THOMAS S. DEE
University of Virginia

BRIAN A. JACOB
University of Michigan

The Impact of No Child Left Behind on Students, Teachers, and Schools

ABSTRACT The controversial No Child Left Behind Act (NCLB) brought test-based school accountability to scale across the United States. This study draws together results from multiple data sources to identify how the new accountability systems developed in response to NCLB have influenced student achievement, school-district finances, and measures of school and teacher practices. Our results indicate that NCLB brought about targeted gains in the mathematics achievement of younger students, particularly those from disadvantaged backgrounds. However, we find no evidence that NCLB improved student achievement in reading. School-district expenditure increased significantly in response to NCLB, and these increases were not matched by federal revenue. Our results suggest that NCLB led to increases in teacher compensation and the share of teachers with graduate degrees. We find evidence that NCLB shifted the allocation of instructional time toward math and reading, the subjects targeted by the new accountability systems.

The No Child Left Behind (NCLB) Act of 2001 is arguably the most far-reaching education policy initiative in the United States over the last four decades. The hallmark features of this legislation compelled states to conduct annual student assessments linked to state standards, to identify schools that are failing to make "adequate yearly progress" (AYP), 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 test performance and linking that performance to the possibility of meaningful sanctions can improve the focus and productivity of public schools.

NCLB has been extremely controversial from its inception. Critics charge that NCLB has led educators to shift resources away from important but nontested subjects, such as social studies, art, and music, and to focus instruction within mathematics and reading on the relatively narrow set of topics that are most heavily represented on the high-stakes tests (Rothstein, Jacobsen, and Wilder 2008, Koretz 2008). In the extreme, some suggest that high-stakes testing may lead school personnel to intentionally manipulate student test scores (Jacob and Levitt 2003). Although there have been hundreds of studies of test-based accountability policies in the United States over the past two decades, the evidence on NCLB is more limited, both because it is a newer policy and because the national scope of the policy makes it extremely difficult to find an adequate control group by which to assess the national policy.

This paper examines the impact NCLB has had on students, teachers, and schools across the country. We investigate not only how NCLB has influenced student achievement, but also how it has affected education spending, instructional practice, and school organization. Given the complexity of the policy and the nature of its implementation, we are skeptical that any single analysis can be definitive. For this reason we present a broad collage of evidence and look for consistent patterns.

Several findings emerge. First, the weight of the evidence suggests that NCLB has had a positive effect on elementary student performance in mathematics, particularly at the lower grades. The benefits appear to be concentrated among traditionally disadvantaged populations, with particularly large effects among Hispanic students. We do not find evidence that the policy has adversely affected achievement at either the top or the bottom end of the test-score distribution. Instead, the policy-induced gains in math performance appear similar across the test-score distribution. However, the available evidence suggests that NCLB did not have a comparable effect on reading performance.

A closer look at the potential mechanisms behind the observed improvement provides some additional insight. For example, we find evidence that NCLB increased average school district expenditure by nearly $600 per pupil. This increased expenditure was allocated both to direct student instruction and to educational support services. We also find that this increased expenditure was not matched by corresponding increases in federal support. The test-score gains associated with these expenditure increases fall short of the ambitious goals enshrined in NCLB. However, we present some qualified evidence suggesting that the size of the gains reflects a reasonable return on investment.

We also discuss evidence on how NCLB may have influenced alternative measures of educational practice and student outcomes. This evidence suggests that NCLB led to an increase in the share of teachers with master's degrees. We also find evidence that teachers responded to NCLB by reallocating instructional time from social studies and science toward key tested subjects, particularly reading. We also present evidence that NCLB led to distinct improvements in a teacher-reported index of student behaviors (which covers, among other things, attendance, timeliness, and intellectual interest) commonly understood as measuring "behavioral engagement" with school.

The paper proceeds as follows. Section I outlines the theoretical underpinnings of school accountability and provides background on the NCLB legislation. Section II examines the impact of NCLB on student achievement, marshaling evidence from a variety of different sources. Section III investigates potential mediating mechanisms, discussing how the policy affected educational expenditure, classroom instruction, and school organization, among other things. Section IV concludes with recommendations for future policy and research.

I. Background on School Accountability and NCLB

NCLB represented a bold new foray into education policy on the part of the federal government. However, the provisions it embodied built on a long history of reforms in standards and accountability at the state and local levels over several decades.

I.A. Theoretical Underpinnings of School Accountability

A basic perception that has motivated the widespread adoption of school accountability policies like NCLB is that the system of public elementary and secondary schooling in the United States is "fragmented and incoherent" (Ladd 2007, p. 2). 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 weak incentives and a lack of accountability among teachers and school administrators. For example, Eric Hanushek and Margaret Raymond (2001, pp. 368–69) write that accountability policies are "premised on an assumption that a focus on student outcomes will lead to behavioral changes by students, teachers, and schools to align with the performance goals of the system"

and that "explicit incentives . . . will lead to innovation, efficiency, and fixes to any observed performance problems."

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

The principal-agent lens is also useful for understanding criticisms of accountability-based reforms. The assumption that the self-interest of teachers and administrators is misaligned 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"—those most likely, with additional attention, to meet the proficiency standard (see, for example, Neal and 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.

I.B. Research on Accountability Reforms Adopted by States before NCLB

School accountability reforms similar to those brought about by NCLB were adopted in a number of states during the 1990s. Several studies have evaluated the achievement consequences of these reforms. Because of the similarities between NCLB and aspects of these pre-NCLB accountability systems, this body of research provides a useful backdrop against which to consider the potential achievement impacts of NCLB. In a recent review of this diverse evaluation literature, David Figlio and Helen Ladd (2007) suggest that three studies (Carnoy and Loeb 2002, Jacob 2005, and Hanushek and Raymond 2005) are the "most methodologically sound" (Ladd 2007, p. 9).

A study by Martin Carnoy and Susanna Loeb (2002), based on state-level achievement data from the National Assessment of Educational Progress (NAEP), found that the within-state improvement 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 the Chicago public schools, relative both to prior trends and to contemporaneous changes in other large urban districts in the region. However, Jacob (2005) also found that younger students did not experience similar gains on a state-administered, low-stakes exam and that teachers responded strategically to accountability pressures (for example, increasing special education placements).

Hanushek and Raymond (2005) evaluated the impact of school accountability policies on state-level NAEP math and reading achievement, as measured by the difference between the performance of a state's eighth-graders and that of fourth-graders in the same state 4 years earlier. This gain-score approach applied to the NAEP data implied that there were two cohorts of state-level observations in both math (1992–96 and 1996–2000) and reading (1994–98 and 1998–2002). Hanushek and Raymond (2005) classified state accountability policies as implementing either "report-card accountability" or "consequential accountability." States with report-card accountability provided a public report of school-level test performance, whereas states with consequential accountability both publicized school-level performance and could attach consequences to that performance. The types of potential consequences were diverse. However, virtually all of the systems in consequential accountability states included key elements of the school accountability provisions later enacted in NCLB (for example, identifying failing schools, replacing principals, allowing students to enroll elsewhere, and taking over, closing, or reconstituting schools). Hanushek and Raymond (2005, p. 307) note that "all states are now effectively consequential accountability states (at least as soon as they phase in NCLB)."

Hanushek and Raymond (2005) find that the introduction of consequential accountability within a state was associated with statistically significant

1. The accountability index constructed by Carnoy and Loeb (2002) ranged from 0 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.

increases in the gain-score measures. The achievement gains implied by consequential accountability were particularly large for Hispanic students and, to a lesser extent, white students. However, the estimated effects for the gain 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 are so similar to the consequential accountability policies that had been adopted in some states.

I.C. Key Features of the NCLB Legislation

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

NCLB also requires that states introduce "sanctions and rewards" relevant to every school and based on their AYP status. It mandates explicit and increasingly severe sanctions (from implementing public-school choice to staff replacement to school restructuring) for persistently low-performing schools that receive Title I aid. According to data from the Schools and Staffing Survey of the National Center for Education Statistics, 54.4 percent of public schools participated in Title I services during the 2003–04 school year. Some states applied these explicit sanctions to schools not receiving Title I assistance as well. For example, 24 states introduced accountability systems that threatened all low-performing schools with reconstitution, regardless of whether they received Title I assistance.[2]

2. Lynn Olson, "Taking Root," *Education Week,* December 8, 2004.

II. The Impact of NCLB on Student Achievement

The overarching goal of NCLB has been to drive broad and substantive improvements in student achievement. This section discusses the available empirical evidence on the achievement effects of NCLB, drawing on a variety of research designs and data sources including national time trends, comparisons between private and public schools, and comparisons across schools and states.

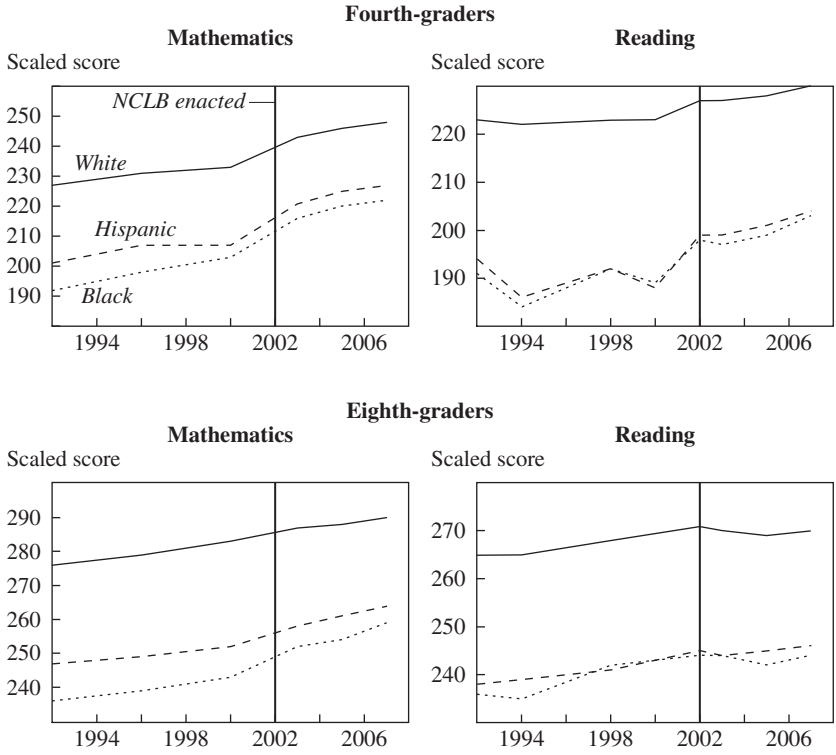
II.A. National Time Trends in Student Achievement

Because NCLB was introduced simultaneously throughout the United States, many observers have turned to state and national time-series trends in student achievement to assess its impact. For example, several studies have noted that student achievement, particularly as measured by state assessment systems, appears to have improved both overall and for key subgroups since the implementation of NCLB (Center on Education Policy 2008b). Others, however, argue that changes in student performance on high-stakes state tests can be highly misleading when states strategically adjust their assessment systems and teachers narrow their instructional focus to state-tested content (Fuller and others 2007).

Figure 1 presents data on national trends in student achievement from 1992 to 2007. These data are from the main NAEP and provide separate trends by grade (fourth and eighth), by subject (math and reading), and by race and ethnicity (white, black, and Hispanic).[3] These trends suggest that NCLB may have increased the math performance of fourth-graders. That is, these NAEP data suggest that fourth-grade math achievement has shifted noticeably higher during the NCLB era and may have also begun trending upward more aggressively. The trend data suggest similar gains in the math performance of black eighth-graders. However, the trends provide no clear suggestion that the onset of NCLB improved performance in

3. There are several different versions of the NAEP. The original NAEP, first administered in the early 1970s, is now called the Long-Term Trend (LTT) NAEP, because the Department of Education has made an effort to keep the content of this examination as consistent as possible over time in order to accurately gauge national trends. The LTT NAEP is administered to a small random sample of 9-, 13-, and 17-year-olds across the country and generally focuses on what many educators now think of as "basic" skills. What is now called the main NAEP was initiated in the early 1990s in an effort both to update the content and format of the national assessment so as test a broader domain of knowledge and skills, and to allow individual states to obtain their own, state-representative estimates. This exam is administered to fourth and eighth graders (and more recently to twelfth-graders).

Figure 1. Mean Scaled Scores on the Main NAEP, by Ethnicity, 1992–2007[a]



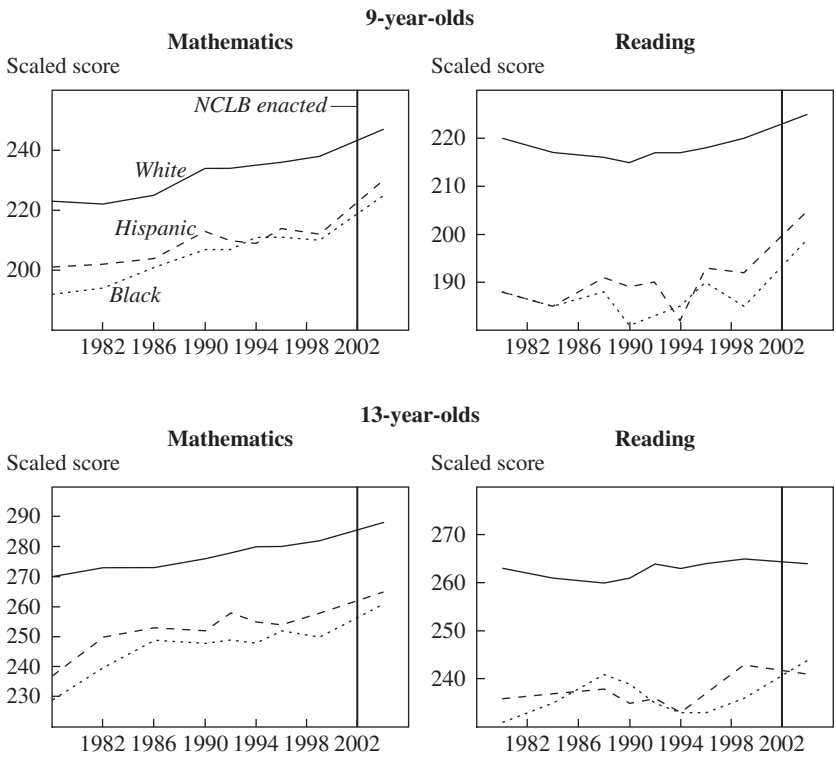
Source: National Center for Education Statistics.
a. Data are for all public schools.

the other three grade-subject combinations. Figure 2 shows achievement growth for 9- and 13-year-olds in math and reading, using data from the Long-Term Trend (LTT) NAEP, which has tracked student performance from the early 1970s. These data similarly suggest that the effects of NCLB on student achievement have been at best limited to certain groups.

