Institute for Policy Research
Northwestern University
Working Paper Series

WP-09-11

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

Manyee Wong
Postdoctoral Fellow, Institute for Policy Research
Northwestern University

Thomas D. Cook
Faculty Fellow, Institute for Policy Research
Professor of Sociology, Psychology, Education and Social Policy
Northwestern University

Peter M. Steiner
Senior Research Associate, Institute for Policy Research
Northwestern University

This work was supported in part by the Institute of Education Sciences
Grant # R305U07003.
Special thanks go Christopher Jencks, Larry V. Hedges, and Vivian Wong for feedback on this paper. All errors are the sole responsibility of the authors.

DRAFT
Please do not quote or distribute without permission.

Abstract

Enacted in January 2002, No Child Left Behind (NCLB) holds schools accountable by testing whether they make adequate yearly progress (AYP) toward state-specified proficiency standards. The law requires failing schools to take specified corrective actions that become increasingly more onerous with the number of years a school has failed to make AYP. This paper evaluates NCLB using National Assessment of Educational Progress (NAEP) data between 1990 and 2009 for 4th grade reading and 4th and 8th grade math. One set of analyses is at the national level and contrasts public schools separately with Catholic and non-Catholic private schools. Other analyses are at the state level. Some of these analyses contrast states whose high or low proficiency standards result in many or few schools implementing NCLB-required changes or fearing they will have to do so. Other analyses factorially combine states whose standards are high or low with states whose pre-2002 accountability system did or did not contain sanctions for failure. Across all these analyses, NCLB consistently improved both 4th and 8th grade math, though 4th grade reading effects were limited to states with both high standards and an accountability system that included sanctions only after NCLB.

Introduction

Although federal dollars still account for less than 8% of the national education budget (Department of Education, 2009a), federal involvement in education has been increasing since 1964 when the Elementary and Secondary Education Act (EASA) was passed. Despite the billions of dollars spent on ESEA and its successors, reports have pointed to disappointingly low levels of performance by American students, especially economically disadvantaged ones (e.g., National Commission on Excellence in Education, 1983: Borman & D'Agostino, 1983;

Rudalevige, 2003), and many leaders in politics and business have called for fundamental educational reform.

Standards-based school reform has been one response. It uses test score results to hold schools accountable for student performance, and then mandates specific reforms that failing schools must adopt in order to improve student learning and to avoid further sanctions. For instance, the 1994 Improving America's Schools Act (IASA) required states to establish curriculum and performance standards and conduct annual reading and math assessments. Any Title I school or district that failed to meet its annual proficiency target was identified for improvement and required to take corrective action (Department of Education, 1994). However, IASA was not strictly enforced and few schools were sanctioned (Sunderman, Kim, and Orfield, 2005).

Nonetheless, President George W. Bush took the basic IASA concept, revised many of its details, devoted new resources to it, and named it No Child Left Behind (NCLB). Passed in 2001 and officially enacted in January 2002, NCLB required states to: (1) conduct regular achievement testing using their own tests aligned to their own curriculum standards, (2) establish a clear time schedule by which an increasing fraction of students should become proficient by each state's own performance standards; (3) impose sanctions on failing schools; (4) require all teachers be highly qualified by 2006-2007; (5) use scientifically based teaching strategies; and (6) have all students be proficient in all basic subjects by 2014.

Central to the program is the notion of sanctions whose consequences increase with the number of years a school has consistently failed to make annual yearly progress (AYP). Failing the first year requires informing parents of this fact. Failing a second year requires giving parents the option to transfer their child to better performing schools. A third year of failure requires

providing supplemental services like tutoring, after-school and summer school programs. A fourth year means replacing staff, hiring consultants and/or implementing a new curriculum. A fifth year means a school can be closed, become a charter school, or become private (Department of Education, 2008; Goertz, 2005).

Supported by Title I money, NCLB applies to any school district accepting federal funds. So it is national in scope, built around Federal conceptions of a failing school and of the reforms such a school must undertake. NCLB also increases state control over educational policy since state agencies are responsible for selecting tests and cutoff values, for setting annual proficiency targets, and for monitoring and otherwise supporting the interventions a failing school has to implement (Sunderman et al., 2005). This system has led some to describe NCLB as "the most revolutionary education policy since EASA", and a reform “that will put American schools on a new path of reform and results" (McDonnell, 2005; Milbank, 2002; Smith, 2005). Others describe it less positively, as "the most elaborate case of federal micromanagement of state policy, local schools, and teachers in the entire history of American education." (Chapman, 2007).

President Obama has repeatedly cited education as a top domestic priority (Dinan, 2009), and the 2009 American Recovery and Reinvest Act (ARRA) provides NCLB with new revenue streams: (1) An additional $10 billion in one-time funding, nearly doubling its 2009 appropriation of $14.5 billion; (2) an additional $3 billion in Title I School Improvement Grants for schools in corrective action; (3) $4.3 billion in Race to the Top Competitive Grants designed to reduce caps on charter schools and allow the use of student test scores in teacher evaluations; and (4) $650 million for Invest in What Works and Innovation grants for which local education agencies and partnering nonprofit groups compete if the state is already narrowing racial

achievement gaps (Department of Education, 2009b and 2009c). The ARRA money requires states to adopt rigorous academic standards, develop high quality assessments, increase teacher effectiveness, distribute effective teachers more equitably, and turn around the lowest performing schools (Department of Education, 2009b). In spending the money, states are encouraged to expand after-school and summer programs, better align preschool and early elementary school practices, develop better curricula, train teachers more effectively, establish systems to track student performance, implement a new teacher evaluation system, and redesign performance pay (Department of Education, 2009d). These plans modify NCLB without eliminating it and so disappoint both guardians of local educational control (Branigin, 2009; Dillion, 2009; Shear 2009) and critics who resent the focus on testing and negative sanctions.

Justifying the present evaluation are the high expenditures for NCLB, the conflicting debates about it in public discourse, its shift towards greater federal and state control over public education, and its likely continuation in modified form after the next re-authorization of Title 1. But also important is the inconclusive nature of past evaluations of NCLB. The congressionally mandated National Assessment of Title I (Stullich, Eisner, and McCrary, 2007) comprehensively described NCLB's implementation. The study presented some NAEP interrupted time series (ITS) data to describe achievement time trends before and after NCLB. However, Stulllich et al (2007) counseled against causal interpretation because any differences apparent to the eye were not evaluated statistically and might anyway be due to other forces co-occurring with NCLB's introduction. The Center on Education Policy (2007) used trend data between 2002 and 2005 from 36 states, each with its own state achievement test, and showed that the percentage of proficient 4^{th} and 8^{th} grade students increased by 1 to 3 percentage points in a majority of states. However, no comparison group data were presented and so the counterfactual is unclear. Other

studies have reported less positive findings. Fuller et al. (2007) suggested that growth on the 4th grade NAEP reading test faded after NCLB and slowed for math after 2002. In perhaps the earliest ITS study, Lee (2006) compared states with high stake testing and strong accountability system prior to NCLB with states that only adopted a stronger accountability system after NCLB, using the accountability index of Lee and Wong (2004) to partition the states. But Lee (2006) found no statistically significant difference in the growth rate of math or reading from pre- to post-NCLB. However, the time series data ended in 2005 and did not have the statistical power of two later evaluations using ITS.

In an unpublished dissertation, Wong (2008) compared differences between states whose proficiency standards for passing AYP were higher or lower and so linked to more or fewer sanctions for failure. Assessing 2002 group differences of differences in both mean and slope, the study found no effect for 4th grade reading but statistically significant effects for both 4th and 8th grade math. It attributed these effects to higher standards causing more schools to fail, thus forcing them to reform because of the provisions of NCLB. However, Wong (2008) did not provide data to show that state variation in standards was associated with more school reform. And she defined states with high and low standards by means of the percent of students rated proficient averaged over 2003 and 2005, thus raising the possibility that post-2002 changes in achievement might have affected the standards a state adopted. Moreover, Wong (2008) did not correct for possible auto-correlation in the ITS data that biased standard errors. Finally, as careful as the study was to rule out alternative interpretations occurring in 2002, one can never be certain of having identified them all. So *the first aim of this paper is to try to replicate the Wong (2008) standards-based findings* using the same basic method but improving on the technical limitations noted above.

The other, a working paper with similar ITS design and focus, is Dee and Jacob (2009). A conceptual replication of Lee (2006), it contrasts states whose accountability standards prior to 2002 already led to sanctions/consequences (the comparison time series) with those states whose accountability system did not require consequences until after NCLB (the treatment time series). Dee and Jacob (2009) improves on Lee (2006) because the later authors 1) had NAEP data through 2009 instead of 2005; 2) opted for a state fixed effects model rather than a random effects one; and 3) grouped accountability states before and after NCLB based largely on Hanushek and Raymond's (2005) consequential accountability index that focus more directly on a key component of a state's accountability strength - sanctions. Dee and Jacob (2009) used "consequential accountability" (CA) to refer to their source of state variation and found a difference in 4^{th} grade math after 2002 and no such effect for 4^{th} (or 8^{th}) grade reading. But, unlike in Wong (2008) where an 8^{th} grade math effect was robust, Dee and Jacob (2009) only detected an 8th grade math effect in some of their analyses of all students. Moreover, it is clear from their table that some states adopted CA before NCLB and were continuing to do so until 2001. So it is not clear whether all states adopting CA in 2002 did so because of NCLB or because they would have done so anyway as part of the system of spontaneous adoption that begun in the early 1990's.

