable results from weighted least squares using as weights the inverse of the variance of the state-year average.

Table 6 shows the effect of NCLB on the achievement distribution by grade and subject. As many have noted, the design of NCLB necessarily focused the attention of schools on helping students attain proficiency. Hence, one would expect NCLB to disproportionately influence achievement in the left tail of the NAEP distribution. We find results roughly consistent with this. However, in contrast with some prior work and many prior concerns, we do not find that the introduction of NCLB harmed students at higher points on the achievement distribution. Indeed, NCLB seemed to increase achievement at higher points on the achievement distribution more than one might have expected. For example, in 4th grade math, the impacts at the 75th percentile were only 3 scale points lower than at the 10th percentile.

5.3 Heterogeneity by Student Subgroup, Subject and Subscale

One of the primary objectives of NCLB was to reduce inequities in student performance by race and socioeconomic status. Indeed, this concern drove the requirement that accountability under the statute be determined by subgroup performance in addition to aggregate school performance. Hence, it is of particular interest to understand the effect of NCLB on specific student subgroups, which is what we do in Tables 7-10. In each table, we present results separately by race, gender and poverty subgroups. Several interesting findings emerge.

In the 4th grade math sample (Table 7), effects are somewhat larger for Black and Hispanic students relative to white students. Interestingly, in the case of Black students, weighting by student enrollment substantially increases the magnitude of the effects. This suggests that NCLB had more positive effects on Black students in states with larger Black populations. Similarly, the effects were substantially larger among students who were eligible

for subsidized lunch (regardless of race) relative to students who were not eligible. The effects were roughly comparable for boys and girls.

In 8th grade math (Table 8), we find extremely large positive effects for Hispanic students and small, only marginally significant for white students. The point estimates for Black students are large but imprecisely estimated, and generally not statistically distinguishable from zero at conventional levels. The effects for free-lunch eligible students are large and statistically significant. Interestingly, the effects are substantially larger for girls, with boys experiencing little if any benefit of the accountability.

The results for the 4th grade reading shown in Table 9 suggest some moderate positive effects for white students and for male students. However, as noted earlier, the trends prior to NCLB were distinctly non-linear and it thus the estimates shown here are likely invalid. We choose to present them primarily for the sake of transparency. In results not shown here, but available upon request, we re-estimated the specifications in this table limiting the sample to the years 1998 through 2007. In these models, the NCLB impact is identified off of deviations from the 1998 to 2002 trend in states with and without prior accountability policies. The results for white students are roughly half as large as those shown in Table 9 and not statistically different than zero. Similarly, the large point estimates for Hispanic students are reduced to close to zero. Indeed, none of the impact estimates in any of the specifications in Table 9 are statistically distinguishable from zero when using the restricted sample. Looking at the 8th grade reading results in Table 10, we see no positive effects. The most surprising finding is that NCLB appeared to have a statistically *reduced* the performance of Black students and students who were eligible for subsidized lunch. We have no good explanation for this finding, and hesitate to over-interpret what might be due to sampling variability.

One concern about NCLB and most other test-based school accountability policies is that because they focus almost exclusively on math and reading performance they will cause schools to neglect other important subjects to the detriment of student learning. To date, the evidence for such resource shifting is mixed. There is some evidence that schools have shifted resources away from subjects other than reading or math. For example, a recent study by the Center on Education Policy (2006) reported that 71 percent of school districts had reduced the elementary-school instructional time in at least one subject so that more instructional time could be spent on reading and mathematics. From a theoretical perspective, however, it is not clear how such shifting will influence student performance in these other areas given that math and reading skills are complementary to student learning in subjects such as science and social studies. The few studies that have examined this issue have not found that school accountability policies substantially reduce student performance in other subjects (Jacob 2005, Winters et al. Forthcoming).

The NAEP data offers some opportunity to test this hypothesis in the context of NCLB. A sizeable number of states administered state-representative NAEP science tests to 8th graders in 1996, 2000 and 2005 (n = 31) and to 4th graders in 2000 and 2005 (n = 36). Using this data, we estimate models similar to equation (1), comparing deviations from predicted achievement in 2005 in states with and without prior school accountability. For the 8th grade sample, we use the 1996 and 2000 data to estimate a prior intercept and trend. In the 4th grade sample where there is only one pre-NCLB observation, we estimate a simple difference-in-difference model. We find no statistically significant effects at either grade level at any point on the achievement distribution (see appendix Table B2). Our standard errors are relatively precise, allowing us to rule out effects larger than roughly 3-4 scale points (about .1 standard deviations). Similarly, we

find no significant effects when looking separately by subgroup (see appendix Tables B3 and B4). Together, these results suggest that NCLB did not have an adverse impact on student performance in science as measured by the NAEP.[14]

Another major concern with test-based accountability, including NCLB, is that it provides teachers an incentive to divert energy towards the types of questions that appear most commonly on the high-stakes test and away from other topics within the tested domain. This resource reallocation within subjects could reduce the validity of inferences based on performance on the high-stakes test. One of the benefits of the analysis presented here is that it relies on student performance on the NAEP, which should be relatively immune from such test score "inflation" since it is not used as a high-stakes test under NCLB (or any other accountability system of which we are aware).

It is still interesting to examine whether NCLB has improved student achievement in any particular topic within math or reading. To explore this, we re-estimate equation (1) using NAEP subscale scores as the dependent variable. The NAEP math exam measures student performance in five specific topic areas: Algebra, Geometry, Measurement, Number Properties and Operations, and Data Analysis, Statistics and Probability. The results shown in Table 11 suggest that NCLB had a positive impact in all math topic areas for the 4th grade sample. The point estimates are somewhat larger in Algebra (0.26 standard deviations), Number Properties (0.26 standard deviations) and Data Analysis (0.22 standard deviations) than in Geometry (0.17 standard deviations) and Measurement (0.16 standard deviations). In the 8th grade sample, NCLB had a moderately large and statistically significant impact within Data Analysis (6.7 scale

[14] The NAEP science exam measures not only factual and conceptual understanding of science topics, but also the ability to integrate science knowledge into a larger context and to use tools, procedures, and reasoning processes in scientific investigation. For example, the science exam includes a hands-on task that requires students to conduct actual experiments using materials provided to them.

points, or 0.16 standard deviations) and marginally significant effects for Number Properties and Geometry (roughly 0.11 standard deviation in both topics). These results are consistent with some earlier work indicating large impacts of accountability in similar areas (Jacob 2005), suggesting that some topics may be more amenable to instruction than others.

The NAEP reading exam measures student competency in several skills related to comprehension: reading for information (i.e., primarily non-fiction reading), reading for literary experience (i.e., primarily fiction reading), and (for 8th grade only) the ability to perform a task (e.g., students apply knowledge from reading bus schedules or directions for repairing something). In results not shown here but available upon request, we find no significant differences in student achievement effects by topic area in reading – that is, NCLB did not appear to have significant effects on student achievement in any of the three reading competencies.

6. Conclusions

NCLB is an extraordinarily influential and controversial policy that, over the last seven years, has brought test-based school accountability to scale at public schools across the United States. The impact of this Federally mandated reform on student achievement is an empirical question of central importance. This study presents evidence on this broad question using state-year panel data on multiple student-outcome measures from the NAEP. We utilize a comparative interrupted time series research design that relies on the changes over time in states that had no prior school-accountability system like those required by NCLB and those that did. Our results suggest that the achievement consequences of NCLB are decidedly mixed. Specifically, we find that NCLB generated large and broad gains in the math achievement of 4th

and (to a somewhat lesser extent) 8th graders. However, our results suggest that NCLB had no impact on reading achievement for 4th or 8th graders.

The mixed results presented here pose difficult but important questions for policymakers questioning whether to "end" or "mend" NCLB. The evidence of substantial and almost universal gains in math is undoubtedly good news for advocates of NCLB. On the other hand, the lack of any effect in reading, and the fact that the policy appears to have generated only modestly larger impacts among disadvantaged subgroups in math (and thus only made minimal headway in closing achievement gaps), suggests that, to date, the impact of NCLB has fallen short of its extraordinarily ambitious, eponymous goals. Some commentators have argued that the failure of NCLB and earlier accountability reforms to close achievement gaps reflects a flawed, implicit assumption that schools alone can overcome the achievement consequences of dramatic socioeconomic disparities. For example, Ladd (2007) argues that school-accountability policies should be complemented by research-informed programs situated outside of schools (e.g., early childhood and health interventions). Ladd (2007) also emphasizes that schools are embedded within systems with district and state-level actors for whom some form of accountability may also be appropriate.

However, an effective redesign of accountability policies like NCLB may also need to pay more specific attention to the processes and practices within schools (Ladd 2007). Along those lines, it is interesting to note that our evidence of treatment heterogeneity by race, ethnicity, grade and subject is broadly similar to the results from evaluations of earlier state-level school-accountability polices (e.g., Hanushek and Raymond 2005). Understanding the sources of this treatment heterogeneity (i.e., the mediating mechanisms by which accountability can be effective) is likely to be a particularly useful policy datum as the proper design and

implementation of school-accountability are evaluated and discussed. For example, the unique effectiveness of NCLB in improving the math skills of younger students could be related to the biological evidence on the age-dependent malleability of specific cognitive and non-cognitive skills (Heckman 2007). On the other hand, it may be due to the specific ways in which schools and teachers have adjusted their instructional practices, perhaps differently for mathematics and reading. For example, recent studies (Whitehurst 2009) suggest that school decisions about curricula (e.g., textbooks, instructional software, and the corresponding pedagogy) can have comparatively large effects on student achievement. Further research that can credibly and specifically explicate how school and teacher responses have contributed to the achievement effects documented here would be a useful next step in identifying effective policies and practices that can reliably improve student outcomes in chronically low-performing schools at scale.

REFERENCES

Angrist, Joshua D. and Jörn-Steffen Pischke. Mostly Harmless Econometrics: An Empiricist's Companion. Princeton University Press, 2009.

Associated Press. "Diocese of Tucson Becomes 2nd to File for Bankruptcy," September 21, 2004.

Ballou, Dale and Matthew G. Springer. "Achievement Trade-offs and No Child Left Behind," working paper, October 2008.

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

Bertrand, M., Duflo, E., Mullainathan, S. How Much Should We Trust Differences-in-Differences Estimates? Quarterly Journal of Economics 2004; 119(1), 249-75.

Bloom, Howard S. (1999). "Estimating Program Impacts on Student Achievement Using Short Interrupted Time Series" Manpower Demonstration Research Corporation, Working Paper.

Bloom, Howard S., Sandra Ham, Laura Melton, Julieanne O'Brien (2001). "Evaluating the Accelerated Schools Approach A Look at Early Implementation and Impacts on Student Achievement in Eight Elementary Schools." Manpower Demonstration Research Corporation.

Braun, Henry, & Qian, Jiahe. (2008). *Mapping state standards to the NAEP scale*. Report No: ETS RR-08-57. Princeton, NJ: Educational Testing Service.

Bryk, Anthony, Valerie E. Lee and Peter B. Holland. *Catholic Schools and the Common Good*. Harvard University Press, 1993.

Carnoy, Martin and Susanna Loeb. "Does External Accountability Affect Student Outcomes? A Cross-State Analysis," Educational Evaluation and Policy Analysis 24(4), Winter 2002, pages 305-331.

Carroll, Matt, Sacha Pfeiffer, Michael Rezendes and Walter V. Robinson. "Church Allowed Abuse by Priest for Years, Aware of Geoghan Record, Archdiocese Still Shuttled Him from Parish to Parish," *The Boston Globe*, January 6, 2002, A1.

Center on Education Policy. From the Capital to the Classroom: Year 4 of the No Child Left Behind Act. Washington, DC, March 2006.

Center on Education Policy. Has Student Achievement Increased Since 2002: State Test Score Trends Through 2006-07. Washington, DC, June 2008.

Dobbs, Michael. “Conn. Stands in Defiance on Enforcing 'No Child'”, Washington Post, Sunday, May 8, 2005

Figlio, D. N., & Ladd, H. (2008). School accountability and student achievement. In H. Ladd & E. Fiske (Eds.), *Handbook of research in education finance and policy* (pp. 166-182). New York and London: Routledge.

Fuller, Bruce, Joseph Wright, Kathryn Gesicki, and Erin Kang. “Gauging Growth: How to Judge No Child Left Behind?” Educational Researcher 36(5), 2007, pages 268-278.

Goertz, M.E., and M.E. Duffy. Assessment and Accountability Systems in the 50 States: 1999-2000. CPRE Research Report RR-046. Consortium for Policy Research in Education, Philadelphia PA, 2001.

Hanushek, Eric A. and Margaret E. Raymond. “The Confusing World of Educational Accountability,” National Tax Journal 54(2), June 2001, pp. 365-384.

Hanushek, Eric A. and Margaret E. Raymond. “Does School Accountability Lead to Improved Student Performance?” Journal of Policy Analysis and Management 24(2), 2005, pages 297-327.

Heckman, James J. “The Economics, Technology and Neuroscience of Human Capability Formation,” Proceedings of the National Academy of Sciences, 104(33): 13250-13255, (2007).

Hess, Frederick M. and Michael J. Petrilli. *No Child Left Behind Primer*. New York: Peter Lang Publishing, 2006.

Hess, Frederick M. and Michael J. Petrilli. “Wrong Turn on School Reform,” Policy Review, February/March 2009.

Jacob, Brian A. (2008). Lecture for the David N. Kershaw Award, Annual Fall Meeting of the Association of Public Policy Analysis, November 2008, Los Angeles, CA.

Jacob, B. (2005). “Accountability, Incentives and Behavior: Evidence from School Reform in Chicago.” *Journal of Public Economics*. 89(5-6): 761-796.

Jacob, B. and Levitt, S. (2003). “Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating.” *Quarterly Journal of Economics*. 118(3): 843-877.

Koretz, Daniel (2008). Measuring Up: What Educational Testing Really Tells Us. Harvard University Press, Cambridge, MA.

Krieg, John M. “Are Students Left Behind? The Distributional Effects of the No Child Left Behind Act,” Education Finance and Policy 3(2), Spring 2008, pages 250–281.

Ladd, Helen F. “Holding Schools Accountable Revisited,” 2007 Spencer Foundation Lecture in Education Policy and Management, Association for Public Policy Analysis and Management, https://www.appam.org/awards/pdf/2007Spencer-Ladd.pdf, Accessed November 8, 2009.

Lee, Jaekyung. “Tracking Achievement Gaps and Assessing the Impact of NCLB on the Gaps: An In-depth Look into National and State Reading and Math Outcome Trends,” The Civil Rights Project, Harvard University, June 2006.

National Center for Education Statistics. *Mapping 2005 State Proficiency Standards Onto the NAEP Scales* (NCES 2007–482). U.S. Department of Education, National Center for Education Statistics, Washington, D.C.: U.S. Government Printing Office.

Neal, Derek and Diane Whitmore Schanzenbach “Left Behind by Design: Proficiency Counts and Test-Based Accountability” *Review of Economics and Statistics*, forthcoming.

Nichols, Sharon L. and David C. Berliner. *Collateral Damage: How High-Stakes Testing Corrupts America's Schools*. Harvard Education Press, 2007.

Olson, Lynn. “States Strive Toward ESEA Compliance,” *Education Week*, December 1, 2002.

Olson, Lynn. “In ESEA Wake, School Data Flowing Forth,” *Education Week*, December 10, 2003.

Olson, Lynn. “Taking Root,” *Education Week*, December 8, 2004.

Palmer, Scott R. and Arthur L. Coleman. “The No Child Left Behind Act: Summary of NCLB Rrequirements and Deadlines for State Action,” Council of Chief State School Officers, November 2003 (http://www.ccsso.org/content/pdfs/Deadlines.pdf, Accessed November 13, 2009).

Rothstein, Richard, Rebecca Jacobsen and Tamara Wilder (2008). “Grading Education: Getting Accountability Right.” Teachers College Press, New York City.

Ravitch, Diane “Time to Kill ‘No Child Left Behind’” Education Week, June 10, 2009.