II.B. Evidence from International Comparisons

Although these national achievement trends are suggestive, they do not necessarily provide the basis for reliable inferences about the impact of NCLB. Simple time-series comparisons may be biased by the achievement consequences of other time-varying determinants, such as the recession

Figure 2. Mean Scaled Scores on the Long-Term Trend NAEP, by Ethnicity, 1978–2004[a]

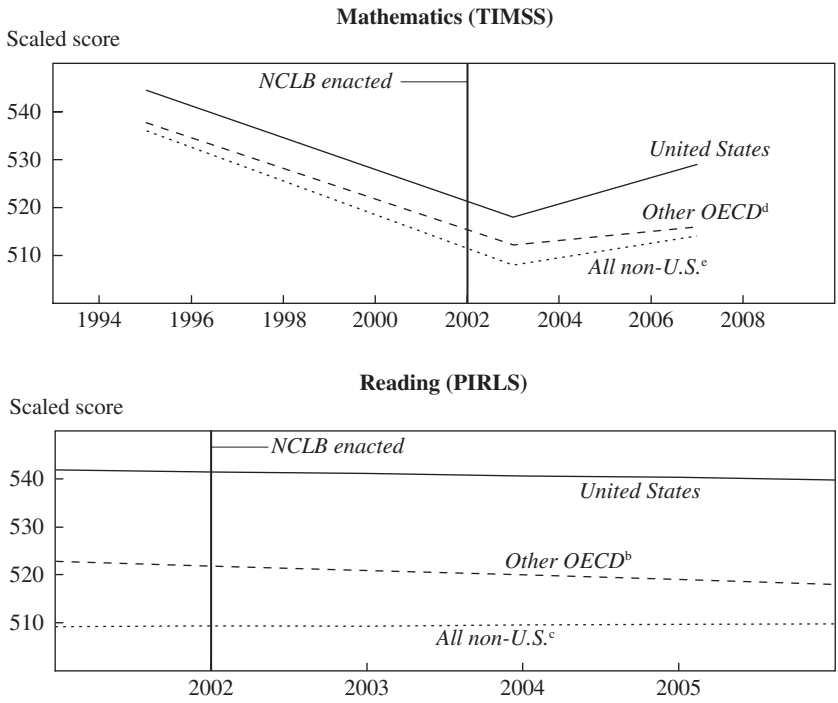


Source: National Center for Education Statistics.
a. Data are for all public schools.

that just preceded the introduction of NCLB. One straightforward way to benchmark the achievement trends observed in the United States is to compare them with the contemporaneous trends in other countries.

Because the time-series evidence in figure 1 suggests that any positive achievement effects from NCLB were likely to have been concentrated in fourth-grade math achievement, the comparative international achievement data from the Trends in International Mathematics and Science Study (TIMSS) are particularly relevant. The TIMSS collected trend data on fourth-grade math achievement for participating countries in 1995, 2003, and 2007. The top panel in figure 3 presents the fourth-grade scale scores in math from the TIMSS for the United States, for the 12 other countries that collected these performance data in each of these

Figure 3. Mean Scaled Scores of Fourth-Graders on the TIMSS and the PIRLS in the United States and Other Countries[a]



Source: National Center for Education Statistics.

a. The TIMSS (Trends in International Mathematics and Science Study) is an international assessment of the mathematics and science knowledge of fourth- and eighth-grade students, administered every 4 years since 1995. The PIRLS (Progress in International Reading Literacy Study) is an international assessment of the literacy achievement of fourth-grade students, administered every 5 years since 2001. Both studies are conducted by the International Association for the Evaluation of Educational Achievement.

b. Australia, England, Hungary, Japan, Netherlands, New Zealand, Norway, and Scotland.

c. Countries in note b plus Iran, Latvia, Singapore, and Slovenia.

d. England, France, Germany, Hungary, Italy, Kuwait, Netherlands, New Zealand, Norway, Scotland, Slovak Republic, and Sweden.

e. Countries in note d plus Bulgaria, Hong Kong, Iran, Israel, Latvia, Lithuania, Macedonia, Moldova, Morocco, Romania, Russia, Singapore, and Slovenia.

three study years, and for the subset of these comparison countries that are members of the Organization for Economic Cooperation and Development (OECD).

These trend data indicate that average math achievement on the TIMSS fell for all sets of countries by roughly equal amounts between the only available pre-NCLB year (1995) and the first academic year in which

NCLB was implemented (2002–03). Without additional years of data, we cannot assess the extent to which these comparative changes deviate from pre-NCLB trends. However, the available TIMSS data indicate that, by 2007, math achievement had comparatively improved in the United States, particularly with respect to the other OECD countries (an improvement of 11 scale points versus 4). These cross-country trends provide suggestive evidence consistent with the hypothesis that NCLB led to improvements in the math performance of younger students in the United States. However, the comparative test-score gain for the United States is relatively modest, amounting to only a 1.35 percent increase in average performance over pre-NCLB scores, and an 8 percent increase relative to the standard deviation in test scores.

Moreover, like the national time-series evidence, international comparisons provide no indication that NCLB improved the reading achievement of young students. The Progress in International Reading Literacy Study (PIRLS) reports data on the reading achievement of fourth-graders across a number of countries both in 2001 and in 2006. The bottom panel of figure 3 presents overall reading scores from the PIRLS by year for the United States, the group of 26 other countries that participated in both surveys, and the OECD members of this comparison group. On average, the United States outperformed these comparison countries. However, over the period that NCLB was implemented, all three sets of countries experienced quite similar and modest changes in PIRLS reading achievement.

Overall, then, the international evidence is at best suggestive. Contemporaneous changes within other countries may make them a poor comparison group for evaluating NCLB. The lack of multiple years of data also makes it difficult to distinguish possible policy effects from other trends or to identify any comparative differences with statistical precision. A subtler shortcoming of national and international time-series comparisons is that the presumption of a common, national effect ignores the possibility of heterogeneous effects of NCLB across particular types of states and schools.

II.C. Evidence from Accountability Risk Studies

However, several recent econometric studies have creatively leveraged this heterogeneity to identify the effects of NCLB. In particular, a widely used approach involves structuring comparisons across schools or students that face different risks of sanctions under NCLB. Derek Neal and Diane Schanzenbach (2010) present evidence that following the introduction of

NCLB in Illinois, the performance of Chicago public-school students near the proficiency threshold (that is, those in the middle of the distribution) improved while the performance of those at the bottom of the distribution remained the same or fell. Using data from the state of Washington, John Krieg (2008) finds that the performance of students in both tails of the distribution is lower when their school faces the possibility of NCLB sanctions.

Dale Ballou and Matthew Springer (2008), using data from a low-stakes exam fielded in seven states over a 4-year period, identify the achievement consequences of NCLB by constructing comparisons across grade-year cells that were included in AYP calculations and those that were not. Their approach takes advantage of the fact that between 2002–03 and 2005–06, states differed with respect to whether particular grades mattered for a school's accountability rating. Hence, their identification strategy leverages the fact that if the math scores of fourth-graders counted toward a school's accountability rating in one year but the math scores of fifth-graders in the same school did not count until the following year, one would expect student achievement to rise more quickly among fourth-graders relative to fifth-graders in the current year. Ballou and Springer find that the presence of AYP accountability modestly increased the math achievement of elementary-school students, particularly lower-performing students.

A recent study by Randall Reback, Jonah Rockoff, and Heather Schwartz (2010) adopts a similar approach, comparing student performance across elementary schools on the margin of making AYP. Using nationally representative data from the Early Childhood Longitudinal Study (ECLS), they find that reading and science scores on low-stakes tests improve by as much as 0.07 standard deviation when a school is on the margin for making AYP, but that the effects on math scores are smaller and statistically insignificant.

These accountability risk studies provide credible evidence on how NCLB-induced pressure influences the level and the distribution of student achievement. However, they have at least three potential limitations with respect to understanding the broad achievement consequences of NCLB. First, most of these studies have limited external validity because they do not rely on national data. Second, some rely on high-stakes assessments, which may not accurately reflect true student ability in the presence of strategic responses to NCLB (such as teaching to the test). Third, and perhaps most important, the treatment contrast in these studies may not approximate the full impact of NCLB because they rely on comparisons

across schools or students, all of whom were observed in the post-NCLB policy regime. To the extent that NCLB had broad effects on public schools (that is, even on students and schools not under the direct threat of sanctions), these comparisons could understate the effects of interest.

II.D. Evidence from a Comparison of States over Time

To address some of the limitations described above and estimate what one might consider the "full" impact of NCLB, we utilize a strategy that compares changes in student performance within states over time (see Dee and Jacob forthcoming). We take advantage of the fact that NCLB was explicitly modeled on an earlier generation of state-level school accountability systems. In the decade before NCLB, about 30 states implemented consequential school accountability policies that were fundamentally similar to NCLB in that they mandated systematic testing of students in reading and math, public reporting of school performance on these exams, and the possibility of meaningful sanctions (including school takeover, closure, or reconstitution, replacing the principal, and allowing students to change schools) based on test-based school performance. In fact, some state officials argued that NCLB needlessly duplicates preexisting state accountability systems.[4]

The existence of these earlier NCLB-like accountability systems establishes natural treatment and control state groups. In our framework, states that adopted NCLB-like accountability before NCLB form our control group. Other states, for which NCLB catalyzed an entirely new experience with consequential school accountability, form our treatment group.[5] Of course, states that adopted accountability programs before NCLB were not randomly distributed. For this reason our "comparative interrupted time series" (CITS) strategy, described in more detail below, relies on within-state variation over time, allowing not only for different levels of

4. Michael Dobbs, "Conn. Stands in Defiance of Enforcing 'No Child.' " *Washington Post,* May 8, 2005.

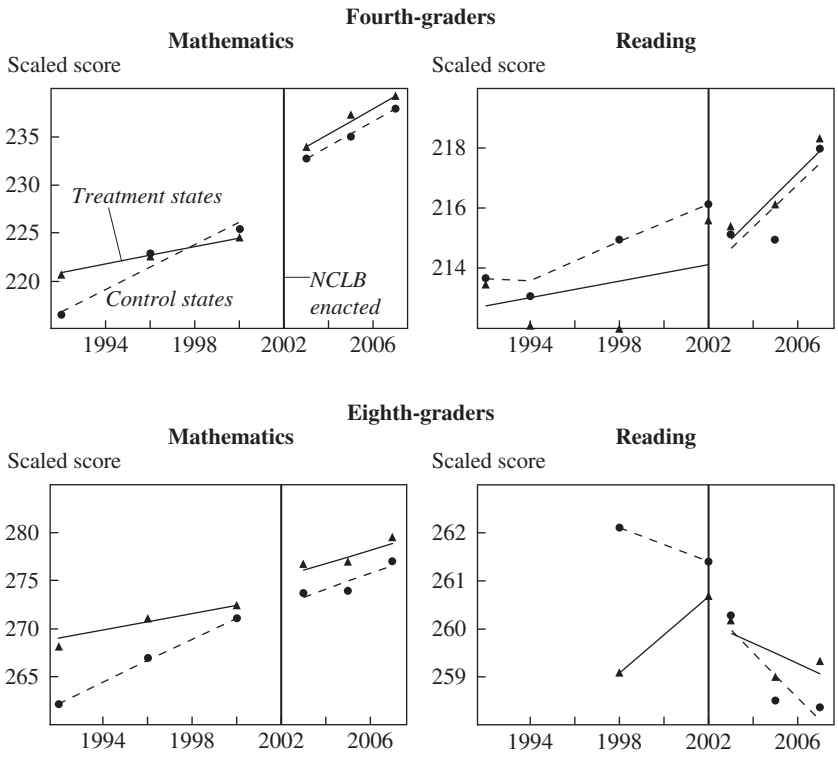
5. We relied on a number of different sources to categorize pre-NCLB accountability policies across states, including prior studies of such policies (for example, Carnoy and Loeb 2002, Lee and Wong 2004, and Hanushek and Raymond 2005) as well as primary sources such as the Quality Counts series put out by *Education Week* ("Quality Counts '99," January 11, 1999, www.edcounts.org/archive/sreports/qc99/), the state-specific "Accountability and Assessment Profiles" assembled by the Consortium for Policy Research in Education (Goertz, Duffy, and Le Floch 2001), annual surveys on state student assessment programs fielded by the Council of Chief State School Officers, information from state education department websites, Lexis-Nexis searches of state and local newspapers, and conversations with academics and state officials in several states.

achievement across states before NCLB but also for different *trends* in achievement across states before NCLB.

GRAPHICAL EVIDENCE. We illustrate the logic of our identification strategy through a series of figures. This graphical evidence has the advantages of transparency and simplicity. We then present regression estimates that more clearly show the magnitude and statistical precision of our findings and allow us to demonstrate that the results are robust to a variety of alternative specifications and several falsification exercises.

Figure 4 shows the trends in NAEP scores for two groups: states that had adopted school accountability by 1998 (control states), and states that

Figure 4. Mean Scaled Scores on the Main NAEP, by Timing of Increased School Accountability, 1992–2007[a]



Source: National Center for Education Statistics and authors' calculations.

a. Data are for public schools only. Treatment states are those that did not adopt consequential school accountability policies before NCLB, and control states those that had adopted such policies before 1998. A small number of states that adopted accountability programs between 1999 and 2001 are excluded.

had not adopted school accountability before NCLB (treatment states).[6] The NAEP data are particularly well suited to this evaluation for several reasons. First, the NAEP is a technically well-developed assessment that covers a broad domain of knowledge and schools. Second, it provides consistent, state-representative measures of student performance for most states over the last two decades. Finally, the exam is a low-stakes exam for students, teachers, and schools.[7] Because teachers have no incentive to "teach to" the NAEP, it is likely to provide the most accurate measure of student achievement (Fuller and others 2007).

The figure plots the simple (unweighted) average scale score of each group of states in all years in which the exam was administered. Years are identified by the spring of the relevant academic year (for example, "1992" refers to the 1991–92 school year). The sample of states is consistent across years (that is, it is a balanced panel), and the state classification is a time-invariant characteristic. Data points to the left of the vertical line that indicates the enactment of NCLB are considered "pre-policy," and those to the right "post-policy."[8] To illustrate the pre- and post-NCLB achievement trends within each group, we also plot the fitted regression line from a simple linear regression conducted separately for each group × period (pre- or post-NCLB).

6. These figures exclude a small number of states that adopted state accountability programs between 1999 and 2001, in order to make a clear distinction between our treatment and comparison groups. However, the regression analysis described in the following section includes these "late adopter" states. Dee and Jacob (forthcoming) show that the inclusion of these late adopters does not change the findings in any substantive way.

7. That is, the NAEP is not used as the basis for student promotion or retention, teacher evaluation, or school accountability. Indeed, the NAEP is administered only to a small, random sample of fourth-, eighth-, and twelfth-grade students in each state.

8. When one dates the start of NCLB is a potentially important issue. NCLB secured final congressional approval on December 18, 2001, and was signed by President George W. Bush on January 8, 2002, both events thus occurring in the middle of the 2001–02 academic year. NCLB is often characterized as having been implemented during 2002–03 because states were required to use testing outcomes from the previous academic year as the starting point for determining whether a school was making adequate yearly progress (Palmer and Coleman 2003; Lynn Olson, "States Strive toward ESEA Compliance," *Education Week,* December 1, 2002). However, one could reasonably conjecture that the discussion and anticipation surrounding the adoption of NCLB would have influenced school performance during the 2001–02 school year. Alternatively, it could also be argued that NCLB should not be viewed as in effect until the 2003–04 academic year, when new state accountability systems were more fully implemented as well as more informed by guidance from and negotiations with the U.S. Department of Education (Lynn Olson, "States Strive toward ESEA Compliance," *Education Week,* December 1, 2002; Olson, "Taking Root," *Education Week,* December 8, 2004). For a more detailed discussion, see Dee and Jacob (forthcoming).