The papers by Wong (2008) and Dee and Jacob (2009) involve different causal contrasts. The former contrasted states with higher (HS) or lower standards (LS) for passing AYP immediately after 2002. Dee and Jacob (2009) contrasted states whose accountability system did or did not include negative consequences (CA) before 2002. Obviously, a state can newly adopt sanctions but have high or low standards that result in sanctions applied to many or few schools.

An accountability system with sanctions does not necessarily result in a system where these sanctions apply widely. So how are the two mechanisms (HS and CA) related?

We correlated the dichotomous CA measure of Dee and Jacob with our continuous standards measure based on student proficiency rates in 2003. (States with higher proficiency rates will tend to have fewer students failing AYP and so have lower standards). The resulting state-level correlation is -0.05. It is only slightly higher (r = -.19) when standards are indexed as the difference between the percent proficient on a state's own tests and NAEP, thereby controlling for state differences in true achievement as they are indexed by NAEP passing rates. So CA is not related to HS or LS, indicating that the states deemed treatment states in Dee and Jacob (2009) are not related to the states deemed treatment states in Wong (2008). This then raises the question of what happens when a state that newly adopts sanctions has higher rather than lower proficiency standards so that these sanctions apply to relatively more schools. *The second aim of this paper is to test the joint impact of standards and consequential accountability.*

One possibility is that the two mechanisms will create some emergent property that is more than the sum of their two parts, thus leading to a positive statistical interaction between CA and HS. The second possibility is of an additive result, implying no such interaction. The third possibility is that the two might countervail in some way and hence result in a negative statistical interaction. The first two patterns of results raise an intriguing possibility. In both Wong (2008) and Dee and Jacob (2009), the 4th grade reading effect was small, in the direction expected for an NCLB effect, but far from statistically significant. If the two mechanisms of CA and HS combine in either additive or positive multiplicative fashion, then the reading effect will be larger and perhaps even statistically significant for the first time in ITS studies of NCLB.

NCLB lets states use their own achievement tests and set their own passing standards. Even so, it is a national program because the Federal government mandates that states must have a proficiency test, set standards on it, link failure on these standards to sanctions, and then monitor and enforce compliance with these sanctions. The most appropriate test of a national program is to assess its effects at the national level where NCLB is almost exclusively a program for public schools. This suggests one might evaluate it by estimating whether public schools come to perform relatively better than private schools after NCLB than before it.

Dee and Jacob (2009) argue that it is not possible to use Catholic schools for this purpose because sex abuse scandals likely decreased enrollments there in 2002. Table 1 presents enrollment time trends for public, Catholic and non-Catholic private schools. It shows a secular trend towards decreasing enrollment in Catholic schools but a likely acceleration of this trend in 2002. However, there is no corresponding shift in the relationship between public and non-Catholic private schools in that year. Moreover, class size and student racial profile data show no reliable changes in 2002 for any kind of school and no differences in change between types of school from before to after NCLB. This implies that the attrition from Catholic schools in 2002 was not selective, at least on these two correlates of achievement if not on other unobserved variables. So while the enrollment data indicate that the contrast of Catholic and public schools may be compromised, the race and class size data suggest it may not be compromised by much. Fortunately no compromise is indicated with the contrast of public and *non-Catholic* private schools. In 2002, the latter had no corresponding sex scandal or any other historical change we have been able to identify. However, some of the children leaving Catholic schools might have transferred into non-Catholic private schools in 2002, two-thirds of these latter schools being sectarian but not Catholic. However, there is no evidence of such a rapid increase in numbers in

non-Catholic private schools in 2002 (Table 1). Much more likely is that a greater percentage of Catholic students are transferring to public schools in and after 2002 than prior years but their small numbers had little effect on means or trends in the much larger public school system. In any event, the non-Catholic private schools provide a better counterfactual than the Catholic schools, though we will provide information about how each kind of private school performs when separately contrasted with public school achievement over time. *The final purpose of this paper is to test whether NCB affected achievement in national tests contrasting public schools with first Catholic and then non-Catholic private schools*.

For all three purposes, Main and sometimes Trend NAEP data are used to examine performance in both math and reading. Three estimates are computed in both the national contrast of public and private schools and in the contrasts of HS and LS states, both alone and also when they are crossed with CA. The three are: (1) whether the treatment and comparison group means are different from what they are predicted to be from the groups' observed pretest means and slopes after NCLB; (2) whether the observed difference in treatment and comparison *slopes* after NCLB is different from the difference in slopes before then; and (3) whether the final difference between treatment and comparison means differs from what is predicted from the pre-NCLB means and slopes. This last effectively adds together the mean and slope differences of differences until 2009 (for math) or until 2007 (for reading), the latest time points when data are available.

Methods

Data. The outcome data come from NAEP, beginning in 1990 or 1992 and continuing until 2009 for Main NAEP math, 2007 for Main NAEP reading, and 2004 for Trend NAEP math and reading. Main NAEP is the principal data source. While no so significant changes were

made to Main NAEP test content when NCLB was introduced, the sampling design was modified in 2002 to reduce the total number of schools participating. The reduction was achieved by randomly drawing schools for the national estimates from those already randomly selected to provide state Main NAEP data (Lazar, 2004, National Center for Education Statistics, 2009).

Fundamental changes had been made in Main NAEP earlier in order to accommodate students with disabilities and special English language needs. To assess the effects of this population change, in 1996 and 2000 for math and in 1998 for reading, a split sample design was used to create two data sets for each year -- one with and one without the new accommodations. We analyzed the achievement data with each of these sets of pre-NCLB values, but they made no difference. So the analyses we present here will use the pre-NCLB data with accommodations since that is what is used for the post-NCLB outcome data.

Unlike Main NAEP, Trend NAEP holds test content constant across years. But Trend NAEP is not collected at the state level; it is not publicly available for non-Catholic private schools; the intervals between waves are generally longer than in Main NAEP; there is a gap in data collection from 1999 to 2004 when the intervention occurred; and accommodations for students with special needs began in 2004. While data were collected using both the old and the new sampling frames in 2004, this was not the case in 2008, making 2004 the last data point comparable to the pre-NCLB time series and so limiting Trend NAEP analysis to differences in mean differences by 2004. A further complication with Trend NAEP is that some functional forms are inexplicably complex prior to 1990. As a result, we limit analyses to the same 1990-2009 time frame that is available for Main NAEP. Although Trend NAEP is less useful than Main NAEP, it is still important because we can replicate any immediate mean changes that

might be observed on Main NAEP in the contrast of public and Catholic schools with Trend NAEP data on the same two kinds of schools.

Federal standards place emphasis on at least 70% of the invited schools actually participating in NAEP. Achieving such participation has not been a problem with public or Catholic schools, but in some years it has been a problem with non-Catholic private schools. We nonetheless use the data for years with participation rates under 70%, noting that variability around the obtained time trends is indeed somewhat greater for the non-Catholic private schools than for the other kinds of schools. Even so, we will later note considerable agreement between short-term achievement results when public schools are contrasted with both Catholic and non-Catholic private schools.

Study Design. We present two basic ITS designs, one at the national and the other at the state level. Each involves a comparison time series and takes the intervention point to be January 2002 when the law came into power.

Design I: Comparing public schools with both non-Catholic and Catholic private schools at the national level. A key assumption of this analysis is that the private schools can function as a no-treatment comparison group. This is largely but not completely true because they have modest treatment overlap with NCLB. The two most popular NCLB programs in Catholic schools are (1) Reading First with roughly 3% of students participating; and (2) Title 1, Part A in which 6% of students participate (Department of Education, 2007a). In total, only 4.7% of all K-12 private school students (Catholic and non-Catholic) receive any type of Title I services (Keigher, 2009). More important perhaps is that private schools are not required to have proficiency standards associated with testing (no HS, therefore), and they do not have to institute

reforms because of test scores (no CA). So private schools are not quite a no-treatment control group, but they are close to it.

Comparisons at the national level pose a statistical challenge since analyses will have at most 16 degrees of freedom (two types of school x up to 8 time points). Somewhat mitigating against this limitation are that national achievement data should be very stable, we can replicate some results across two math grades and two independent data sets (Main and Trend NAEP), and we can test for larger effects due to combining differences of differences in both means and slopes after 2002.