Shadish, W.R., Cook, T.D., & Campbell, D.T. (2002). Experimental and Quasi-Experimental Designs for Generalized Causal Inference. Boston: Houghton-Mifflin.

Springer, Matthew G. “The Influence of an NCLB Accountability Plan on the Distribution of Student Test Score Gains,” Economics of Education Review 27, 2008, pages 556-563.

Stullich, Stephanie, Elizabeth Eisner, Joseph McCrary, and Collette Roney. *National Assessment of Title I Interim Report to Congress: Volume I: Implementation of Title I*, Washington, DC: U.S. Department of Education, Institute of Education Sciences, 2006.

U.S. Department of Education. "No Child Left Behind is Working," December, 2006, http://www.ed.gov/nclb/overview/importance/nclbworking.html, Accessed July 29, 2009.

U.S. Department of Education, Office of Planning, Evaluation and Policy Development, Policy and Program Studies Service, Private School Participants in Federal Programs Under the *No Child Left Behind Act* and the *Individuals with Disabilities Education Act,* Washington, D.C., 2007.

Whitehurst, Grover. "Don't Forget Curriculum," Brown Center Letters on Education #3, October 2009.

Winters, Marcus A., Julie R. Trivitt, and Jay P. Greene (Forthcoming). "The Impact of High-Stakes Testing on Student Proficiency in Low-Stakes Subjects: Evidence from Florida's Elementary Science Exam." *Economics of Education Review*.

Appendix A - Catholic Schools as an NCLB Comparison Group

In earlier versions of this research (e.g., Jacob 2008), we also presented results based on using Catholic schools as a comparison group for evaluating NCLB. The basic logic of this complementary research design was that, though private-school students are eligible to participate in a number of major programs under the Elementary and Secondary Education Act (ESEA), the NCLB reauthorization of ESEA left these prior provisions "largely intact" (U.S. Department of Education 2007). This implies that the NCLB reforms were largely, though not completely, irrelevant for Catholic schools.

Figures A1 and A2 show the comparative achievement trends for public and Catholic-school students from the national NAEP. In Figures A1 we see that students in Catholic schools outperformed their counterparts in public schools over the entire period 1990-2007. While both groups showed increasing achievement during the pre-NCLB period, public school students (particularly in 4th grade) experienced a shift in achievement in 2003 and continued at roughly the same slope afterwards. Students in Catholic schools, by contrast, experienced no such shift and achievement trends appeared to flatten for this group after 2003. These comparisons appear to be broadly consistent with the results based on comparing the achievement changes across states with and without school accountability prior to NCLB. That is, they suggest a modest positive impact for 4th grade math and a potential (and smaller) effect for 8th grade math. Figure A2 suggests a similar pattern for reading – potentially positive impacts in 4th grade, but no evidence of impacts at 8th grade.

However, upon further examination, we view public-Catholic comparisons as a deeply suspect approach to evaluating the effects of NCLB. The key issue is that the implementation of NCLB during the 2002-03 school corresponded closely with widespread, nationwide attention to the sex abuse scandal in Catholic schools. Beginning in January of 2002, the Boston Globe published investigative reporting based on access to previously sealed court documents and Church documents related to the prosecution of abusive Catholic priests in the Boston Archdiocese (Carroll et al. 2002). These documents revealed that church officials had frequently reassigned priests known to have been abusive

to different parishes where they were allowed to continue working with children. The high-profile nationwide coverage of this evidence in the spring of 2002 led to similarly incriminating investigations in Catholic dioceses across the United States. Over the subsequent years, these inquiries resulted in high-profile resignations as well as civil lawsuits and large cash settlements. Several Catholic dioceses have also declared bankruptcy since 2002 to protect themselves from the financial repercussions of these lawsuits (e.g., Associated Press 2004).

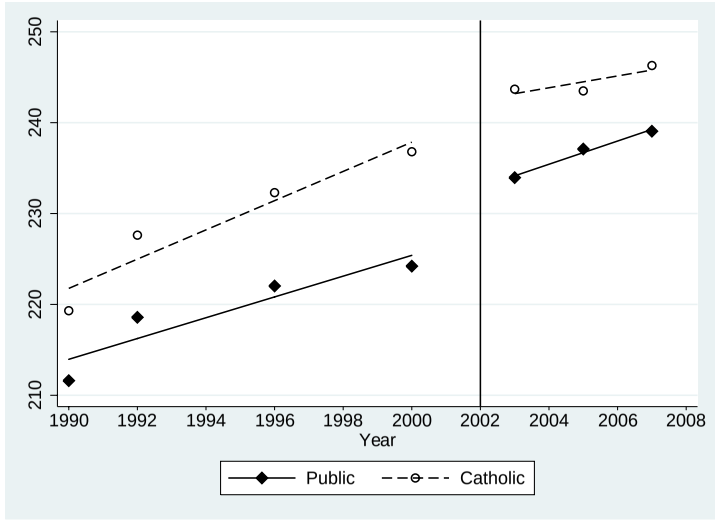
Figure A3 , which shows the comparative elementary-school enrollment trends in Catholic and public schools, strongly suggests that this wide-ranging sex-abuse scandal had a substantial effect on Catholic schools. To facilitate interpretation of the trends, the y-axis in this figure measures the natural logarithm of enrollment, demeaned by the initial year (1992) value so that both trends are zero in 1992 by construction. The trends thus reflect percent changes relative to 1992 in each sector. Catholic enrollment declined slightly prior to NCLB, but then dropped by nearly 10 percent between 2002 and 2004, and fell an additional 7 percent between 2004 and 2006. In contrast, public school enrollment increased steadily prior to NCLB, and leveled off following 2002. Figure A4 suggests that the dramatic enrollment decline in Catholic schools led to a noticeable decline in pupil-teacher ratios in Catholic schools relative to public schools. Pupil-teacher ratios in public schools appeared to increase modestly in absolute terms (relative to steady decline in prior years) after the implementation of NCLB while ratios in Catholic schools dropped relative to prior trends.

The enrollment-driven change in pupil-teacher ratios within Catholic schools that occurred simultaneously with the implementation of NCLB is one factor that complicates using Catholic schools as an NCLB control group. However, a more direct concern is the possibility of confounding bias due to the non-random attrition of students from Catholic schools because of the abuse scandal (and, possibly, as a response to the ongoing recession). To examine the empirical relevance of such non-random attrition, we collected data on the comparative trends in student and parent observables across public and Catholic schools. Figure A5 shows the comparative trends in the percent of public and

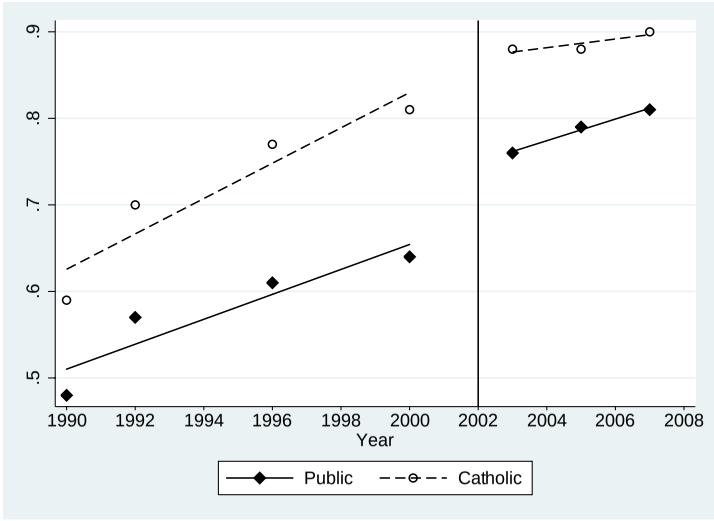
Catholic-school students that are black and Hispanic. Using data available from NAEP surveys, Figure A6 shows the comparative trends in the educational attainment of parents whose children attend Catholic and public schools. The data on the racial and ethnic composition of Catholic and public schools do not suggest that the sharp enrollment drop had noticeable consequences. In contrast, the data on parental education suggests that, after 2002, there was a noticeable comparative increase in the educational attainment of parents whose kids attended Catholic schools. One possible explanation for this pattern is that financial pressure on dioceses which were compelled to respond to civil litigation may have led to tuition increases that led more poorly educated parents to withdraw their children from Catholic schools. Overall, these data provide at best suggestive evidence for non-random attrition from Catholic schools.

Nonetheless, the dramatic enrollment decline that coincided with the abuse scandal and the implementation of NCLB suggests to us that Catholic schools are problematic as a convincing control group. For example, even in the absence of non-random attrition, the scandal may have improved Catholic school quality by lowering class sizes or, alternatively, lowered it by degrading the social trust and sense of community that is often characterized as a key dimension of Catholic school quality (Bryk, Lee, and Holland 1993). In contrast, the Catholic abuse scandal should not confound the identifications strategy based on comparing the achievement trends across states that did and did not have school accountability prior to NCLB. The influx of Catholic school students into public schools would, in all likelihood, have empirically negligible effects on the measured achievement of public school students. Furthermore, the cross-sectional variation in this student sorting should be unrelated to the identifying variation based on a state's pre-NCLB accountability policies.

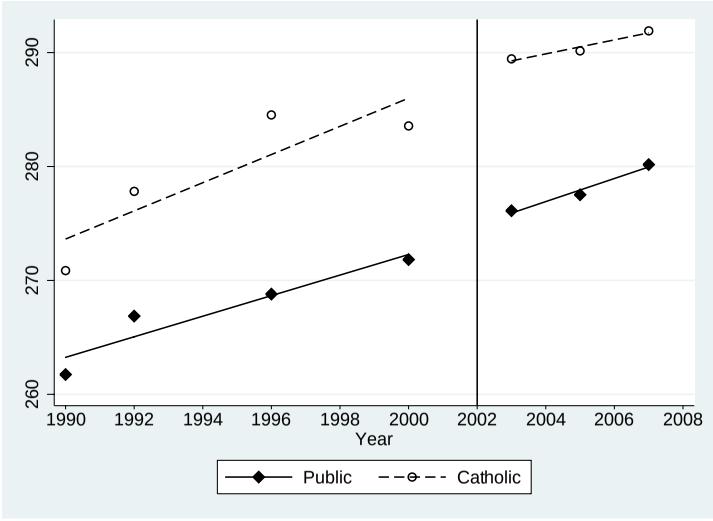
Figure A1: Main NAEP Mathematics Achievement Trends in Public versus Catholic Schools



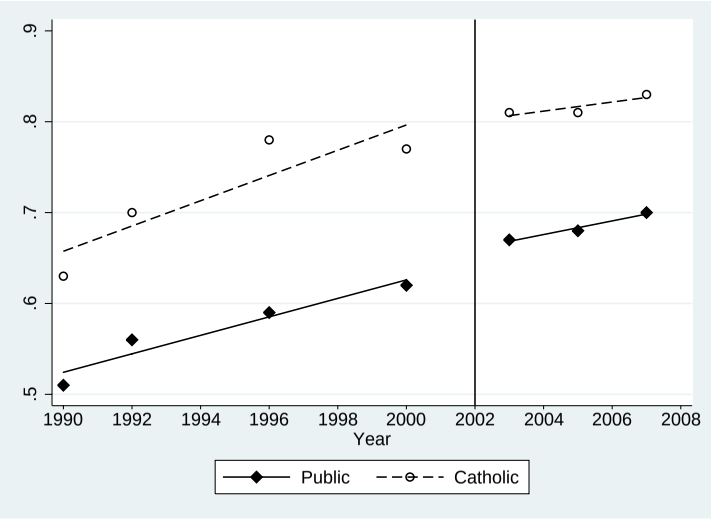
(a) 4th grade average scale score



(b) 4th grade percent meeting "basic" standard

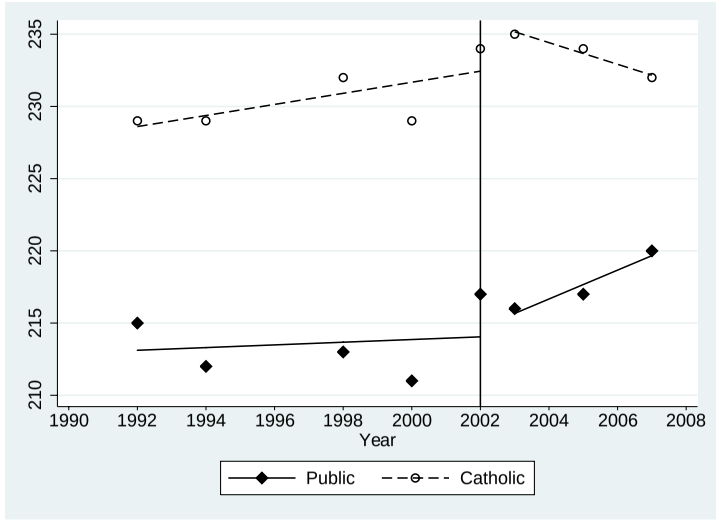


(c) 8th grade average scale score

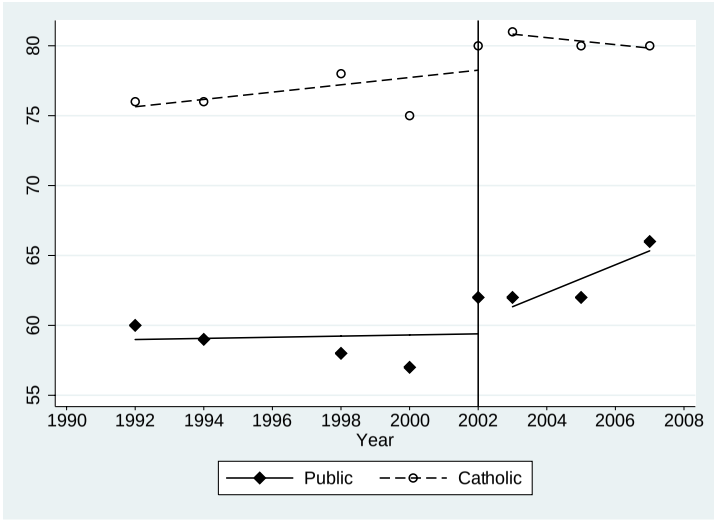


(d) 8th grade percent meeting "basic" standard

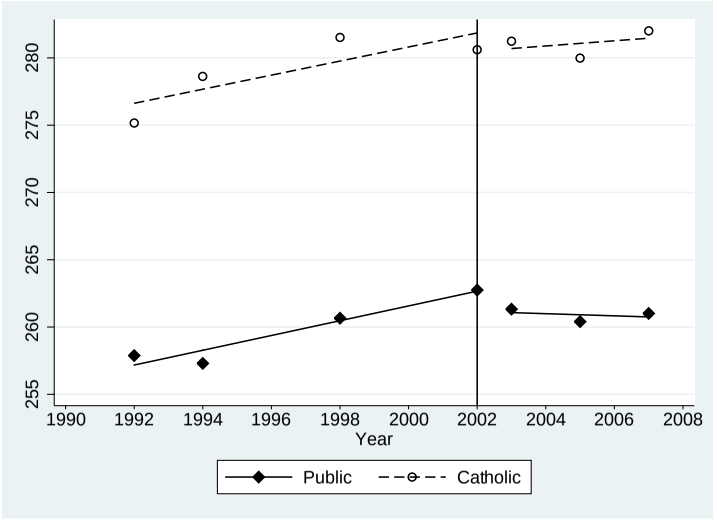
Figure A2: Main NAEP Reading Achievement Trends in Public versus Catholic Schools