The top left panel of figure 4, which plots trends in fourth-grade math achievement, shows that in 1992, states that did not adopt accountability until NCLB scored roughly 5 scale points (0.18 standard deviation) higher on average than states that adopted school accountability policies by 1998. Although both groups of states made modest gains between 1992 and 2000, the group that adopted accountability policies before 1998 experienced more rapid improvement during this period.[9]

If the NCLB accountability provisions had an impact on student performance, one would expect achievement to increase more after 2002 in states with no prior accountability than in states with prior accountability. It is possible that NCLB led to a level shift in student achievement, which would be manifest as a shift in the intercept after NCLB. It is also possible that NCLB changed the *rate* of achievement growth, which would be manifest as a change in the *slope* of the achievement trend after NCLB.[10] Whether one considers a shift in the intercept or a change in the slope, our identification strategy relies on a comparison of treatment versus control states that accounts not only for the pre-NCLB levels of achievement in each group but also for the pre-NCLB achievement *trends* in each group.

The top left panel of figure 4 shows that the mean level of math achievement jumped noticeably in 2003 for both groups of states. However, relative to prior trends, this shift was larger among the "no prior accountability" group (the treatment states). Interestingly, there was little noticeable change in the growth *rate* across periods for the states with prior accountability (the control states): the slope of the achievement trend for this group is roughly the same before and after 2002. In contrast, achievement rose more rapidly in states with no prior accountability 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 trends suggest that NCLB had a positive impact on fourth-grade math achievement.

9. This visual evidence is consistent with the earlier evaluation literature that studied pre-NCLB state accountability reforms (for example, Carnoy and Loeb 2002, Jacob 2005, and Hanushek and Raymond 2005).

10. The rate of achievement growth might have increased after NCLB for several reasons. First, it may take states time to implement new curricula, instructional strategies, or other support services for students. Second, later cohorts of students will have been "exposed" to NCLB for a larger fraction of their school careers than earlier cohorts. Without imposing additional assumptions, we cannot cleanly distinguish between these effects. For this reason we focus on the "net" impact of NCLB in different years after the legislation was passed.

The trends for eighth-grade math (bottom left panel of figure 4) are similar to those for fourth-grade math, although less pronounced. The pattern for fourth-grade reading (top right panel of figure 4) is much less clear. The pre-NCLB reading trends for both groups are much noisier than the math trends. In particular, the two groups both experienced a decline in achievement in 1994 and then diverged in 1998, but both had made very large gains by 2002.[11] The states with prior accountability experienced a drop in achievement from 2002 to 2003, both in absolute terms and relative to trend. The other group experienced very little increase from 2002 to 2005. Perhaps most important, however, visual inspection of the data in these plots indicates that the earlier achievement trend was not linear, which is a central assumption of the linear CITS model. Similarly, the bottom right panel of the figure provides no evidence of an NCLB effect on eighth-grade reading achievement.

ESTIMATION STRATEGY. Perhaps the most straightforward approach to estimating the impact of NCLB in the framework described above is a simple difference-in-differences framework in which one compares the achievement levels of treatment and control states before and after the introduction of NCLB. However, a fundamental assumption of this model is that any preexisting trends in the outcome variables are equivalent across treatment and control groups. Figure 4 clearly showed that the control states (those that implemented consequential accountability before NCLB) realized more rapid improvements during the pre-NCLB period. For this reason we estimate a more flexible specification that allows for preexisting trends to differ across groups. Our model is the following:

$$
\begin{aligned}
(1) \qquad Y_{st} = {} & \beta_0 + \beta_1 YEAR_t + \beta_2 NCLB_t + \beta_3 \left(YR\_SINCE\_NCLB_t\right) \\
& + \beta_4 \left(T_s \times YEAR_t\right) + \beta_5 \left(T_s \times NCLB_t\right) \\
& + \beta_6 \left(T_s \times YR\_SINCE\_NCLB_t\right) + \beta_7 \mathbf{X}_{st} + \boldsymbol{\mu}_s + \varepsilon_{st},
\end{aligned}
$$

where Y_{st} is a measure of student achievement for state s in year t, $YEAR_t$ is a trend variable (defined as the year of the test minus 1989 so that it starts with a value of 1 in 1990), and $NCLB_t$ is a dummy variable equal to 1 for observations starting in the academic year 2002–03. $YR\_SINCE\_NCLB_t$ is defined as the year of the test minus 2002, so that this variable takes on a value of 1 for the 2002–03 year, which corresponds to the 2003 NAEP testing. $\mathbf{X}_{st}$ represents a vector of state × year covariates. In the main

11. The graph is scaled to accentuate what are really quite small absolute changes from year to year.

specification the only state-year covariates included are the fraction and its square of students who were tested but excluded from official reporting because of limited English proficiency or some type of learning disability. The variables $\boldsymbol{\mu}_s$ 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. In the most basic setup, T_s could be specified as a dummy variable, with a value of 1 indicating that a given state did not institute consequential accountability before NCLB. This is the approach implicitly taken in figure 4. However, it is more accurate to view the "treatment" provided by the introduction of NCLB in the framework of a dosage model. Slightly more than half of the states that introduced consequential school accountability before NCLB did so within the 4 years before NCLB's implementation. The simple binary definition of T_s above could lead to attenuated estimates of the NCLB effect, because the control group would include some states for which the effects of prior state policies and NCLB are closely intertwined.

For this reason we instead define T_s as the number of years during our panel period that a state did *not* have school accountability. Specifically, we define the treatment as the number of years *without* prior school accountability between the 1991–92 academic year and the onset of NCLB. Hence, states with no prior accountability have a value of 11. Illinois, which implemented its policy during the 1992–93 school year, has a value of 1; Texas has a value of 3, since its policy started in 1994–95; and Vermont has 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 the potential impact of NCLB.

This regression specification allows for an NCLB effect that can be reflected in both a level shift in the coefficient on the outcome variable (β_5) and a shift in the coefficient on the achievement trend variable (β_6), each of which varies with treatment status, T_s. Specifications based on alternative functional forms generate results similar to those based on this canonical CITS design.[12] For the sake of parsimony, the impact estimate we report is the effect of NCLB by 2007 for states with no prior accountability relative to states that adopted school accountability in 1997 (the mean adoption

12. For example, we get similar results when we allow for a separate NCLB "effect" unique to each post-NCLB year. We also find similar results when we measure treatment status with multiple dummy variables, allowing the trend and shift variables to differ across groups of states that were early, middle, or late adopters of pre-NCLB accountability.

year among states that adopted accountability before NCLB).[13] For all models we present standard errors clustered by state to account for serial correlation and other forms of heteroskedasticity.

The primary threat to causal inference in our CITS design is the existence of time-varying unobservable factors that are coincident with the introduction of NCLB, affect treatment and control states differently, and independently affect student performance. One example is endogenous student mobility, such as might occur if NCLB caused families to leave or return to the public schools. Another problematic scenario would be one where either the treatment or the control states recovered from the 2001 recession more quickly. As discussed below, we examine the empirical relevance of these concerns in several ways and find no evidence that our findings are biased.

Other threats to the causal validity of this state-based research design are closely linked to exactly how the NCLB impact estimates from equation 1 should be interpreted. For example, our estimates will capture the impact of the accountability provisions of NCLB but *not* the effects of other NCLB provisions such as Reading First or the "highly qualified teacher" provision, which were unique nationwide. Second, under the maintained assumption that NCLB was effectively irrelevant in states with prior consequential accountability systems, our estimates will identify the effects of NCLB-induced school accountability provisions for a particular subgroup of states (those without prior accountability policies). To the extent that one believes that those states expecting to gain the most from accountability policies adopted them before NCLB, the results we report would understate the average treatment effects of school accountability. Similarly, our estimates will also understate the general effects of school accountability if NCLB amplified the effects of school accountability within the states that already had it. An alternative concern is that the accountability systems within control states may have been weakened as they were adjusted in response to NCLB. To the extent this occurred, our CITS approach would instead overstate the effects of NCLB. We suspect this concern is not empirically relevant because the school reporting and performance sanctions occasioned by NCLB (such as the possibility of school reconstitution or closure) were strong relative to prior state accountability policies. There is also direct empirical evidence

13. Specifically, the effect as of 2007 would be calculated as $\beta_5 + \beta_6(5)$ in the simple case where T_s is binary, but as $\beta_5(6) + \beta_6(6 \times 5)$ in our preferred specification where T_s is allowed to vary across states and the NCLB effect is identified relative to a state that implemented school accountability in 1997. As a practical matter, both approaches generate similar results (Dee and Jacob forthcoming, table 3).

consistent with this assumption: Dee and Jacob (forthcoming) find that states with preexisting school accountability systems did not change their proficiency thresholds after the onset of NCLB.

RESULTS. Table 1 presents estimates of regressions, based on equation 1, of the impact of NCLB on student performance in mathematics and reading. Overall, the results suggest that NCLB had uniformly positive effects on math performance among elementary students, particularly fourth-graders. The mean impact of 7.2 score points for fourth-grade math translates to an effect size of 0.23 standard deviation. The effects are even larger toward the left of the ability distribution. These estimates suggest that NCLB increased the proportion of fourth-graders reaching the basic level on NAEP by 10 percentage points, or a 16 percent increase relative to the control mean of 64 percent. Although the mean effects for eighth-graders are not statistically significant at conventional levels (a 0.10-standard-deviation effect, with a p-value of 0.12), the effects at the bottom tail are stronger. NCLB increased the fraction of eighth-graders reaching the basic level in math by 5.9 percentage points (9 percent).

Although we find that NCLB had larger impacts on the mathematics performance of lower-achieving students, we do not find any evidence that the introduction of NCLB harmed students at higher points on the achievement distribution. In contrast to some prior work within individual districts and states, we find that NCLB seems to have increased achievement at higher points on the achievement distribution more than one might have expected. For example, in fourth-grade math the impacts at the 75th percentile were only 3 scale points lower than those at the 10th percentile.

In contrast to the mathematics results, we do not find consistent evidence that NCLB influenced student achievement in reading. The NCLB impact estimates for the reading measures are smaller and, in most cases, statistically indistinguishable from zero. The one notable exception is the finding that NCLB improved the reading performance of higher-achieving fourth-graders (those at the 75th and the 90th percentiles) modestly but significantly. However, as noted earlier, a caveat to the reading results is the suggestive evidence that the pre-NCLB trends in reading achievement, which are noisy and nonlinear, poorly match the assumptions of the CITS design. Furthermore, the capacity of this research design to detect effects on the reading achievement of eighth-graders is attenuated by the fact that only 2 years of pre-NCLB NAEP data are available for this grade-subject combination.

To test the sensitivity of our results to some of the potentially time-varying unobservable factors described above, we conducted a series of falsification exercises in which we reestimated equation 1 with a variety of

Table 1. Regressions Estimating the Effect of NCLB on Fourth- and Eighth-Grade NAEP Mathematics and Reading Scores[a]

	Mathematics				*Reading*			
	Fourth grade (39 states, N = 227)		*Eighth grade (38 states, N = 220)*		*Fourth grade (37 states, N = 249)*		*Eighth grade (34 states, N = 170)*	
Dependent variable	*Estimated effect*	*Mean pre-NCLB outcome in states without prior accountability*	*Estimated effect*	*Mean pre-NCLB outcome in states without prior accountability*	*Estimated effect*	*Mean pre-NCLB outcome in states without prior accountability*	*Estimated effect*	*Mean pre-NCLB outcome in states without prior accountability*
Mean NAEP score	7.244**	224	3.704	272	2.297	216	−2.101	261
	(2.240)		(2.464)		(1.441)		(2.070)	
Percent of pupils achieving at or above basic level	10.090**	64	5.888**	64	2.359	61	−3.763	73
	(3.145)		(2.680)		(1.592)		(2.561)	
75th-percentile NAEP score	6.634**	244	4.340**	296	2.258**	240	1.289	282
	(1.902)		(2.189)		(0.938)		(2.249)	
90th-percentile NAEP score	5.205**	259	2.537	314	2.097**	258	1.172	299
	(1.916)		(2.404)		(0.805)		(2.897)	

Source: Authors' regressions.

a. Each reported coefficient is from a separate regression specified as in equation 1 in the text and is the sum of coefficients β_5 and β_6. Effects are as of 2007 for states with no prior accountability relative to states that adopted school accountability in 1997. See the text for details. Standard errors clustered by state are in parentheses. Asterisks indicate statistical significance at the ***1 percent, **5 percent, or *10 percent level.

Table 2. Regressions Estimating the Effect of NCLB on Fourth- and Eighth-Grade NAEP Mathematics Scores, by Ethnicity and Eligibility for Free School Lunch Program[a]

	Whites		*Blacks*	
Dependent variable	*Estimated effect*	*Mean pre-NCLB outcome in states without prior accountability*	*Estimated effect*	*Mean pre-NCLB outcome in states without prior accountability*
Fourth-graders				
Mean NAEP score				
OLS	5.953**	232	4.582	203
	(1.990)		(5.436)	
WLS	5.074**	233	15.378**	202
	(2.159)		(3.710)	
Percent of pupils achieving at or above basic level				
OLS	7.278**	76	8.431	35
	(3.016)		(6.693)	
WLS	7.597**	77	22.690**	33
	(3.531)		(6.199)	
Eighth-graders				
Mean NAEP score				
OLS	2.863	281	9.261	241
	(2.561)		(6.774)	
WLS	1.828	282	8.826	242
	(3.680)		(8.999)	
Percent of pupils achieving at or above basic level				
OLS	4.740*	74	9.977	28
	(2.639)		(7.886)	
WLS	4.253	76	10.004	28
	(3.134)		(11.955)	

Source: Authors' regressions.

a. Each reported coefficient is from a separate regression and estimates the effect of NCLB as of 2007. See table 1 and the text for details. OLS = ordinary least squares; WLS = weighted least squares (weighting by student enrollment). Standard errors clustered by state are in parentheses. Asterisks indicate statistical significance at the ***1 percent, **5 percent, or *10 percent level.

alternative outcome measures, including state-year poverty rates, median household income, employment-population ratios, and the fraction of students in the public schools. Across 40 regressions (10 models for each of the four grade × subject combinations), we find only one estimate significant at the 5 percent level and three estimates significant at the 10 percent level. These largely null findings suggest that the assumptions required for identification are indeed met. In Dee and Jacob (forthcoming), we also show that the results presented in table 1 are robust to a host of alternative specifications, including the inclusion of a variety of state-year covariates,

Hispanics		*Eligible for free school lunch*		*Not eligible for free school lunch*	
Estimated effect	*Mean pre-NCLB outcome in states without prior accountability*	*Estimated effect*	*Mean pre-NCLB outcome in states without prior accountability*	*Estimated effect*	*Mean pre-NCLB outcome in states without prior accountability*
12.409**	204	6.934*	212	3.916	232
(4.540)		(3.604)		(3.102)	
11.625**	204	9.734**	212	2.603	234
(1.572)		(2.836)		(2.907)	
12.499*	40	11.186*	49	5.388	76
(6.334)		(5.769)		(4.435)	
25.883**	36	17.256**	49	6.832**	78
(2.779)		(4.986)		(3.118)	
20.031**	246	10.702*	257	2.199	279
(5.766)		(6.155)		(3.924)	
8.219**	247	15.761**	256	0.992	281
(4.135)		(5.631)		(4.171)	
22.006**	36	12.773*	47	3.152	72
(4.618)		(7.328)		(4.045)	
18.692**	36	23.432**	46	2.392	74
(4.666)		(6.398)		(3.478)	

the inclusion of state-specific time trends, the inclusion of a full set of year fixed effects, and weighting the data by the number of students enrolled in that state and year.[14] Moreover, the inclusion of 2009 NAEP data does not change the basic pattern of results presented here.