Design IIa: Comparing States that vary in Proficiency Standards. We measure each state's proficiency standards by averaging its student proficiency rate in 2003 across two grades (4^{th} and 8^{th}) and two subject areas (reading and math).[1] This aggregation increases reliability and provides a continuous measure of state proficiency ranging from 26 percent to 85 percent of students passing. A trichotomous variable was also constructed to reduce dependence on functional form assumptions and also to present results more intuitively. States where fewer than 50 percent of students met proficiency standards were assigned to the LS group (N = 13) including the District of Columbia (hereafter treated as though it were a state); states with 75 percent or more were assigned to the HS group (N = 11); and the rest became the medium proficiency standards group (N = 25). Thus, the difference between them represents a variation in dosage rather than a treatment/no-treatment contrast.

State variation in proficiency standards may reflect true achievement rather than a state's strategy of standards setting. But that is not the case. Table 2 shows the percentage proficient in HS and LS states on both NAEP and state tests. The NAEP difference is modest (33% proficient

[1] We exclude from the analysis New York because it uses its own state proficiency rate scale that is not based on the 0 to 100 percent proficiency scale that all other states use. We also exclude Vermont because it has no state assessment data for the years we examined for group assignment.

vs. 27%), but the difference on the states' own tests is very large (79% vs. 40%). While true achievement differences between HS and LS states exist, they are quite small when compared to the differences on states' own tests (Department of Education, 2007b; Dillon, 2007; Fuller et.al., 2006; Fuller, et. al., 2007; Kingsbury, et.al., 2007; Skinner, 2005). Moreover, the analysis we conduct controls for whatever pre-NCLB state differences are observed in reading and math means and slopes.

Using 2003 state proficiency data to define standards runs the risk that disappointing achievement levels in 2002 may have caused some states to re-set their standards in 2003. Table 3 provides data on state proficiency rates over time and Table 4 shows the correlation of the percent of student proficient on state tests before and after NCLB. They are all over 0.80. There is some evidence that the between-year correlations are lower around the 2002 intervention point than at other times when they are over 0.90. We identified six states that significantly lowered their standards after NCLB (see those in bold in Table 3) and only one state, Hawaii, that significantly increased them. So most states did not change their standards in response to NCLB and those that did seem to have lowered them to cushion the law's impact rather than increase them and take on a new challenge. While these few changes probably contributed to the slightly lower correlation between 2002 and 2003 than for other adjacent years, all the correlations are high and reflect a system of standards that was quite stable and largely invariant even around 2002. We choose 2003 as the year for defining standards, not just because the year chosen will not affect the results much, but also because standards can only determine how much reform a state has to undertake after NCLB. Before NCLB, standards are not necessarily linked either to specific reform acts or to passing schools seeking to improve out of fear of failure in the future.

But do HS states actually undertake more fundamental educational reforms? Relevant data are available from Consolidated State Performance Reports (CSPRs), but not earlier than 2007. Data from that year are in Table 5, and analyses with schools as the unit of analysis show that HS states have indeed undertaken more reform. Although most schools in most states are not failing AYP, 13 percent more schools in HS states are classified as "in need of improvement" for failing to make AYP in at least two consecutive years. Sixteen percent more students are eligible for school choice and 8 percent more for supplemental services. Of schools that failed AYP, more are taking on corrective and restructuring actions in HS states. Thus, 4 percent more of their schools have instituted a new curriculum and 9 percent more have taken alternative types of restructuring actions. Schools can choose among reform activities, using more than one if they want. If all schools took a single corrective action, the difference between HS and LS states is 9 percent in both corrective action and restructuring. If schools made all the suggested reforms simultaneously, then the difference is 5 percent in corrective action and 8 percent in restructuring. So HS states have more failing schools, undertake more reforms, and these reforms engage more fundamental aspects of school life. Some of the differences may seem modest, but note that (1) most schools in the nation are not failing; (2) failing schools can chose among various reform options so that analyzing single reform activities will impose a low ceiling on how much change can be observed; and (3) NCLB is supposed to induce passing schools to do better so that they will not fail in the future – a mechanism not captured by data limited to what failing schools do.

NCLB analyses at the state level have two advantages over the national level. First, the use of 49 state-level time series increases the total degrees of freedom. There are 352 in the state variation design with 8 time points, somewhat less than the 392 possible (49 states x 8 time

points) due to haphazardly missing data at some times in some states. (However, power also depends on inter-temporal variability, and single data points are likely to be less stable at the state than the national level). Second, adding state-level ITS comparisons should reduce some threats to internal validity because some sources of bias in the public vs. private school contrasts do not apply in the state-level contrasts. For example, if a school reform initiative other than NCLB were introduced just into public schools in 2002 or just into both kinds of private schools, what are the chances that the same reform would be differentially introduced at exactly the same time into HS versus LS states?

Design IIb: Comparing States that simultaneously vary in both Proficiency Standards and Consequential Accountability. Dee and Jacob (2009) compared states that implemented CA either pre- or post-NCLB. Their CA and our standards measure are essentially orthogonal so we further divide the HS and LS states into those with and without CA before NCLB, thus creating four groups of states. One group has both treatments (HS & CA after NCLB). Another has no treatment (LS & CA before NCLB). The third has HS alone – viz, combined with CA before NCLB. And the fourth has CA alone – viz., combined with LS.

Dee and Jacob (2009)'s sample for analysis include 24 states with CA before NCLB and 14 states with it after, excluding states with missing data for some crucial pre-NCLB years. Based on this set of states, we further partition them by our high and low standards cut off. This results in a total of only 19 states -- 5 in the combined treatment category, 6 in the no-treatment category, 6 in the HS alone category, and only 2 in the CA alone category. While these sample definitions respect the contrasts presented earlier in this paper, they reduce power for hypothesis tests. So we replicated the analyses using a median split for standards and including states excluded by Dee and Jacob (2009). This results in 10 states in the combined treatment category,

15 states in the no-treatment category, 15 states in the HS alone category, and 10 states in the CA alone category. These 48 states led to 311, 301, and 343 degrees of freedom for 4th grade math, 4th grade reading and 8th grade math respectively. While this increased power due to a much greater sample size, it also reduced power by creating smaller treatment contrasts.

Analytic Models. A linear regression model was used to analyze the public vs. private school contrasts and a fixed effects model to analyze the contrast of states. Each model includes a dummy variable for the relevant contrast at the national or state level and interactions of this dummy with the change in average test score and growth rate from before to after NCLB. So each analysis has two time-series segments. One represents the mean and growth in achievement scores prior to NCLB, and the other, the corresponding means and slopes after 2002 (Raudenbush, 2002). We selected 2002 as the year for implementing NCLB because it became law in January 2002 and most school years begin in August or September, suggesting that districts will not have implemented NCLB until later in 2002. Moreover, NAEP tests are conducted in January through March of each testing year, and so the 2002 NAEP reading score precedes the school-level implementation of NCLB. This means we consider 2002 NAEP reading scores to be in the pretest period. For math, the first data point after 2001 is in 2003, clearly a post-NCLB year.

For the contrast of public schools with either Catholic or non-Catholic private schools we estimate the following regression model:

$$Y_{tj} = \beta_0 + \beta_1(year)_{tj} + \beta_2(group)_j + \beta_3(policy)_{tj} + \beta_4(year \times group)_{tj} + \beta_5(policy \times year)_{tj} + \beta_6(policy \times group)_{tj} + \beta_7(policy \times year \times group)_{tj} + \varepsilon_{tj}, \quad (1)$$

where Y_{tj} is the outcome on Main NAEP at $t = 1, \ldots, 8$ time points for reading and math, while for Trend NAEP, both outcomes are at $t = 1, \ldots, 6$ time points for public ($j = 1$) and private

schools ($j = 0$). $Year$ is a continuous variable indicating the year of measurement; $group$ is a dichotomous variable representing public ($group = 1$) versus private schools ($group = 0$); $policy$ is another dichotomous variable indicating pre- and post- NCLB period (1= post-NCLB). Hypothesis tests of differences in mean and slope changes between groups require the inclusion of all two- and three-way interactions. Hence, the regression coefficient β_6 of the $policy \times group$ interaction gives the differences in the mean change in 2002. The three-way interaction effect β_7 tests whether public and private schools differ in their post-NCLB slope changes. The error term ε_{tj} is assumed to be independent and identically distributed according to a normal distribution with a mean of zero and variance σ_ε^2. No covariates are included, given the limited degrees of freedom for the public vs. private contrast.

