The Impact of No Child Left Behind on Student Achievement

Thomas S. Dee
Brian Jacob

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 fourth graders (effect size = 0.23 by 2007) as well as improvements at the lower and top percentiles. There is also evidence of improvements in eighth-grade math achievement, particularly among traditionally low-achieving groups and at the lower percentiles. However, we find no evidence that NCLB increased fourth-grade reading achievement. © 2011 by the Association for Public Policy Analysis and Management.

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 failing to make "adequate yearly progress" (AYP) toward the stated goal of having all students achieve proficiency in reading and math by 2013–2014 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 & 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.,

Journal of Policy Analysis and Management, Vol. 30, No. 3, 418–446 (2011)

Published by Wiley Periodicals, Inc. View this article online at wileyonlinelibrary.com/journal/pam
DOI: 10.1002/pam.20586

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, Jacobsen, Wilder, 2008; Koretz, 2008). In the extreme, some suggest that high-stakes testing may lead school personnel to intentionally manipulate student test scores (Jacob & 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, e.g., Fuller et al., 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 eight 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 (CITS) 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 not only consider 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 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 fourth graders (effect size = 0.23 by 2007). These gains occurred at multiple points in the achievement distribution and were concentrated among white and Hispanic students as well as among students eligible for subsidized lunches. We also find that NCLB led to more moderate and targeted improvements in the math achievement of eighth graders (e.g., low-performing students). However, we did not find consistent and reliable evidence that NCLB improved the reading achievement of fourth 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 elementary school math is undoubtedly good news for advocates of NCLB and school accountability. On the other hand, these gains are more modest when compared to NCLB's statutory goal of universal proficiency. Furthermore, the lack of similarly large and broad effects on reading achievement, 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 & Petrilli, 2009). The organization of the paper proceeds as follows. The second section briefly reviews the literature on prior school accountability policies and NCLB and situates the contributions of this study within that literature. The third and fourth sections discuss the methods and data used in this study. The fifth section summarizes the key results and robustness checks. The sixth section concludes with suggestions for further research and thoughts on the contemporary policy implications of these results.

NCLB, SCHOOL ACCOUNTABILITY, AND STUDENT ACHIEVEMENT

The NCLB legislation reauthorized the Elementary and Secondary Education Act (ESEA) in a way that dramatically expanded the historically limited scope and scale of federal involvement in K–12 schooling. In particular, NCLB required states to introduce school accountability systems that applied to *all* public schools and students in the state. These accountability systems had to include annual testing of public school students in reading and mathematics in grades 3 through 8 (and at least once in grades 10 through 12) and ratings of school performance, both overall and for key subgroups, with regard to whether they are making adequate yearly progress (AYP) toward their state's proficiency goals. NCLB required that states introduce sanctions and rewards relevant to every school and based on their AYP status. NCLB mandated 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). However, some states introduced accountability systems that threatened all low-performing schools with explicit sanctions (e.g., reconstitution), regardless of whether they received Title I assistance (Olson, 2004).

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" (pp. 368–369).

However, 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 toward the "bubble kids" who are most likely, with additional attention, to meet the proficiency standard (e.g., Neal & Schanzenbach, 2010). 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.

Studies of the NCLB-like school accountability systems adopted in several states during the 1990s provide evidence on these questions and 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 & Loeb, 2002; Jacob, 2005; Hanushek & 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 eighth grade.[1]

Similarly, Jacob (2005) found that, following the introduction of an accountability policy, math and reading achievement increased in Chicago Public Schools, relative to both the prior trends and 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 teachers responded strategically to accountability pressures (e.g., increasing special education placements).

Hanushek and Raymond (2005) evaluated the impact of within-state variation in school accountability policies on state-level NAEP math and reading achievement growth. Hanushek and Raymond (2005) classified state accountability policies as either "report-card accountability" or "consequential accountability." Report-card states provided a public 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)" (p. 307). They find that the introduction of consequential accountability within a state was associated with statistically significant increases in the gain-score measures, particularly 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.

More recent studies of the achievement effects attributable to NCLB have focused on careful scrutiny of national trends. For example, in a report commissioned by the U.S. Department of Education's Institute of Education Sciences (IES), Stullich et al. (2006, p. v) 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 that reading and math achievement measures based on state assessments have increased in most states since 2002 and 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."

[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.

Other studies have taken a less sanguine view of these achievement gains. For example, Fuller et al. (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" (p. 275). 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.

Other recent studies identify the achievement effects of NCLB by leveraging the variation in sanction risk faced by particular students and schools. For example, Neal and Schanzenbach (2010) 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 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. Similarly, using ten years of student-level longitudinal data from North Carolina, Ladd and Lauen (2010) find that, conditional on student, school, and year fixed effects, school-level accountability pressure leads to relative achievement *gains* for students well below the proficiency threshold.

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 and 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 did not 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 1). Furthermore, this 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.

RESEARCH DESIGN

The national time trends in student achievement are a natural point of departure for considering the impact of NCLB on student achievement. Figure 1 presents national trends on the main NAEP from 1990 to 2007. The solid horizontal line in 2002 visually identifies the point just prior to the implementation of NCLB. The

[2] In fact, like our study, Lee (2006, Table C-7) finds evidence for a positive NCLB effect on math scores among fourth graders. Lee (2006, p. 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.

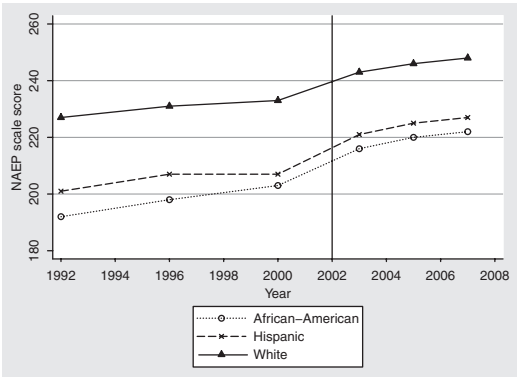
Table 1. States with consequential accountability prior to NCLB.

Implementation		Hanushek and Raymond (2005)	Carnoy and Loeb (2002)	Lee and Wong (2004)
State	Year	Accountability Type (Year)	School Repercussions (1999 to 2000)	Accountability Type (1995 to 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

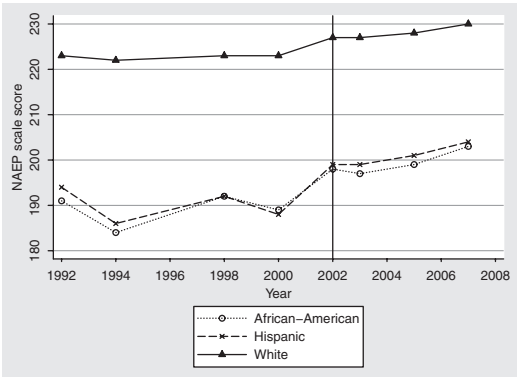
Additional sources: CPRE Assessment and Accountability Profiles, *Education Week* (1999), CCSSO annual surveys, state Department of Education Web sites, and Lexis-Nexis searches of state and local newspaper archives.

trends shown in these figures suggest that NCLB may have had some positive effects on fourth-grade math achievement but provide little suggestion of impacts in the other grade–subject combinations.[3] However, the other nationwide changes in social, economic, and educational factors over this period make it difficult to credibly identify causal inferences from these trends. 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

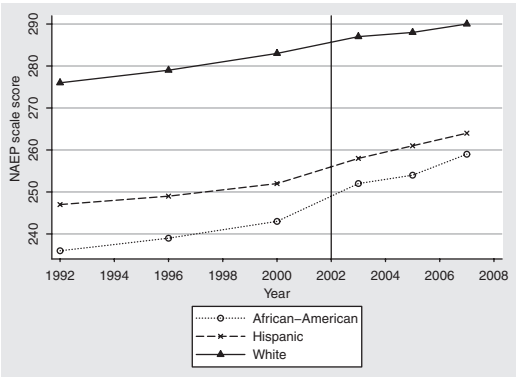
[3] One exception is a noticeable improvement in eighth-grade math scores among African Americans. Data from the NAEP Long-Term Trend Assessment tell a similar story for 9- and 13-year-olds in math and reading.



(a) Grade 4 mathematics



(b) Grade 4 reading



(c) Grade 8 mathematics

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

this period. Similarly, the Reading Excellence Act of 1999 (the precursor to the Reading First program within NCLB) provided hundreds of millions of dollars to states and local education agencies (LEA) to adopt scientifically based instructional practices and professional development activities (Manzo, 1999).

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, e.g., the discussion in Bloom, 1999, and Shadish, Cook, & Campbell, 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).