Table 2 reports regression estimates separately by subgroup, both unweighted and weighted by student enrollment. Interestingly, the positive

14. Dee and Jacob (forthcoming) show that these results are also robust to measuring the intensity of the treatment imposed by NCLB in terms of the stringency of the proficiency standards imposed by the state. Wong, Cook, and Steiner (2009) find this as well.

effects are particularly large among lower-income and minority students. For example, among fourth-graders, NCLB increased the math achievement of black and Hispanic students. Interestingly, the enrollment-weighted estimates are systematically larger than the unweighted estimates for low-income and black students. For example, the weighted estimate for African-American students is 15.4 points (roughly 0.5 standard deviation) compared with the unweighted estimate of 4.6 points. Taken at face value, this suggests an important source of treatment-effect heterogeneity, namely, that NCLB had a more positive effect on disadvantaged students in states with a greater number of such children (for example, NCLB was more effective for black students in Alabama than for black students in South Dakota). However, given the relatively small number of treatment states with large populations of black students, the possibility that this heterogeneity reflects other state-specific traits cannot be discounted. The effect of NCLB on Hispanic students was also quite large (roughly 12 points) and did not vary with weighting. The weighted impact on students eligible for subsidized lunches was 9.7 points (roughly 0.3 standard deviation).

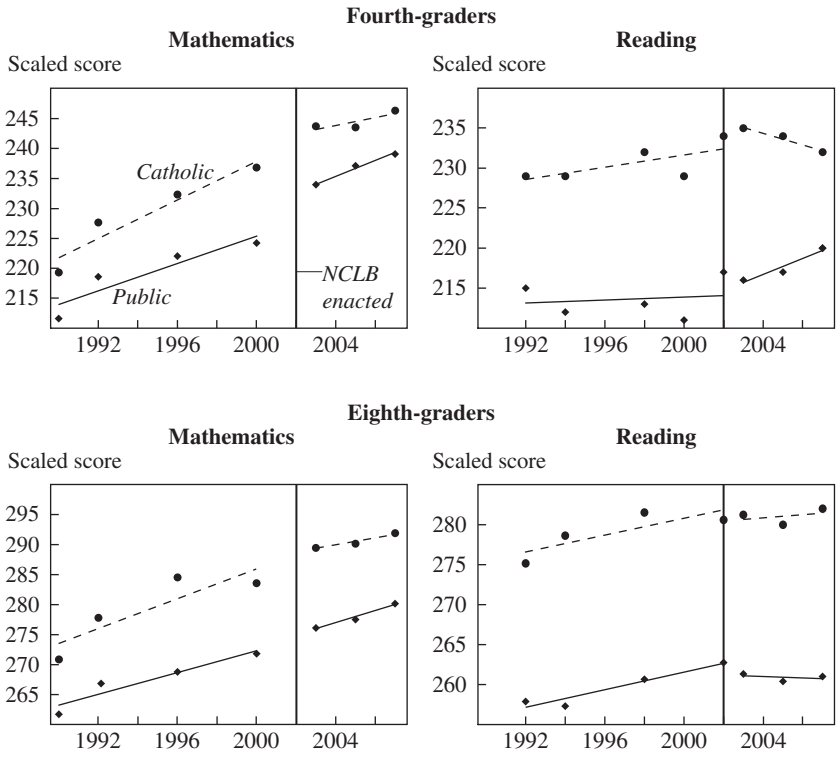
II.E. Evidence from Public- and Private-School Comparisons

The above comparison of trends in student performance within states over time suggests that NCLB had a substantial impact on math achievement, particularly among disadvantaged students in fourth grade. As with any nonexperimental design, however, the findings rest on assumptions that cannot be fully tested. For this reason we present results from a complementary analysis that makes use of an alternative control group.

In this approach we assess the impact of NCLB by comparing trends over time in student performance in public versus Catholic schools.[15] Students in private schools are eligible to participate in a number of major programs under the ESEA, and NCLB's reauthorization of ESEA left these prior provisions largely intact (U.S. Department of Education 2007), implying that the NCLB reforms were comparatively irrelevant for private schools. The use of Catholic schools in this analysis improves upon international comparisons by providing a within-nation control group. However, as with the national and international time-series evidence, this approach

15. In earlier work (Dee and Jacob forthcoming) we identify several potential concerns with using Catholic schools to identify the impact of NCLB.

Figure 5. Mean Scaled Scores on the Main NAEP in Public and Catholic Schools, 1990–2007



Source: National Center for Education Statistics.

also conflates the effects of NCLB across states and schools where its impact was heterogeneous.

Figure 5, which follows the same structure as figure 4 in comparing treatment and control states, shows pre- and post-NCLB achievement trends across public and Catholic schools. Although the performance of both public and Catholic students trended upward during the sample period, the latter consistently outperformed their public-school counterparts. However, following the implementation of NCLB, the math performance of public-school students converged somewhat toward that in the Catholic schools and entered a period of somewhat stronger trend growth. This comparative convergence is particularly pronounced for fourth-graders and is consistent with the other time-series evidence suggesting that NCLB

improved math achievement, particularly among younger students. The reading achievement trends of eighth-graders are quite similar across public and Catholic schools, suggesting the absence of a meaningful NCLB impact. However, the reading achievement of public-school fourth-graders trended upward during the NCLB era, particularly relative to that of Catholic-school fourth-graders, which began a distinctive downward trend during the NCLB era.

These public-Catholic comparisons are broadly consistent with the state-based comparisons, suggesting that NCLB led to substantial gains in the mathematics achievement of fourth-graders and possibly of eighth-graders as well. These particular cross-sector comparisons also suggest that NCLB increased the reading achievement of fourth-graders. A recent study by Manyee Wong, Thomas Cook, and Peter Steiner (2009) includes regression estimates based on public-Catholic comparisons of this sort and draws similar conclusions. They also find similar, although less precisely estimated, results in comparisons of public schools and non-Catholic private schools.

II.F. Summary of Achievement Effects

Given the national scope of the policy, assessing the causal impact of NCLB on student performance is not straightforward. However, the body of evidence presented above seems to suggest that the federal school accountability policy did improve the math achievement of elementary students, particularly among socioeconomically disadvantaged groups. Comparable evidence that NCLB generated meaningful improvements in reading achievement is lacking, however. Moreover, the analysis presented above focuses exclusively on elementary schools. NCLB also requires AYP determinations for high schools, but here relatively little is known about NCLB's effects, in part because of data limitations: for example, no state-level data for secondary-school math achievement on the main NAEP are available after 2000.

What is the relevance for policy of the overall gains in math achievement that NCLB appears to have brought about? One way to benchmark a 7.2-point (0.23-standard-deviation) gain in fourth-grade math achievement is to compare this effect with achievement gaps that are of interest. For example, a test-score gain of this size is equivalent to approximately 24 percent of the black-white test-score gap observed in the 2000 NAEP data. Furthermore, because NCLB appears to have been more effective among disadvantaged subgroups, it may have contributed to closing some achievement gaps. For example, the effect of NCLB on the math achievement of

Hispanic fourth-graders was roughly 6 points larger than the corresponding effects on white students, implying that NCLB closed the white-Hispanic achievement gap by 19 percent.

III. Impact of NCLB on the Organization and Practice of Education

Given the encouraging effects on math achievement and the somewhat puzzling lack of effects for reading, it is natural to ask how NCLB affected the organization and practice of elementary education across the country. Such evidence on potential mediating mechanisms could not only guide revisions to the NCLB legislation, but also shed light on the education production function in ways that would inform other school reforms. To provide some coherence to the discussion that follows, we group nonachievement outcomes from a variety of sources into several broad categories: changes in educational resources, changes in instructional focus or methods or both, and changes in school organization, climate, or culture.

III.A. Impact on Education Expenditure

The direct costs of managing an accountability system are quite small on a per-pupil basis (Hoxby 2002). However, standards-based reforms have often been presented to the public as a trade: greater resources and flexibility for educators in exchange for greater accountability. One of the most strident criticisms of NCLB is that it failed to deliver on this bargain. However, there is surprisingly little research on the relationship between school accountability and spending, despite an extensive literature on education finance more generally.

One notable exception is an analysis of district-level expenditure data from 1991–92 to 1996–97 by Jane Hannaway, Shannon McKay, and Yasser Nakib (2002). Examining four states that implemented comprehensive accountability programs in the 1990s—Kentucky, Maryland, North Carolina, and Texas—they find that only two (Texas and Kentucky) increased educational expenditure more than the national average (but those two did so substantially). Hannaway and Maggie Stanislawski (2005) present evidence that the major pre-NCLB accountability reforms in Florida were associated with increased expenditure for instructional staff support and professional development, particularly in low-performing schools. Of course, it is difficult to determine whether the accountability policy caused the increased expenditure or whether both were merely parts of a broader

reform agenda. Overall, the extant literature offers at best suggestive evidence on how accountability reforms may have influenced school spending.

To provide new evidence on how NCLB influenced local school finances, we pooled annual, district-level data on revenue and expenditure from U.S. Census surveys of school district finances (the F-33 Annual Survey of Local Government Finances) over the period from 1994 to 2008 (Dee, Jacob, and Schwartz 2010). Our analytical sample consists of all operational, unified school districts nationwide (roughly 10,000) for each survey year. To identify the effects of NCLB accountability on district finances, we utilize the same cross-state trend analysis described above, comparing within-state changes in school finance measures across states with and without pre-NCLB accountability programs.

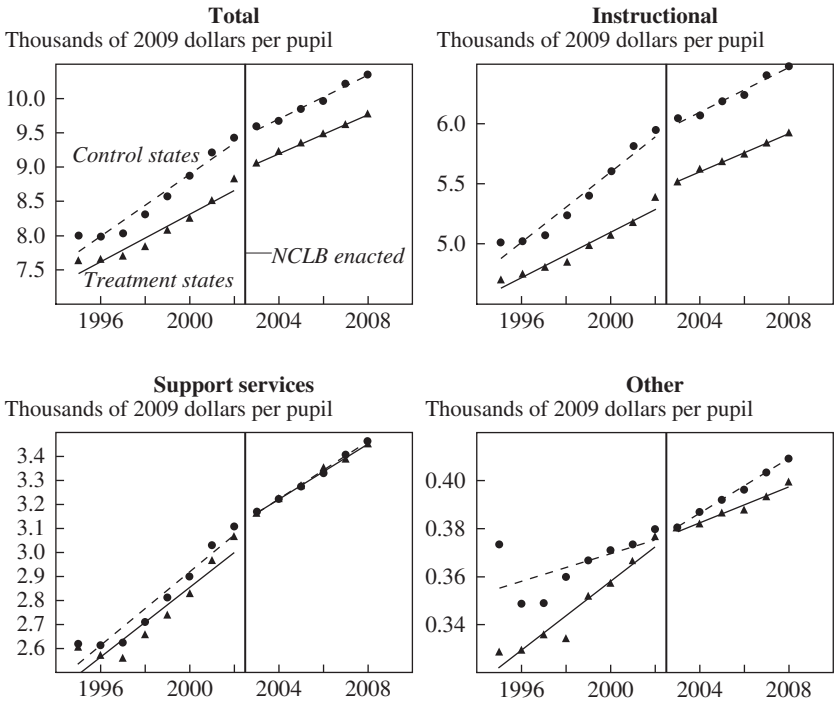
Figure 6 shows trends in district expenditure over time separately for states that adopted consequential accountability before NCLB and those that did not. All results are reported in 2009 dollars and are weighted by district enrollment. As in the earlier figures, the trend lines are fitted linear regression lines.[16] The top left panel of figure 6 shows that total per-pupil expenditure rose more quickly from 1994 to 2002 in states that adopted pre-NCLB accountability policies. But following the introduction of NCLB, spending grew more slowly in these early-adopting states, suggesting that NCLB increased expenditure. The top right and bottom left panels of the figure show comparable results for the two largest categories of total expenditure, instructional and support service spending.

Table 3 presents regression estimates based on the model in equation 1, with the inclusion of the following district-year controls: enrollment, enrollment squared, the fraction of the student population that is black or Hispanic, the poverty rate (based on 2000 census data), the poverty rate squared, and the interaction between the poverty rate and the fraction black or Hispanic. As in earlier models, we present standard errors clustered by state. We report estimates of the impact of NCLB as of 2008 for states that did not have consequential accountability before NCLB relative to states that adopted consequential accountability in 1997.

The results indicate that NCLB increased total current expenditure by $570 per pupil, or by 6.8 percent from the 1999–2000 mean of $8,360. The increased expenditure was allocated to direct instruction and support ser-

16. Also as before, the figures omit states that adopted school accountability programs between 1999 and 2001, because the impacts of these state programs might be confounded with the introduction of NCLB in 2002. In the regression estimates discussed below, however, we include all states.

Figure 6. Expenditure per Pupil by Timing of Increased School Accountability, 1995–2008[a]



Source: Authors' calculations using data from the Common Core of Data's Local Education Agency (School District) Finance Survey.

a. All data are for elementary and secondary school expenditure. Sample is composed of all noncharter, unifed local education agency school districts, excluding Hawaii, the District of Columbia, and zero-enrollment districts. Estimates are weighted by district enrollment. Treatment and control states are defined as in figure 4.

vices in proportions roughly equivalent to average spending patterns, with effects of $430 (8.3 percent) and $155 (5.6 percent), respectively. Results presented in the bottom two rows of the table reveal that the increased expenditure was not matched by corresponding increases in federal support, consistent with allegations that NCLB constitutes an unfunded mandate. (However, the increase in spending on student support is not statistically significant at conventional levels.) In results not shown here, we find that the effects were fairly similar across districts with different baseline levels of student poverty, suggesting that NCLB did not meaningfully influence distributional equity. Moreover, in results reported elsewhere, we demonstrate that these findings are robust to the same falsification

Table 3. Regressions Estimating Effects of NCLB on Education Expenditure by Function and Revenue by Source

Constant (2009) dollars per pupil

Dependent variable	*Mean for 1999–2000 school year*	*Estimated impact of NCLB*[a]
Expenditure		
Total current expenditure, K-12	8,360	570**
Instructional	(2,061)	(237)
	5,209	430***
Support services	(1,428)	(137)
	2,786	155
Other	(772)	(112)
	365	–0.015
	(111)	(25)
Revenue		
Federal	660	42
	(473)	(31)
State and local	9,155	448
	(2,250)	(288)

Source: Authors' regressions.

a. Each reported coefficient is from a separate regression, based on roughly 140,000 district-year observations, that identifies the effect of NCLB as of 2008. See table 1 and the text for details. Standard deviations or standard errors clustered by state are in parentheses. Asterisks indicate statistical significance at the ***1 percent, **5 percent, or *10 percent level.

exercises and alternative specifications described earlier for the achievement analysis (Dee, Jacob, and Schwartz 2010).[17]

In light of the achievement effects discussed in the previous section, a natural and policy-relevant question is to ask how the monetized benefits of those test-score gains compare with the corresponding expenditure increases presented here. On the basis of prior estimates that a 1-standard-deviation increase in elementary math scores is associated with an 8 percent increase in adult earnings (Krueger 2003), the 0.23-standard-deviation impact of NCLB would translate into a lifetime earnings boost of 1.8 percent. Assuming a 3 percent discount rate, the present discounted value as of age 9 of such an increase beginning at age 18 is at least $13,300.[18] Hence, even if we assume that the increased expenditure due to NCLB is

17. As discussed in related work, we do not find substantial impacts on class size, suggesting that the increase in instructional expenditure due to NCLB may have been allocated to other functions (Dee, Jacob, and Schwartz 2010).