For the categorical contrast of states with high, medium and low proficiency standards, we estimate a model with state and year fixed effects. The outcomes at time points $t = 1, \ldots,7$ for 4th grade reading and math, $t = 1, \ldots,8$ for 8th grade math, and for states $i = 1,\ldots,49$ are modeled as:

$$\begin{aligned} Y_{ti} = {} & \beta_0 + \beta_1 (policy)_{ti} + \beta_2 (policy \times year)_{ti} + \beta_3 (year \times group\_h)_{ti} + \beta_4 (year \times group\_m)_{ti} \\ & + \beta_5 (policy \times group\_h)_{ti} + \beta_6 (policy \times group\_m)_{ti} \\ & + \beta_7 (policy \times year \times group\_h)_{ti} + \beta_8 (policy \times year \times group\_m)_{ti} \\ & + \beta_9 (percent\_free\_lunch)_{ti} + \beta_{10} (pupil\_teacher\_ratio)_{ti} + \mu_i + \tau_t + \varepsilon_{ti}, \end{aligned} \quad (2)$$

where $group\_h$ and $group\_m$ in the interaction terms are dummy variables indicating high and medium performance standard states, respectively. μ_i are the state fixed effects, τ_t the year fixed effects, and $\varepsilon_{ti} \sim N(0, \sigma_\varepsilon^2)$ the independent and identically distributed error term for $t = 1,\ldots,7$ or 8 depending on outcome and $i = 1, \ldots, 49$. Again, hypothesis tests of differences in mean and slope changes between groups are represented by two- and three-way interactions. The

regression coefficient β_5 of the $policy \times group\_h$ interaction gives the differences in the mean change in 2002 for the high vs. low proficiency standards group. The three-way interaction effect β_7 tests whether high and low proficiency standards group differ in their post-NCLB slope changes. The model also controls for time-varying covariates assessing the percentage of students eligible for free lunch and the pupil-to-teacher ratio $(percent\_free\_lunch, pupil\_teacher\_ratio)$.

We examined other potential time-varying covariates at both the state and national level, including school level expenditures, family income and various student demographics. They all correlate quite highly with the percentage of students eligible for free lunch or pupil-to-teacher ratio and do not change treatment effects when included. Therefore, the final model is restricted to the latter two covariates that are largely independent of each other yet are highly correlated with the achievement outcomes.[2]

The main null hypotheses for all models are:

Null Hypothesis 1: Group mean differences do not differ from before to after NCLB.

$H_1: \beta_6 = 0$ for the public vs. either private school contrast (equation (1))

$H_1: \beta_5 = 0$ for the high vs. low proficiency standard states contrast (equation (2))

Null Hypothesis 2: Group slope differences do not differ from before to after NCLB.

$H_2: \beta_7 = 0$ for the public vs. either private school contrast (equation (1))

$H_2: \beta_7 = 0$ for the high vs. low proficiency standard states contrast (equation (2))

[2] The models were not weighted by student population or the inverse sampling variance of NAEP estimates because the design uses state as the unit of analysis and little variation results when the standard deviation is examined separately across states and years.

Null Hypothesis 3: Group mean differences in 2009 (math) or 2007 (reading) do not differ from what the pre-2002 group mean and slope trends predict them to be. This total impact hypothesis combines the first two hypotheses about differences in mean and slope.

H_3: $\beta_6 + (\beta_7 \cdot k) = 0$ for the public vs. either private school contrast

H_3: $\beta_5 + (\beta_7 \cdot k) = 0$ for the high vs. low proficiency standard states contrast

with $k = 7$ for 4th and 8th grade mathematics and $k = 5$ for 4th grade reading.

For the categorical contrast of states that differ in some combination of HS and CA, we estimate the same model as above except that we now have three dummies instead of two, resulting in three additional interaction terms. The model is as follow:

$$\begin{aligned} Y_{ti} = {} & \beta_0 + \beta_1(policy)_{ti} + \beta_2(policy \times year)_{ti} \\ & + \beta_3(year \times group\_hc)_{ti} + \beta_4(year \times group\_h)_{ti} + \beta_5(year \times group\_c)_{ti} \\ & + \beta_6(policy \times group\_hc)_{ti} + \beta_7(policy \times group\_h)_{ti} + \beta_8(policy \times group\_c)_{ti} \\ & + \beta_9(policy \times year \times group\_hc)_{ti} + \beta_{10}(policy \times year \times group\_h)_{ti} \\ & + \beta_{11}(policy \times year \times group\_c)_{ti} \\ & + \beta_{12}(percent\_free\_lunch)_{ti} + \beta_{13}(pupil\_teacher\_ratio)_{ti} + \mu_i + \tau_t + \varepsilon_{ti}, \end{aligned} \tag{3}$$

where $group\_hc$, $group\_h$, and $group\_c$ in the interaction terms are dummy variables indicating respectively HS & post-NCLB CA states, HS states only, and CA states only. Again, hypothesis tests of differences in mean and slope changes between groups are represented by two- and three-way interactions. Regression coefficients β_6 of the $policy \times group\_hc$ interaction, β_7 of the $policy \times group\_h$ interaction, and β_8 of $policy \times group\_c$ interaction give the differences in the mean change in 2002 for each of the three treatment groups when compared to the control group. Coefficients β_9, β_{10}, and β_{11} test whether the groups differ in their post-NCLB slope changes. The hypotheses of interest are the same as above, but with coefficients from model (3).

Scaling. Since it is difficult to understand the meaning of causal effects in NAEP points we also compute effect sizes as individual-level standard deviation units. For this we use NAEP-provided grade- and subject-specific standard deviations (SD) from individual student test score data. We also compute percentile rank gains - the number of ranks a state would have risen relative to other states by virtue of its NCLB gains. Gains in percentile rank are based on the distribution of state rank in 2002. Finally, we translate the study's obtained effect sizes to months of learning. Analyses of nationally normed tests by Hill (2007) show that the average annual test score gain in effect size from 4th to 5th grade is roughly 0.40 standard deviation units for reading and 0.56 for math. A much smaller effect size of 0.22 is observed for the average test score gain from 8th grade to 9th grade math. So an obtained effect size of 0.20 SD in 4th grade reading translates to 6 months worth of learning based on the benchmark effect size of 0.40 (i.e., 0.20/0.40 x12 months). But the same effect size will translate into many more months of learning in 8th grade because of smaller benchmark effect sizes. Since none of these transformations is perfect, we present all of them in the tables. However, for simplification reasons alone, the text presents the marginal math or reading gain in standardized effect sizes and calendar months with the above proviso about why 8th grade results will seem larger in the latter metric.

Serial Autocorrelation. ITS data run the risk of biased standard errors due to serial autocorrelation. Like Dee and Jacob (2009), we use the robust standard errors using the CLUSTER option in STATA as one way to guard against this (Rogers, 1993). Better would be ARMA modeling (Box & Jenkins, 1970), but the time series is too short to do that well. We did do it, nonetheless, and found the best approximation to be a second order autoregressive model. Results hardly differed by the way in which we corrected for serial autocorrelation, and so we report here clustered standard errors.

Results

National Contrasts of Public vs. Private Schools. Figures 1a-c present visual results from Trend NAEP data for each outcome comparing public schools to Catholic schools. Figures 2a-b through Figures 4a-b present the visual results for each outcome comparing public schools to Catholic and non-Catholic private schools using Main NAEP. Raw time series means are presented with best fitting regression lines.

For the pre-intervention time period, all figures indicate: (1) a high degree of stability since the regression lines fit the data very well, as they do after NCLB also; (2) all the functional forms seem reasonably linear, with the only possible exception being for 4th and 8th grade math observations in 1990; (3) there is no evidence of a sudden change in test scores just before 2002, thus ruling out regression to the mean as a possible alternative interpretation; (4) the pre-intervention means always favor the comparison group, whether Catholic or not; and (5) pre-intervention slope differences suggest that achievement was changing at a faster rate in Catholic than public schools, but not in non-Catholic private schools relative to public ones. To be effective, NCLB would have to reverse or reduce these initial differences favoring the comparison groups and so reduce the gap between them.

Table 6 and 7 report on statistical tests and effect size estimates for the differences in differences in mean (H_1), slope (H_2) and total change (H_3). There is no statistical evidence of a reading effect, though almost all coefficients trend in the hypothesized direction.

However, there is evidence of math effects. For the public/Catholic contrast, the pre- and post-NCLB mean differences in math differ at the .05 level for both 4th and 8th grade math when Trend NAEP is used. For 4th grade math, the gain by 2004 is 10.9 points (Table 6) or 7.2 months of learning (Table 7) and for 8th grade math, the gain is 7.3 points or 13.0 months. The Main

NAEP math results are also all in the hypothesized direction, but only one is marginally significant – a total 4th grade math gain by 2009 of 11.0 points or 8.9 months.

As for the public vs. non-Catholic private school contrast using Main NAEP, all the math differences are in the required direction. However, none reached statistical significance unless the outlier 1990 data point was removed. Then, a reliable 8th grade total math effect was observed.

These small sample contrasts of public and private schools show reliable effects on math when Trend NAEP is used; and effects are consistently in the right direction when Main NAEP is used, though only the total effects that come from combining mean and slope changes are reliable. It seems reasonable to conclude therefore, that something happened in the nation's public schools in 2002. It also seems reasonable to conclude that this force reduced the 4th and 8th grade gaps initially favoring non-Catholic private schools over public ones by about a half. It also reduced the 4th grade gap initially favoring Catholic over public schools by similar amount.