As discussed in more detail below, there are several 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.

Consequential Accountability Prior to NCLB

The central challenge for any CITS design is to identify plausible comparison groups that were not affected by the intervention under study. In the case of NCLB, this is a seemingly intractable problem. As noted earlier, the policy was signed into law on January 8, 2002, and implemented nationwide during the 2002–2003 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. This study implements a CITS design by relying on the observation that the relevance of these federal school accountability mandates was highly heterogeneous across states. The intuition behind this approach is straightforward. NCLB catalyzed entirely new experiences with consequential school accountability in states that had not already implemented such systems prior to 2002. In contrast, NCLB's requirements were comparatively, if not totally, irrelevant in states that had previously instituted a similar form of school accountability. To the extent that NCLB-like accountability had either positive or negative effects on measured student achievement, we would expect to observe those within-state changes most distinctly in states that had not previously introduced similar policies.[4]

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 and attempted to block 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 & Loeb, 2002; Lee & Wong, 2004; Hanushek & 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.[5] 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 & Duffy, 2001), annual surveys on state assessment programs fielded by the

[4] Another seemingly attractive approach would be to rely on comparisons across public schools and Catholic schools, for which NCLB was largely irrelevant (Jacob, 2008; Wong, Cook, & Steiner 2009). However, in our online Appendix A, we discuss potential internal-validity concerns with this approach. All appendices are available at the end of this article as it appears in JPAM online. See the complete article at wileyonlinelibrary.com.

[5] States that publicize information on school performance without attaching sanctions to that performance are categorized as having report-card accountability (Hanushek & Raymond, 2005).

Council of Chief State School Officers (CCSSO), information from states' 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.[6] 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 to school performance (e.g., ratings, takeover, closure, reconstitution, replacing the principal, and 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. To the extent that NCLB-driven reforms did differ at all from the first generation of state-level accountability policies, our review suggested that they constituted a stronger form of accountability in that they combined reporting and school ratings with at least the possibility of more severe and statutorily explicit sanctions. This possibility suggests that the treatment contrast leveraged in our study would provide a lower bound on the overall impact of NCLB. However, it is also possible that our comparison states weakened their preexisting school accountability systems with the onset of NCLB, thus creating a treatment contrast that instead overstates the effects of NCLB.

We examined this issue, in a more explicitly quantitative manner, by pooling data from several recent studies (Braun & Qian, 2008; National Center for Education Statistics, 2007; Bandeira de Mello, Blankenship, & McLaughlin, 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). We find that states compelled by NCLB to introduce school accountability did lower their proficiency standards somewhat from 2003 to 2007.[7] In contrast, we find that the comparison states that had pre-NCLB accountability policies did *not* lower proficiency standards from 2003 to 2007. This pattern is consistent with the claim that NCLB was effectively irrelevant in the comparison states. And it suggests that our impact estimates are not due to the possible weakening of accountability standards in the comparison states during the NCLB era. However, one limitation to this evidence is that the NAEP equivalence measures are available for only slightly more than half of the states in 2003 and 2005.[8]

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

[6] However, there are also a few notable distinctions between our classification of consequential accountability states (Table 1) and the coding reported by Hanushek and Raymond (2005). These discrepancies are discussed more fully in the online Appendix B. All appendices are available at the end of this article as it appears in JPAM online. See the complete article at wileyonlinelibrary.com.

[7] For example, our point estimates suggest that states without prior accountability lowered proficiency standards in fourth-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.

[8] Only about 20 states have NAEP equivalence measures prior to NCLB because of a combination of reasons, including (1) many states did not administer state-representative NAEP prior to 2003, (2) many states did not report proficiency levels as part of their state testing regime, (3) many states did not test the two grades tested in NAEP (i.e., grades 4 and 8), and (4) 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 fourth and eighth graders, and (b) there are were a handful of states that did not have sufficient NAEP data in certain grade-year-subject cells to justify the equivalence exercise.

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 underscoring exactly how the treatment contrast leveraged in this CITS design should be interpreted. 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, under the assumption that NCLB was effectively irrelevant in consequential accountability states, our estimates will identify the impact of NCLB-induced school accountability provisions specific to those states without prior accountability policies. To the extent that states expecting to gain the most from accountability policies adopted them prior to NCLB, the results we present can be viewed as an underestimate of the average treatment effect of school accountability. Similarly, as noted above, if the comparison states were, to some extent, influenced by NCLB, the implied treatment contrast would be attenuated and our approach would understate the impact of school accountability.

Estimation

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

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

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 1 for observations from the NCLB era. For the majority of our analysis, we assume the NCLB era begins in the academic year 2002–2003, the first year of full implementation after the legislation was signed into law in January 2002. $YR\_SINCE\_NCLB_t$ is defined as $YEAR_t$–2002, so that this variable takes on a value of 1 for the 2002–2003 year, which corresponds to the 2003 NAEP testing. X_{st} represents covariates varying within states over time (e.g., per pupil expenditures and NAEP test exclusion rates). 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 4 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.

To address this concern, our preferred alternative is to 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 1991–1992 academic year and the onset of NCLB. Hence, states with no school accountability at all prior to NCLB would have the highest value for the treatment measure, T_s (i.e., 11). In contrast, Illinois, which implemented its policy in the 1992–1993 school year, would have a value of only 1. Texas would have a value of 3 since its policy started in 1994–1995, and Vermont would have a value of 8 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. In specifications based on this construction of T_s, we define the impact of NCLB as of 2007 and relative to a state that introduced consequential accountability in 1997 (i.e., $\hat{\beta}_5 + 30 \times \hat{\beta}_6$).

Robustness Checks

Arguably, the most 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 the 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 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, 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).

We also explore the robustness of our results through an alternative definition of the treatment intensity imposed by NCLB. More specifically, we note that NCLB may have represented a more substantial treatment in states that had adopted relatively weaker accountability provisions during the 1990s. To address this possibility, we considered 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. Higher values of this measure imply that a state had weaker pre-NCLB accountability standards. If a state did not have any 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. This approach may be underpowered relative to our preferred definition of T_s (i.e., years without prior school accountability) because of the downward bias implied by late-adopting comparison states and because not all states collected proficiency information on students prior to NCLB.[9] Nonetheless, estimates based on this measure provide a useful check on our main results.

Another important robustness check involves how the implementation of NCLB is dated. Our preferred approach is to view NCLB as first in effect during the academic year following its final authorization (i.e., AY 2002–2003). NCLB is often characterized as having been implemented during this year, in part because states were required to use testing outcomes from the prior 2001–2002 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 & Coleman, 2003; Olson, 2002). Furthermore, state data collected during the 2002–2003 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, our test-score results are also consistent with this conventional definition of NCLB's start date in that this implementation year witnessed a trend break unique to states that had no prior experience with consequential school accountability.

However, one could reasonably conjecture that the discussion and anticipation surrounding the adoption of NCLB would have influenced school performance during the 2001–2002 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 & Petrilli, 2006). Alternatively, it could also be argued that NCLB should not be viewed as being in effect until the 2003–2004 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. Regardless, we find broadly similar results when NCLB is considered first in effect during either the 2001–2002 or 2003–2004 school years (see Table C2 in the online Appendix[10]). However, dating the implementation of NCLB one year later does reduce the impact estimate for grade-4 math to a smaller but statistically significant 0.10 standard deviation.

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

[9] 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.

[10] All appendices are available at the end of this article as it appears in JPAM online. See the complete article at wileyonlinelibrary.com.

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 construct validity concerns related to accountability-driven test-score inflation (Jacob, 2005; Fuller et al., 2007; Koretz, 2009). A third benefit of the NAEP is that the available data make it possible to identify changes at different points in the test score distribution as well as effects on specific subject matter competencies.

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: (1) They may encourage schools to test students with special needs who had previously been completely excluded from testing, which may lower scores on average; (2) 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. However, our robustness checks show that our results are not sensitive to using data from the alternative administration or to using data from all administrations.

All states administered NAEP in the spring of 2003, 2005, and 2007. However, 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.[11] Because so few states administered the eighth-grade math exam in 1990, when looking at math we focus on the pre-NCLB NAEP data from the spring of 1992, 1996, and 2000. For fourth-grade reading, we focus on the NAEP data from the spring of 1992, 1994, 1998, and 2002. Because eighth-grade reading was assessed in only two pre-NCLB years, we do not include these limited NAEP data in our primary analysis. However, results based on these data (Table C2 in the online Appendix[12]) suggest the absence of a detectable NCLB effect.

Our final sample includes 39 states (227 state-by-year observations) for fourth-grade math, 38 states (220 state-by-year observations) for eighth-grade math, and 37 states (249 state-by-year observations) for fourth-grade reading. A complete list of states in each NAEP sample can be found in Table C1 in the online Appendix. Because 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. Table 2 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.