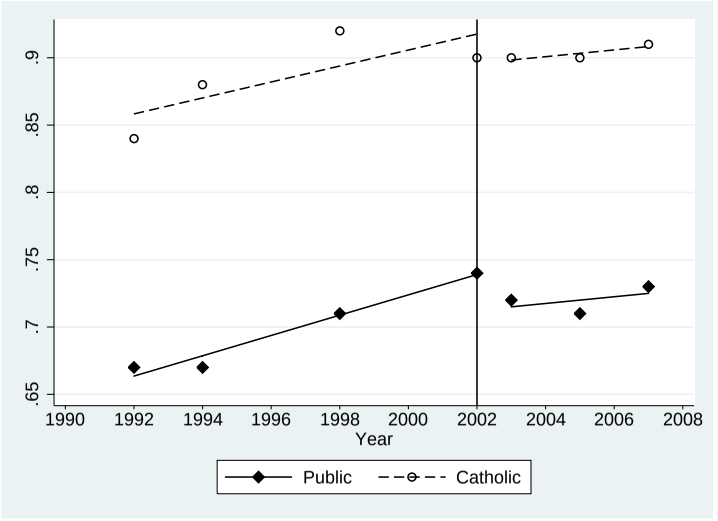
(a) 4th grade average scale score



(b) 4th grade percent meeting "basic" standard

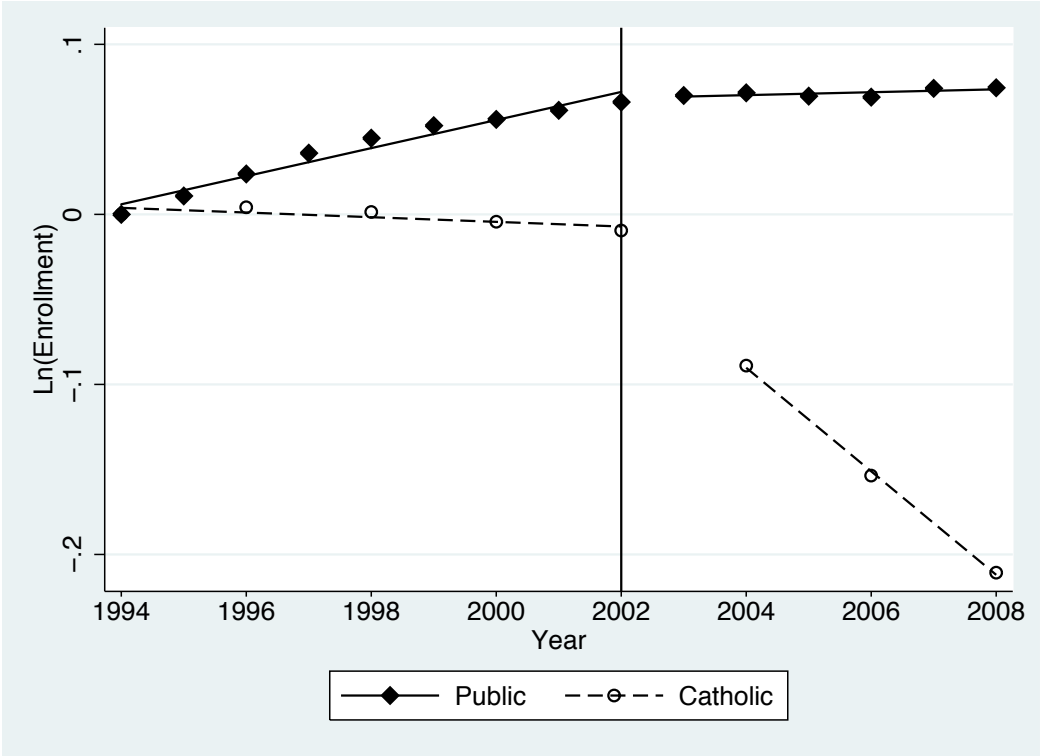


(c) 8th grade average scale score



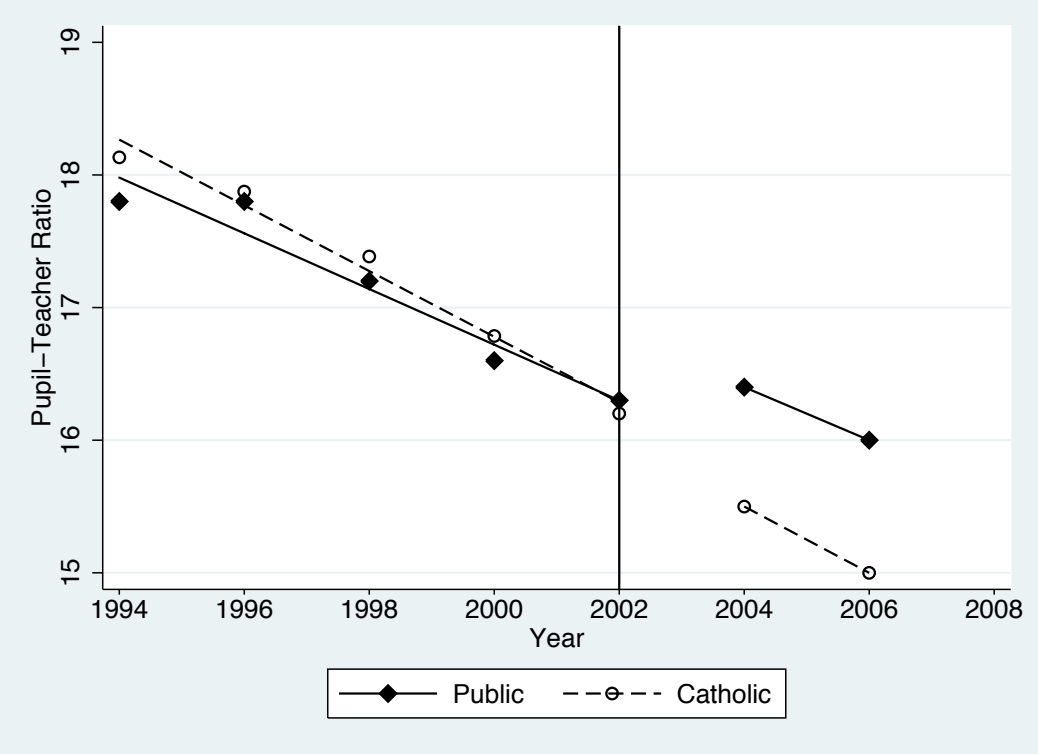
(d) 8th grade percent meeting "basic" standard

Figure A3: Student Enrollment Trends in Public versus Catholic Schools



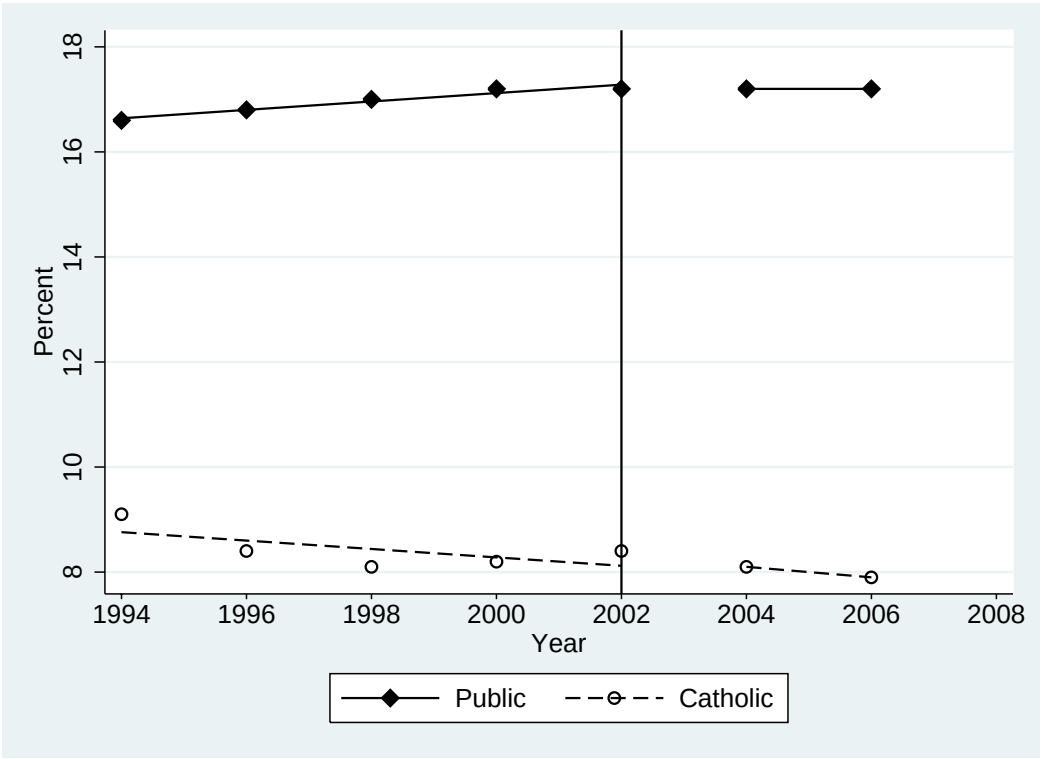
(a) Log elementary school enrollment, no pre-K, relative to 1994

Figure A4: Pupil-Teacher Ratio Trends in Public versus Catholic Schools

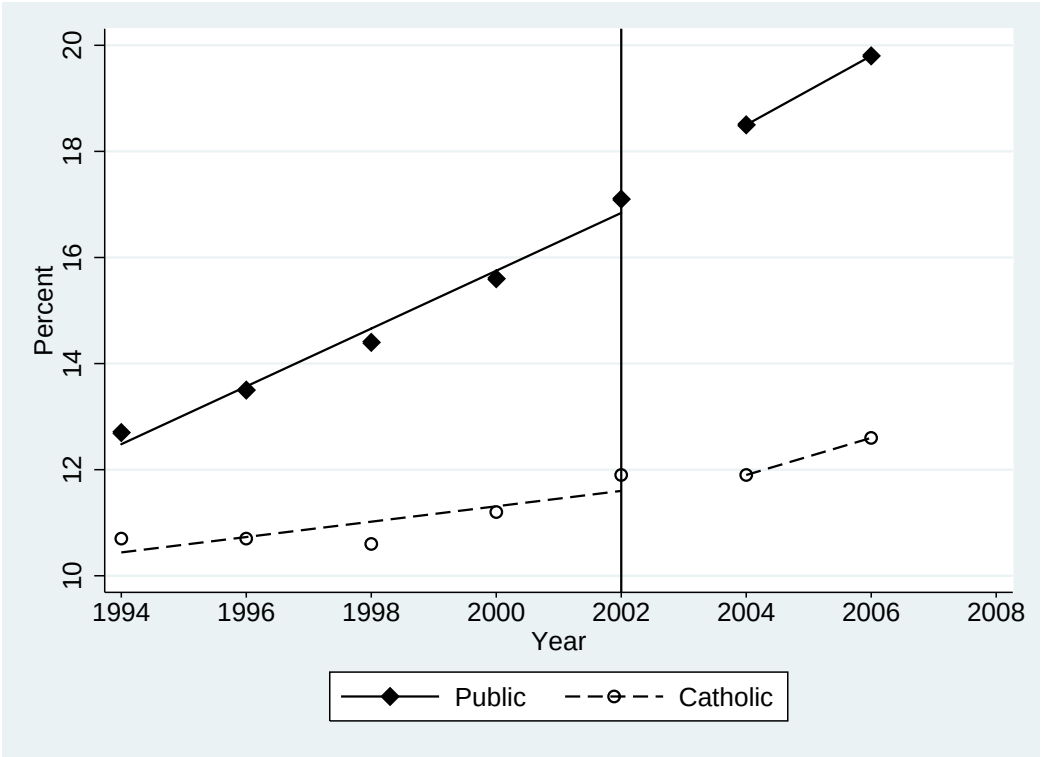


(a) Pupil-teacher ratio

Figure A5: Student Composition Trends in Public versus Catholic Schools

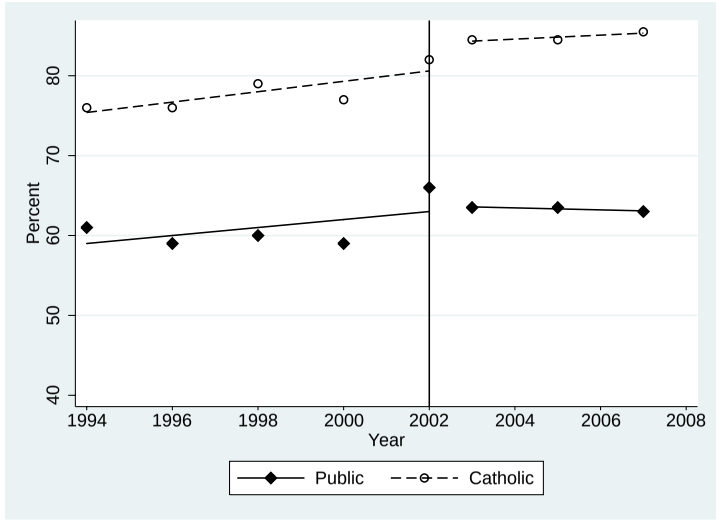


(a) Percent black

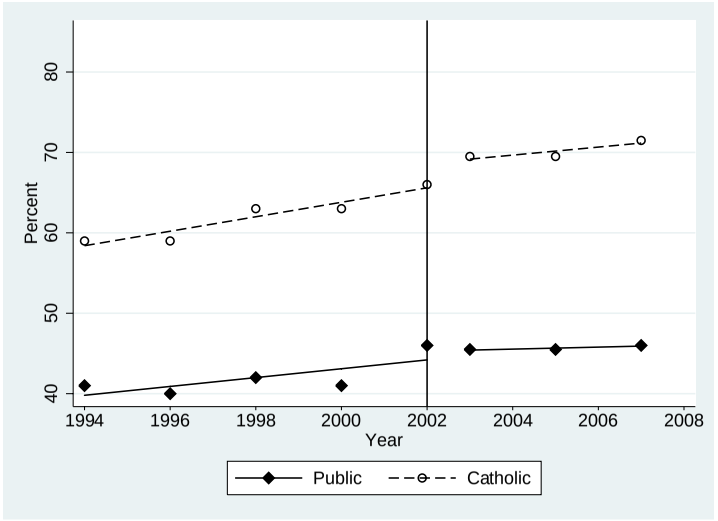


(b) Percent hispanic

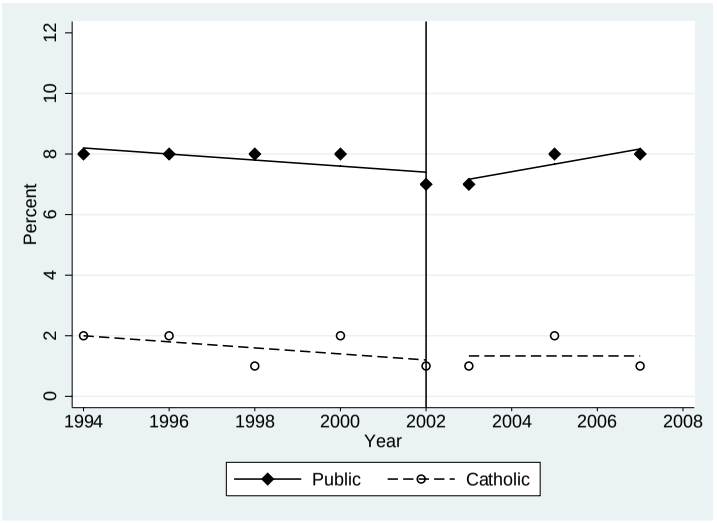
Figure A6: Parental Education Trends in Public versus Catholic Schools