These 2002 differences of differences can only be attributed to NCLB if we can show they were not due to a higher fraction of better performing students entering public schools in 2002 relative to private schools. To explore this, we examined data on the national composition of the three types of schools over time on observables usually correlated with achievement. Information for public schools comes from Common Core Data (CCD) and for private schools, the Private School Universe Survey (PSUS). Common variables are available for overall enrollment, enrollment by grades, race, and student-to-teacher ratio. Table 8 presents coefficients for the change in mean, slope and total change by 2006 for all common variables. No reliable changes in composition were observed after 2002 between these types of schools, thus reducing the plausibility of internal validity threats based on differential compositional shifts around 2002.

It is also necessary to rule out alternative interpretations based on events occurring differentially in public and private schools around 2002 that would affect achievement in math (but not reading). The sex scandals in Catholic schools do not apply to non-Catholic private schools. It is not likely that changes in the sampling design of Main NAEP in 2002 affected the national results because they would presumably have affected both public and private schools and they could not have affected Trend NAEP, which showed similar short-term results to Main NAEP. Even so, national contrasts are bound to be vulnerable to unidentifiable forces differentially operating between public and private schools in 2002.

Contrast of States varying in Proficiency Standards. State-level analyses are limited to public schools and so hold constant all forces that differentiate public from private schools unless these forces also happen to vary at the state level in ways that are correlated with our definition of HS and LS states.

No state-level Trend NAEP data exist and so we present only Main NAEP results. Figures 5a-c present the relevant time series and Tables 9 through 10 give the relevant statistical results.[3] They show no statistically significant effects for reading (though all are in the hypothesized direction) but many for math at both grade levels. The discussion that follows focuses mostly on HS and LS states, though the statistical analyses included the 25 states with medium-level standards.

For 4th grade math, the difference in mean change immediately after 2002 is significant at the .05 level, with HS states doing best. All differences in slope change are also in the required direction. Combining the differences of differences in both means and slopes leads to 4th grade

[3] The scatter plots are shown with a linear fitted line (even though our model specified year fixed effects) to provide ease of interpretation and comparison with other graphs.

means in 2009 that are significantly different from what pre-NCLB mean and slope differences projected them to be.

As for 8th grade math, the initial difference in mean differences is in the required direction but not reliable. However, the difference of slope differences favor HS states and is statistically significant, as is also the total change observed relative to the difference in initial baselines. HS states had initially lower test scores than LS states but changed more after NCLB and so narrowed the prior gap. The size of the total 4th grade math gain by 2009 is 7.4 points (Table 9) or 5.7 months of learning (Table 10). For 8th grade math, it is 7.0 points or 10.6 months of learning.

These state-level findings are important and so require particularly critical scrutiny. The cutoff values used to define high, medium, and low proficiency standards are arbitrary. So we reanalyzed the data using the continuous measure of state proficiency standards in 2003 (Table 9). Again, the total effect estimates were non-significant for reading but significant for math. All *t*-values are largely similar to those from analyses with three categories representing different levels of proficiency standards. We also contrasted states with *medium* and low proficiency standards. All coefficients for the total effect by 2009 are in the hypothesized direction, but are understandably smaller than for the HS vs. LS contrast and none is statistically significant (results not shown but available upon request). This suggests that states with medium standards may have improved more than LS states but less than HS states, exactly the intermediate change status we would expect. Another specification test reclassified states according to the difference between their state and NAEP scores in 2003, thus controlling for true state differences in achievement. Adding the NAEP test to the description of the causal agent again made no difference to the basic pattern of results, though effects did tend to be smaller and fewer tests are

statistically significant. We defined HS in 2003, and critics might argue that endogeneity could be a problem. So we reclassified HS as state values in 2001 rather than 2003 and re-ran all the achievement analyses. The correlation between standards in 2001 and 2003 is high ($r = 0.85$), and the achievement differences due to assessing HS at these different times is negligible. Changing treatment dates is not consequential.

We are cognizant that bias in statistical tests can occur when many tests are conducted that inflate the overall type I error rate. So our analytic strategy weights *independent replication* more than null hypothesis tests, and the same pattern of results basically emerged in analyses across 4^{th} and 8^{th} grade math, at the state and national levels, in both Catholic and non-Catholic private schools and, where testable, when both Trend and Main NAEP data were used. So the present results are robust by replication.

Although something happened in 2002 that led HS states to begin out-performing LS states, causal interpretation requires ruling out whether student populations suddenly changed around 2002 in ways that advantaged schools in HS states. We again examined time series data from Common Core Data (CCD) on variables such as enrollment, the percent of students eligible for free lunch, per pupil expenditures, pupil-to-teacher ratios, percent Black, White and Hispanic students, and percent 4^{th} and 8^{th} graders. We also examined the percentage of elementary teachers (out of the total number of teachers) and the student to guidance counselor ratio to see if more resources were reallocated to HS states in or about 2002. Graphs and corresponding ITS analyses (available upon request) show no visually large or statistically detectable changes in or around 2002 on any of these variables.

The types of students taking the NAEP test could also have differentially changed by type of state about 2002. For instance, if relatively fewer educationally disabled or limited English

proficiency students were represented in NAEP testing in HS states. Table 11 provides data on the percentage of student identified with a disability (SD) or as English Language Learners (ELL). There is a general increase over time, but no systematic differences in the classification of students around 2002 in HS relative to LS states. Table 12 shows the percentage of SD and ELL students from the total student count who were excluded from NAEP testing immediately around 2002 and 2003 as well as for the entire period before and after NCLB for which we have data. It shows that the *actual* exclusion rate has dropped since 2002 and generally more so for students from HS states. However, differentially removing more of the lowest scoring students from NAEP testing in HS states would presumably under- and not over-estimate NCLB effects.

Did any other events co-occur with NCLB that might have *differentially* affected math (but not reading) in HS states? We are not able to obtain systematic data on math curricula changes around 2002 and so cannot test whether they are different in schools with higher standards for reasons that have nothing to do with NCLB. The National Council of Teachers of Mathematics (NCTM) updated its math standards in 2000 and later claimed that this was responsible for the subsequent steady improvements in NAEP math scores nationally (National Council of Teachers Mathematics, 2008). But it is not clear that the NCTM standards were adopted more or better in HS states, that they can even raise test scores, or that they do so with a three year causal lag. Indeed, one evaluation concluded that all states adopted NCTM standards and with little variation, and that the NCTM standards did not improve student math learning and may even have hurt it (Fordham, 2005). Another possibility is that the IASA reforms of 1994 were eventually implemented better or more frequently in HS states around 2002. Given the conceptual and operational overlap of IASA and NCLB, it is not clear whether this would constitute an alternative interpretation of NCLB or a restatement of it. But in any event, the

IASA explanation requires assuming that the program's implementation was delayed and reached a critical mass only in 2002 and then more in HS than LS states -- a growing concatenation of necessary events of ever diminishing total probability.

Combining higher State Standards in 2003 and Consequential Accountability in 2002. We examined four groups of states – HS states that either did or did not have CA pre-2002, and LS states that either did or did not have CA before 2002.

Table 13 reports on group differences and the HS x CA interaction effects when the high and low standards groups were used together with the states Dee and Jacob classified as CA pre- or post-2002. When reporting estimated group differences, the LS & pre-NCLB CA states serve as the reference group since they are the closest approximation to a no-treatment comparison group of states. This means we estimate the extent to which these reference group states are outperformed (1) by states with both HS & CA after 2002, (2) by states with HS but CA prior to 2002, and (3) by states with LS & CA after 2002.

The interaction effects were obtained from a re-parameterized model and in both specifications are negative in sign for both 4^{th} and 8^{th} grade math but positive for reading. While the power of these interaction tests is presumably low, their absolute values generally seem large and point to the low likelihood of an additive model for math but a higher likelihood for reading. For 4^{th} grade math, all three groups do better than the no-treatment group, with most of the effects being statistically significant. While each mechanism made a difference, combining them adds nothing. The coefficients for their joint influence have approximately the same value as the coefficients representing just HS or just CA. For 8^{th} grade math, the interaction is again negative and the combined HS and CA effect is again no larger than the separate (and usually not reliable)

individual HS and CA effects. It is as though the CA and HS mechanisms are substitutable as to effects, even though they are quite different sets of states.

The 4th grade reading results are quite different. First, the interaction effects are positive in sign, though smaller and never statistically significant. Still, the absence of a negative interaction suggests that the two mechanisms might well be additively related. Indeed, combining HS and CA always leads to a statistically significant total effect size by 2007 that is larger than for either HS or CA alone. Thus, in the largest sample test, the CA reading effect in NAEP points is 2.05, the HS effect is 1.96, and the combined effect is of 4.19, and only the last is significantly different from the reference group of states. So it seems that combining two small and unreliable reading effects for two different NCLB mechanisms generates the first reading effect ever attributed to NCLB, albeit a contingent one.