[11] 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 robustness checks (see Table C2 in the online Appendix), our results are not sensitive to this sample restriction. All appendices are available at the end of this article as it appears in JPAM online. See the complete article at wileyonlinelibrary.com.

[12] All appendices are available at the end of this article as it appears in JPAM online. See the complete article at wileyonlinelibrary.com.

Table 2. Descriptive statistics, national data and state-based analysis samples (1992 to 2007).

Variable	Nation	State-based analysis samples		
		4th-Grade Math	8th-Grade Math	4th-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%
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, fourth grade	6%			7.06%
NAEP exclusion rate, eighth grade	5%			
Poverty rate	12.10%			12.43%
Pupil teacher ratio	16.20			16.57
Current per pupil expenditures	$8,259			$9,252
Percent free lunch	28.81%			33.41%
Percent of students white	60.30%			55.60%
Percent of students black	17.20%			18.13%
Percent of students Hispanic	17.10%			19.85%
Percent of students Asian	5.60%			4.16%
Number of states		39	38	37
Sample size		227	220	249

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

RESULTS

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 status (Figures 2–4). These figures illustrate the intuition underlying our research design and provide tentative evidence with regard to the achievement effects of NCLB. In each case, we present trends for two groups: (1) states that adopted school accountability between 1994 and 1998 and (2) states that did not adopt school accountability prior to NCLB. The dots reflect the simple mean for each group-by-year cell, and the connecting lines show the predicted trends from the model described above.

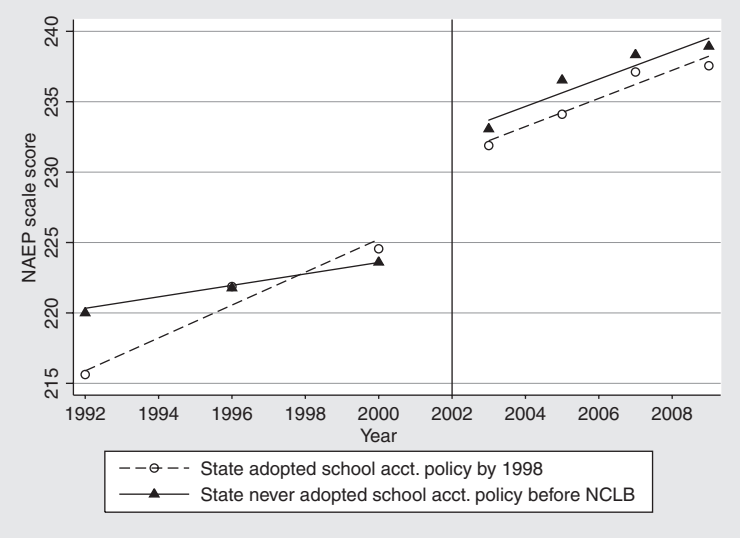


Figure 2. Trends in grade 4 mathematics achievement in the main NAEP by timing of accountability policy.

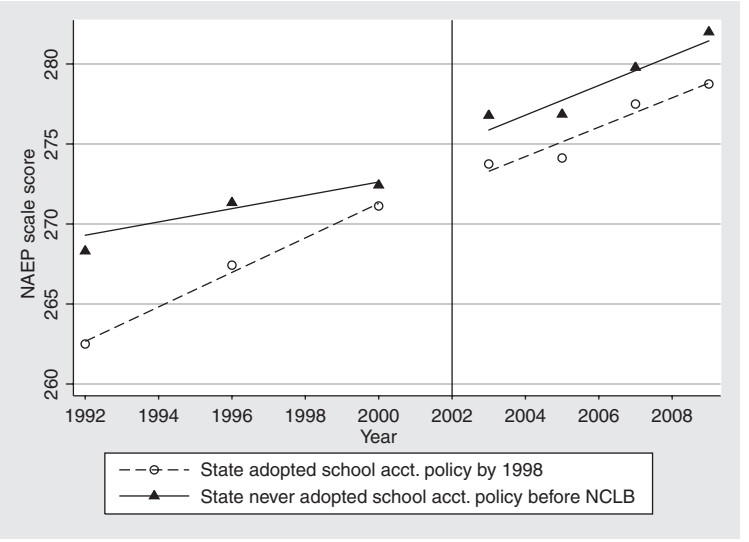


Figure 3. Trends in grade 8 mathematics achievement in the main NAEP by timing of accountability policy.

Consider first Figure 2, which shows trends in fourth-grade math achievement. We see that in 1992 states that never adopted accountability scored roughly 5 scale points (0.18 standard deviation) 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,

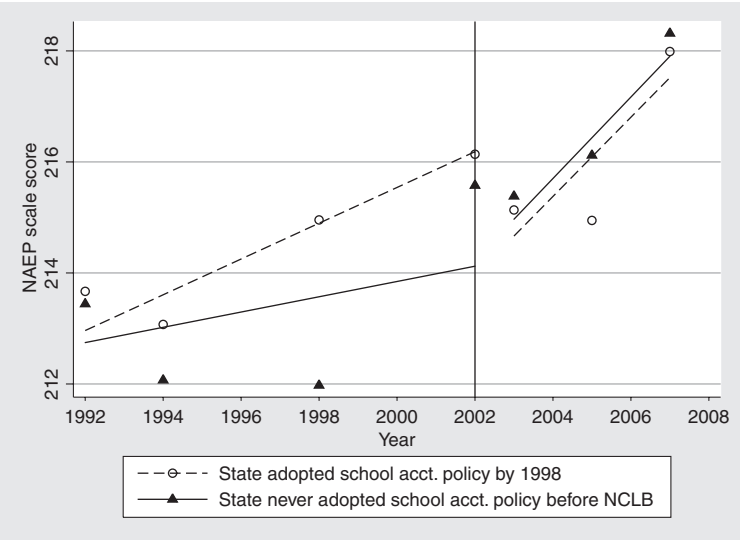


Figure 4. Trends in grade 4 reading achievement in the main NAEP by timing of accountability policy.

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 the 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.

These comparative changes in the achievement levels and trends suggest that NCLB had a positive impact on fourth-grade math achievement. The comparative trends for eighth-grade math (Figure 3) are similar to those for fourth-grade math, though possibly somewhat less clear in showing a positive achievement effect. That is, following the introduction of NCLB, both the level and the trend in the math achievement of eighth graders grew within the no prior accountability states relative to the contemporaneous changes within the states that did have prior experience with school accountability.

The pattern for fourth-grade reading is much less clear (Figure 4). 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 very large gains in 2002.[13] 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).

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-by-year cell.

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

Table 3. The estimated effects of NCLB on mean NAEP scores.

Independent variables	Grade 4 Math (1)	Grade 8 Math (2)	Grade 4 Reading (3)
Panel A: T_s = no prior accountability, excludes 1998 to 2001 adopters			
$NCLB_t \times T_s$	4.438**	2.602*	1.851
	(1.261)	(1.346)	(1.205)
$NCLB_t \times T_s \times$ (years since $NCLB)_t$	0.755*	0.530	−0.086
	(0.405)	(0.359)	(0.330)
Total effect by 2007	8.212**	5.253**	1.420
	(2.318)	(2.457)	(1.531)
Number of states	24	23	21
Sample size	139	132	140
Panel B: T_s = years without prior school accountability, no sample exclusions			
$NCLB_t \times T_s$	0.647**	0.273	0.307**
	(0.212)	(0.194)	(0.148)
$NCLB_t \times T_s \times$ (years since $NCLB)_t$	0.112*	0.069	0.015
	(0.058)	(0.060)	(0.046)
Total effect by 2007 relative to state with school accountability starting in 1997	7.244**	3.704	2.297
	(2.240)	(2.464)	(1.441)
Number of states	39	38	37
Sample size	227	220	249
Mean of Y before NCLB in states without prior accountability	224	272	216
Student-level standard deviation prior to NCLB	31	38	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.

*** $p < 0.01$; ** $p < 0.05$; * $p < 0.1$.

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 and exclude late-adopting comparison states that implemented school accountability between 1999 and 2001. Hence, the results shown in Panel A correspond to the trends shown in Figures 2–4. In this specification, the impact of NCLB on fourth-grade math achievement is roughly 8.2 scale points (0.26 standard deviation) and the effect on eighth-grade math is 5.3 scale points (0.14 standard deviation). The sample restrictions used in these specifications limit the bias that would otherwise exist when the sample includes comparison states that implemented school accountability just prior to NCLB. However, this approach also reduces the precision of our estimates and relies on a somewhat arbitrary decision of which states to exclude.