18. This calculation uses an age-earnings profile of 18- to 65-year-olds taken from the March 2007 Current Population Survey. Allowing for reasonable productivity-related growth in earnings of 2 percent a year increases the monetized benefit of the test-score gains due to NCLB to roughly $25,500.

sustained for all eight elementary-school years, the economic benefits of the corresponding test-score gains are at least twice as large. It should be stressed, however, that this exercise turns on multiple unstated assumptions. In particular, this back-of-the-envelope calculation ignores certain socially relevant benefits (such as the externalities of human capital improvements) and costs (such as the deadweight losses associated with raising government revenue to pay for the added spending). More generally, it is not clear that these expenditure increases were even a relevant mediating mechanism behind NCLB's achievement effects. Nonetheless, this calculation provides suggestive evidence that the achievement gains attributable to NCLB may compare favorably with the corresponding spending increases.

III.B. Impact on Teachers and Classrooms

One of the most prominent issues raised by NCLB concerns the intended and unintended ways in which it may have influenced classroom practice. In particular, test-based accountability policy creates a strong incentive for educators to focus on tested content and skills. Indeed, according to many, this is precisely the point of the reform. But at the same time, critics have worried that such incentives may cause schools to neglect important but nontested subjects, or to change instructional practice in a way that prioritizes narrow test preparation over broader learning. In this section we discuss the available evidence on how school accountability programs, including NCLB, influence classroom instruction.

The most consistent and compelling finding with regard to school accountability and classroom instruction involves the allocation of instructional time. A number of studies have documented that test-based accountability programs cause educators to reallocate instructional time toward tested subjects, to reallocate time within tested subjects toward specific content and skills covered on the exam, and to increase time devoted to narrow test preparation activities that may have little broader value (Hannaway and Hamilton 2008).

In 2001, for example, researchers at the National Board on Educational Testing and Public Policy surveyed a nationally representative sample of teachers, asking them a series of questions about how state-mandated testing programs influenced their practice (Pedulla and others 2003). Teachers in states where the exam results were used to hold teachers or schools accountable reported shifting instruction toward tested subjects more than did teachers in states where the exam results were used primarily for informational purposes. For example, 34 percent of teachers working in high-stakes testing regimes, but only 17 percent of teachers in moderate-stakes

regimes, reported that the testing program had led them to increase the time spent in tested areas "a great deal." In addition, teachers in states with school accountability programs reported spending more time on a variety of activities designed to improve student test-taking skills, such as taking practice tests (Pedulla and others 2003). In states where the tests had important consequences for the schools, roughly 36 percent of elementary teachers reported spending more than 30 hours per year on test preparation activities, compared with only 12 percent of teachers in states where tests had few consequences for schools.[19]

Recent studies that focus on NCLB itself find similar results. In 2005, for example, researchers at the RAND Corporation collected data from teachers, principals, and superintendents in three states (California, Georgia, and Pennsylvania) to examine how they were responding to the introduction of NCLB (Hamilton and others 2007). Educators reported a narrowing of the curriculum and an emphasis on test preparation, particularly for "bubble kids" near the proficiency cut score for their state assessment system. In addition, educators responded to NCLB by increasing the alignment between the curriculum and state standards.

Studies of earlier school accountability programs found a similar increase in alignment. The programs led teachers to shift the content of their instruction within subjects (Stecher and others 1998, Koretz and others 1996, Jacob 2005, Koretz and Hamilton 2006). This literature emphasizes that the format and structure of the test itself can influence instruction. For example, Grace Taylor and others (2003) find that testing programs with short, open-ended items lead teachers to focus greater attention on problem-solving skills.

The Center on Education Policy (CEP) has studied the implementation and impact of NCLB since its inception (CEP 2006, 2007, 2008a, 2008b). As part of its work, CEP not only surveyed a nationally representative sample of school districts in 2005–06 and again in 2006–07, but also conducted more intensive case studies of selected school districts. District officials report that NCLB led them to increase the instructional time their schools devote to math or English language arts (ELA) or both. About 62 percent of districts reported that between 2001–02 and 2006–07 they increased instruction in these subjects in elementary schools, with the

19. Ladd and Zelli (2002) found similar results in a survey study of school principals in North Carolina during the period when the state was introducing its school accountability program. Principals reported devoting more resources to the high-stakes subjects of math, reading, and writing.

largest increases in districts with more schools in need of improvement (CEP 2007) and in urban and high-poverty districts (CEP 2006). Moreover, the reallocation reported by officials appears substantial. For example, 80 percent of districts that reported increasing ELA time did so by at least 75 minutes per week, and 54 percent reported doing so by at least 150 minutes per week (CEP 2008). Most districts that reported increased time for ELA or math reported cuts in time for other subjects or periods (such as social studies, art, music, gym, recess, or lunch) rather than increases in total time in school (CEP 2008).[20]

The CEP studies also suggest that NCLB influenced classroom practice in ways that may have attenuated teacher autonomy. For example, CEP (2006, 2007) reports that schools made a concentrated effort to align their curriculum with state standards in the wake of NCLB, thus changing the focus of their curriculum to put greater emphasis on tested content and skills. Many districts also became more prescriptive during this period about what and how teachers were supposed to teach (CEP 2006).

It is worth noting that the costs and benefits of these instructional changes depend on one's objectives and are not always clear even for a given objective. For example, many observers applaud the increasing emphasis on math and reading instruction, while others lament the decreasing attention on subjects such as art and music (Rothstein and others 2008).

Although these studies paint a consistent picture, they need to be interpreted with some caution. All of the research described thus far relies on self-reports from teachers or administrators. Moreover, the information is based on questions that ask respondents to retrospectively assess whether certain practices have changed over time. For this reason, one might be worried about the reliability and validity of the data (Bradburn and Sudman 1988).

Few studies have implemented regression-based research designs that attempt to isolate the effects of school accountability policies on district, school, and classroom practices from the potentially confounding effects of other determinants. One prominent exception is a recent study by Cecilia Elena Rouse and others (2007), which used a regression-discontinuity

20. A 2009 Government Accountability Office study based on teacher survey data (and supplemental interviews with state officials) finds that 90 percent of elementary teachers reported no change in instructional time for arts education between 2004–05 and 2006–07. At the same time, a larger fraction of teachers in schools identified as needing improvement under NCLB reported a decline in art instruction, relative to teachers in other schools. This study used data from the Department of Education's National Longitudinal Study of No Child Left Behind (NLS-NCLB).

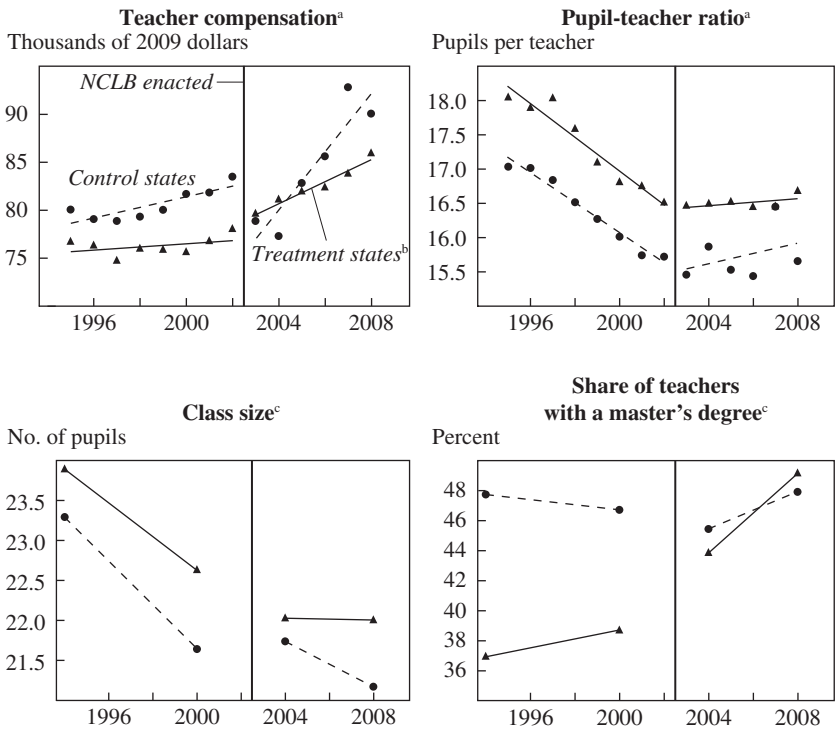
design and data from surveys of principals in Florida to examine how schools responded to pressure from the state's accountability system. They find that accountability pressure leads to an increased emphasis on low-performing students (through grade retention, summer school, and tutoring, for example), increased overall instructional time, and reorganized school days. They also find suggestive evidence that accountability reduced principal control and increased the resources available to teachers. Furthermore, the school policies influenced by school accountability explain a meaningful fraction of student test-score gains, suggesting that schools responded to accountability pressure in specific ways that improved student achievement.

Although the work just summarized addresses some of the concerns raised in previous work, it has its own set of limitations. It does not address NCLB per se, and it estimates what one might describe as the partial impact of the Florida accountability system, comparing schools more or less affected by accountability pressure. However, it is possible that the accountability system in Florida, or NCLB more generally, led to changes across all schools.

In recent work we present new evidence on these issues using data from the nationally representative Schools and Staffing Survey (SASS) and the state-based CITS research design (Dee, Jacob, and Schwartz 2010). The SASS is a nationally representative survey of teachers and school administrators that has been conducted periodically since the early 1990s (in 1994, 2000, 2004, and 2008).[21] We use teacher responses from the survey to construct a variety of measures of classroom instruction and school organization. These data allow us to compare changes in teacher responses over time rather than rely on the teachers' retrospective judgments. They also provide more objective measures of some of the constructs: for example, the time-use questions ask about the actual number of hours per week a teacher devotes to math, rather than asking teachers to characterize their emphasis on math as "big" or "small" or whether it is greater or less than it was a certain number of years ago.

21. Because the pooled SASS data contain data from only two pre-NCLB periods, Dee, Jacob, and Schwartz (2010) also examined the robustness of the SASS-based models to the use of conventional difference-in-differences specifications. The results were quite similar to the CITS results, with the modest exception of one result discussed below. That paper also presents falsification exercises similar to those presented for the NAEP and F-33 models, the results of which generally suggest that the CITS specification based on the SASS data generates internally valid estimates.

Figure 7. School Resources by Timing of Increased School Accountability, 1994–2008



Source: Authors' calculations using data from the Common Core of Data (top panels) and the Schools and Staffing Survey (bottom panels).

a. Sample is composed of all noncharter, unifed local education agency school districts, excluding Hawaii, the District of Columbia, and zero-enrollment districts. Estimates are weighted by district enrollment. Teacher compensation includes the value of noncash employee benefits.

b. Treatment and control states are defined as in figure 4.

c. Sample consists of full-time elementary- and middle-school teachers with a main assignment in mathematics, English language arts, or general elementary instruction.

The top left panel of figure 7 shows the comparative trends in real teacher compensation by year across treatment and control states. As in the previous figures, we show the trends separately for states that did and did not adopt school accountability programs before NCLB. These district-level data indicate that, after the introduction of NCLB, average annual teacher compensation increased distinctly, from roughly $75,000 to over $80,000, in states that did not have prior school accountability. However, this graph also suggests an NCLB effect: this compensation growth was particularly large relative to the corresponding changes in

states that had school accountability regimes before NCLB. The bottom right panel of figure 7 shows changes over time in the fraction of elementary- and middle-school teachers with a master's degree (based on the pooled SASS data) and similarly suggests an NCLB effect. In states with prior accountability programs, roughly 47 percent of teachers had a master's degree in 1994, compared with 37 percent of teachers in other states. Following the introduction of NCLB, the fraction of teachers with a master's degree jumped notably in states without prior accountability, so that in 2004 the rates were approximately equal across both groups of states. By 2008, teachers in states without prior accountability were slightly more likely to have a master's degree than their counterparts in other states. In contrast, the top right and bottom left panels of figure 7 provide no clear indication that NCLB influenced class sizes or pupil-teacher ratios, respectively (see also Dee, Jacob, and Schwartz 2010).

Table 4 presents regression estimates of these effects based on the CITS model in equation 1, adding a variety of controls for teacher, school, and district observed traits (Dee, Jacob, and Schwartz 2010). As above, standard errors are clustered by state. The results indicate that NCLB increased average teacher compensation by over $5,000, or by roughly 8 percent relative to the pre-NCLB mean of $79,577. The table also indicates that by 2008, NCLB had increased the fraction of teachers with a master's degree by roughly 0.056, from a baseline of 0.41, an increase of roughly 14 percent. Given that many districts require teachers to have a master's degree for permanent certification, it is possible that this effect reflects the response of states to the NCLB provision requiring schools to have "highly qualified" teachers in every classroom. The fact that states with prior accountability policies also had a substantially larger fraction of teachers with a master's degree suggests that programs adopted by states before NCLB may have contained some provisions regarding teacher qualifications.

The top right and bottom left panels of figure 8 show trends in time use for our sample of elementary-school teachers and principals for states that did and those that did not adopt school accountability programs before NCLB. The top right panel shows the amount of instructional time (in hours per week) that teachers report for core academic subjects. The bottom left panel shows the fraction of time during the week that self-contained teachers (those who provide instruction in multiple subjects to a single group of students) teach math and ELA, where the denominator is total time spent on the four core subjects (math, ELA,

Table 4. Estimated Effects of NCLB on School Resources, Allocation of Instructional Time, and Educational Climate

Dependent variable	*Mean for 1999–2000 school year*	*Estimated impact of NCLB*[a]
School resources		
Teacher compensation (dollars)[b]	79,577	5,067*
	(20,338)	(2,888)
Pupil-teacher ratio[b]	16.986	–0.151
	(2.692)	(0.491)
Class size	22.120	–0.328
	(4.990)	(0.500)
Fraction of teachers with master's degree	0.412	0.056**
	(0.492)	(0.028)
Instructional time		
Hours per week spent on core academic subjects[c]	21.758	–0.307
	(6.445)	(0.684)
Fraction of total hours spent on math and English	0.737	0.036***
	(0.130)	(0.012)
Fraction of total hours spent on English	0.476	0.023*
	(0.156)	(0.013)
Educational climate		
Fraction of schools where principal places highest priority on academic excellence or basic skill acquisition	0.875	–0.003
	(0.331)	(0.037)
Teachers' perceptions of school discipline	–0.003	0.074
	(0.989)	(0.115)
Teachers' perceptions of student engagement	0.059	0.220***
	(0.990)	(0.056)

Source: Authors' regressions.

a. Each reported coefficient is from a separate regression and identifies the effect of NCLB as of 2008. See table 1 and the text for details. Except where noted otherwise, estimates use pooled data from the Schools and Staffing Survey. Standard deviations or standard errors clustered by state are in parentheses. Asterisks indicate statistical significance at the ***1 percent, **5 percent, or *10 percent level.