Discussion

Summary. This study has three main findings. First, it constitutes a technical improvement over Wong (2008) and reveals the same effects of NCLB on both 4th and 8th grade math when the causal contrast is specified as how high a state's standards are in 2003. Dee and Jacob (2009) have also demonstrated the 4th grade math effect and hints at an 8th grade one, but using a different treatment specification – whether consequences (sanctions) were or were not included in a state's accountability system before 2002. The present analyses make the Wong (2008) claims and her 8th grade NAEP math finding more secure.

Second, this study shows that the same two math effects are apparent in national level analyses when both Trend and Main NAEP data are used and when the total change occurring after 2002 in public schools is compared to the total change separately occurring in Catholic and in non-Catholic private schools. Such results are important. They reflect the fact that NCLB is

national in both content and reach; and they do not require imperfect measures of how high standards are or when sanctions became part of an accountability system. Although sources of state-level variation like CA and HS speak to mechanisms within NCLB, there is little evidence that many states changed their standards *because of NCLB* or that all the states that adopted sanctions in 2002 did so *because of NCLB.* Sanctions were being regularly adopted by states up to 2001, and in some cases this process might have continued into 2002 even without NCLB while state standards were relatively constant from before to after NCLB. Clearer is that NCLB affected public schools in 2002 more than either Catholic or non-Catholic private schools,

Third, reading effects are apparent for the first time in time series studies of NCLB, though they are more contingent than the math results. That is, they are only detectable when a state adopting sanctions in 2002 also has high standards. The two mechanisms are orthogonal, and each by itself has only a small and statistically non-detectable reading effect. However, when combined the two create something reliable and larger (Table 14 shows the effect to be of about .11 standard deviation units). But while the mechanisms of CA and HS are at least additively related for reading, this is not the case for math where scores are improved either by sanctions or by higher standards.

Math Effects. By most conventional standards, the math effects are large at the national or state levels. By 2009, math achievement had increased by about .30 standard deviation units in 4th grade on both the national and state tests, corresponding to a gain of about six to seven months. In 8th grade the difference is closer to .15 units. In the comparison of public and non-Catholic private schools, for both 4th and 8th grade math the achievement gap favoring private schools before NCLB is reduced by half by 2009.

The math results seem to be attributable to both an immediate mean change in 2002 and a slope change thereafter. These two separate effects were not statistically detectable in all analyses, but each is visually apparent in all graphs and some were statistically detectable by the usual criteria. The slope differences are perhaps more interpretable since the most serious school-level changes only occur after many years of failure to make AYP. Less understandable in program design terms is the immediate math increase, given the usual pitfalls of immediate implementation and the fact that the law's most fundamental provisions for school change are linked to *successive* years of failure.

We have no definitive explanation for the immediate effect. One possibility follows from the law passing Congress in June 2001 and being widely discussed beforehand. So schools could have focused on raising test scores in anticipation of the law's passage. This implies many passing schools may have changed out of fear of being publicly stigmatized for possible future failure rather than because of the reforms implemented as a consequence of years of repeated failure. To borrow language from criminology, was the mechanism for the immediate math effects school improvement efforts designed to deter future punishment, while the mechanism for the slope effect was the school improvements required as punishments for repeated failure? Or did the two operate in some as yet unidentified package of influences?

The immediate math effect makes one suspect that some artifact occurred in 2002 and raised achievement. The change in the Main NAEP test sampling design will not suffice as an explanation for spuriousness, since the Trend NAEP did not change its sampling design in 2002 and also showed an immediate math effect. The adoption of NCTM math standards in 2000 probably cannot account for the math effects either. No relevant theory specifies a causal lag of two years, and available evidence suggests the standards were not very effective anyway. Catholic schools experiencing a sex scandal cannot provide an adequate explanation either since the same result is obtained for non-Catholic private schools; and it is not evident that a greater percentage of better performing children left Catholic schools. Table 8 shows no evidence that the migration from Catholic schools was systematic with respect to race or changes in class size. Widespread fraud in testing after NCLB is also possible, but not very plausible. NAEP math is a low stakes test when compared to state achievement measures. Also, 12th grade math time series (not reported here) do not indicate any evidence of immediate gains (Stullich et al., 2007). Why should fraud affect elementary and middle school math but not high school math? Then there is the smaller and contingent reading effect. Why should there be more fraud in math than reading? The sudden large increases in math achievement after 2002 surprised us. But no obvious alternative interpretations can withstand the multiple falsifications built into generally replicated results from tests varying: (a) two different kinds of comparison schools in national tests; (b) two different kinds of causal agent in state-level tests – HS and CA; (c) two grades for math – 4th and 8th; and (d) two different achievement tests – Trend and Main NAEP.

On all metrics except one (months of individual learning gain), the effect sizes for 4th grade math are larger than for 8th grade math. This research was not designed to explain such a finding, and so we can only speculate. One possibility is that mathematics is particularly

cumulative so that failure to master earlier material inhibits learning later material. If this were the case, math would become especially more difficult as a student advances through school.

Reading Effect. This is the first study to claim a reading effect associated with NCLB. (Actually, it is better labeled a comprehension effect since the *NAEP 4th grade reading* test assumes mastery of the mechanics of reading and only assesses understanding of text). Previous work has not succeeded in discovering statistically significant reading results from NCLB. The present study did, presumably because it combined two causal mechanisms whereas Wong (2008) and Dee and Jacob (2009) each examined only one.

The reading effect is more contingent than the math ones. States did best in reading if their accountability system had both high standards and sanctions newly added in 2002. So neither HS nor CA alone was sufficient for a detectable reading effect with the power of this study or of Dee and Jacob. Also worth noting is that the reading effect is smaller than either math effect, being about .10 standard deviation units when both mechanisms are combined. In months of gain over five years, the combined effect is of the order of 3.5 to 4.5 months depending on whether standards are scaled as median splits with nearly all states or as more extreme contrasts including only the HS and LS states that were also in Dee & Jacob.

The reading effect emboldens us to reconsider past NCLB-related reading findings, the more so because all the non-significant reading effects in this paper were in the direction indicating an NCLB effect in contrasts of both public versus private schools, of states with high versus low standards, and of states with CA before or after 2002. The main reading program in NCLB is Reading First. It was evaluated in a quantitative synthesis of 17 small regression-discontinuity studies and one randomized experiment (Gamse, 2008). The study concluded that Reading First failed to increase reading. However, the estimate was in the required direction and

statistically significant at the .10 level in an only modestly powered cumulative test across sites. Had conventional null hypothesis testing criteria been used in a more flexible way, the conclusion might well have been that the results about a reading effect were indeterminate. So putting this marginal finding together with the robust but contingent reading results reported here suggests that, in the nation at large, NCLB may well have increased reading.

To gain almost a month a year in math and half a month a year in reading from 2002 to 2007 (for reading) or 2009 (math) is not trivial. We have used highly aggregated data representing millions of students and have no responsible idea what the national consequences are of such average gains over so many persons. Moreover, many sources of potential treatment heterogeneity were not estimated here and probably cannot be estimated with a high level of confidence since we do not have individual-level data. Yet it is plausible to contend that NCLB could have heterogeneous effects. For instance, if it led to targeting "bubble" students (Neal & Schanzenbach, in press) or those most economically disadvantaged.

Conclusion

From a policy perspective, the most important implication of this study is that NCLB made a difference, not just to 4th grade math, but also to 8th grade math and probably also 4th grade reading. The next most important conclusion is that standards matter when they are linked to consequences. Newly adopting consequences in a state also mattered for math, but now that every state has to have CA it has no future leverage as a policy tool. However, leverage is still possible from raising standards and making them more uniform across states. $350 million of the ARRA money is to be spent to support state efforts to develop common academic standards (Department of Education, 2009c), and the current findings predict that this could be money well spent if the uniform standards are high ones.

No research report can do everything, and this one offers only a partial evaluation of NCLB. We have not provided the level of detail on implementation that one finds in the congressionally mandated evaluation of Title 1 (Stullich et al., 2007). Nor have we tested whether effects are larger in states with higher proficiency standards *because* more and more stringent reform activities are undertaken there. Nor have we waited until 2014 when current legislation mandates that all children are to be proficient by the standards of a given state, though this endpoint will certainly change at the next Congressional re-authorization. Nor have we examined NCLB's effects on housing prices, teacher behavior, fraud and abuse, student targeting, social-behavioral or affective outcomes or even high school achievement.

Instead, this partial evaluation of NCLB focuses on a single central issue: Has the program raised achievement test scores among younger children in the nation at large, particularly in those states where the program has more teeth because higher proficiency standards led many more schools to fail and thus have to change and even led some passing schools to change how they did business out of fear of the consequences of future failure? We conclude that NCLB has had positive effects that are largest for 4th grade math, next largest for 8th grade math, and smallest and most contingent for 4th grade reading. We also conclude that leverage for positive results can still be found in the program by raising standards and making them more uniform across states.