Panel B presents results in which the treatment variable is instead 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 adopted prior to NCLB). The results suggest moderate positive effects for fourth-grade math (7.2 scale points or 0.23 standard deviation) and smaller effects for eighth-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 estimated effect for fourth-grade reading is smaller in magnitude and statistically indistinguishable from zero.

In an effort to test our identifying assumption, Table 4 examines the association between NCLB and a variety of variables other than student achievement. The specification we use is identical to the one shown in Panel B from Table 3. Column 1 shows the estimated 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.[14] The results of our auxiliary regressions indicate that there are no statistically significant relationships between NCLB and any of these state-year observables.[15]

Column 5 reports the key results of an additional robustness check based on the fraction of students in each state-year observation enrolled in public schools. These measures are based on grade-specific enrollment data from the Common Core and the Private School Universe survey. We find an extremely small, marginally significant association in the fourth-grade math sample suggesting that NCLB reduced the fraction of students attending public school by roughly 1 percent. Columns 6 and 7 show the results for the fraction of public school students 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 association between the identifying variation and the fraction of students eligible for free lunch. Moreover, student-reported parental education data from the NAEP (available only for eighth graders) indicates that NCLB was associated with an *increase* in parental education (p-value = 0.12; results available upon request). Hence, there is some evidence of small changes in student composition associated with NCLB. However, these results do not suggest a large or consistent pattern of selection. Furthermore, the small number of statistically significant effects in Table 4 are consistent with the Type I errors that would be expected when conducting multiple hypothesis tests. Finally, we find that the inclusion of these variables as time-varying covariates in the CITS does not meaningfully change our results.[16]

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 had been the case, the impacts we document could misleadingly reflect lagged policy changes at the state level—namely, students who started attending preschool or full-day (as opposed to half-day) kindergarten in the late 1990s and entered elementary school better prepared, which is reflected on state NAEP tests.

[14] For poverty, median household income, and employment rates, we use the state-by-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.

[15] 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 by 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. In addition, our main results are robust to conditioning on state-year unemployment rates.

[16] Table C2 in the online Appendix presents this and other sensitivity analyses that examine the robustness of our findings. These include conditioning on year fixed effects, controlling for state-specific linear trends, using enrollments as weights, utilizing alternative coding schemes for consequential accountability and treatment intensity, and alternative constructions of the analytical sample based on the available of pre-NCLB NAEP data. All appendices are available at the end of this article as it appears in JPAM online. See the complete article at wileyonlinelibrary.com.

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	3.4	11.6	38796	51.4	91.4	13.1	6.1	26.6	45.0	42.3
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	3.2	11.9	38631	51.2	91.9	12.6	5.4	27.1	33.2	34.4
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	6.2	11.8	41229	49.7	90.6	15.7	6.9	29.5	48.9	45.5

Notes: Each cell is a separate regression as in Panel B 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 fourth-grade students. Standard errors are robust to clustering at the state level.

*** $p < 0.01$; ** $p < 0.05$; * $p < 0.1$.

Table 5 shows the estimated effects of NCLB defined for different points of the achievement distribution as well as by grade and subject. As many have noted, the design of NCLB necessarily focused the attention of schools on helping students attain state-specific proficiency standards. These state standards are generally set at lower levels than the proficiency standard defined for the NAEP. In particular, only one state (Massachusetts) had a mathematics standard for fourth graders that exceed the proficiency standard in NAEP (Bandeira de Mello, Blankenship, & McLaughlin, 2009, Figure 3). Hence, we might expect NCLB to disproportionately influence achievement in the left tail of the NAEP distribution. We find results roughly consistent with this. More specifically, the results in Table 5 show that NCLB had particularly large effects on grades 4 and 8 math achievement at the 10th percentile and on the proportion of these students meeting NAEP's basic math standard. Interestingly, state standards for fourth-grade achievement in math are generally well above the 10th percentile of NAEP achievement.[17] So the positive effect of NCLB at this point of the achievement distribution suggests the absence of broad triage effects in the left tail of the achievement distribution. Similarly, in contrast with some prior research and 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 by a surprisingly large amount. For example, in fourth-grade math, the impacts at the 90th percentile were only 4 scale points lower than at the 10th percentile. Similarly, Table 5 indicates that NCLB increased the share of students meeting NAEP's proficiency standard for fourth-grade math, even though this standard exceeds virtually all state standards.

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 evaluate the achievement effects of NCLB among specific student subgroups. In Table 6, we present results separately by race, gender, and poverty subgroups. Several interesting findings emerge.

In the fourth-grade math sample, the impact of NCLB is generally 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 grade 4 math gains attributable to NCLB were substantially larger among students who were eligible for subsidized lunch (regardless of race) relative to students who were not eligible. However, the NCLB effects were roughly comparable for boys and girls.

In eighth-grade math, we find extremely large positive effects for Hispanic students and small, only marginally significant positive effects 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 eighth-grade girls, with their male peers experiencing little, if any, achievement benefit from NCLB.

The results for the fourth-grade reading shown in Table 6 suggest some moderate positive effects for white students and for male students. However, as noted earlier,

[17] The pretreatment, comparison-group mean of fourth-grade NAEP achievement in math at the 10th percentile is 186 scale points (Table 5). However, no state has a corresponding standard for fourth-grade achievement in math that is below roughly 200 NAEP-equivalent scale points (Bandeira de Mello, Blankenship, & McLaughlin, 2009, Table 2).

Table 5. The estimated effects of NCLB on achievement distributions by grade and subject.

Grade−Subject Sample	Mean (1)	Percent Basic (2)	Percent Proficient (3)	10th Percentile (4)	90th Percentile (5)
4th-grade math (39 states, n = 227)					
Total effect by 2007	7.244**	10.090**	5.590**	9.046**	5.205**
	(2.240)	(3.145)	(1.891)	(3.767)	(1.916)
Mean of Y before NCLB in states without prior accountability	224	64	21	186	259
8th-grade math (38 states, n = 220)					
Total effect by 2007	3.704	5.888**	1.286	5.598*	2.537
	(2.464)	(2.680)	(2.055)	(3.236)	(2.404)
Mean of Y before NCLB in states without prior accountability	272	64	24	228	314
4th-grade reading (37 states, n = 249)					
Total effect by 2007	2.297	2.359	2.542**	3.611	2.097**
	(1.441)	(1.592)	(1.035)	(2.804)	(0.805)
Mean of Y before NCLB in states without prior accountability	216	61	29	171	258

Notes: Each cell is a separate regression as in Panel B 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 6. The estimated effects of NCLB on NAEP scores by grade, subject, and subgroup.

	Grade 4 Math		Grade 8 Math		Grade 4 Reading	
Subgroup	OLS (1)	WLS (2)	OLS (3)	WLS (4)	OLS (5)	WLS (6)
White						
Total effect by 2007	5.817**	4.855**	2.863	1.828	4.854**	5.362**
	(1.679)	(2.047)	(2.561)	(3.680)	(1.231)	(1.201)
Baseline mean of Y	232	233	281	282	226	225
Sample	39 states, n = 227		37 states, n = 214		37 states, n = 249	
Black						
Total effect by 2007	4.931	14.573**	9.261	8.826	−1.873	−0.871
	(5.342)	(3.731)	(6.774)	(8.999)	(3.698)	(2.569)
Baseline mean of Y	203	202	241	242	200	195
Sample	30 states, n = 176		27 states, n = 158		32 states, n = 214	
Hispanic						
Total effect by 2007	11.429**	9.793**	20.031**	8.219**	6.094	0.242
	(4.242)	(1.411)	(5.766)	(4.135)	(4.835)	(4.805)
Baseline mean of Y	204	204	246	247	199	193
Sample	19 states, n = 108		16 states, n = 90		22 states, n = 140	
Male						
Total effect by 2007	7.408**	7.612**	1.678	−1.702	3.399**	2.241*
	(2.368)	(3.545)	(2.488)	(4.024)	(1.578)	(1.287)
Baseline mean of Y	224	227	273	276	212	214
Sample	39 states, n = 227		38 states, n = 220		37 states, n = 249	

(*Continued*)

Table 6. (*Continued*)

Subgroup	Grade 4 Math		Grade 8 Math		Grade 4 Reading	
	OLS (1)	WLS (2)	OLS (3)	WLS (4)	OLS (5)	WLS (6)
Female						
Total effect by 2007	7.365**	7.426**	6.300**	6.436	1.395	0.741
	(2.258)	(2.480)	(2.664)	(4.459)	(1.535)	(1.697)
Baseline mean of Y	223	225	272	274	220	222
Sample	39 states, $n = 227$		38 states, $n = 220$		37 states, $n = 249$	
Free lunch–eligible						
Total effect by 2007	5.487*	8.011**	10.702*	15.761**	0.567	2.482
	(3.294)	(2.631)	(6.155)	(5.631)	(4.235)	(4.296)
Baseline mean of Y	212	212	257	256	205	206
Sample	36 states, $n = 180$		34 states, $n = 170$		37 states, $n = 185$	
Not free lunch–eligible						
Total effect by 2007	3.027	1.385	2.199	0.992	1.355	−4.790
	(2.568)	(2.508)	(3.924)	(4.171)	(3.042)	(5.073)
Baseline mean of Y	232	234	279	281	225	227
Sample	36 states, $n = 180$		34 states, $n = 170$		37 states, $n = 185$	

Notes: Each cell is a separate regression as in Panel B 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. The baseline mean of Y is the mean test score before NCLB in states without prior accountability. Standard errors are clustered at the state level.