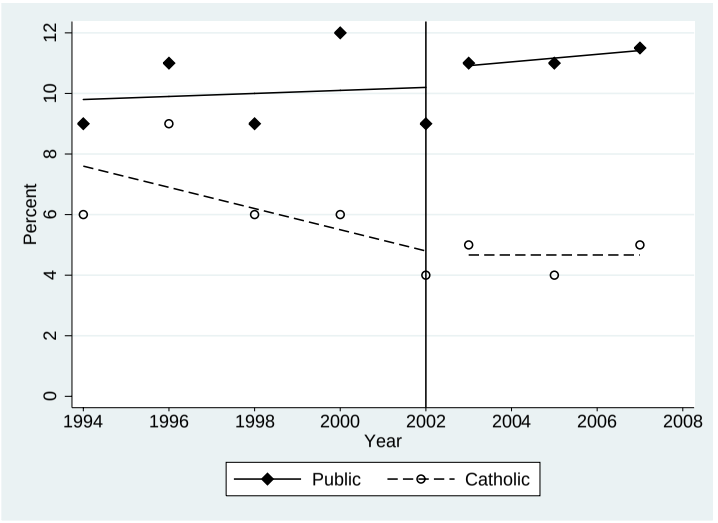
(a) At least some college



(b) College degree



(c) Less than high school



(d) Don't know

Table B1 - States included in NAEP analysis samples

State	Subject-Grade					
	Grade 4 Math	Grade 8 Math	Grade 4 Read	Grade 8 Read	Grade 4 Science	Grade 8 Science
Alabama	1	1	1	1	1	1
Alaska	0	0	0	0	0	0
Arizona	1	1	1	1	1	1
Arkansas	1	1	1	1	1	1
California	1	1	1	1	1	1
Colorado	0	0	0	0	0	0
Connecticut	1	1	1	1	1	1
Delaware	0	0	1	1	0	0
District of Columbia	1	1	1	1	0	0
Florida	0	0	1	1	0	0
Georgia	1	1	1	1	1	1
Hawaii	1	1	1	1	1	1
Idaho	1	1	0	0	1	0
Illinois	0	1	0	0	1	0
Indiana	1	1	0	0	1	1
Iowa	1	0	1	0	0	0
Kansas	0	0	1	1	0	0
Kentucky	1	1	1	1	1	1
Louisiana	1	1	1	1	1	1
Maine	1	1	1	1	1	1
Maryland	1	1	1	1	1	1
Massachusetts	1	1	1	1	1	1
Michigan	1	1	1	0	1	1
Minnesota	1	1	1	0	1	1
Mississippi	1	1	1	1	1	1
Missouri	1	1	1	1	1	1
Montana	1	1	1	1	1	1
Nebraska	1	1	0	0	0	0
Nevada	1	0	1	1	1	0
New Hampshire	0	0	0	0	0	0
New Jersey	0	0	0	0	0	0
New Mexico	1	1	1	1	1	1
New York	1	1	1	1	0	0
North Carolina	1	1	1	1	1	1
North Dakota	1	1	0	0	1	1
Ohio	1	1	0	0	1	0
Oklahoma	1	1	1	1	1	0
Oregon	1	1	1	1	1	1
Pennsylvania	0	0	0	0	0	0
Rhode Island	1	1	1	1	1	1
South Carolina	1	1	1	1	1	1
South Dakota	0	0	0	0	0	0
Tennessee	1	1	1	1	1	1
Texas	1	1	1	1	1	1
Utah	1	1	1	1	1	1
Vermont	1	1	0	0	1	1
Virginia	1	1	1	1	1	1
Washington	0	0	1	1	0	0
West Virginia	1	1	1	1	1	1
Wisconsin	0	0	0	0	0	0
Wyoming	1	1	1	1	1	1
Total	39	38	37	34	36	31

Notes: Our analysis samples consist of states that have 1996 and 2000 NAEP scores in mathematics, 1998 and 2002 scores in reading, and 2000 and 2005 scores in science. NAEP achievement data are not available for racial-ethnic subgroups within all participating state-year observations.

Table B2 - The Estimated Effects of NCLB on NAEP Science Scores

	Grade 4 Science				Grade 8 Science			
	Mean	% Basic	10th percentile	90th percentile	Mean	% Basic	10th percentile	90th percentile
Independent variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: T_s = no prior accountability, no sample exclusions								
$NCLB_t \times T_s$	-0.968	-1.257	-1.778	-0.912	-2.017	-2.489	-2.603	-1.709
	(1.314)	(1.472)	(2.303)	(1.020)	(1.602)	(1.717)	(2.787)	(1.051)
Number of states	36	36	36	36	31	31	31	31
Sample size	72	72	72	72	93	93	93	93
Panel B: T_s = no prior accountability, excludes 1998-2001 adopters								
$NCLB_t \times T_s$	0.883	0.805	1.188	0.261	1.752	1.226	2.934	1.235
	(1.714)	(1.961)	(2.853)	(1.442)	(1.384)	(1.477)	(2.356)	(1.124)
Number of states	22	22	22	22	31	31	31	31
Sample size	44	44	44	44	93	93	93	93
Panel C: T_s = Years without prior school accountability, no sample exclusions								
$NCLB_t \times T_s$	0.168	0.184	0.216	0.055	-0.105	-0.192	-0.086	-0.113
	(0.291)	(0.337)	(0.431)	(0.248)	(0.302)	(0.319)	(0.511)	(0.218)
Total effect relative to state with school accountability starting in 1997	1.008	1.107	1.297	0.332	-0.628	-1.152	-0.514	-0.678
	(1.746)	(2.020)	(2.585)	(1.489)	(1.810)	(1.912)	(3.066)	(1.311)
Number of states	36	36	36	36	31	31	31	31
Sample size	72	72	72	72	93	93	93	93
Mean of Y before NCLB in states without prior accountability	152	71	113	188	151	63	109	190
Student-level standard deviation prior to NCLB	35				36			

Notes: Each column within a panel is a separate regression. All specifications include state fixed effects and linear and quadratic exclusion rates. Standard errors are clustered at the state level. ***p<0.01, ** p<0.05, * p<0.1.

Table B3 - The Estimated Effects of NCLB on NAEP 4th Grade Science Scores, by Subgroup

	OLS				WLS			
	Mean	% Basic	10th percentile	90th percentile	Mean	% Basic	10th percentile	90th percentile
Subgroup	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
White (36 states, n=72)								
Total effect	1.095	1.141	1.666	-0.056	1.976	2.262	2.418	1.015
	(1.613)	(1.757)	(2.330)	(1.437)	(2.181)	(2.279)	(3.026)	(1.869)
Mean of Y before NCLB in states without prior accountability	158	79	124	191	160	80	126	192
Black (28 states, n=56)								
Total effect	3.372	2.606	5.261*	2.368	4.557**	5.656**	5.575	5.296**
	(2.360)	(3.668)	(2.957)	(2.618)	(2.197)	(2.776)	(3.472)	(1.433)
Mean of Y before NCLB in states without prior accountability	126	38	85	163	124	33	87	160
Hispanic (19 states, n=38)								
Total effect	-0.433	-0.612	-1.756	-0.887	2.183	1.273	1.639	1.040
	(3.274)	(4.059)	(5.448)	(2.238)	(4.743)	(5.440)	(5.436)	(4.653)
Mean of Y before NCLB in states without prior accountability	126	39	82	166	119	31	72	163
Male (36 states, n=72)								
Total effect	0.876	0.594	1.518	0.439	2.436	1.781	3.217	2.382
	(1.565)	(1.791)	(2.708)	(1.321)	(2.408)	(2.336)	(4.115)	(1.850)
Mean of Y before NCLB in states without prior accountability	154	73	114	191	153	72	112	190
Female (36 states, n=72)								
Total effect	1.112	1.627	0.969	0.413	1.299	2.081	0.631	0.946
	(1.989)	(2.299)	(2.972)	(1.913)	(2.609)	(2.727)	(3.641)	(2.755)
Mean of Y before NCLB in states without prior accountability	149	68	112	185	149	67	111	186
Free Lunch Eligible (36 states, n=72)								
Total effect	0.487	0.397	1.932	-0.022	1.256	0.807	2.317	1.042
	(2.764)	(3.512)	(4.038)	(2.230)	(3.831)	(4.376)	(5.734)	(2.861)
Mean of Y before NCLB in states without prior accountability	139	55	99	176	137	52	96	174
Not Free Lunch Eligible (36 states, n=72)								
Total effect	1.065	1.305	2.072	0.313	1.884	2.185	2.750	1.580
	(1.591)	(1.673)	(2.196)	(1.629)	(1.962)	(2.017)	(2.754)	(2.147)
Mean of Y before NCLB in states without prior accountability	160	81	127	193	161	82	127	194

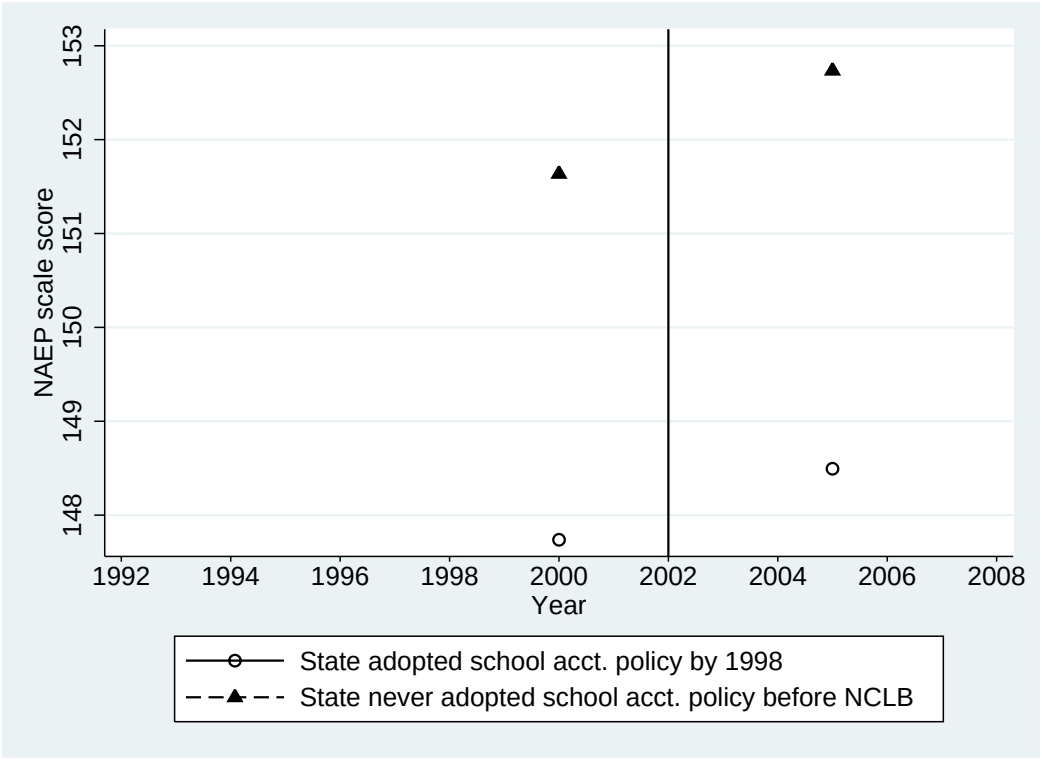
Notes: Each cell is a separate regression as in Panel C of Table 3. The total NCLB effect by 2007 is relative to a state with school accountability starting in 1997. All specifications include state fixed effects and a quadratic in the exclusion rate. Standard errors are clustered at the state level. ***p<0.01, ** p<0.05, * p<0.1.

Table B4 - The Estimated Effects of NCLB on NAEP 8th Grade Science Scores, by Subgroup

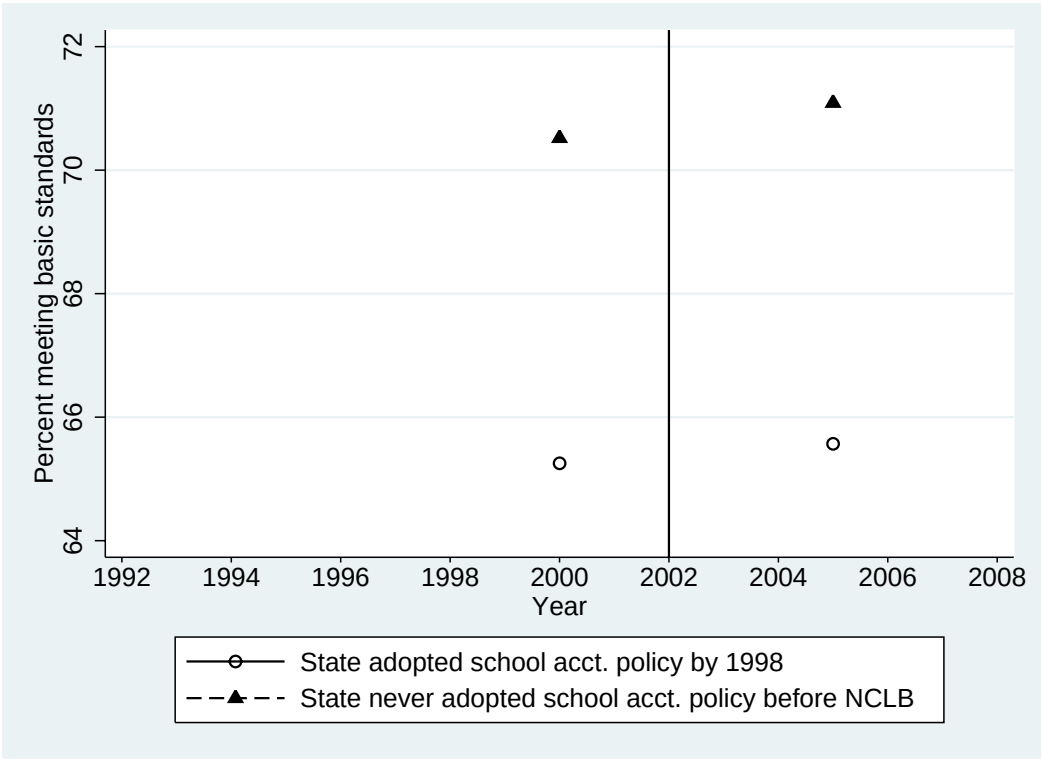
	OLS				WLS			
Subgroup	Mean (1)	% Basic (2)	10th percentile (3)	90th percentile (4)	Mean (5)	% Basic (6)	10th percentile (7)	90th percentile (8)
White (31 states, n=93)								
Total effect	0.098	0.050	0.612	-0.419	0.911	1.074	1.713	0.177
	(1.561)	(1.869)	(2.543)	(1.415)	(1.232)	(1.451)	(2.116)	(1.407)
Mean of Y before NCLB in states without prior accountability	158	71	119	194	159	72	120	195
Black (21 states, n=63)								
Total effect	1.383	1.139	1.007	0.560	0.527	-1.077	2.474	-1.140
	(3.189)	(2.885)	(5.671)	(3.177)	(2.745)	(3.276)	(3.806)	(3.022)
Mean of Y before NCLB in states without prior accountability	120	23	82	160	117	20	79	156
Hispanic (10 states, n=30)								
Total effect	-2.081	-2.175	-2.512	-1.602	2.068	0.564	2.116	1.617
	(3.557)	(3.074)	(6.520)	(2.185)	(7.019)	(6.216)	(7.985)	(6.256)
Mean of Y before NCLB in states without prior accountability	133	40	90	172	126	32	84	164
Male (31 states, n=93)								
Total effect	-0.615	-0.994	-0.913	-0.749	0.005	-0.662	0.854	-0.970
	(1.910)	(2.079)	(3.130)	(1.264)	(1.715)	(1.725)	(2.877)	(1.429)
Mean of Y before NCLB in states without prior accountability	153	65	109	193	153	64	108	193
Female (31 states, n=93)								
Total effect	-0.716	-1.402	-0.231	-0.336	0.059	-0.507	0.613	0.460
	(2.044)	(2.192)	(3.291)	(1.685)	(1.873)	(1.978)	(3.128)	(1.250)
Mean of Y before NCLB in states without prior accountability	149	61	109	186	148	60	107	186
Free Lunch Eligible (31 states, n=93)								
Total effect	-2.570	-4.574	-0.412	-3.717**	-0.428	-2.791	2.384	-1.368
	(2.582)	(2.653)	(4.161)	(1.848)	(2.620)	(2.421)	(3.587)	(2.286)
Mean of Y before NCLB in states without prior accountability	138	46	94	179	135	43	91	177
Not Free Lunch Eligible (31 states, n=93)								
Total effect	0.635	0.896	2.076	-0.444	0.586	0.990	2.097	-0.684
	(1.590)	(1.767)	(2.457)	(1.473)	(1.607)	(1.695)	(2.473)	(1.635)
Mean of Y before NCLB in states without prior accountability	158	71	119	194	158	71	119	194

Notes: Each cell is a separate regression as in Panel C of Table 3. The total NCLB effect by 2007 is relative to a state with school accountability starting in 1997. All specifications include state fixed effects and a quadratic in the exclusion rate. Standard errors are clustered at the state level. ***p<0.01, ** p<0.05, * p<0.1.