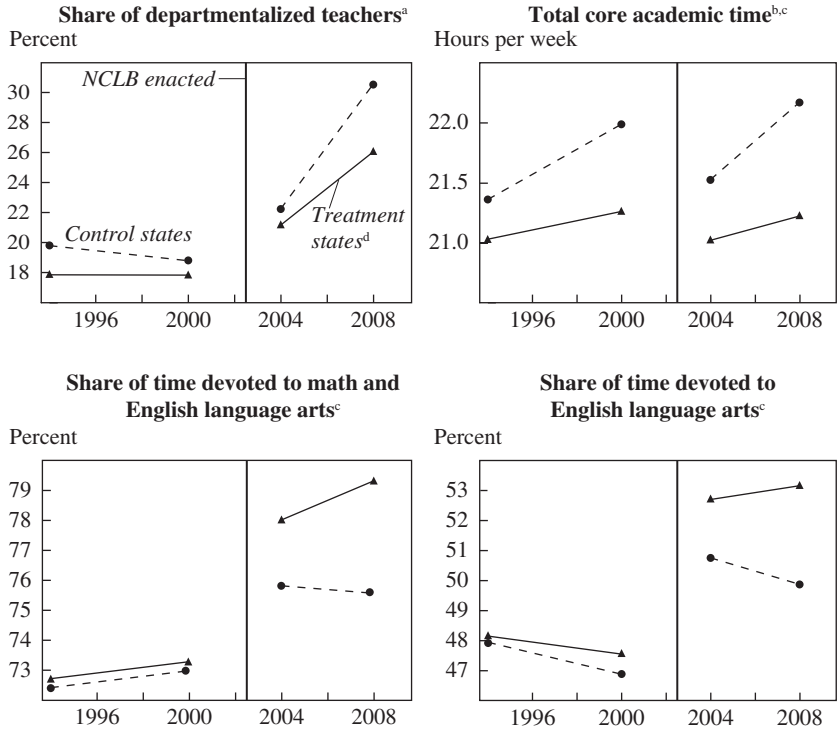
b. Estimates use data from the Common Core of Data's Local Education Agency (School District) Finance Survey from the National Center for Education Statistics. Teacher compensation includes the value of noncash employee benefits.

c. Mathematics, English language arts, social studies, and science.

social studies, and science). The bottom right panel of the figure shows this ratio for ELA alone.

These figures suggest that NCLB did *not* lead to meaningful increases in the total amount of instructional time devoted to core subjects, but that instructional time allocated to math and ELA increased following the introduction of NCLB. Moreover, the effects seem to be larger in states that had not previously instituted school accountability, consistent with NCLB leading to this change.

Figure 8. School Allocation of Time, by Timing of Increased School Accountability, 1994–2008



Source: Authors' calculations using data from the Schools and Staffing Survey.

a. Departmentalized teachers (those who instruct several classes of different students, usually in the same subject) as a share of all (departmentalized, self-contained, and team) teachers.

b. Time devoted to mathematics, English language arts, social studies, and science.

c. Sample consists of classes taught by self-contained and team teachers only.

d. Treatment and control states are defined as in figure 4.

Estimates reported in the middle panel of table 4 indicate that NCLB increased the fraction of time that teachers spend on math and ELA by 3.6 percentage points, relative to a baseline of 74 percent. This implies an additional 45 minutes per week of math and ELA instruction by teachers who spend 20 instructional hours on these two subjects. It appears that this increase was driven primarily by an increase in time devoted to ELA: table 4 indicates that NCLB increased the share of instructional time devoted to ELA by a weakly significant 2.3 percentage points. In contrast, we do not find that NCLB had statistically significant effects

on the fraction of time devoted to math. This is particularly interesting given that we find substantial achievement effects in math but not in reading.

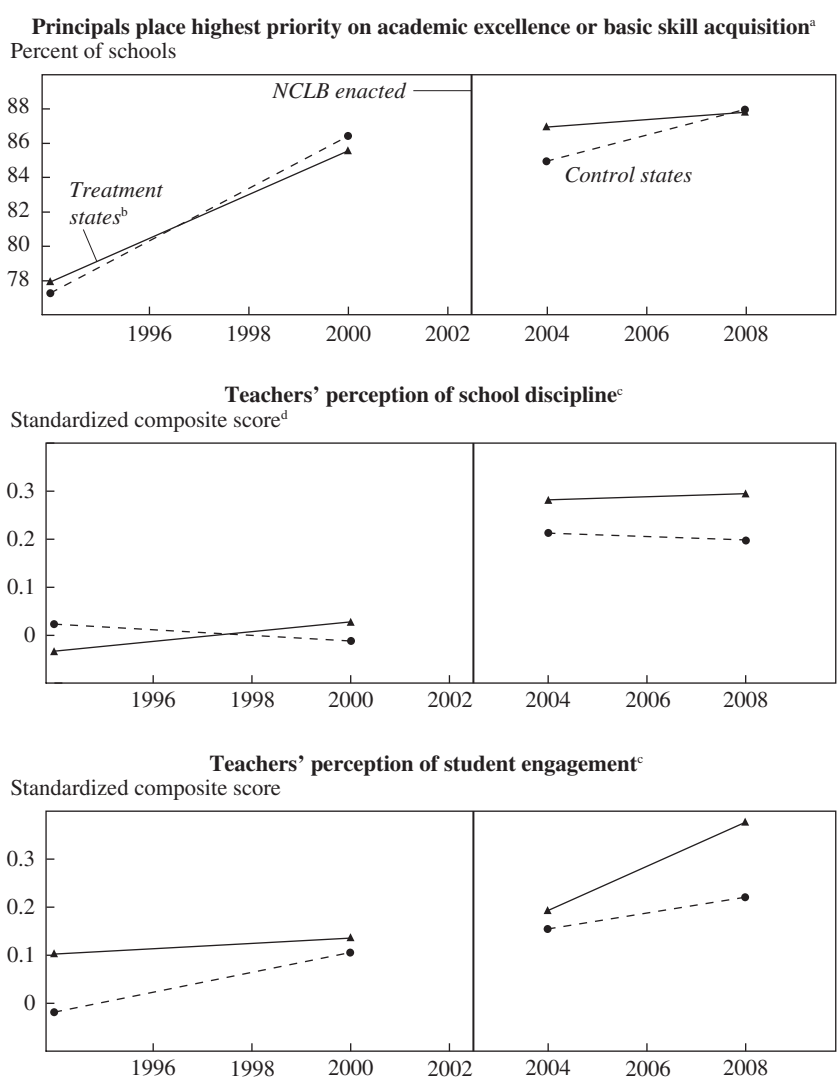
III.C. Impacts on School Organization, Climate, and Culture

The literature provides some evidence that test-based accountability policies, including NCLB, have spurred other useful changes in school-wide instructional practice. In the RAND study cited above (Hamilton and others 2007), for example, school and district administrators reported that NCLB increased the use of diagnostic assessment (exams used by teachers to determine a student's areas of strength and weakness) as an instructional tool and increased the technical assistance and professional development opportunities offered to schools. In earlier survey work, researchers found that teachers in high-stakes environments found test results more useful, and were more likely to use test information to inform their practice, than colleagues in low-stakes environments (Pedulla and others 2003). Similarly, teachers in the RAND study reported that their state's accountability system under NCLB led them to search for more effective teaching practices and, in nearly all cases, had led to positive changes in their schools (Hamilton and others 2007). Interestingly, for example, teachers reported that teaching practices and the general focus on student learning "changed for the better" under accountability. District officials in the CEP study reported an increase in the use of data to guide instruction (CEP 2006).

Unfortunately, the SASS has not routinely collected data on many of the school and teacher practices that are of interest, and this limits our capacity to isolate the effects of NCLB on some of these outcomes. However, the SASS has collected consistent data on several relevant school-level traits. For example, the principals who responded to the SASS were asked to choose their top three priorities from a list of nine educational goals. The top panel of figure 9 shows the comparative trend data for the share of principals who indicated that either academic excellence or acquisition of basic skills was their top goal. States with and without prior accountability did not converge on this measure of instructional focus after NCLB. This pattern suggests that NCLB did not generate a detectable increase in instructional focus, a result confirmed by the regression results in table 4.

Teachers surveyed in the SASS provided scaled responses to questions about whether principals and fellow teachers enforced rules for student conduct. The middle panel of the figure shows the comparative trends for the standardized responses to this question and suggests the lack of an

Figure 9. School Culture Outcomes by Timing of Increased School Accountability, 1994–2008



Source: Authors' calculations using data from the Schools and Staffing Survey.
a. Sample is limited to full-time elementary- and middle-school principals.
b. Treatment and control states are defined as in figure 4.
c. Sample is defined as in figure 7.
d. Higher scores indicate greater enforcement of rules or greater engagement.

NCLB effect (again confirmed by the results in table 4). The bottom panel shows the trend data for a standardized measure of how teachers viewed their students' school-relevant behavior and attitudes. This measure, an index of "behavioral engagement" with school, standardized to have a mean of zero and a standard deviation of 1, reflects the extent to which teachers feel that traits such as tardiness, absenteeism, and apathy are not a problem within their school. Here there seems to have been comparative improvement in this measure of student engagement for states that introduced school accountability because of NCLB. The regression results in table 4 indicate that the size of this increase is 0.22 base-year (1994) standard deviation.

This estimated effect on student engagement is over twice as large in high-poverty schools. However, Dee, Jacob, and Schwartz (2010) find that NCLB's estimated effects on this engagement measure are noticeably smaller (an effect size of 0.094) in a difference-in-differences specification, which does not condition on pretreatment trends unique to treatment status. We cannot definitively establish whether the CITS or the difference-in-differences specifications generate more accurate point estimates in this context. However, the differences in pre-NCLB student engagement trends across treatment and control states depicted in figure 9 are consistent with the motivation for the CITS specification. A very different caveat to this result is that NCLB's apparent effect on teacher-reported student engagement could simply be due to policy-induced changes in teacher expectations. For example, to the extent that NCLB increased the expectations for academic achievement in states without prior school accountability policies, it is possible that teachers simultaneously chose to benchmark the behavioral engagement of their students with school against a more lax standard. If this is true, our estimate of the impact on behavioral engagement is biased upward.

III.D. Summary

The evidence presented above suggests that NCLB has had both desirable and undesirable effects on school district spending, teachers, classroom practice, and school culture. Unfortunately, the lack of objective measures for several important instructional practices limits our ability to examine many of the most plausible mechanisms through which accountability may have operated to improve student achievement. Moreover, the analysis here does not allow us to identify which, if any, of the factors we identify as improving (for example, per-pupil spending, student engagement, teacher qualifications, or instruction time devoted to English) might

explain the achievement effects we document. Nonetheless, this analysis provides important evidence on the policy-relevant, nonachievement consequences of NCLB (for example, its fiscal effects) as well as guideposts to the intended and unintended ways in which NCLB has shaped the available measures of educational practice.

IV. Conclusions

Eight years have passed since No Child Left Behind dramatically expanded the federal role in public schooling. Given the national scope of the policy, it is difficult to reach definitive conclusions about its impact. Nonetheless, evidence from a variety of data sources utilizing several plausible comparison groups suggests that NCLB has had a positive effect on elementary student performance in mathematics, particularly at the lower grades. The benefits appear to be concentrated among traditionally disadvantaged populations, with particularly large effects among Hispanic students. On the other hand, the existing evidence suggests that NCLB has not had a comparable effect on reading performance.

We find compelling evidence that NCLB increased per-pupil school district expenditure, particularly on direct instruction, a mediating mechanism that may explain the corresponding achievement gains. By 2008, for example, the policy appears to have increased annual spending per pupil by nearly $600 in states that did not have any school accountability program before NCLB; these increased outlays were not supported by corresponding increases in federal revenue. We also presented evidence that these expenditure increases may be modest relative to the present discounted value of the corresponding test-score gains. We also discussed evidence suggesting that NCLB influenced teachers and schools in several potentially important ways. It appears that NCLB has led elementary schools to increase instructional time devoted to math and reading, although the majority of evidence on this point comes from teacher and administrator survey data that are subject to potential bias. Similarly, teachers report that NCLB has encouraged schools to spend time on narrow test preparation activities. However, we also found evidence that NCLB led to increases in teacher-reported measures of students' behavioral engagement with school. Unfortunately, a lack of a richly detailed dataset that lends itself to credible identification strategies makes it difficult to assess whether NCLB influenced curriculum and instructional practices in more fundamental ways.

Nonetheless, the extant body of evidence can provide some guidance to the ongoing debate over the proposed reauthorization of NCLB. In March

2010 the Obama administration released an NCLB "blueprint" that outlined proposed features of a reauthorization (Klein and McNeil 2010). This proposal calls for continued annual reporting of school-level, test-based student assessments but allows for flexibility in how states calculate school effectiveness. The blueprint also calls for the use of nontest accountability indicators, especially measures of college and career readiness (such as attendance, course completion, and school climate). Another potentially critical feature of this proposal involves changing how measures of school performance are linked to consequences.

The blueprint also proposes to give states increased flexibility in how they might intervene in low-performing schools, mandating specific consequences only for the very lowest-performing schools and those with persistently large achievement gaps. It is not clear how states would respond to this added flexibility. However, the literature on pre-NCLB accountability policies suggests that simply reporting accountability measures that were unconnected to explicit consequences did not drive improvements in student achievement (Hanushek and Raymond 2005). This suggests that the targeted achievement gains attributable to NCLB could be at risk under state reforms that decouple performance measures from meaningful consequences.

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 Brookings Panel conference, a conference at the Harris School of Public Policy Studies at the University of Chicago, and at the NCLB: Emerging Findings Research Conference at the CALDER Center of the Urban Institute for helpful comments. An earlier version of this work was presented by Brian Jacob as the David N. Kershaw Lecture at the annual meeting of the Association of Public Policy Analysis and Management, November 2008. All errors are our own.

The authors report no relevant conflicts of interest.

References

Ballou, Dale, and Matthew Springer. 2009. "Achievement Trade-offs and No Child Left Behind." Washington: Urban Institute.

Bradburn, Norman M., and Seymour Sudman. 1988. *Polls and Surveys: Understanding What They Tell Us.* San Francisco: Jossey-Bass.

Carnoy, Martin, and Susanna Loeb. 2002. "Does External Accountability Affect Student Outcomes? A Cross-State Analysis." *Educational Evaluation and Policy Analysis* 24, no. 4 (Winter): 305–31.

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

———. 2007. "Choices, Changes, and Challenges: Curriculum and Instruction in the NCLB Era." A report in the series From the Capital to the Classroom: Year 5 of the No Child Left Behind Act. Washington.

———. 2008a. "Instructional Time in Elementary Schools: A Closer Look at Changes for Specific Subjects." A report in the series From the Capital to the Classroom: Year 5 of the No Child Left Behind Act. Washington.

———. 2008b. "Has Student Achievement Increased since 2002? State Test Score Trends through 2006–07." Washington.

Dee, Thomas, and Brian Jacob. Forthcoming. "The Impact of No Child Left Behind on Student Achievement." *Journal of Policy Analysis and Management.*

Dee, Thomas S., Brian A. Jacob, and Nathaniel L. Schwartz. 2010. "The Effect of No Child Left Behind on School Finance, Organization and Practice." Working paper. University of Virginia and University of Michigan.

Figlio, David N., and Helen F. Ladd. 2007. "School Accountability and Student Achievement." In *Handbook of Research in Education Finance and Policy,* edited by Helen F. Ladd and Edward B. Fiske. New York and London: Routledge.

Fuller, Bruce, Joseph Wright, Kathryn Gesicki, and Erin Kang. 2007. "Gauging Growth: How to Judge No Child Left Behind?" *Educational Researcher* 36, no. 5: 268–78.