*** $p < 0.01$; ** $p < 0.05$; * $p < 0.1$.

the trends prior to NCLB were distinctly nonlinear (Figure 4), raising doubts about the validity of the CITS approach for this particular outcome. To provide evidence on the robustness of these reading results, we reestimated the specifications in this table limiting the sample to the years 1998 through 2007 (see Table C3 in the online Appendix[18]). In these models, the NCLB impact is identified by the comparative 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 6 and not statistically different from 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 6 are statistically distinguishable from zero when using the restricted sample. In light of these issues, we view the impact of NCLB on reading achievement as uncertain.

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 science or social studies (Jacob, 2005; Winters, Trivitt, & Greene, in press).

The NAEP data offer some opportunity to test this hypothesis in the context of NCLB. A sizable number of states administered state-representative NAEP science tests to eighth graders in 1996, 2000, and 2005 ($n = 31$) and to fourth graders in 2000 and 2005 ($n = 36$). Using these 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 eighth-grade sample, we use the 1996 and 2000 data to estimate a prior intercept and trend. In the fourth-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 Table C4 in the online Appendix). Our standard errors are relatively precise, allowing us to rule out effects larger than roughly 3 to 4 scale points (about 0.1 standard deviation). Similarly, we find no significant effects when looking separately by subgroup (see the online Appendix Tables C5 and C6[19]). Together, these results suggest that NCLB did not have an adverse impact on student performance in science as measured by the NAEP.[20]

Another major concern with test-based accountability, including NCLB, is that it provides teachers an incentive to divert energy toward 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

[18] All appendices are available at the end of this article as it appears in JPAM online. See the complete article at wileyonlinelibrary.com.

[19] All appendices are available at the end of this article as it appears in JPAM online. See the complete article at wileyonlinelibrary.com.

[20] 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.

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). Another way to examine this issue is consider whether NCLB has improved student achievement in any particular topic within math or reading. To explore this, we reestimate 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 the online Appendix[21] (Table C7) suggest that NCLB had a positive impact in all math topic areas for the fourth-grade sample. The point estimates are somewhat larger in algebra (0.26 standard deviation), number properties (0.26 standard deviation), and data analysis (0.22 standard deviation) than in geometry (0.17 standard deviation) and measurement (0.16 standard deviation). In the eighth-grade sample, NCLB had a moderately large and statistically significant impact within data analysis (6.7 scale points, or 0.16 standard deviation) 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 fourth-grade NAEP reading exam measures student competency in two skills related to comprehension: reading for information (i.e., primarily nonfiction reading) and reading for literary experience (i.e., primarily fiction reading). We do not find robust or statistically significant evidence that NCLB influenced either of these reading competencies (Table C8 in the online Appendix).

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 NAEP data and a research design that leverages the fact NCLB catalyzed entirely new experiences with school accountability in some states while simply reconstituting preexisting school accountability systems in others. Our results suggest that the achievement consequences of NCLB are decidedly mixed. Specifically, we find that the new school accountability systems brought about by NCLB generated large and broad gains in the math achievement of fourth graders and, to a somewhat lesser extent, eighth graders. However, we find no consistent evidence that NCLB influenced the reading achievement of fourth graders.

The evidence of substantial and almost universal gains in fourth-grade math achievement is undoubtedly good news for advocates of NCLB and school accountability. On the other hand, critics of NCLB can point to the lack of similarly robust effects on reading and the reform's limited contributions to reducing achievement gaps. Similarly, NCLB's contributions to math achievement appear more modest when benchmarked to the legislation's ambitious requirement of universal proficiency by 2014. For example, NCLB increased grade 4 math proficiency by nearly 27 percent (Table 5, column 3). Nonetheless, more than 60 percent of fourth graders still fail to meet the math proficiency standard defined by NAEP.

[21] All appendices are available at the end of this article as it appears in JPAM online. See the complete article at wileyonlinelibrary.com.

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 research-informed programs situated outside of schools (e.g., early childhood and health interventions) should complement school accountability policies. 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. An effective redesign of accountability policies like NCLB may also need to pay more specific attention to how accountability interacts with specific policies and practices in schools and districts (Ladd, 2007). For example, recent research based on the CITS research design used in this study (Dee, Jacob, & Schwartz, 2011) provides evidence that NCLB compelled meaningful increases in available resources (e.g., increased per pupil spending funded largely by state and local rather than federal sources). The evidence from this research design also suggests that NCLB increased teacher compensation and the share of teachers with graduate degrees, which implies that new resources linked to teacher quality are one possible mediator of NCLB's effects. Further research that can credibly and specifically explicate how educational practices (e.g., spending, teacher quality, and classroom practice) have contributed to the achievement effects documented here would be a useful next step in guiding both sensible revisions to NCLB and effective strategies for improving student outcomes at scale more generally.

The Obama administration recently advanced this discussion by releasing a "blueprint" that outlined proposed features of a reauthorization of NCLB (Klein & McNeil, 2010). This proposal calls for the continued reporting of school-level, test-based student performance, suggests that states be given increased flexibility in how they judge school effectiveness (e.g., using achievement growth rather than achievement levels), and encourages the reporting of non-test outcomes (e.g., high school graduation and college enrollment rates). Based on our understanding of the extant literature, we see little reason to doubt that modifications of this sort can sustain or enhance the effectiveness of the legislation. However, this blueprint also calls for limiting explicit and mandatory school-level consequences to "Challenge" schools (i.e., the very lowest-performing schools and those with large and persistent achievement gaps) while allowing "local flexibility to determine the appropriate improvement and support strategies for most schools" (U.S. Department of Education, 2010, p. 8). This recommendation appears to imply that, for most schools, poor performance may no longer lead to meaningful consequences. The evidence from pre-NCLB state reforms suggests that such report-card accountability (i.e., performance reporting without sanctions) is ineffective (Hanushek & Raymond, 2005). To the extent that the reauthorization of NCLB reduces consequential school accountability, the targeted achievement gains we document in this study may be at risk.

THOMAS S. DEE is Professor of Public Policy and Economics and Research Professor of Education in the Department of Economics, Frank Batten School of Leadership and Public Policy, University of Virginia, Charlottesville, VA.

BRIAN A. JACOB is the Walter H. Annenberg Professor of Education Policy, Professor of Economics, and Director of the Center on Local, State and Urban Policy (CLOSUP) at the Gerald R. Ford School of Public Policy, University of Michigan, 735 South State Street, Ann Arbor, MI 48109.

ACKNOWLEDGMENTS

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, the NCLB: Emerging Findings Research Conference, the University of Wisconsin, the University of Virginia,

and the World Bank 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.

REFERENCES

Angrist, J. D., & Pischke, J. (2009). Mostly harmless econometrics: An empiricist's companion. Princeton, NJ: Princeton University Press.

Associated Press. (2004, September 21). Diocese of Tucson becomes 2nd to file for bankruptcy.

Ballou, D., & Springer, M. G. (2008). Achievement trade-offs and No Child Left Behind. Unpublished manuscript.

Bandeira de Mello, V., Blankenship, C., & McLaughlin, D. H. (2009). Mapping state proficiency standards onto NAEP scales: 2005–2007. NCES 2010-456. Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education.

Bertrand, M., Duflo, E., & Mullainathan, S. (2004). How much should we trust differences-in-differences estimates? Quarterly Journal of Economics, 119, 249–275.

Bloom, H. S. (1999). Estimating program impacts on student achievement using short interrupted time series, New York: Manpower Demonstration Research Corporation.

Bloom, H. S., Ham, S., Melton, L., & O'Brien, J. (2001). Evaluating the accelerated schools approach: A look at early implementation and impacts on student achievement in eight elementary schools. New York: Manpower Demonstration Research Corporation.