References

Borman, G.D., D'Agostino, J.V. (1996). Title I and Student Achievement: A Meta-analysis of Federal Evaluation Results. Educational Evaluation and Policy Analysis, 18(4): 309-326.

Box, G. E. P., & Jenkins, G. M. (1970). *Time series analysis: Forecasting and control*. San Francisco: Holden-Day.

Branigin, W. (2009, July 24). Obama Launches 'Race' for $4 Billion in Education Funds. *The Washington Post*. Retrieved August 29, 2009 at http://www.washingtonpost.com/wp-dyn/content/article/2009/07/24/AR2009072402203.html?hpid=topnews

Carnoy, M. and Loeb, S. (2002). Education Evaluation and Policy Analysis. 24: 305-331.

Center on Education Policy. (2007). Answering the Question that Matters Most: Has Student Achievement Increased Since No Child Left Behind.

Center on Education Policy. 2008. Many States Have Taken a "Backloaded" Approach to No Child Left Behind Goal of All Students Scoring "Proficient".

Chapman, Laura H. 2007. "An Update on No Child Left Behind and National Trends in Education." *Arts Education Policy Review* 109 (1): 25-36.

Cook, T. D., Wong, V. C., Steiner, P. M.,Taylor, J., Gandhi, A., Kendziora, K., Choi, K., et al. (2009). Impacts of School Improvement Status on Students with Disabilities: Feasibility Report. Washington, DC: American Institutes for Research

Dee and Jacob (2009). "The Impact of No Child Left Behind on Student Achievement. National Bureau of Economic Research: Cambridge, MA. Working Paper 15531.

Department of Education. (1994). Improving America's Schools Act of 1994. Archived Information. Retrieved June 8, 2008 at http://www.ed.gov/legislation/ESEA/toc.html.

Department of Education. (2002). Private Schools: A Brief Portrait. (Washington, D.C.: Department of Education)

Department of Education. (2007a). Private School Participants in Federal Program Under the No Child Left Behind Act and the Individuals with Disabilities Education Act. (Washington, D.C.: Department of Education)

Department of Education (2007b). Mapping 2005 State Proficiency Standards Onto the NAEP Scales. (Washington, D.C.: U.S. Department of Education)

Department of Education. (2008). No Child Left Behind Act of 2001, Public Law print of PL 107-110. Retrieved June 8, 2008 at http://www.ed.gov/policy/elsec/leg/esea02/107-110.pdf.

Department of Education (2009a). The Federal Role in Education. Retrieved March 1, 2009 at http://www.ed.gov/about/overview/fed/role.html

Department of Education, (2009b). The American Recovery and Reinvestment Act of 2009: Saving and Creating Jobs and Reforming Education. Retrieved March 7, 2009 at http://www.ed.gov/policy/gen/leg/recovery/implementation.html

Department of Education, (2009c). Race to the Top Fund, 84.395A: Notice of Proposed Priorities, Requirements, Definitions, and Selection Criteria. Last Retrieved August 29, 2009 at http://www.ed.gov/legislation/FedRegister/proprule/2009-3/072909d.html

Department of Education, (2009d). American Recovery and Reinvestment Act of 2009: Using ARRA Funds to Drive School Reform and Improvement. Retrieved August 29, 2009 at www.ed.gov/policy/gen/leg/recovery/guidance/uses.doc

Dillion, S. (2009, August 17). Dangling Money, Obama Pushes Education Shift. *The New York Times*. Retrieved August 29, 2009 at http://www.nytimes.com/2009/08/17/education/17educ.html?_r=1

Dillon, S. (2007). New Study Finds Gains Since No Child Left Behind. *New York Times*. Retrieved June 10, 2009 at http://www.nytimes.com/2007/06/06/education/06report.html

Dinan, S. (2009, March 10). Obama to Demand 'Rigorous' School Standards. *The Washington Times*. Retrieved June 10, 2008 at http://www.washingtontimes.com/news/2009/mar/10/obama-demand-rigorous-school-standards/

Education Next. (2009). The Future of No Child Left Behind: End It or Mend It. Education Next. Summer 2009, p.49-56.

Fordham Foundation (2005). The State of State Math Standards. The Fordham Foundation: Washington, D.C.

Fuller, B., Gesicki, K., Erin, K., Wright, J. (2006). Is the No Child Left Behind Act Working?: The Reliability of How States Track Achievement (Berkeley, CA: Policy Analysis for California Education)

Fuller, B., Wright, J., Gesicki, K and Kang, E. (2007). Gauging Growth: How to Judge No Child Left Behind. Educational Researcher. 36(5): 268-278.

Gamse, B.C., Bloom, H.S., Kemple, J.J., Jacob, R.T. (2008). Reading First Impact Study: Interim Report (Washington, D.C.: Department of Education, Washington, DC)

Goertz, M.E. (2005). Implementing the No Child Left Behind Act: Challenges for the States. Peabody Journal of Education, 80(2): 73-89.

Hansen, B.B. and Klopfer, S.O. (2006). Optimal full matching and related designs via network flows, Journal of Computational and Graphical Statistics, 15, 609-627.

Hanushek and Raymond (2005). "Does School Accountability Lead to Improved Student Performance?" Journal of Policy Analysis and Management, 24(2), 297-327.

Hollister, R. (2009). Reply to Point/Counterpoint statements on "The role of random assignment in social policy research". Journal of Policy Analysis and Management, 28(1), 178-180.

Hill, C., Bloom, H., Black, A., and Lipsey, M. (2007). Empirical Benchmarks for Interpreting Effect Sizes in Research, (New York, New York: Manpower Demonstration Research Corporation)

Keigher, A. (2009). Characteristics of Public, Private, and Bureau of Indian Education Elementary and Secondary Schools in the United States: Results from the 2007-08 Schools and Staffing Survey (NCES 2009-321). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.

Kingsbury, G.G., Olson, A., Cronin, J., Hauser, C., and Hauser, R. (2007). The State of the State Standards: Research Investigating Proficiency Levels in Fourteen States" (Lake Oswego, OR: Northwest Evaluation Association)

Lazer, S. (2004). The Nation's Report Card: Evolution and Perspectives. L. Jones, I. Olkin, Eds. (Bloomington, IN: Phi Delta Kappa Educational Foundation and American Educational Research Association)

Lee, Jaekyung. (2006). Tracking Achievement Gaps and Assessing the Impact of NCLB on the Gaps: An In-Depth Look into National and State Reading and Math Outcome Trends. Cambridge, MA: Harvard Civil Rights Project

McDonnell, L.M. (2005). No Child Left Behind and the Federal Role in Education: Evolution or Revolution? Peabody Journal of Education, 80(2): 19-38.

Milbank, D. (2002, January 9). With Fanfare, Bush Signs Education Bill. *The Washington Post*, p.A3

National Center for Education Statistics. (2009). How the Samples of Schools and Students Are Selected for the Main Assessments (State and National). Retrieved March 8, 2008 at http://nces.ed.gov/nationsreportcard/about/nathow.asp

National Commission on Excellence in Education (1983). A Nation at Risk: the Imperative for Educational Reform: A Report to the Nation and the Secretary of Education. (Washington, D.C. : U.S. Department of Education)

National Council of Teachers Mathematics. 2008. "Rise in NAEP Math Scores Coincides with NCTM Standards." *NCTM News Bulletin* 2008 (January/February): 1-2.

Neal, D. & Schanzenback, D.W. (in press). "Left Behind by Design: Proficiency Counts and Test-Based Accountability" *Review of Economics and Statistics*.

R Development Core Team (2009). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Raudenbush, S. and Bryk, A. Hierarchial Linear Models: Applications and Data Analysis Methods (Thousand Oaks, CA: Sage Publications)

Rogers, W. H. (1993). Regression standard errors in clustered samples. *Stata Technical Bulletin* 13: 19–23. Reprinted in *Stata Technical Bulletin Reprints*, vol. 3, 88–94.

Rudalevige, A. (2003). "No child Left Behind: Forging a Congressional Compromise." Pp. 23 - 54 in *No Child Left Behind? The Politics and Practice of School Accountability*, P.E. Peterson and M.R. West (The Brookings Institute: Washington, D.C.)

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

Shear, M.D. and Anderson, N. (2009, July 23). President Obama Discusses New "Race to the Top" Program. *The Washington Post.* Retrieved August 29, 2009 at http://www.washingtonpost.com/wp-dyn/content/article/2009/07/23/AR2009072302938.html?hpid=topnews

Skinner, R. A. (2005). State of the States. Education Week, p.77-80.

Smith, Emma. (2005). Raising Standards in American Schools: The Case of No Child Left Behind. Journal of Educational Policy, 20(4): 507-524.