Figure B1: Trends in Grade 4 Science Achievement in the Main NAEP by Timing of Accountability Policy

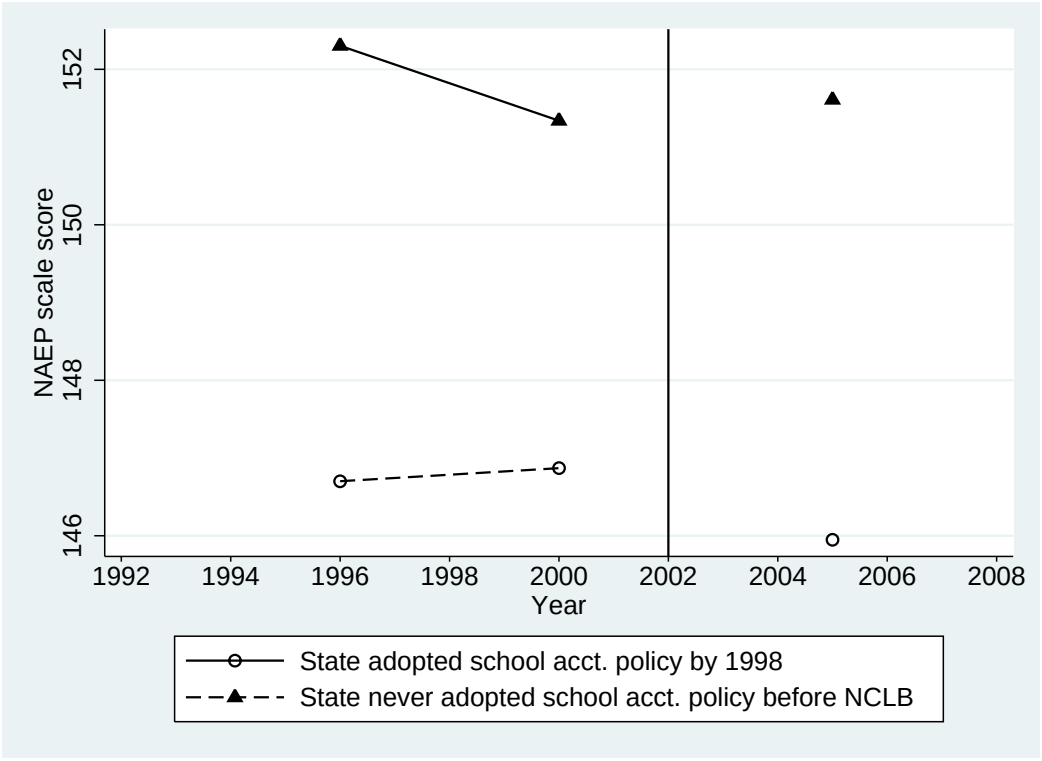


(a) Average scale score

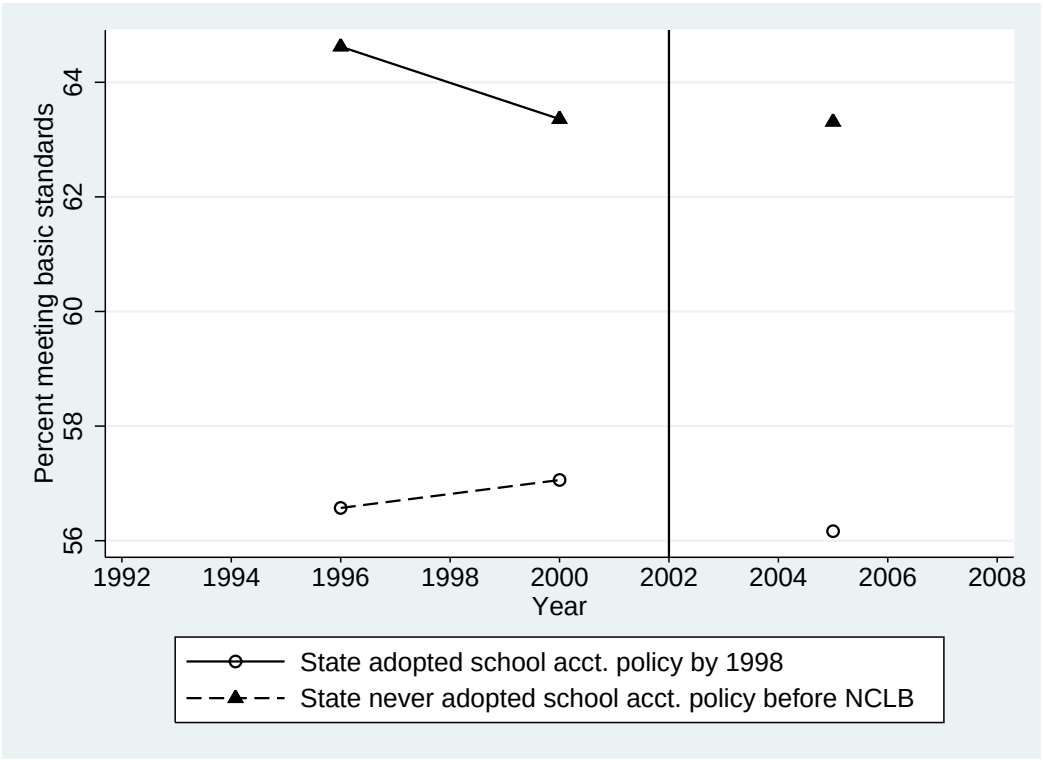


(b) Percent meeting "basic" standard

Figure B2: Trends in Grade 8 Science Achievement in the Main NAEP by Timing of Accountability Policy



(a) Average scale score



(b) Percent meeting "basic" standard

Appendix C - Discrepancies in Accountability Coding

To begin, we reviewed a small number of states that were not included in the study by Hanushek and Raymond (2005) and identified two (i.e., Illinois and Alaska) that implemented consequential accountability in advance of NCLB (i.e., in 1992 and 2001, respectively). Our review also suggested that the timing of consequential-accountability policies differed from that reported by Hanushek and Raymond (2005) in four states: Connecticut, New Mexico, North Carolina and Tennessee. We identified Connecticut as implementing consequential accountability in 1999 (i.e., with the adoption of Public Act 99-288) rather than in the early 1990s. While Connecticut reported on school performance in the early 1990s, it only rated schools that were receiving Title I schools and schools for which a district made a request during this period. We also identified New Mexico as implementing school accountability (i.e., rating school performance and providing financial rewards as well as the threat of possible sanctions) with the 1998 implementation of the Incentives for School Improvement Act rather than in 2003. We identified North Carolina as implementing school accountability in 1996 under the "ABCs of Public Education" rather than in 1993. We identified Tennessee as implementing consequential school accountability in the fall of 2000 rather than in 1996. While Tennessee did begin reporting school performance in 1996, it did not rate schools, identify low performers or attach other school-level consequences until the State Board of Education approved a new accountability system in 2000.

Finally, there are four additional states (Indiana, Kansas, Wisconsin and Virginia), which are identified as having consequential accountability in our baseline coding but could be viewed as marginal cases. Hanushek and Raymond (2005) identified both Wisconsin and Virginia as having consequential accountability prior to NCLB. However, in both Wisconsin and Virginia,

the available state sanctions appear to have been clearly limited to school ratings. For example, Education Week (1999) notes, “Wisconsin law strictly limits the state's authority to intervene in or penalize failing schools.” Similarly, Virginia began identifying low-performance schools through an accreditation system that became effective during the 1998-99 school year. However, because of limited state authority, the loss of accreditation was not clearly tied to the possibility of other explicit school sanctions (e.g., school closure). Hanushek and Raymond (2005) also identify Indiana and Kansas as introducing report-card, rather than consequential, accountability prior to NCLB (i.e. in 1995). However, in addition to school-level performance reporting, Kansas had an accreditation process that rated schools and could culminate in several possible sanctions for low-performing schools (e.g., closure). Furthermore, Education Week (1999) indicated that, in addition to rating schools, Indiana rewarded high performing schools and state officials viewed vague state statutes as suggesting they could also close low-performing schools. In our baseline coding, we identify all four of these states as having consequential accountability prior to NCLB. However, we also report the results of a robustness check in which these designations are switched.

Table 1 - Descriptive Statistics, National Data and State-Based Analysis Samples (1992-2007)

Variable	Nation	State-based Analysis Samples			
		4th Grade Math	8th grade Math	4th grade Reading	8th grade Reading
Pre-NCLB NAEP Performance					
4th grade math - 2000 average	224	224			
4th grade math -Percent change, 1992 to 2000	2.28%	3.53%			
8th grade math - 2000 average	272		271		
8th grade math - Percent change, 1992 to 2000	1.87%		2.35%		
4th grade reading - 2002 average	217			216	
4th grade reading -Percent change, 1994 to 2002	2.36%			3.41%	
8th grade reading - 2002 average	263				260
8th grade reading -Percent change, 1998 to 2002	0.77%				0.32%
Observed traits in 2000					
NAEP Exclusion rate, 4th Grade	4%	4.47%			
NAEP Exclusion rate, 8th Grade	4%		4.40%		
Poverty rate	11.30%	12.54%	12.47%		
Pupil teacher ratio	16.40	16.43	16.42		
Current per pupil expenditures	$7,394	$8,773	$8,844		
Percent free lunch	26.92%	31.88%	31.86%		
Percent of students white	62.10%	59.82%	62.08%		
Percent of students black	17.20%	17.78%	16.66%		
Percent of students Hispanic	15.60%	16.39%	15.41%		
Percent of students asian	5.20%	4.06%	4.44%		
Observed traits in 2002					
NAEP Exclusion rate, 4th Grade	6%			7.06%	
NAEP Exclusion rate, 8th Grade	5%				5.91%
Poverty rate	12.10%			12.43%	12.79%
Pupil teacher ratio	16.20			16.57	16.58
Current per pupil expenditures	$8,259			$9,252	$9,174
Percent free lunch	28.81%			33.41%	34.39%
Percent of students white	60.30%			55.60%	56.84%
Percent of students black	17.20%			18.13%	16.75%
Percent of students Hispanic	17.10%			19.85%	19.56%
Percent of students asian	5.60%			4.16%	5.18%
Number of states		39	38	37	34
Sample size		227	220	249	170

Notes: State data are weighted by state-year public-school enrollment.

Table 2 - States with Consequential Accountability prior to NCLB

		Hanushek and Raymond (2005)	Carnoy and Loeb (2002)	Lee and Wong (2004)
State	Implementation Year	Accountability Type (Year)	School Repercussions (1999-2000)	Accountability Type (1995-2000)
IL	1992	n/a	Moderate	Strong
WI	1993	Consequential (1993)	Weak to Moderate	Moderate
TX	1994	Consequential (1994)	Strong	Strong
IN	1995	Report Card (1993)	Moderate	Strong
KS	1995	Report Card (1993)	Weak	Moderate
KY	1995	Consequential (1995)	Strong	Strong
NC	1996	Consequential (1993)	Strong	Strong
NV	1996	Consequential (1996)	Weak	Moderate
OK	1996	Consequential (1996)	Weak	Moderate
AL	1997	Consequential (1997)	Strong	Strong
RI	1997	Consequential (1997)	Weak implementation	Moderate
WV	1997	Consequential (1997)	Strong	Moderate
DE	1998	Consequential (1998)	None	Weak
MA	1998	Consequential (1998)	Implicit only	Weak
MI	1998	Consequential (1998)	Weak	Moderate
NM	1998	Consequential (2003)	Moderate to strong	Strong
NY	1998	Consequential (1998)	Strong	Strong
VA	1998	Consequential (1998)	Weak to Moderate	Moderate
AR	1999	Consequential (1999)	None	Weak
CA	1999	Consequential (1999)	Strong	Moderate
CT	1999	Consequential (1993)	Weak	Moderate
FL	1999	Consequential (1999)	Strong	Strong
LA	1999	Consequential (1999)	Moderate	Strong
MD	1999	Consequential (1999)	Strong	Strong
SC	1999	Consequential (1999)	Moderate	Moderate
VT	1999	Consequential (1999)	Weak	Moderate
GA	2000	Consequential (2000)	None	Moderate
OR	2000	Consequential (2000)	Weak to Moderate	Moderate
TN	2000	Consequential (1996)	Weak	Moderate
AK	2001	n/a	None	Weak

Additional sources: CPRE Assessment and Accountability Profiles, Education Week (1999), CCSSO annual surveys, state Department of Education websites and Lexis-Nexis searches of state and local newspaper archives.

Table 3 - The Estimated Effects of NCLB on Mean NAEP Scores

Independent variables	Grade 4 Math (1)	Grade 8 Math (2)	Grade 4 Read (3)	Grade 8 Read (4)
Panel A: T_s = no prior accountability, no sample exclusions				
$NCLB_t \times T_s$	1.538	0.177	1.053	0.104
	(1.209)	(1.350)	(0.869)	(1.035)
$NCLB_t \times T_s \times (Years\ since\ NCLB)_t$	0.649**	0.100	0.354	-0.217
	(0.266)	(0.268)	(0.222)	(0.394)
Total effect by 2007	4.782**	0.677	2.824**	-0.982
	(1.952)	(2.304)	(1.242)	(1.931)
Number of states	39	38	37	34
Sample size	227	220	249	170
Panel B: T_s = no prior accountability, excludes 1998-2001 adopters				
$NCLB_t \times T_s$	4.438**	2.602*	1.851	-0.287
	(1.261)	(1.346)	(1.205)	(1.260)
$NCLB_t \times T_s \times (Years\ since\ NCLB)_t$	0.755*	0.530	-0.086	-0.386
	(0.405)	(0.359)	(0.330)	(0.487)
Total effect by 2007	8.212**	5.253**	1.420	-2.219
	(2.318)	(2.457)	(1.531)	(2.404)
Number of states	24	23	21	19
Sample size	139	132	140	95
Panel C: T_s = Years without prior school accountability, no sample exclusions				
$NCLB_t \times T_s$	0.647**	0.273	0.307**	-0.074
	(0.212)	(0.194)	(0.148)	(0.215)
$NCLB_t \times T_s \times (Years\ since\ NCLB)_t$	0.112*	0.069	0.015	-0.055
	(0.058)	(0.060)	(0.046)	(0.074)
Total effect by 2007 relative to state with school accountability starting in 1997	7.244**	3.704	2.297	-2.101
	(2.240)	(2.464)	(1.441)	(2.070)
Number of states	39	38	37	34
Sample size	227	220	249	170
Mean of Y before NCLB in states without prior accountability	224	272	216	261
Student-level standard deviation prior to NCLB	31	38	36	34

Notes: Each column within a panel is a separate regression. All specifications include state fixed effects and linear and quadratic exclusion rates. Standard errors are clustered at the state level. ***p<0.01, ** p<0.05, * p<0.1.