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

Bryk, A., Lee. V. E., & Holland, P. B. (1993). Catholic schools and the common good. Cambridge, MA: Harvard University Press.

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

Carroll, M., Pfeiffer, S., Rezendes, M., & Robinson, W. V. (2002, January 6). Church allowed abuse by priest for years, aware of Geoghan record, archdiocese still shuttled him from parish to parish. Boston Globe, p. A1.

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

Center on Education Policy. (2008). Has student achievement increased since 2002: State test score trends through 2006–2007. Washington, DC: Author.

Dee, T. S., Jacob, B. A., & Schwartz, N. L. (2011). The effects of NCLB on school resources and practices. Unpublished manuscript.

Dobbs, M. (2005, May 8). Conn. stands in defiance on enforcing "No Child." Washington Post. Retrieved April 1, 2011, from http://www.washingtonpost.com/wp-dyn/content/article/2005/05/07/AR2005050700973.html.

Education Week. (1999). Accountability in context. Retrieved April 4, 2011, from http://www.edcounts.org/archive/sreports/qc99/ac/tables/ac-tnotes.htm.

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: Routledge.

Fuller, B., Wright, J., Gesicki, K., & Kang, E. (2007). Gauging growth: How to judge No Child Left Behind? Educational Researcher, 36, 268–278.

Goertz, M. E., & Duffy, M. E. (2001). Assessment and accountability systems in the 50 states: 1999–2000. CPRE Research Report RR-046. Philadelphia: Consortium for Policy Research in Education.

Hanushek, E. A., & Raymond, M. E. (2001). The confusing world of educational accountability. National Tax Journal, 54, 365–384.

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

Hess, F. M., & Petrilli, M. J. (2006). No Child Left Behind primer. New York: Peter Lang Publishing.

Hess, F. M., & Petrilli, M. J. (2009). Wrong turn on school reform. Policy Review, February/March, pp. 55–68.

Jacob, B. A. (2005). Accountability, incentives and behavior: Evidence from school reform in Chicago. Journal of Public Economics, 89, 761–796.

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

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

Klein, A., & McNeil, M. (2010, March 17). Administration unveils ESEA reauthorization blueprint. Education Week, 29, 19.

Koretz, D. (2008). Measuring up: What educational testing really tells us. Cambridge, MA: Harvard University Press.

Krieg, J. M. (2008). Are students left behind? The distributional effects of the No Child Left Behind Act. Education Finance and Policy, 3, 250–281.

Ladd, H. F. (2007, November). Holding schools accountable revisited. 2007 Spencer Foundation Lecture in Education Policy and Management, Association for Public Policy Analysis and Management Retrieved November 8, 2009, from https://www.appam.org/awards/pdf/2007Spencer-Ladd.pdf.

Ladd, H. F., & Lauen, D. L. (2010). Status versus growth: The distributional effects of school accountability policies. Journal of Policy Analysis and Management, 29, 426–450.

Lee, J. (2006). 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. Boston: Civil Rights Project, Harvard University.

Lee, J., & Wong, K. K. (2004). The impact of accountability on racial and socioeconomic equity: Considering both school resources and achievement outcomes. American Educational Research Journal 41, 797–832.

Manzo, K. K. (1999). Reading-achievement program is off to a quiet start. Education Week, 18, January 13, 21–25.

National Center for Education Statistics. (2007). Mapping 2005 state proficiency standards onto the NAEP Scales. NCES 2007-482. Washington, DC: U.S. Government Printing Office.

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

Nichols, S. L., & Berliner, D. C. (2007). Collateral damage: How high-stakes testing corrupts America's schools. Cambridge, MA: Harvard Education Press.

Olson, L. (2002, December 1). States strive toward ESEA compliance. Education Week, 22, p. 1, 18–19.

Olson, L. (2003, December 10). In ESEA wake, school data flowing forth. Education Week, 23, pp. 1, 16–18.

Olson, L. (2004, December 8). Taking root. Education Week, 24, pp. S1, S3, S7.

Palmer, S. R., & Coleman, A. L. (2003). The No Child Left Behind Act: Summary of NCLB requirements and deadlines for state action. Council of Chief State School Officers. Retrieved November 13, 2009, from http://www.ccsso.org/content/pdfs/Deadlines.pdf.

Ravitch, D. (2009, June 10). Time to kill "No Child Left Behind." Education Week, 28, 30–36.

Rothstein, R., Jacobsen, R., & Wilder, T. (2008). Grading education: Getting accountability right. New York: Teachers College Press.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). Experimental and quasi-experimental designs for generalized causal inference. Boston: Houghton-Mifflin.

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

Stullich, S., Eisner, E., McCrary, J., & Roney, C. (2006). National assessment of Title I interim report to Congress, Vol. 1: Implementation of Title I. Washington, DC: U.S. Department of Education, Institute of Education Sciences.

U.S. Department of Education. (2006). No Child Left Behind is working. Retrieved July 29, 2009, from http://www.ed.gov/nclb/overview/importance/nclbworking.html.

U.S. Department of Education. (2007). Private school participants in federal programs under the No Child Left Behind Act and the Individuals with Disabilities Education Act. Washington, DC: Author.

U.S. Department of Education. (2010). A blueprint for reform: The reauthorization of the Elementary and Secondary Education Act. Washington, DC: Author.

Winters, M. A., Trivitt, J. R., & Greene, J. P. (in press). The impact of high-stakes testing on student proficiency in low-stakes subjects: Evidence from Florida's Elementary Science Exam. Economics of Education Review.

Wong, M., Cook, T. D., & Steiner, P. M. (2009). No Child Left Behind: An interim evaluation of its effects on learning using two interrupted time series each with its own non-equivalent comparison series. Working Paper Series, WP-09-11. Evanston, IL: Institute for Policy Research, Northwestern University.

APPENDIX A

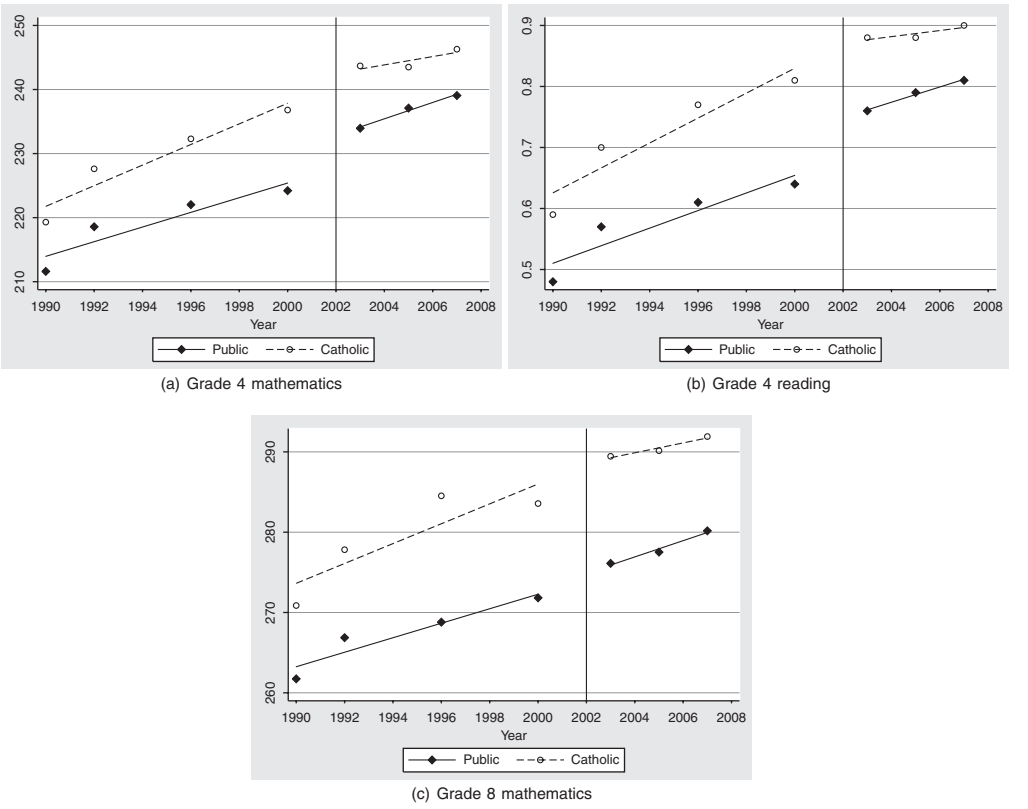
Catholic Schools as an NCLB Comparison Group

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.

The three panels in Figures A1 show the comparative achievement trends for public and Catholic school students from the national NAEP. In these figures we see that students in Catholic schools outperformed their counterparts in public schools over the entire period, 1990 to 2007. While both groups showed increasing achievement during the pre-NCLB period, public school students (particularly in fourth 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 fourth-grade math and a potential (and smaller) effect for eighth-grade math.