Goertz, Margaret E., Mark C. Duffy, and Kerstin Carlson Le Floch. 2001. "Assessment and Accountability Systems in the 50 States: 1999–2000." CPRE Research Report no. RR-046. Philadelphia: Consortium for Policy Research in Education.

Hamilton, Laura S., Brian M. Stecher, Julie A. Marsh, Jennifer Sloan McCombs, Abby Robyn, Jennifer Lin Russell, Scott Naftel, and Heather Barney. 2007. *Standards-Based Accountability under No Child Left Behind: Experiences of Teachers and Administrators in Three States.* Santa Monica, Calif.: RAND Corporation.

Hannaway, Jane, and Laura Hamilton. 2008. "Performance-Based Accountability Policies: Implications for School and Classroom Practices." Washington: Urban Institute and RAND Corporation.

Hannaway, Jane, and Maggie Stanislawski. 2005. "Responding to Reform: Florida School Expenditures in the 1990s." Urban Institute and Colorado State University.

Hannaway, Jane, Shannon McKay, and Yasser Nakib. 2002. "Reform and Resource Allocation: National Trends and State Policies." In *Developments in School Finance, 1999–2000,* edited by William J. Fowler Jr. Washington: U.S. Department of Education, National Center for Education Statistics.

Hanushek, Eric A., and Margaret E. Raymond. 2001. "The Confusing World of Educational Accountability." *National Tax Journal* 54, no. 2: 365–84.

———. 2005. "Does School Accountability Lead to Improved Student Performance?" *Journal of Policy Analysis and Management* 24, no. 2: 297–327.

Hoxby, Caroline M. 2002. "The Cost of Accountability." In *School Accountability,* edited by Williamson M. Evers and Herbert J. Walberg. Stanford, Calif.: Hoover Institution Press.

Jacob, Brian A. 2005. "Accountability, Incentives and Behavior: The Impact of High-Stakes Testing in the Chicago Public Schools." *Journal of Public Economics* 89, no. 5–6: 761–96.

Jacob, Brian A., and Steven D. Levitt. 2003. "Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating." *Quarterly Journal of Economics* 118, no. 3: 843–77.

Klein, Alyson, and Michele McNeil. 2010. "Administration Unveils ESEA Reauthorization Blueprint." *Education Week* (March 17).

Koretz, Daniel. 2008. *Measuring Up: What Educational Testing Really Tells Us.* Harvard University Press.

Koretz, Daniel M., and Laura S. Hamilton. 2006. "Testing for Accountability in K-12." In *Educational Measurement,* 4th ed., edited by Robert L. Brennan. Westport, Conn.: American Council on Education/Praeger.

Koretz, Daniel M., Sheila Barron, Karen J. Mitchell, and Brian M. Stecher. 1996. *The Perceived Effects of the Kentucky Instructional Results Information System (KIRIS).* MR-792-PCT/FF. Santa Monica, Calif.: RAND Corporation.

Krieg, John M. 2008. "Are Students Left Behind? The Distributional Effects of the No Child Left Behind Act." *Education Finance and Policy* 3, no. 2: 250–81.

Krueger, Alan B. 2003. "Economic Considerations and Class Size." *Economic Journal* 113: F34–F63.

Ladd, Helen F. 2007. "Holding Schools Accountable Revisited." 2007 Spencer Foundation Lecture in Education Policy and Management. Washington: Association for Public Policy Analysis and Management. www.appam.org/awards/pdf/2007Spencer-Ladd.pdf (accessed November 8, 2009).

Ladd, Helen F., and Arnaldo Zelli. 2002. "School Based Accountability in North Carolina: The Responses of School Principals." *Education Administration Quarterly* 38, no. 4: 494–529.

Lee, Jaekyung, and Kenneth K. Wong. 2004. "The Impact of Accountability on Racial and Socioeconomic Equity: Considering Both School Resources and Achievement Outcomes." *American Educational Research Journal* 41, no. 4: 797–832.

Neal, Derek, and Diane Whitmore Schanzenbach. 2010. "Left Behind by Design: Proficiency Counts and Test-Based Accountability." *Review of Economics and Statistics* 92, no. 2: 263–83.

Palmer, Scott R., and Arthur L. Coleman. 2003. "The No Child Left Behind Act of 2001: Summary of NCLB Requirements and Deadlines for State Action." Washington: Council of Chief State School Officers (November). events.ccsso.org/content/pdfs/Deadlines.pdf (accessed November 13, 2009).

Pedulla, Joseph J., Lisa M. Abrams, George F. Madaus, Michael K. Russell, Miguel A. Ramos, and Jing Miao. 2003. *Perceived Effects of State-Mandated Testing Programs on Teaching and Learning: Findings from a National Survey of Teachers.* Chestnut Hill, Mass.: National Board on Educational Testing and Public Policy.

Reback, Randall, Jonah E. Rockoff, and Heather L. Schwartz. 2010. "The Effects of No Child Left Behind on School Services and Student Outcomes." Working paper. Barnard College and Columbia University.

Rothstein, Richard, Rebecca Jacobsen, and Tamara Wilder. 2008. "Grading Education: Getting Accountability Right." New York: Teachers College Press.

Rouse, Cecilia Elena, Jane Hannaway, Dan Goldhaber, and David Figlio. 2007. "Feeling the Florida Heat? How Low-Performing Schools Respond to Voucher and Accountability Pressure." Cambridge, Mass.: National Bureau of Economic Research.

Stecher, Brian M., Sheila Barron, Tessa Kaganoff, and Joy Goodwin. 1998. "The Effects of Standards-Based Assessment on Classroom Practices: Results of the 1996–97 RAND Survey of Kentucky Teachers of Mathematics and Writing." CSE Technical Report 482. National Center for Research on Evaluation, Standards, and Student Testing, University of California, Los Angeles.

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

Taylor, Grace, Lorrie Shepard, Freya Kinner, and Justin Rosenthal. 2003. "A Survey of Teachers' Perspectives on High-Stakes Testing in Colorado: What Gets Taught, What Gets Lost." CSE Technical Report 588. National Center for Research on Evaluation, Standards, and Student Testing, University of California, Los Angeles.

U.S. Department of Education, Office of Planning, Evaluation and Policy Development, Policy and Program Studies Service. 2007. "Private School Participants in Federal Programs under the No Child Left Behind Act and the Individuals with Disabilities Education Act: Private School and Public School District Perspectives." Washington.

U.S. Government Accountability Office. 2009. "Access to Arts Education." Report to Congressional Requesters. GAO-09-286. Washington (February).

Wong, Manyee, Thomas D. Cook, and Peter M. Steiner. 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 no. WP-09-11. Institute for Policy Research, Northwestern University.

Comments and Discussion

COMMENT BY

CAROLINE M. HOXBY Thomas Dee and Brian Jacob provide a review of existing empirical studies on how No Child Left Behind (NCLB) has affected student achievement. They also present original findings based on a difference-in-differences comparison of states that implemented school accountability only with NCLB and those that had implemented it previously. The difference-in-differences work relies on test scores from the National Assessment of Educational Progress (NAEP), the only examination regularly administered to samples of students representative of each state.

In their review of existing studies, Dee and Jacob concisely, yet accurately, summarize the most credible research on NCLB and the state accountability laws that preceded it. Much of the existing evidence is lacking in rigor or partial in nature, and the authors do an excellent job of differentiating between stronger and weaker results. However, Dee and Jacob's difference-in-differences work is the meat of their paper, and thus these comments focus on it.

Of all the areas within education in which the federal government plays a role, that in primary and secondary education is by far the most minor. Whereas each of the states has a responsibility for education written into its constitution and has a long history of financing and overseeing education, the federal government has traditionally confined itself to funding a small percentage of the education of poor and disabled students and those with limited knowledge of English. At no time before the Obama administration did the federal government account for more than 10 percent of public spending on primary and secondary education, and it has no formal role in the governance of any regular public school. Simply put, the federal

government is a very junior partner in education policy; states and local school districts occupy the driver's seat.

This structure of financing and control has two important implications for all federal education policy. First, it is state governments and large school districts that set the pace and frontier in education policy. Second, the federal government has little ability to enforce its policy will: even when federal law mandates some education policy, it is states and districts that must implement it. And they are quite capable of implementing only the letter of the law, and none of its intent. These two implications are glaringly obvious in the history of NCLB.

Starting in the late 1980s, a number of states began to enact laws designed to hold their schools accountable for overall student outcomes. They instituted mandatory statewide exams and published the results prominently in the media and in "school report cards" sent to parents. They devised grading systems for their schools, which gave better grades to schools with higher test scores and graduation rates. The more sophisticated state grading systems incorporated value-added calculations for schools based on test scores, regression adjustments for sociodemographics, and weights on a variety of complex outcomes beyond scores. Schools that earned poor grades experienced interventions, sometimes welcome and sometimes not, from state departments of education. Chronically failing schools were reconstituted by some states. The states that led the accountability movement created powerful databases that allowed them to track students and teachers longitudinally, certified teachers through nontraditional routes (such as proficiency testing), rewarded teachers and schools they deemed successful, and established very modest forms of school choice.

The accountability movement was led by an informal but like-minded group of governors and chief state school officers, prominent among whom were Governor George W. Bush of Texas and Governor Jeb Bush of Florida. Both Bushes would later point to education accountability as their main legacy to the states they governed. Thus, it should come as no surprise that when NCLB was drafted as one of the first major policies of the George W. Bush presidency, the first draft looked a lot like Texas' and Florida's accountability policies. The final version, however, was a very dilute law, reflecting numerous political compromises and the federal government's negligible powers of enforcement.

NCLB was written in such a way that states antagonistic to school accountability could easily evade every aspect of the law. For instance, by choosing a test on which nearly all students did well from the start, a state could ensure that all its schools met the proficiency standard and thereby

escaped all consequences of failure. (One state, Nebraska, got away with choosing no statewide test at all.) A state could comply with NCLB's reporting provisions by putting its school report cards on an obscure website with little or no functionality. Some websites deliberately prevented parents and journalists from comparing schools' reports. States could ensure that all their teachers met the "highly qualified" mandates of NCLB, which were meant to ensure subject area knowledge, simply by setting standards that automatically made all their existing teachers "highly qualified." For instance, some states declared all teachers who were experienced or certified to be highly qualified regardless of their proficiency or area of teaching. Districts that wished to evade the (very modest) school choice provisions of NCLB could refrain from notifying parents that their children were eligible to transfer to another public school and from telling parents that they could use their federal dollars for private tutoring services.

In short, NCLB was a peculiar law. It could have little effect on states antagonistic to school accountability. Rather, it was a nudge to states that supported accountability enough to welcome a nudge but not enough to have enacted accountability before NCLB. States that had already adopted pro-accountability measures viewed NCLB as a drag on their more ambitious, more sophisticated policies. For instance, Jeb Bush's Florida wrestled with the U.S. Department of Education, trying to get waivers from the federal school grading system, which was manifestly inferior to its own. Several states, especially Massachusetts, argued that NCLB gave them strong incentives to reduce the quality of their tests.

This is the backdrop to Dee and Jacob's analysis, and it is one that causes their difference-in-differences estimates to be quite unreliable. Their estimates almost certainly understate the effects in which policymakers are interested. My logic on this has three different components. First, Dee and Jacob's "control" states did not hold their policies constant, and so do not amount to true counterfactuals. Second, NCLB could be expected to have heterogeneous treatment effects depending on a state's willingness to implement it, and the difference-in-differences analysis does not account for this. Third, policymakers are interested in population-average effects, which the difference-in-differences analysis substantially understates. Let me take each of these in turn.

Dee and Jacob describe as "treated by NCLB" those states that had no school accountability program before the 2001 law. Their "control" states are those that had implemented their own accountability programs before 2001. Their difference-in-differences method identifies the effects of NCLB by comparing the pre-versus-post-NCLB achievement change in

treated states with the pre-versus-post achievement change in control states. There are a few more details to the equation they estimate, but this comparison is the key to their identification strategy.

In a credible difference-in-differences exercise, the control states must reveal what would have occurred in the treated states had NCLB not been enacted. This they do not do in the Dee and Jacob exercise. The control states are the pro-accountability states that had already enacted more ambitious programs than NCLB and that continued to extend and improve their programs in the wake of its enactment. These states rolled out improved school grading systems, estimated the value added of individual teachers, and instituted rewards for individual students to excel (such as scholarships) and remedies for students who did not (such as summer school and grade retention). Many of the control states actually increased accountability more in the immediate aftermath of NCLB than did the treated states. Thus, the control states' achievement reflects not only the lagged effects of their pre-NCLB policies (since student achievement reacts gradually to policy) but also the effects of their post-NCLB increases in accountability. This makes the control states a very poor counterfactual for the treated states, most of which would have enacted no or only weak accountability programs had NCLB not existed. There is no guarantee that removing the pre-NCLB trend in the achievement of control states, as Dee and Jacob do, fixes the difference-in-differences strategy: such a fix would work only if the control states' pre- and post-NCLB policy changes just happened to produce a constant rate of change in achievement. This is possible but seems unlikely. Since the dynamics of how achievement reacts to accountability are unknown, one cannot even say whether removing a linear time trend over- or understates the true counterfactual.

Difference-in-differences strategies estimate a treatment effect that is local to the sort of states being "treated." If the effect of the NCLB was heterogeneous, which it surely was given states' variation in enthusiasm for implementing the law, the analysis requires treatment and control groups that are balanced in terms of their susceptibility to the treatment. Otherwise, the estimated effect does not reveal the population-average effect, which is largely what interests policymakers. That is, policymakers wish to know what the effect of NCLB would be in a state of average enthusiasm. Few if any policymakers have voiced a wish to know the effect of NCLB on *just* those states that balked at implementing accountability. Although a local effect of this kind might be interesting to a few people, it is not what policymakers think they are getting when they ask what the effect of school accountability is. In practice, estimation of a

treatment effect that is local to the balky states almost certainly substantially understates the population-average effect. It is the econometrician's responsibility to insist that readers *not* interpret a local treatment effect as a population-average one when it is clear that the control and treatment groups were poorly balanced in terms of susceptibility to treatment.

In my experience, what interests voters is not even the population-average treatment effect that NCLB would have had. What interests them is the effect of a school accountability policy when implemented fairly faithfully by leaders who believe in it. The voters' choice is between, on the one hand, candidates who run on a platform of accountability and who would therefore attempt to implement it as intended, and on the other, candidates who run against accountability and would not implement it. So far, we have seen no candidates declare that they will implement an accountability program in a way that deliberately evades its intention.

Before summing up, I must say something about the mismatch between the empirical strategy the authors describe and the equation they use to implement it, which is

$$\begin{aligned} Y_{st} = {} & \beta_0 + \beta_1 YEAR_t + \beta_2 NCLB_t + \beta_3 \left(YR\_SINCE\_NCLB_t\right) \\ & + \beta_4 \left(T_s \times YEAR_t\right) + \beta_5 \left(T_s \times NCLB_t\right) \\ & + \beta_6 \left(T_s \times YR\_SINCE\_NCLB_t\right) + \beta_7 \mathbf{X}_{st} + \boldsymbol{\mu}_s + \varepsilon_{st}. \end{aligned}$$

The authors describe their empirics in terms of *dosage:* essentially, they want to allow states that had no accountability program before NCLB to get the "full dose" of the 2001 law. They want to allow states that had already taken a full dose of accountability before 2001 to get little or nothing out of the law. This is reasonable as an explanation, but it suggests that they will measure dosage by indicators of the degree to which the state had already, by 2001, implemented the key features of NCLB (statewide testing, publication of school grades in report cards, highly qualified teachers, modest forms of school choice). One might also include enthusiasm for implementation as part of a state's dosage calculation.