Table 4 - The Estimated Effects of NCLB on Other Outcomes

Grade-Subject Sample	Exclusion Rate (1)	Poverty Rate (2)	Median household income (3)	Employment-population ratio*100 (4)	Fraction in public schools (5)	% Black (6)	% Hispanic (7)	% Free Lunch (8)	Fraction of Cohort in Full-Day K (9)	Fraction of Cohort Attending Any Pre-K (10)
4th Grade Math (39 states, n=227)										
Total effect by 2007	0.766	-2.016	992	0.262	-1.221*	2.151	-2.786	1.868	0.842	3.166
	(1.156)	(2.204)	(2,158)	(0.922)	(0.664)	(1.503)	(2.137)	(1.781)	(6.221)	(4.314)
Mean of Y before NCLB in states without prior accountability	[3.400]	[11.613]	[38,796]	[51.402]	[91.397]	[13.133]	[6.133]	[26.597]	[45.012]	[42.339]
8th Grade Math (38 states, n=220)										
Total effect by 2007	0.912	-2.364	2,535	-0.097	-2.707	2.765**	-0.596	1.535	3.735	-1.366
	(1.298)	(1.990)	(1,995)	(0.723)	(1.985)	(1.380)	(0.907)	(1.669)	(5.506)	(6.899)
Mean of Y before NCLB in states without prior accountability	[3.214]	[11.914]	[38,631]	[51.189]	[91.907]	[12.643]	[5.429]	[27.119]	[33.208]	[34.398]
4th Grade Reading (37 states, n=249)										
Total effect by 2007	1.868	-1.665	3,187*	-0.225	-0.093	2.387**	-0.121	-0.075	-2.585	3.962
	(1.751)	(1.407)	(1,909)	(0.731)	(1.007)	(0.888)	(1.493)	(1.914)	(6.078)	(4.714)
Mean of Y before NCLB in states without prior accountability	[6.167]	[11.792]	[41,229]	[49.667]	[90.612]	[15.667]	[6.917]	[29.450]	[48.913]	[45.504]
8th Grade Reading (34, states, n=170)										
Total effect by 2007	2.072	-1.979	4,331*	-0.762	1.716	1.057	1.693	-1.437	-2.547	1.739
	(2.158)	(1.802)	(2,248)	(0.875)	(4.238)	(1.498)	(1.823)	(2.946)	(7.810)	(8.092)
Mean of Y before NCLB in states without prior accountability	[4.900]	[12.670]	[40,110]	[48.715]	[89.852]	[16.200]	[6.600]	[31.526]	[46.036]	[35.813]

Notes: Each cell is a separate regression as in Panel C of Table 3. The total NCLB effect by 2007 is relative to a state with school accountability starting in 1997. All specifications include state fixed effects. Parental education not available for 4th grade students. Standard errors are robust to clustering at the state level. ***p<0.01, ** p<0.05, * p<0.1.

Table 5 -The Estimated Effects of NCLB on Mean NAEP Scores, Sensitivity Analyses

Grade-Subject Sample	Baseline (1)	State-year covariates (2)	Enrollment weighted least squares (3)	Alternative coding for VA, WI, IN, KS (4)	Treatment intensity measure (Panel B specification) (5)	Full set of year fixed effects (6)	State-specific trends (7)
4th Grade Math							
Total effect by 2007	7.244**	6.651**	7.162**	5.423**	10.953**	7.254**	7.242**
	(2.240)	(2.266)	(2.818)	(2.532)	(5.010)	(2.251)	(2.529)
8th Grade Math							
Total effect by 2007	3.704	3.893*	1.729	2.363	5.516	3.785	3.255
	(2.464)	(2.162)	(4.408)	(2.554)	(3.734)	(2.476)	(2.996)
4th Grade Reading							
Total effect by 2007	2.297	1.881	1.462	1.807	3.321	2.343*	1.688
	(1.441)	(1.580)	(1.478)	(1.401)	(2.477)	(1.371)	(1.623)
8th Grade Reading							
Total effect by 2007	-2.101	-1.848	-2.112	-1.986	-1.969	-2.076	-1.880
	(2.070)	(1.715)	(1.841)	(2.197)	(2.600)	(2.069)	(2.645)

	No state fixed effects (8)	States with 1+ Pre-NCLB Test Score (9)	States with 3+ Pre-NCLB Test Scores (10)	Alternate accom. coding (11)	NCLB starting in 2002 (12)	NCLB starting in 2004 (13)	Including 2009 math scores (14)
4th Grade Math							
Total effect by 2007	7.533**	5.862**	8.093**	8.571**	7.244**	2.867**	7.227**
	(2.720)	(2.061)	(2.445)	(2.598)	(2.240)	(1.440)	(2.353)
8th Grade Math							
Total effect by 2007	4.930	3.741*	2.714	5.334**	3.704	1.710	4.206*
	(3.351)	(2.093)	(2.765)	(2.621)	(2.464)	(1.459)	(2.285)
4th Grade Reading							
Total effect by 2007	3.657	2.006*	2.158	2.490*	5.942*	0.802	n/a
	(2.337)	(1.139)	(1.439)	(1.467)	(3.307)	(1.160)	
8th Grade Reading							
Total effect by 2007	0.734	-1.104	n/a	-1.430	n/a	-1.189	n/a
	(3.356)	(1.816)		(2.078)		(1.806)	

Notes: Each cell is a separate regression as in Panel C of Table 3. The total NCLB effect by 2007 is relative to a state with school accountability starting in 1997. Specifications include state fixed effects and a quadratic in the exclusion rate, except where indicated otherwise. In addition to the exclusion rate, column 6 includes the fraction of students receiving free lunch, fraction black, fraction hispanic, fraction white, parental education level for 8th grade specifications, the poverty rate and poverty rate squared, the unemployment rate and unemployment rate squared, pupil-teacher ratio, and log per-pupil expenditures in 2007 dollars, all at the state-year level. Column 10 uses the assessment administration that began allowing accommodations in 2000 for math and 1998 for reading rather than using the administration allowing no accommodations in all years. Standard errors are clustered at the state level. ***p<0.01, ** p<0.05, * p<0.1.

Table 6 - The Estimated Effects of NCLB on Achievement Distributions by Grade and Subject

Grade-Subject Sample	Mean (1)	Percent Basic (2)	Percent Proficient (3)	Percent Advanced (4)	10th percentile (5)	25th percentile (6)	50th percentile (7)	75th percentile (8)	90th percentile (9)
4th Grade Math (39 states, n=227)									
Total effect by 2007	7.244**	10.090**	5.590**	0.599	9.046**	8.393**	8.004**	6.634**	5.205**
	(2.240)	(3.145)	(1.891)	(0.458)	(3.767)	(2.728)	(2.282)	(1.902)	(1.916)
Mean of Y before NCLB in states without prior accountability	224	64	21	2	186	205	225	244	259
8th Grade Math (38 states, n=220)									
Total effect by 2007	3.704	5.888**	1.286	-0.397	5.598*	5.065*	3.890*	4.340**	2.537
	(2.464)	(2.680)	(2.055)	(0.914)	(3.236)	(2.745)	(2.216)	(2.189)	(2.404)
Mean of Y before NCLB in states without prior accountability	272	64	24	4	228	251	275	296	314
4th Grade Reading (37 states, n=249)									
Total effect by 2007	2.297	2.359	2.542**	1.102**	3.611	2.244	2.251*	2.258**	2.097**
	(1.441)	(1.592)	(1.035)	(0.396)	(2.804)	(1.969)	(1.366)	(0.938)	(0.805)
Mean of Y before NCLB in states without prior accountability	216	61	29	6	171	194	218	240	258
8th Grade Reading (34, states, n=170)	-2.101	-3.763	1.977	0.057	-5.341*	-3.462	-1.022	1.289	1.172
Total effect by 2007	(2.070)	(2.561)	(2.273)	(0.686)	(3.199)	(2.703)	(2.135)	(2.249)	(2.897)
Mean of Y before NCLB in states without prior accountability	261	73	28	2	219	241	263	282	299

Notes: Each cell is a separate regression as in Panel C of Table 3. The total NCLB effect by 2007 is relative to a state with school accountability starting in 1997. All specifications include state fixed effects and a quadratic in the exclusion rate. Standard errors are clustered at the state level. ***p<0.01, ** p<0.05, * p<0.1.

Table 7 - The Estimated Effects of NCLB on 4th Grade NAEP Math Scores by Subgroup

	OLS				WLS			
	Mean	% Basic	10th percentile	90th percentile	Mean	% Basic	10th percentile	90th percentile
Subgroup	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
White (39 states, n=227)								
Total effect by 2007	5.817**	8.802**	6.503**	4.126**	4.855**	8.203**	6.371*	3.743**
	(1.679)	(2.722)	(3.268)	(1.580)	(2.047)	(3.478)	(3.769)	(1.769)
Mean of Y before NCLB in states without prior accountability	232	76	197	265	233	77	198	265
Black (30 states, n=176)								
Total effect by 2007	4.931	8.061	3.648	4.288	14.573**	22.221**	16.043**	11.631**
	(5.342)	(6.548)	(7.870)	(6.089)	(3.731)	(6.316)	(5.752)	(3.362)
Mean of Y before NCLB in states without prior accountability	203	35	168	238	202	33	169	235
Hispanic (19 states, n=108)								
Total effect by 2007	11.429**	10.800*	21.159**	3.895	9.793**	24.896**	5.633*	10.115**
	(4.242)	(6.373)	(8.772)	(3.440)	(1.411)	(3.634)	(3.283)	(2.401)
Mean of Y before NCLB in states without prior accountability	204	40	164	242	204	36	168	240
Male (39 states, n=227)								
Total effect by 2007	7.408**	9.182**	8.314*	4.667**	7.612**	10.835**	7.665*	4.945**
	(2.368)	(3.244)	(4.417)	(2.047)	(3.545)	(5.217)	(4.664)	(2.978)
Mean of Y before NCLB in states without prior accountability	224	65	186	261	227	68	189	264
Female (39 states, n=227)								
Total effect by 2007	7.365**	10.466**	9.205**	5.245**	7.426**	11.216**	8.503**	6.328**
	(2.258)	(3.244)	(3.345)	(1.927)	(2.480)	(4.622)	(3.350)	(1.734)
Mean of Y before NCLB in states without prior accountability	222.757	64.032	186.234	257.659	225	67	189	259
Free Lunch Eligible (36 states, n=180)								
Total effect by 2007	5.487*	9.196*	7.080	2.103	8.011**	15.053**	10.528**	2.761
	(3.294)	(5.463)	(4.618)	(2.962)	(2.631)	(5.347)	(3.384)	(2.162)
Mean of Y before NCLB in states without prior accountability	212	49	175	248	212	49	176	248
Not Free Lunch Eligible (36 states, n=180)								
Total effect by 2007	3.027	5.418	5.011	0.359	1.385	5.508**	3.792*	-1.741
	(2.568)	(3.863)	(3.229)	(2.631)	(2.508)	(2.758)	(1.964)	(3.915)
Mean of Y before NCLB in states without prior accountability	232	76	197	266	234	78	199	266

Notes: Each cell is a separate regression as in Panel C of Table 3. The total NCLB effect by 2007 is relative to a state with school accountability starting in 1997. All specifications include state fixed effects and a quadratic in the exclusion rate. Standard errors are clustered at the state level. ***p<0.01, ** p<0.05, * p<0.1.

Table 8 - The Estimated Effects of NCLB on 8th Grade NAEP Math Scores by Subgroup

	OLS				WLS			
	Mean	% Basic	10th percentile	90th percentile	Mean	% Basic	10th percentile	90th percentile
Subgroup	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
White (37 states, n=214)								
Total effect by 2007	2.863	4.740*	4.045	1.803	1.828	4.253	3.859	-0.943
	(2.561)	(2.639)	(2.792)	(3.040)	(3.680)	(3.134)	(3.466)	(4.402)
Mean of Y before NCLB in states without prior accountability	281	74	240	320	282	76	242	321
Black (27 states, n=158)								
Total effect by 2007	9.261	9.977	10.644	5.424	8.826	10.004	12.414	7.306
	(6.774)	(7.886)	(7.481)	(7.219)	(8.999)	(11.955)	(8.929)	(8.829)
Mean of Y before NCLB in states without prior accountability	241	28	198	284	242	28	200	283
Hispanic (16 states, n=90)								
Total effect by 2007	20.031**	22.006**	17.773**	20.754**	8.219**	18.692**	2.464	9.230**
	(5.766)	(4.618)	(8.349)	(6.084)	(4.135)	(4.666)	(5.996)	(3.236)
Mean of Y before NCLB in states without prior accountability	246	36	200	291	247	36	204	292
Male (38 states, n=220)								
Total effect by 2007	1.678	3.721	2.565	0.050	-1.702	1.943	1.239	-4.986
	(2.488)	(2.369)	(2.944)	(2.847)	(4.024)	(3.387)	(3.953)	(4.857)
Mean of Y before NCLB in states without prior accountability	273	64	226	316	276	67	229	319
Female (38 states, n=220)								
Total effect by 2007	6.300**	7.690**	8.340**	5.052*	6.436	8.442*	9.272	2.541
	(2.664)	(3.191)	(4.058)	(2.592)	(4.459)	(5.072)	(6.548)	(3.422)
Mean of Y before NCLB in states without prior accountability	271.977	64.108	228.979	312.511	274	67	231	315
Free Lunch Eligible (34 states, n=170)								
Total effect by 2007	10.702*	12.773*	16.808	8.116**	15.761**	23.432**	20.328*	12.690**
	(6.155)	(7.328)	(10.388)	(4.027)	(5.631)	(6.398)	(11.063)	(3.899)
Mean of Y before NCLB in states without prior accountability	257	47	211	300	256	46	210	300
Not Free Lunch Eligible (34 states, n=170)								
Total effect by 2007	2.199	3.152	1.318	3.158	0.992	2.392	2.063	-1.393
	(3.924)	(4.045)	(4.015)	(4.791)	(4.171)	(3.478)	(3.882)	(6.157)
Mean of Y before NCLB in states without prior accountability	279	72	238	320	281	74	240	320

Notes: Each cell is a separate regression as in Panel C of Table 3. The total NCLB effect by 2007 is relative to a state with school accountability starting in 1997. All specifications include state fixed effects and a quadratic in the exclusion rate. Standard errors are clustered at the state level. ***p<0.01, ** p<0.05, * p<0.1.

Table 9 - The Estimated Effects of NCLB on 4th Grade NAEP Reading Scores by Subgroup

	OLS				WLS			
	Mean	% Basic	10th percentile	90th percentile	Mean	% Basic	10th percentile	90th percentile
Subgroup	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
White (37 states, n=249)								
Total effect by 2007	4.854**	4.509**	6.626**	3.472**	5.362**	5.679**	8.252**	3.513**
	(1.231)	(1.379)	(1.938)	(1.132)	(1.201)	(1.393)	(2.249)	(0.888)
Mean of Y before NCLB in states without prior accountability	226	73	184	265	225	72	183	264
Black (32 states, n=214)								
Total effect by 2007	-1.873	-2.294	-2.090	-0.608	-0.871	-4.266*	0.794	0.866
	(3.698)	(3.689)	(6.895)	(2.439)	(2.569)	(2.484)	(4.248)	(1.558)
Mean of Y before NCLB in states without prior accountability	200	43	154	244	195	36	151	238
Hispanic (22 states, n=140)								
Total effect by 2007	6.094	5.597	6.464	5.177	0.242	-0.095	5.488	1.421
	(4.835)	(4.948)	(6.902)	(3.754)	(4.805)	(5.079)	(4.887)	(3.769)
Mean of Y before NCLB in states without prior accountability	199	43	154	244	193	37	144	241
Male (37 states, n=249)								
Total effect by 2007	3.399**	3.510**	4.601	2.953**	2.241*	0.894	3.079	3.490**
	(1.578)	(1.756)	(3.352)	(1.073)	(1.287)	(1.668)	(2.906)	(1.054)
Mean of Y before NCLB in states without prior accountability	212	58	166	254	214	60	167	256
Female (37 states, n=249)								
Total effect by 2007	1.395	1.280	1.762	1.325	0.741	0.607	2.404	-0.057
	(1.535)	(1.811)	(2.605)	(1.249)	(1.697)	(1.827)	(2.761)	(1.472)
Mean of Y before NCLB in states without prior accountability	220	65	176	561	222	68	177	263
Free Lunch Eligible (37 states, n=185)								
Total effect by 2007	0.567	1.278	-0.287	1.895	2.482	2.993	-0.256	5.942
	(4.235)	(4.050)	(6.859)	(3.313)	(4.296)	(4.475)	(5.382)	(5.078)
Mean of Y before NCLB in states without prior accountability	205	49	160	248	206	50	161	249
Not Free Lunch Eligible (37 states, n=185)								
Total effect by 2007	1.355	1.248	-1.851	1.674	-4.790	-4.892	-7.998	-4.390
	(3.042)	(3.328)	(4.201)	(3.299)	(5.073)	(4.761)	(5.818)	(4.771)
Mean of Y before NCLB in states without prior accountability	225	72	184	264	227	74	186	265

Notes: Each cell is a separate regression as in Panel C of Table 3. The total NCLB effect by 2007 is relative to a state with school accountability starting in 1997. All specifications include state fixed effects and a quadratic in the exclusion rate. Standard errors are clustered at the state level. ***p<0.01, ** p<0.05, * p<0.1.