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 2002–2003 school year 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 A2, 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 A3 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 declines in prior

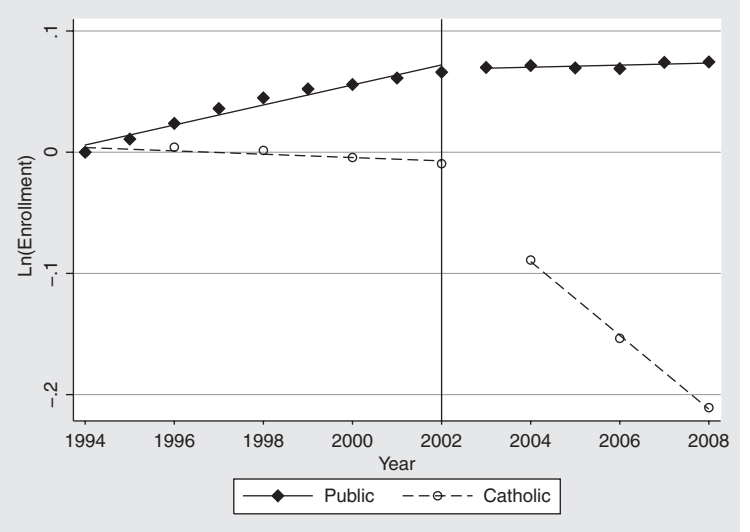


(a) Grade 4 mathematics

(b) Grade 4 reading

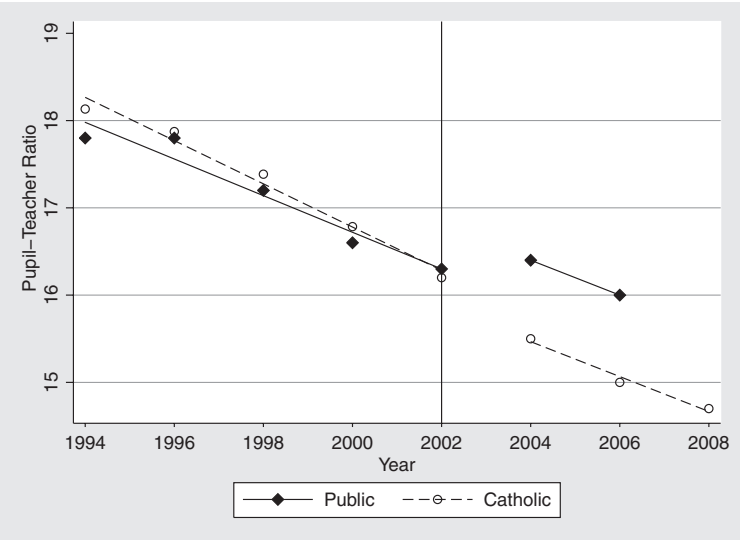
(c) Grade 8 mathematics

Figure A1. Mean scaled score on the main NAEP in public versus Catholic schools.



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

Figure A2. Student enrollment trends in public versus Catholic schools.



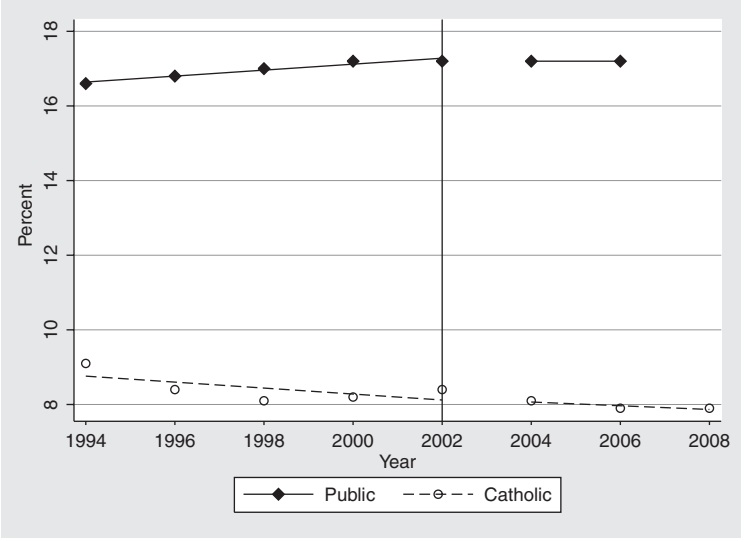
(a) Pupil–Teacher ratio

Figure A3. Pupil–teacher ratio trends in public versus Catholic schools.

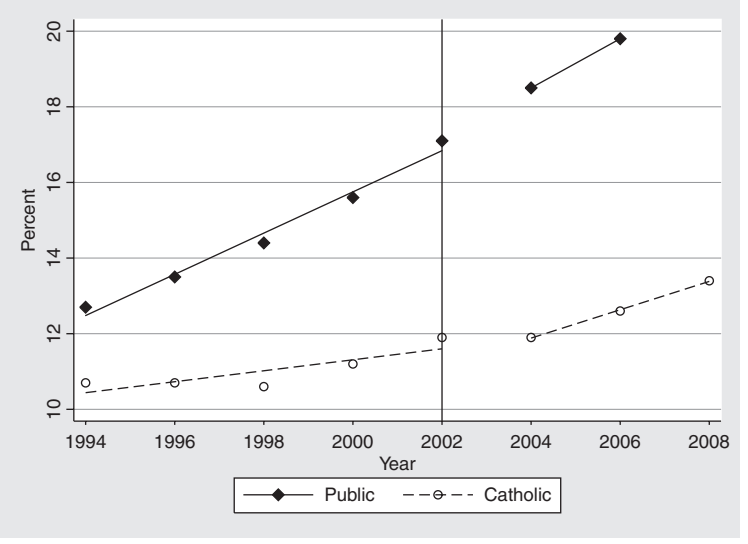
years) after the implementation of NCLB, while ratios in Catholic schools dropped relative to prior trends.

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 nonrandom 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 nonrandom attrition, we collected data on the comparative trends in student and parent observables across public and Catholic schools. Figure A4 shows the comparative trends in the percent of public and Catholic school students who are black and Hispanic. Using data available from NAEP surveys, Figure A5 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 nonrandom 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 nonrandom 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, & 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



(a) Percent Black



(b) Percent Hispanic

Figure A4. Student composition trends in public versus Catholic schools.

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.

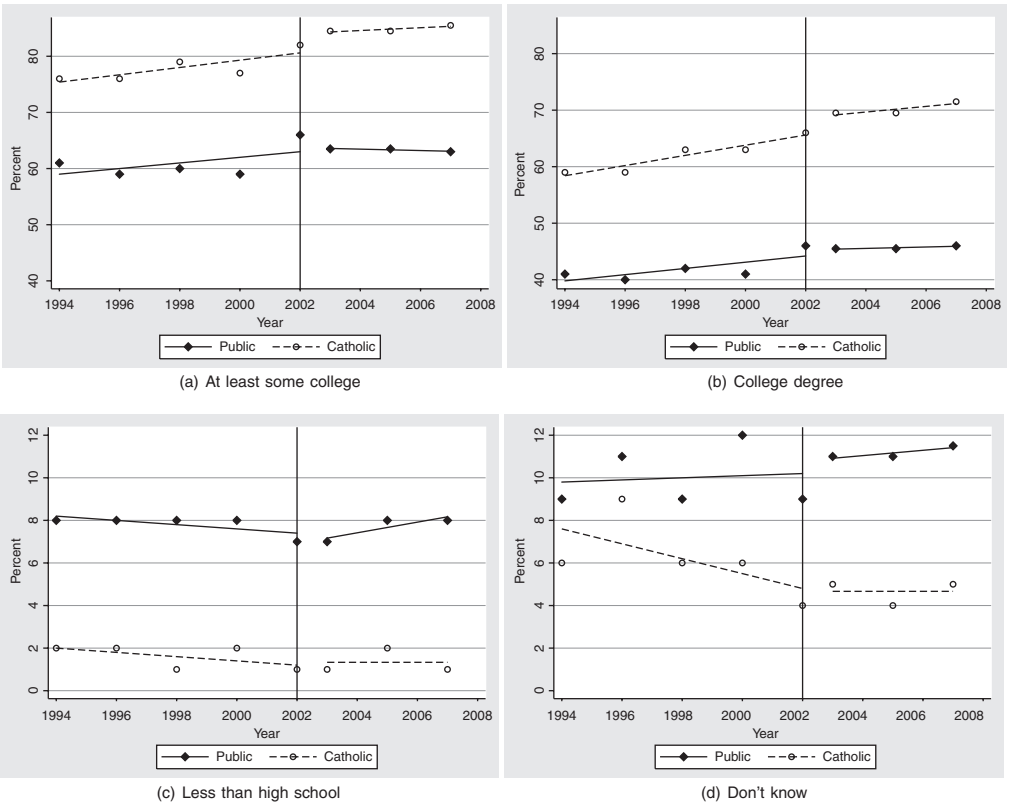


Figure A5. Parental education trends in public versus Catholic schools.