Unfortunately, the estimating equation proxies dosage with T_s, the number of years a state was without school accountability between the 1991–92 academic year and the onset of NCLB. That is, the authors posit that dosage was linear in year of implementation, so that if a state's accountability program was implemented one year earlier, its NCLB dosage decreases by one unit. This specification does not match up with reality. First, states that implemented accountability earlier did not consistently have more ambitious accountability laws. For instance, most commentators would rate

Florida's program as easily the most ambitious, but it was not implemented until 1999–2000, making it a later-than-average implementer among the control states. Second, making the dose linear in year of implementation is very restrictive—so restrictive that the estimating equation is not plausibly eliminating differences in preexisting trends for control states based on their true dosage, which we have seen is crucial for avoiding bias in the difference-in-differences estimates. It would be preferable to have a proxy for dosage that is actually based on measures of the variety and stringency of the policies implemented.

NCLB, like most laws implemented nationwide, is very difficult to evaluate because no natural control group exists. What one would need to evaluate it perfectly is a "twin" United States without the law. Not having such a twin at their disposal, Dee and Jacob make a valiant effort to assess the effect of NCLB on student achievement. However, what they end up estimating is, even under the very optimistic assumption that their controls for preexisting trends and dosage work perfectly, an unusual parameter: the effect of school accountability on states that balk at implementing such laws and can evade them fairly easily. This effect may interest some, but it understates the effect that school accountability has with an average degree of rigor in implementation and probably greatly understates the effect with faithful implementation.

COMMENT BY

HELEN F. LADD Determining how the 2002 reauthorization of the federal Elementary and Secondary Education Act, commonly called No Child Left Behind (NCLB), has affected students, teachers, and schools poses a significant challenge because the program was implemented in all states at the same time. The most straightforward approaches, such as comparing trends in test scores before and after the introduction of NCLB, or comparing trends in the United States with those in other countries, generate conclusions that are at best suggestive because of the potential for confounding changes.

This paper by Thomas Dee and Brian Jacob makes an important contribution to the literature on NCLB. The authors have made a creative and valiant, if not completely successful, effort to use an innovative strategy to investigate the effects of NCLB not only on test scores but also on a number of mediating mechanisms such as spending and instructional practices. The authors' main conclusions are that NCLB generated some positive gains in mathematics, especially among disadvantaged fourth-graders, but

no gains in reading, and that it induced higher state and local spending and a shift of attention toward math and reading within schools. These conclusions are generally plausible. At the same time, however, some perplexing timing patterns emerge that deserve further attention. In general, a stronger framing and a richer discussion of the policy implications would have made an excellent paper even more useful to the current policy debate.

THE STUDY DESIGN—AND A PERPLEXING FINDING FOR FOURTH-GRADE MATH. Central to the paper is the authors' interpretation of which states received the NCLB "treatment." Their innovation is to include in the treatment group only those states that did not have a prior, state-level accountability system at the time NCLB was introduced. The control group is then composed of the states that were early adopters of accountability systems, and the authors pay appropriate attention to when precisely these systems were adopted. The underlying logic is clear: only in states without existing accountability systems similar to that of NCLB would the introduction of NCLB represent a new treatment. Although this strategy is creative, it is not immune from criticism. Among my concerns are that the authors may have overstated the similarity of NCLB and the pre-NCLB accountability systems in many of the control states, and that in the post-NCLB period, such states may have responded to NCLB in ways that would render their outcome trends less than ideal measures of what the trends would have looked like in the treatment states in the absence of NCLB. (As an aside, I also wonder why the authors did not apply their treatment logic to their comparison of test scores between public schools and Catholic schools over time.) In any case, the authors' main analytical strategy is plausible—and better than most alternative approaches—but not perfect.

Using this strategy, the authors find the strongest test score gains for fourth-graders in math. The authors' analysis of test results is based on scores on the National Assessment of Educational Progress (NAEP), which is appropriate because the low stakes on this test make the scores far less subject to manipulation through teaching to the test than would be the case with the states' own high-stakes tests. The reliance on NAEP scores, however, means the authors are working with a small dataset: it includes only 39 states that had NAEP scores both before and after NCLB, and only a few years of data because the tests are not given every year. Their figure 4 shows a very big jump in fourth-grade NAEP scores from 2000 to 2003 in the treatment group relative to the control group, as well as some differential increase in the growth rates over time, and these patterns are confirmed by the authors' regression analysis. Such a large jump over

that period seems highly implausible to me, however, given that 2003, which represents the 2002–03 school year, is the first year after the introduction of NCLB.

More framing is needed to help the reader evaluate the plausibility of this big jump and to provide more of a foundation for the rest of the paper. One perspective is that the accountability mechanism of NCLB was simply intended to reduce teacher shirking. Such a view is consistent with the brief principal-agent framing that the authors provide early in the paper. Within the context of such a perspective, teachers could conceivably raise student achievement quite quickly. Once teachers are put on notice that they are accountable for student test scores, for example, they might immediately stop shirking, and test scores might quickly rise. An alternative view is that NCLB is best interpreted as a catalyst for a variety of changes that may ultimately raise test scores but not immediately. Such changes might include, for example, increased state and local spending, more professional development for teachers, changes in instructional processes, or shifts in school schedules so that teachers can devote more effort to the tested subjects. All of these changes are likely to take some time to play out. This catalytic view is consistent with other findings in the paper, including, for example, the finding that states raised spending, but it is inconsistent with the empirical finding of a big gain in fourth-grade math scores in the first year of the program.

Two other considerations are also relevant to the expected timing of any effects. To the extent that education is a cumulative process, learning in fourth grade, for example, depends in part on how much children learned in earlier grades. For that reason, even if the teachers in the early grades understand the importance of their teaching for student test scores in subsequent grades and respond to NCLB by changing their practices in positive ways, it would take some time for those changes to show up in the achievement levels of fourth-graders. Moreover, the complexity of the implementation process should also have led to delayed effects on test scores. In response to NCLB, many states phased in the required testing and set up requirements in a way that backloaded the gains necessary to meet the 2013–14 proficiency goal, and sanctions were designed to be minimal at first and to increase over time. Such timing considerations make me suspicious of the estimated early gains in fourth-grade math test scores that emerge from Dee and Jacob's estimation strategy. Instead I would have expected at most very small effects in the initial year, with the effects increasing over time.

INTERPRETING THE MAGNITUDES. Even if we accept Dee and Jacob's estimates of the NCLB effects on test scores, it is reasonable to ask how one

Table 1. Gains in NAEP Test Scores in Massachusetts and Dee and Jacob's Estimated Effects of NCLB
Score points

	Fourth grade	*Eighth grade*
	Mathematics	
Massachusetts		
Average score, 2000	233	279
Average score, 2007	252	298
Change	+19	+19
National NCLB effect (Dee and Jacob)	+7.2	+3.7 (NS)[a]
	Reading	
Massachusetts		
Average score, 1998	223	269
Average score, 2007	236	273
Change	+13	+4
National NCLB effect (Dee and Jacob)	+2.3 (NS)	−2.1 (NS)

Sources: National Center for Education Statistics; Dee and Jacob, this volume, table 1.
a. NS = not statistically significant.

should interpret the magnitudes. The answer the authors themselves give is that the fourth-grade math effects are large enough to be policy relevant. They note, for example, that the estimated average gain of 7.2 scale points is about 24 percent of the black-white gap (but why that is a relevant comparison is not clear in this context), and that the gains for Hispanics relative to those for whites imply a 19 percent reduction in the Hispanic-white gap. An alternative, policy-motivated answer is that the gains are tiny relative to the gains needed to get all fourth-graders to 100 percent proficiency.

A third possible answer is that the effects are small relative to what might have been possible with an alternative, more comprehensive policy. This answer is given support by my table 1, which compares gains in NAEP scores in Massachusetts between 2000 and 2007 with the average NCLB "effect" estimated by Dee and Jacob on fourth- and eighth-grade math and reading scores. I have selected Massachusetts for the comparison because of the ambitiousness of its 1993 reform package and its highly touted success in raising test scores. The table shows that Dee and Jacob's estimated NCLB effects are far smaller than the actual gains achieved in Massachusetts. The estimated 7-point NCLB effect in fourth-grade math falls far short of the 19-point gain in Massachusetts, and the statistically insignificant 3.3-point effect in eighth-grade math is even further below the corresponding (also 19-point) gain in Massachusetts. In reading, Massachusetts experienced gains at both grade levels, whereas the NCLB effects are statistically indistinguishable from zero.

The nature of the Massachusetts reform package provides a possible explanation for the different patterns. In contrast to the NCLB reform, which was narrowly focused on test-based accountability, accountability was only one small part of a far more comprehensive reform effort in Massachusetts. This state's reform package included a substantial increase in funding (more than doubling in 10 years), new learning standards, revised student assessments based on clear curriculum frameworks, revised teacher licensing and professional development programs, and early childhood programs, as well as parental choice and the creation of new charter schools. Such a package is far closer to what in the education literature has been called standards-based reform than what was implemented under NCLB. Although NCLB is an outgrowth of the standards movement at the national level, it in fact incorporates only one part—the accountability part—of what was intended to be a far more positive, constructive, and comprehensive approach to raising the achievement of all students.

CONCLUSION AND IMPLICATIONS FOR POLICY. I draw several conclusions from Dee and Jacob's paper. First, any effects of NCLB on test scores are small at best. The positive effect of NCLB on average math scores that emerges from this study occurs too soon relative to what might have been reasonably predicted, and for that reason may be overstated. Moreover, even the reported effects are small relative to what Massachusetts has shown to be feasible. Further, whether there are positive effects on eighth-grade math is not clear, and no effects on reading emerge at either grade level. This latter fact is reported in the paper, but its implications are not discussed. On a brighter note, there may be some positive effects for black students in fourth-grade (although not in eighth-grade) math and for Hispanic students in math at both grade levels.

Among the authors' other findings is that NCLB appears to have induced additional state and local spending on education. My interpretation of this finding is that there is no free lunch. Stated differently, it is a mistake to believe that accountability systems can, by themselves and without associated funding, generate gains in student achievement. In addition, as most policy analysts would have predicted, NCLB appears to have shifted attention and resources away from other subjects toward math and reading. The thorny question, but one that an empirical study of this type cannot answer, is whether that shift is desirable or undesirable.

So what do such findings imply for policy? The answer is not fully clear, but my take differs quite sharply from the thoughts the authors present in the final section of the paper. First, the null findings for reading

indicate to me that to the extent that higher reading scores are an important goal for the country, NCLB is clearly not the right approach. That raises the obvious follow-up question: what is? One possible answer, but one that goes far beyond the subject of this paper, is that policymakers need to pay attention to what goes on outside of schools as well as within schools. A second policy conclusion arises from the suggestive evidence that I have included here on Massachusetts, namely, that states may be in a better position to promote student achievement than the federal government. That raises the question of how the federal government can best encourage the states to engage in significant comprehensive reform efforts. My reading of Dee and Jacob's paper is that NCLB is far from the best approach for the federal government to pursue.

GENERAL DISCUSSION Brian Knight observed that although limited information is available on the paper's control group, the analysis would be improved by making the control and treatment groups as comparable as the data allow. He wondered whether there was any additional variation that could be exploited in choosing these groups. Knight was also struck by the reported increase in spending per pupil after NCLB and was curious whether there were heterogeneous effects, particularly at the state and local levels. He was interested in knowing more about the incidence of the spending increase, and in particular whether it fell mostly on state taxpayers or was paid for by cutting back other state spending.

Christopher Jencks found one of the paper's results, the effect of NCLB on fourth-grade mathematics scores in the first year of the policy change, to be implausibly large. Although math scores should have increased during this period, for the result to be convincing it was necessary to rule out other causes, such as differences in the content of the test over time. James Hines, however, found the results on fourth-grade math scores plausible and argued that it was not necessary to impose a linear trend.

Kristin Forbes found the different observed impacts of NCLB on math and reading scores interesting and wondered whether there was any theory that would have predicted this difference. Have any other countries undertaken similar reforms and observed this type of differential effect? An analysis of why NCLB has had a stronger impact on math than on reading could be important in understanding how NCLB has worked and what other types of educational reforms are likely to be the most effective.

Erik Hurst wondered what the proper way to specify the production function for human capital would be in this context. He was also interested

in the effects of NCLB on other outcomes besides test scores, such as graduation rates and the incomes of graduates.

Caroline Hoxby spoke in defense of NCLB, noting that the law was intended to be similar to those previously implemented in reformist states. She thought most people would like to know what would have been the effect of NCLB on the average state—that is, on a state that had a typical amount of enthusiasm for implementing school accountability. This effect could be approximated by averaging two estimates: first, the estimated effect of school accountability laws in states that implemented them before NCLB (the enthusiastic implementer effect), and second, the estimated effect of NCLB in states that had no accountability law before NCLB (the unenthusiastic implementer effect). She observed further that math and reading are very different subjects, and that math is almost entirely taught at school whereas parents have a much bigger impact on reading, so that the disparate effects of NCLB on the two subjects were plausible.

Annette Vissing-Jorgensen noted that migration rates have differed across booms and recessions. These could affect the results for the impact of NCLB among Hispanic students. She also wondered about possible cohort effects, noting that such an effect could be occurring in the math results presented in the paper's figure 4. Students included at any year in the top panel would enter the bottom panel 4 years later.

Christopher Jencks agreed with the authors that NCLB is difficult to evaluate statistically since it was implemented all at once. However, some states set the standard for student proficiency quite high, so many students would have to learn more for a school's performance to be judged acceptable, while other states set low standards so that relatively few students would have to learn more for a school's performance to be judged acceptable. One would expect NCLB to have less of an impact in states where schools did not have to improve most students' performance. On the difference in results between math and reading, given that kids do not learn math on their own before going to school, as Hoxby had noted, one might have expected that family background would have more of an effect on reading than on math, but that prediction does not hold up in the authors' results.

Helen Ladd wondered whether the different results for reading than for math might have less to do with the differences between the subjects themselves than with the tests used to measure proficiency. State-level results using state tests have shown gains in reading that do not show up in the NAEP. If math curriculums are closer to what is tested on the NAEP than reading curriculums, the results might not come through as well for reading. Ladd also thought it worth noting that test scores are only part of

the picture in evaluating NCLB. Costs are also a consideration, as are other measures of success besides achievement, such as graduation rates and college attendance, as Hurst had noted.

Adele Morris remarked that although the paper found no evidence that NCLB harmed students at the high end of the achievement scale, it might be worth taking a second look, particularly at the very highest level. One current debate about NCLB is the degree to which it could reduce the resources that schools devote to the education of gifted learners. To explore this, the authors could examine the dependent variable at the 95th percentile or higher for the NAEP score, as well as explore the effect of NCLB on expenditure on programs for the gifted.

Melissa Kearney commented that many observers have questioned whether the magnitude of the effect found for NCLB on fourth-grade math test scores is credible. Presumably there is a sizable empirical literature estimating the effects of earlier experimentation with accountability at the state level. Were the effects the authors found in line with those estimates? Or are previous estimates so small that there would have to be large general equilibrium effects in order for the paper's estimates to be right?