Table 10 - The Estimated Effects of NCLB on 8th Grade NAEP Reading Scores by Subgroup

	OLS				WLS			
	Mean	% Basic	10th percentile	90th percentile	Mean	% Basic	10th percentile	90th percentile
Subgroup	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
White (33 states, n=165)								
Total effect by 2007	0.890	-2.111	-1.699	1.523	2.126	-0.287	0.925	2.178
	(2.220)	(2.303)	(3.678)	(3.127)	(2.142)	(2.068)	(3.324)	(4.004)
Mean of Y before NCLB in states without prior accountability	226	73	184	265	225.116	72.202	183.453	263.904
Black (27 states, n=135)								
Total effect by 2007	-14.513**	-16.658**	-20.449**	-8.678	-11.261**	-16.002**	-15.195**	-4.811
	(4.061)	(5.156)	(4.424)	(6.357)	(2.714)	(3.493)	(5.422)	(5.269)
Mean of Y before NCLB in states without prior accountability	245	54	205	282	244	53	205	280
Hispanic (20 states, n=100)								
Total effect by 2007	6.463	8.349	3.476	15.614	-1.831	-2.856	-9.639	7.476**
	(6.609)	(6.645)	(9.900)	(11.650)	(2.262)	(3.246)	(6.246)	(3.793)
Mean of Y before NCLB in states without prior accountability	243	53	196	285	243	51	196	285
Male (34 states, n=170)								
Total effect by 2007	-2.104	-4.375	-5.908	3.362	-3.745*	-7.095**	-11.065**	6.001*
	(2.305)	(3.209)	(4.092)	(2.320)	(2.138)	(2.876)	(4.731)	(3.120)
Mean of Y before NCLB in states without prior accountability	256	67	213	294	258	70	215	296
Female (34 states, n=170)								
Total effect by 2007	-1.753	-3.314	-5.333*	-0.956	-0.427	-1.472	-4.697**	-0.359
	(2.372)	(2.191)	(3.118)	(3.111)	(2.582)	(1.992)	(2.228)	(3.817)
Mean of Y before NCLB in states without prior accountability	266	78	226	304	268	79	227	306
Free Lunch Eligible (34 states, n=170)								
Total effect by 2007	-4.770*	-6.330*	-7.339	-4.105	-6.447**	-8.050**	-11.568**	-4.551
	(2.856)	(3.719)	(4.737)	(3.470)	(2.375)	(2.927)	(5.009)	(3.906)
Mean of Y before NCLB in states without prior accountability	250	61	207	291	250	61	207	291
Not Free Lunch Eligible (34 states, n=170)								
Total effect by 2007	0.899	-1.538	-4.699	4.247	3.486	0.286	-0.714	4.925
	(2.600)	(2.707)	(3.674)	(4.322)	(3.107)	(2.337)	(3.536)	(4.076)
Mean of Y before NCLB in states without prior accountability	268	80	229	304	270	82	231	306

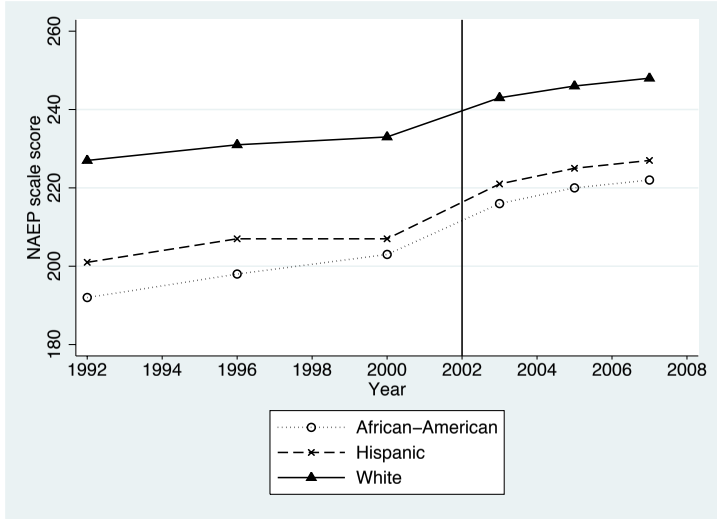
Notes: Each cell is a separate regression as in Panel C of Table 3. The total NCLB effect by 2007 is relative to a state with school accountability starting in 1997. All specifications include state fixed effects and a quadratic in the exclusion rate. Standard errors are clustered at the state level. ***p<0.01, ** p<0.05, * p<0.1.

Table 11 - The Estimated Effects of NCLB on Math Achievement by Subscale

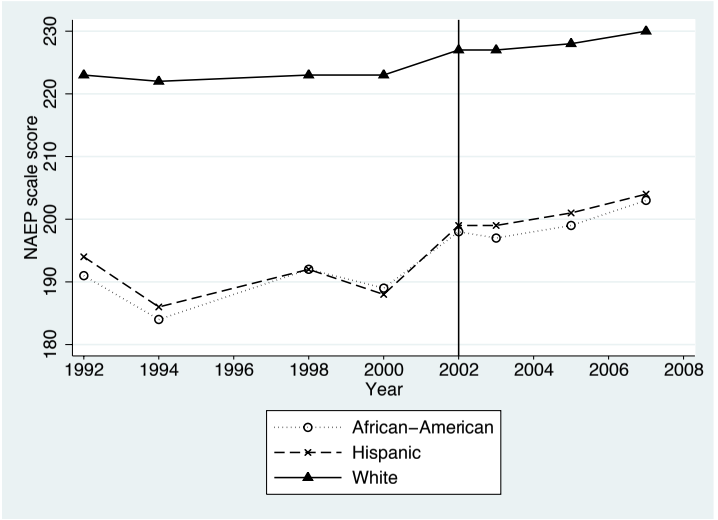
	4th Grade			8th Grade		
	Mean	10th percentile	90th percentile	Mean	10th percentile	90th percentile
Subscale Category	(1)	(2)	(3)	(4)	(5)	(6)
Algebra						
Total effect by 2007	7.725**	13.111**	1.326	3.141	7.202*	-0.730
	(2.213)	(5.069)	(2.540)	(2.798)	(3.979)	(2.660)
Mean of Y before NCLB in states without prior accountability	228	189	264	274	226	319
Student-level standard deviation prior to NCLB	30			36		
Geometry						
Total effect by 2007	5.243**	7.635	2.445	3.943*	8.689**	0.251
	(2.534)	(5.098)	(2.367)	(2.377)	(4.176)	(3.141)
Mean of Y before NCLB in states without prior accountability	225	187	262	270	226	312
	31			34		
Measurement						
Total effect by 2007	5.914**	11.565**	0.751	2.430	9.760	-5.859
	(2.248)	(5.320)	(2.851)	(3.362)	(7.017)	(4.613)
Mean of Y before NCLB in states without prior accountability	224	178	269	272	209	330
Student-level standard deviation prior to NCLB	36			49		
Number Properties and Operations						
Total effect by 2007	8.604**	13.126**	3.619*	4.122*	11.848**	-2.966
	(2.575)	(4.882)	(1.930)	(2.172)	(4.411)	(2.149)
Mean of Y before NCLB in states without prior accountability	220	178	261	273	224	319
Student-level standard deviation prior to NCLB	33			37		
Data Analysis, Statistics, and Probability						
Total effect by 2007	7.306**	16.947**	-1.539	6.767**	16.355**	-2.236
	(2.499)	(5.719)	(3.006)	(2.623)	(6.313)	(4.329)
Mean of Y before NCLB in states without prior accountability	225	183	266	273	217	326
Student-level standard deviation prior to NCLB	32			42		

Notes: Each cell is a separate regression as in Panel C of Table 3. The total NCLB effect by 2007 is relative to a state with school accountability starting in 1997. All specifications include state fixed effects and a quadratic in the exclusion rate. Standard errors are clustered at the state level. ***p<0.01, ** p<0.05, * p<0.1.

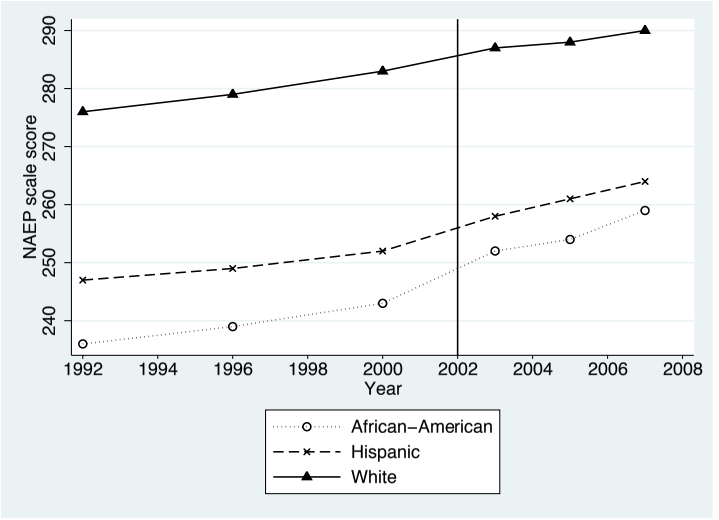
Figure 1: Mean scaled score on the main NAEP for all public schools



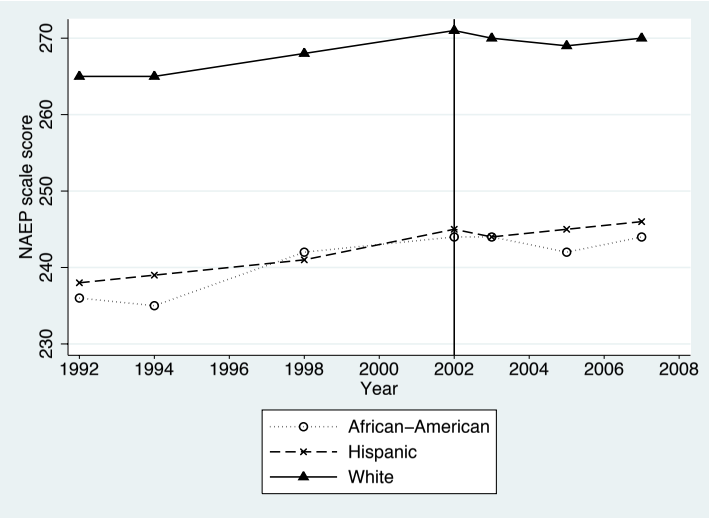
(a) Grade 4 mathematics



(b) Grade 4 reading

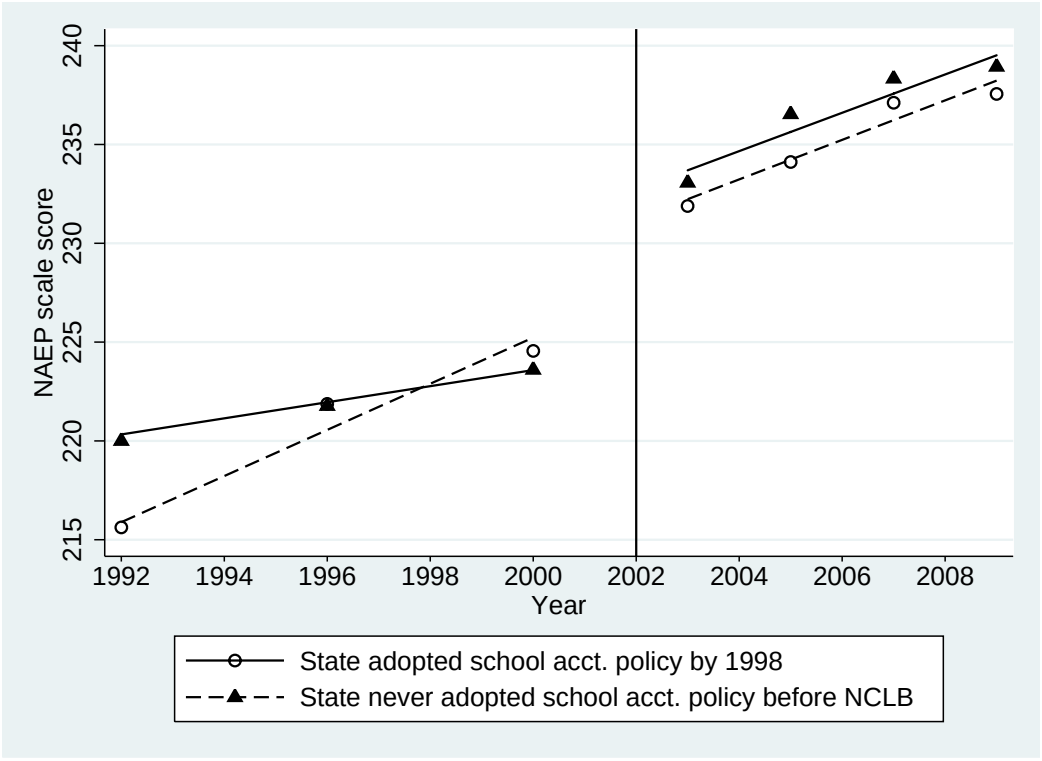


(c) Grade 8 mathematics

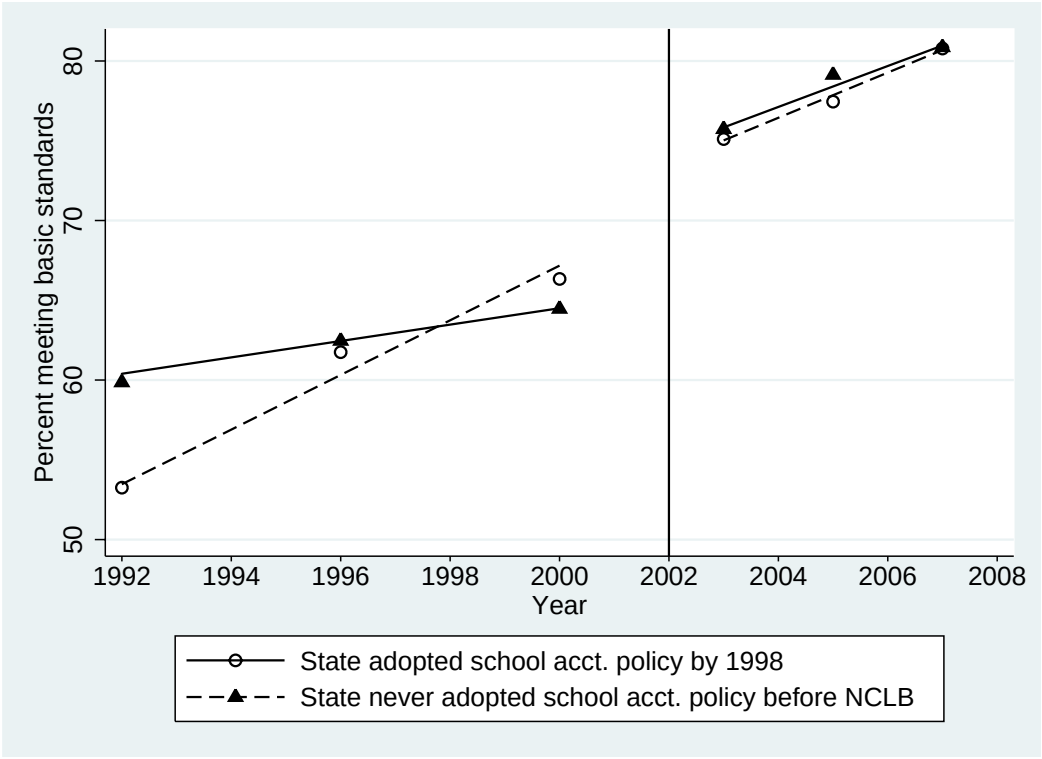


(d) Grade 8 reading

Figure 2: Trends in Grade 4 Mathematics Achievement in the Main NAEP by Timing of Accountability Policy