APPENDIX B

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 (Illinois and Alaska) that implemented consequential accountability in advance of NCLB (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 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."[22] Similarly, Virginia began identifying low-performance schools through an accreditation system that became effective during the 1998–1999 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 applied vague state statutes to suggest 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.

[22] http://www.edcounts.org/archive/sreports/qc99/ac/tables/ac-tnotes.htm.

APPENDIX C

Additional NAEP Results

Table C1. States included in NAEP analysis samples.

	Subject–Grade		
State	Grade 4 Math	Grade 8 Math	Grade 4 Reading
Alabama	1	1	1
Alaska	0	0	0
Arizona	1	1	1
Arkansas	1	1	1
California	1	1	1
Colorado	0	0	0
Connecticut	1	1	1
Delaware	0	0	1
District of Columbia	1	1	1
Florida	0	0	1
Georgia	1	1	1
Hawaii	1	1	1
Idaho	1	1	0
Illinois	0	1	0
Indiana	1	1	0
Iowa	1	0	1
Kansas	0	0	1
Kentucky	1	1	1
Louisiana	1	1	1
Maine	1	1	1
Maryland	1	1	1
Massachusetts	1	1	1
Michigan	1	1	1
Minnesota	1	1	1
Mississippi	1	1	1
Missouri	1	1	1
Montana	1	1	1
Nebraska	1	1	0
Nevada	1	0	1
New Hampshire	0	0	0
New Jersey	0	0	0
New Mexico	1	1	1

(*Continued*)

Table C1. (*Continued*)

State	Subject–Grade		
	Grade 4 Math	Grade 8 Math	Grade 4 Reading
New York	1	1	1
North Carolina	1	1	1
North Dakota	1	1	0
Ohio	1	1	0
Oklahoma	1	1	1
Oregon	1	1	1
Pennsylvania	0	0	0
Rhode Island	1	1	1
South Carolina	1	1	1
South Dakota	0	0	0
Tennessee	1	1	1
Texas	1	1	1
Utah	1	1	1
Vermont	1	1	0
Virginia	1	1	1
Washington	0	0	1
West Virginia	1	1	1
Wisconsin	0	0	0
Wyoming	1	1	1
Total	39	38	37

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 C2. 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 A 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)

Grade–Subject Sample	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**	7.272**	2.966**	7.227**
	(2.720)	(2.061)	(2.445)	(2.598)	(2.223)	(1.407)	(2.353)
8th-grade Math							
Total effect by 2007	4.930	3.741*	2.714	5.334**	4.082*	1.563	4.206*
	(3.351)	(2.093)	(2.765)	(2.621)	(2.360)	(1.442)	(2.285)
4th-grade reading							
Total effect by 2007	3.657	2.006*	2.158	2.490*	5.901*	1.040	n/a
	(2.337)	(1.139)	(1.439)	(1.467)	(3.216)	(1.137)	
8th-grade reading							
Total effect by 2007	0.734	−1.104	n/a	−1.430	n/a	−1.184	n/a
	(3.356)	(1.816)		(2.078)		(1.814)	

Notes: Each cell is a separate regression as in Panel B 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 eighth-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 C3. The estimated effects of NCLB on 4th-grade NAEP reading scores by subgroup using only 1998 to 2007 data.

	OLS				WLS			
Subgroup	Mean (1)	% Basic (2)	10th Percentile (3)	90th Percentile (4)	Mean (5)	% Basic (6)	10th Percentile (7)	90th Percentile (8)
White (37 states, n = 185)								
Total effect by 2007	2.202	1.636	1.359	2.239	0.174	0.858	0.358	1.431
	(2.342)	(2.387)	(4.207)	(1.975)	(2.385)	(2.245)	(3.516)	(1.969)
Mean of Y before NCLB in states without prior accountability	226	73	184	265	225	72	183	264
Black (32 states, n = 160)								
Total effect by 2007	−5.069	−2.640	−6.297	−2.589	−0.092	1.381	1.673	1.211
	(5.083)	(3.733)	(9.285)	(3.939)	(4.707)	(3.744)	(7.649)	(3.460)
Mean of Y before NCLB in states without prior accountability	200	43	154	244	195	36	151	238
Hispanic (22 states, n = 114)								
Total effect by 2007	0.991	1.555	4.845	0.984	−1.013	0.274	−0.501	3.875
	(5.744)	(5.797)	(10.676)	(6.468)	(5.888)	(5.179)	(9.422)	(2.522)
Mean of Y before NCLB in states without prior accountability	199	43	154	244	193	37	144	241
Male (37 states, n = 185)								
Total effect by 2007	−0.073	0.536	−1.386	1.782	0.787	0.526	0.359	2.779
	(3.180)	(3.108)	(5.904)	(2.567)	(2.511)	(2.617)	(4.205)	(2.280)
Mean of Y before NCLB in states without prior accountability	212	58	166	254	214	60	167	256
Female (37 states, n = 185)								
Total effect by 2007	−0.775	−1.147	−2.103	0.986	−4.289	−5.091	−3.319	−2.798
	(2.809)	(3.110)	(5.461)	(2.121)	(3.899)	(4.000)	(4.672)	(3.183)
Mean of Y before NCLB in states without prior accountability	220	65	176	561	222	68	177	263

(*Continued*)

Table C3. (*Continued*)

Subgroup	OLS				WLS			
	Mean (1)	% Basic (2)	10th Percentile (3)	90th Percentile (4)	Mean (5)	% Basic (6)	10th Percentile (7)	90th Percentile (8)
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 B 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 C4. The estimated effects of NCLB on NAEP science scores.

	Grade 4 Science				Grade 8 Science			
Independent Variables	Mean (1)	% Basic (2)	10th Percentile (3)	90th Percentile (4)	Mean (5)	% Basic (6)	10th Percentile (7)	90th Percentile (8)
Panel A: T_s = no prior accountability, excludes 1998 to 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 B: 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 C5. The estimated effects of NCLB on NAEP 4th-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 (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

(*Continued*)

Table C5. (*Continued*)

Subgroup	OLS				WLS			
	Mean (1)	% Basic (2)	10th Percentile (3)	90th Percentile (4)	Mean (5)	% Basic (6)	10th Percentile (7)	90th Percentile (8)
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 B 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 C6. 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

(*Continued*)

Table C6. (*Continued*)

Subgroup	OLS				WLS			
	Mean (1)	% Basic (2)	10th Percentile (3)	90th Percentile (4)	Mean (5)	% Basic (6)	10th Percentile (7)	90th Percentile (8)
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 B 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 C7. The estimated effects of NCLB on math achievement by subscale.

	Fourth Grade			Eighth Grade		
Subscale category	Mean (1)	10th Percentile (2)	90th Percentile (3)	Mean (4)	10th Percentile (5)	90th Percentile (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.

*** $p < 0.01$; ** $p < 0.05$; * $p < 0.1$.

Table C8. The estimated effects of NCLB on reading achievement by subscale.

	Fourth Grade			Eighth Grade		
Subscale category	Mean (1)	10th Percentile (2)	90th Percentile (3)	Mean (4)	10th Percentile (5)	90th Percentile (6)
Gain information						
Total effect by 2007	2.724*	4.239	2.105*	−0.977	−6.061*	2.495
	(1.487)	(3.005)	(1.079)	(2.141)	(3.689)	(3.732)
Mean of Y before NCLB in states without prior accountability	211	163	257	260	216	302
Student-level standard deviation prior to NCLB	39			36		
Literary experience						
Total effect by 2007	2.290	3.544	1.284	−1.988	−8.011**	3.380
	(1.493)	(2.894)	(0.881)	(2.486)	(3.970)	(3.328)
Mean of Y before NCLB in states without prior accountability	220	174	262	262	218	303
Student-level standard deviation prior to NCLB	36			35		
Perform a task						
Total effect by 2007				−3.619	−12.341*	4.657
				(2.923)	(6.715)	(4.041)
Mean of Y before NCLB in states without prior accountability				260	211	306
Student-level standard deviation prior to NCLB				42		

Notes: Each cell is a separate regression as in Panel B 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.

*** $p < 0.01$; ** $p < 0.05$; * $p < 0.1$.