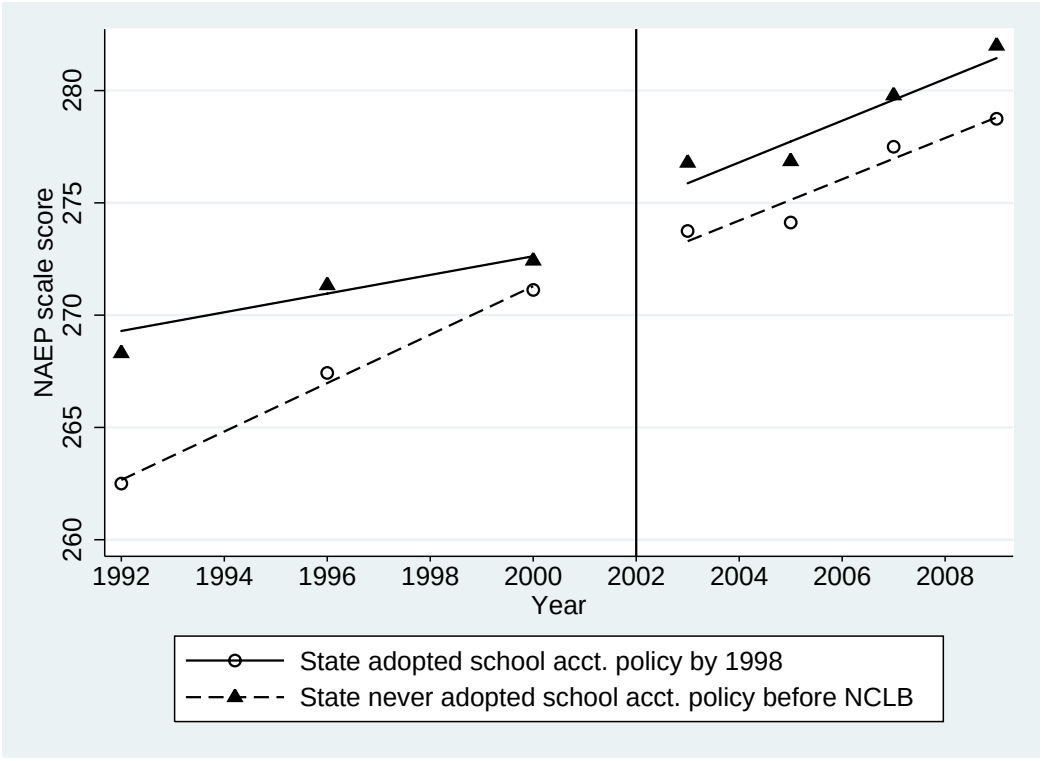


(a) Average scale score

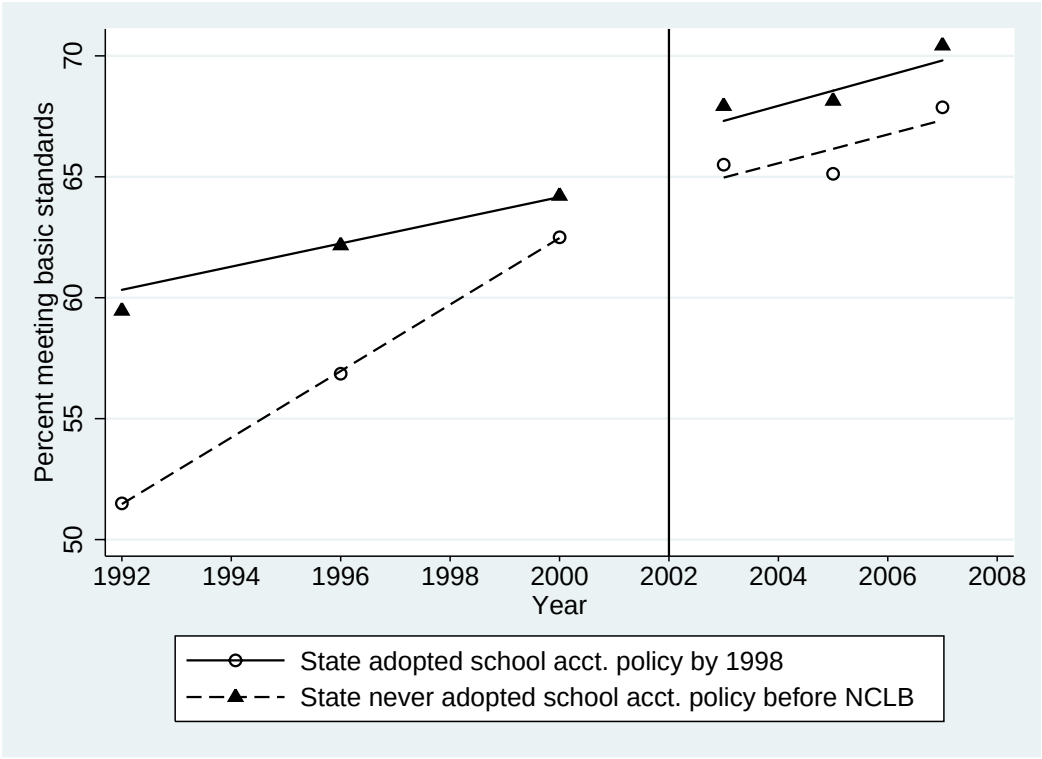


(b) Percent meeting "basic" standard

Figure 3: Trends in Grade 8 Mathematics Achievement in the Main NAEP by Timing of Accountability Policy

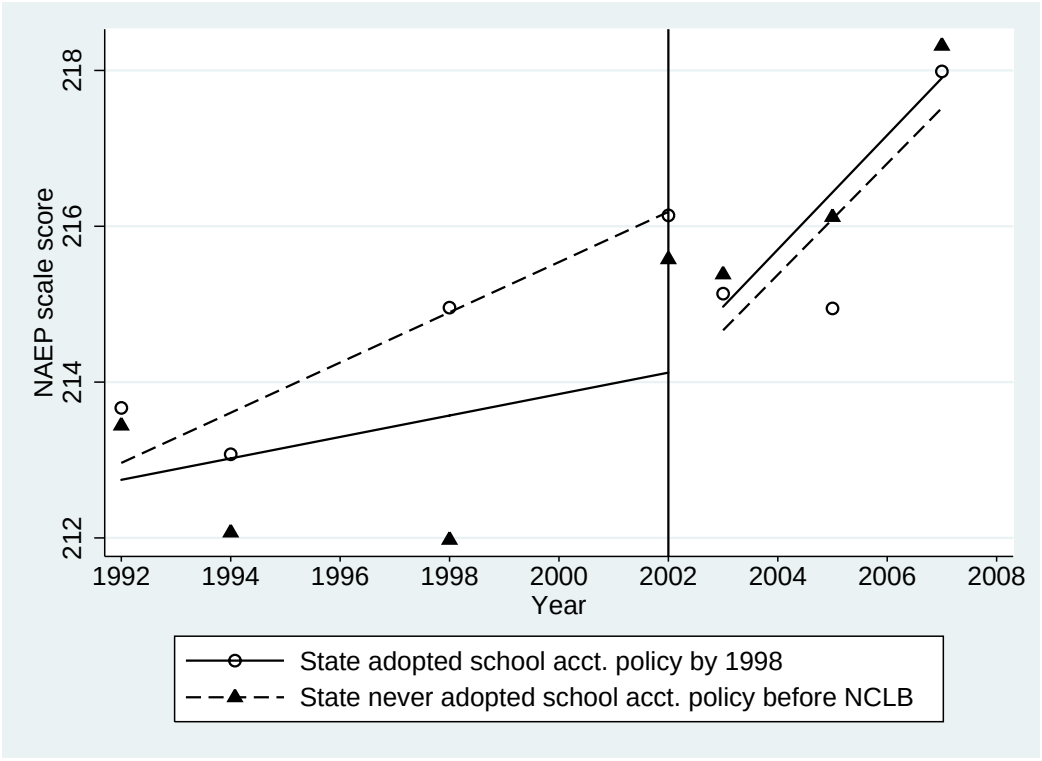


(a) Average scale score

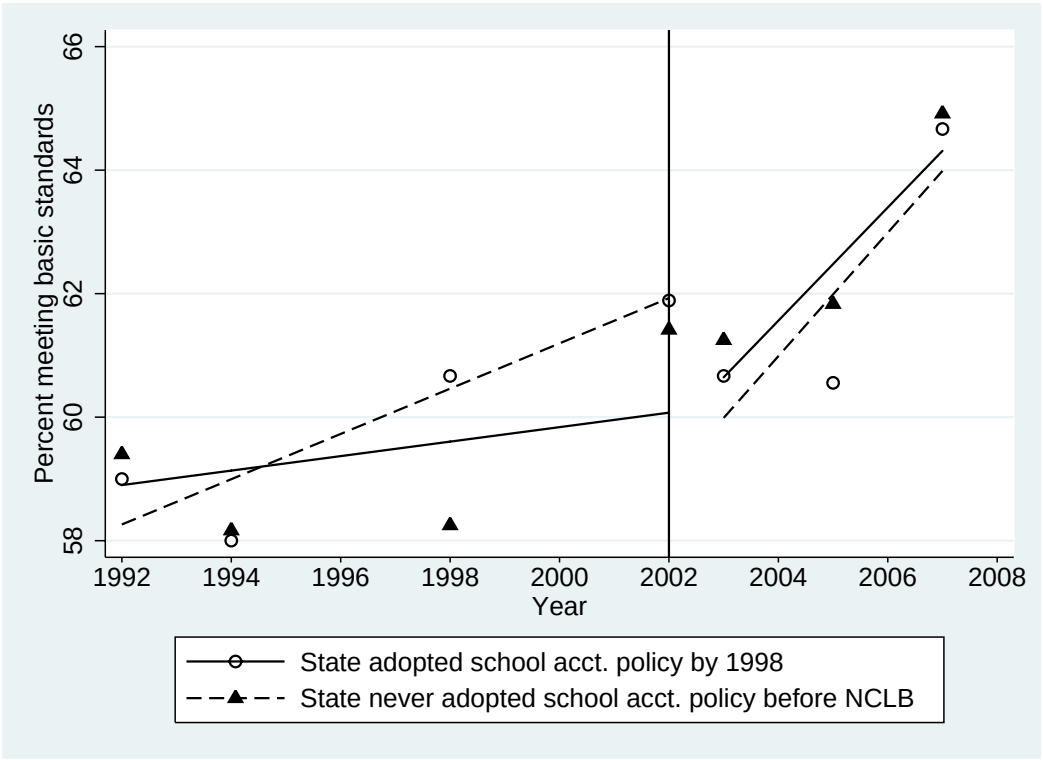


(b) Percent meeting "basic" standard

Figure 4: Trends in Grade 4 Reading Achievement in the Main NAEP by Timing of Accountability Policy

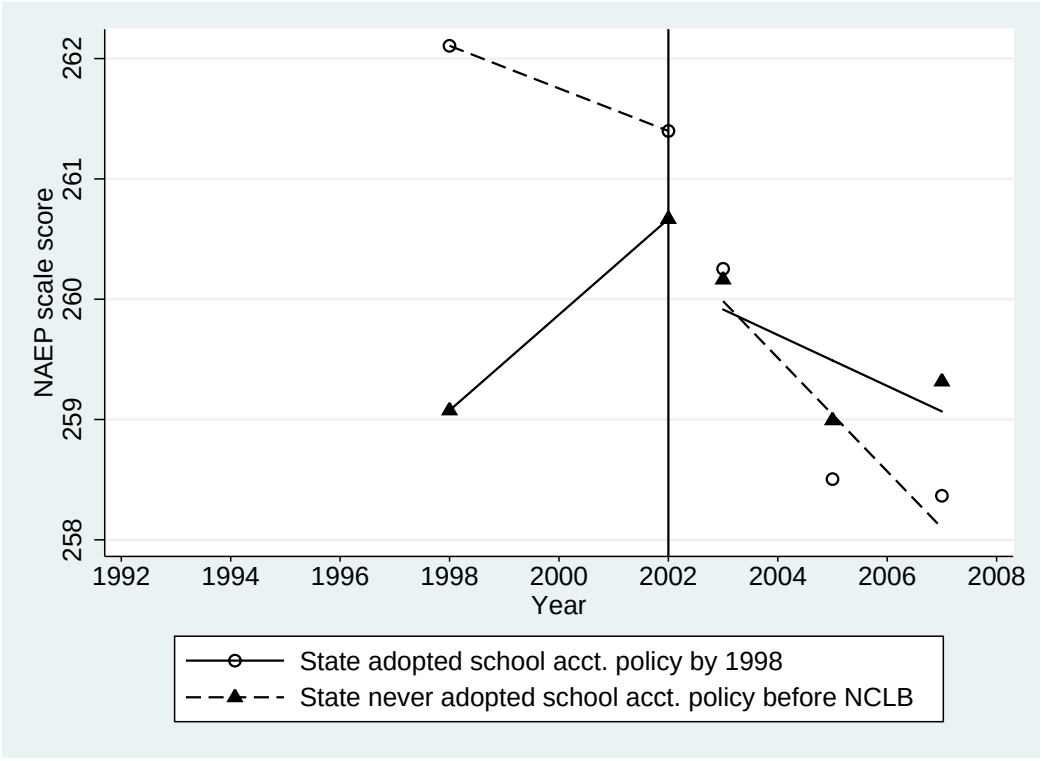


(a) Average scale score

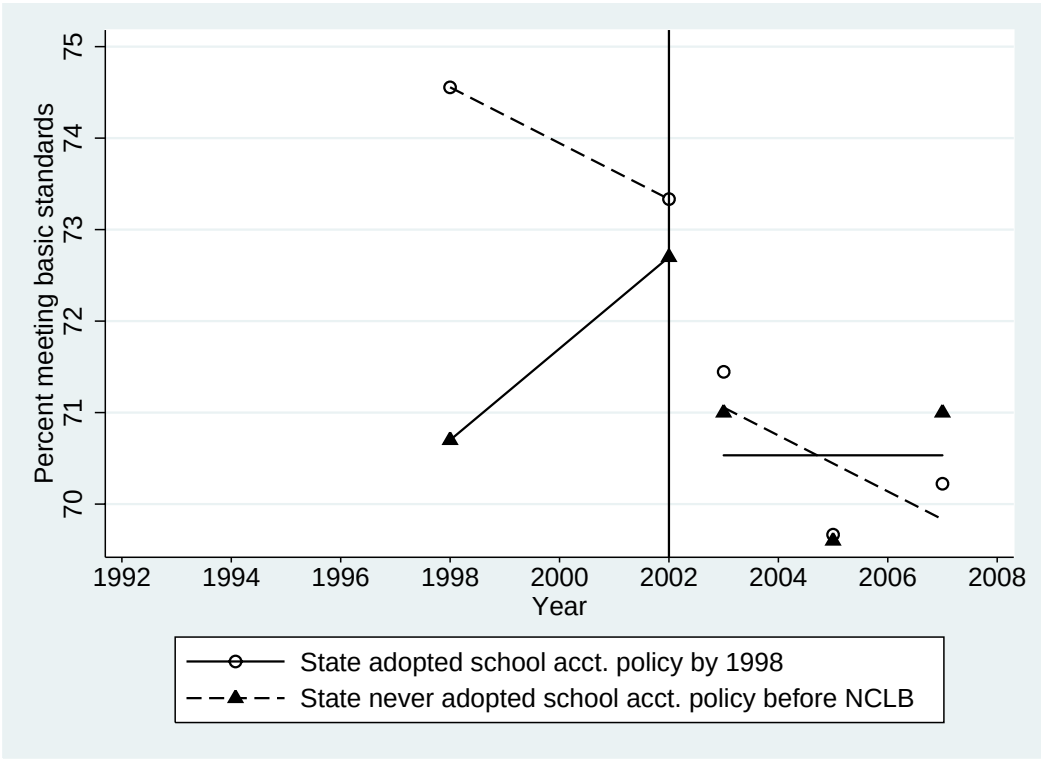


(b) Percent meeting "basic" standard

Figure 5: Trends in Grade 8 Reading Achievement in the Main NAEP by Timing of Accountability Policy



(a) Average scale score



(b) Percent meeting "basic" standard