Stullich, Stephanie, Eisner, E. and McCrary, J. (2007). *National Assessment of Title I: Final Report*. U.S. (Washington, D.C.: Department of Education)

Sunderman, G.L., Kim, J.S., Orfield, G. (2005). NCLB Meets School Realities: Lessons from the Field. (Thousand Oaks, CA: Corwin Press)

Wong, Manyee. (2008). Studies of Educational Change from Three Disciplinary Perspective: Sociology, Policy, and Human Development. Ph.D. Dissertation, Northwestern University, United States – Illinois. Retrieved March 5, 2010, from Dissertations & Theses @ CIC Institutions.(Publication No. AAT 3336462).

Figure 1a. Trend NAEP 4th grade reading scores by year:
Public and Catholic schools

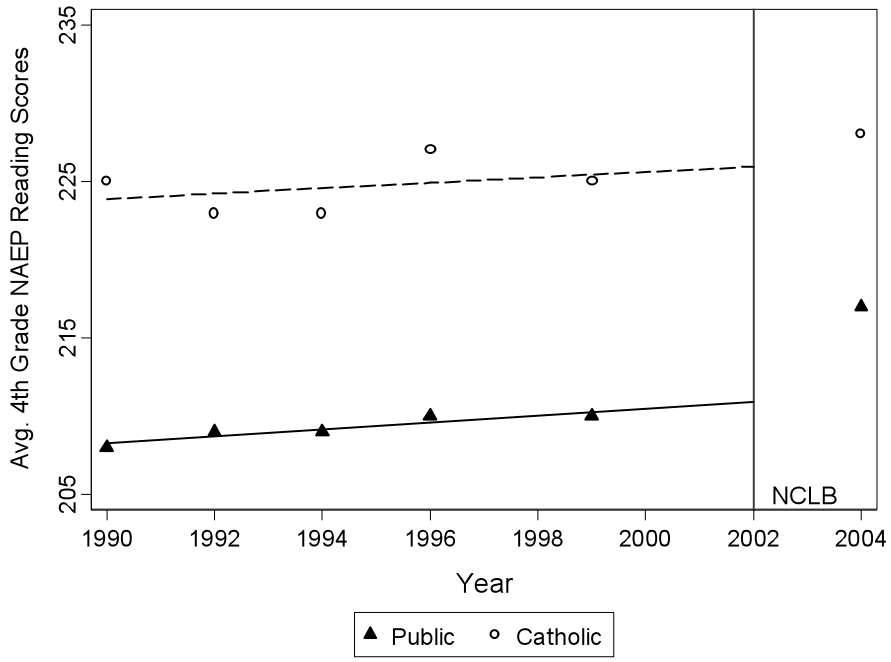

Figure 1b. Trend NAEP 4th grade math scores by year:
Public and Catholic schools

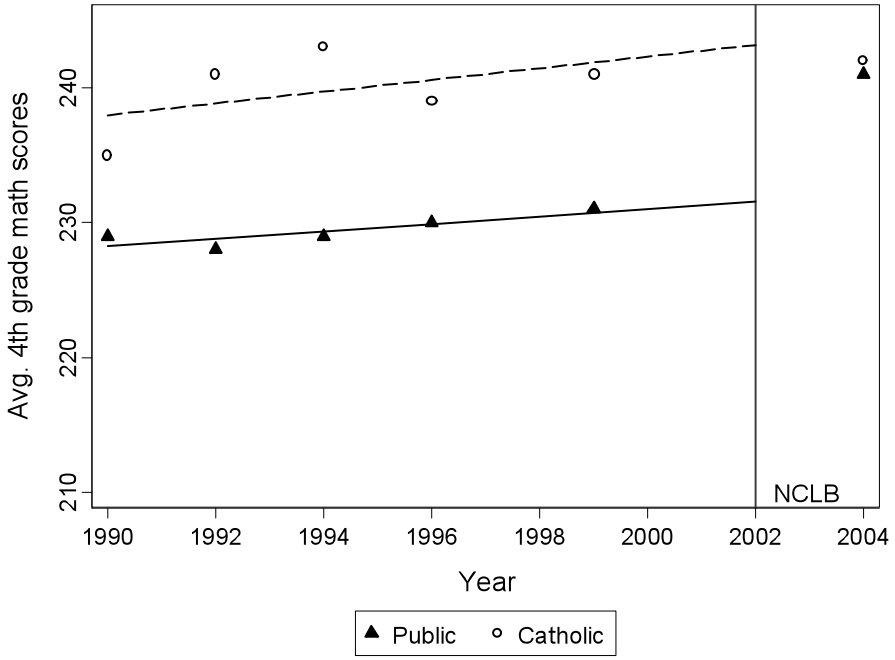

Figure 1c. Trend NAEP 8th grade math scores by year:
Public and Catholic schools

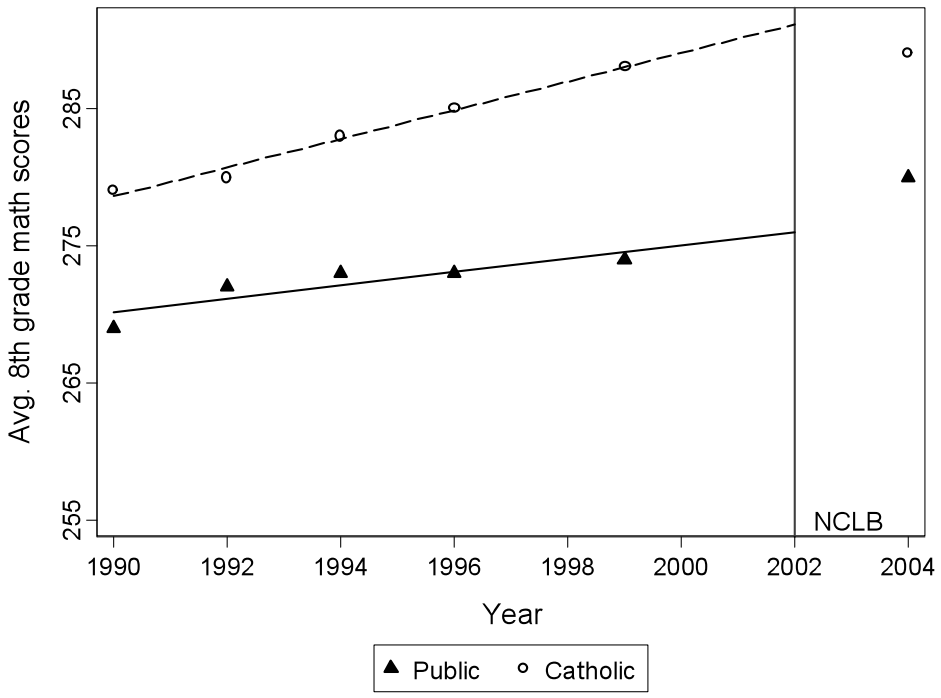

Figure 2a. Main NAEP 4th grade reading scores by year:
Public and Catholic schools

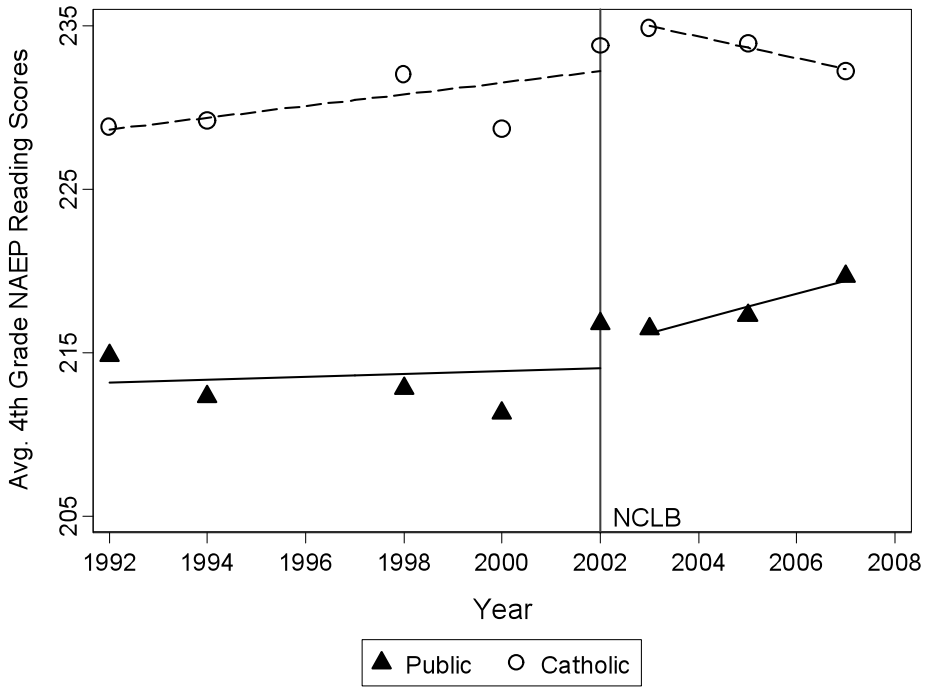

Figure 2b. Main NAEP 4th grade reading scores by year:
Public and non-Catholic private schools

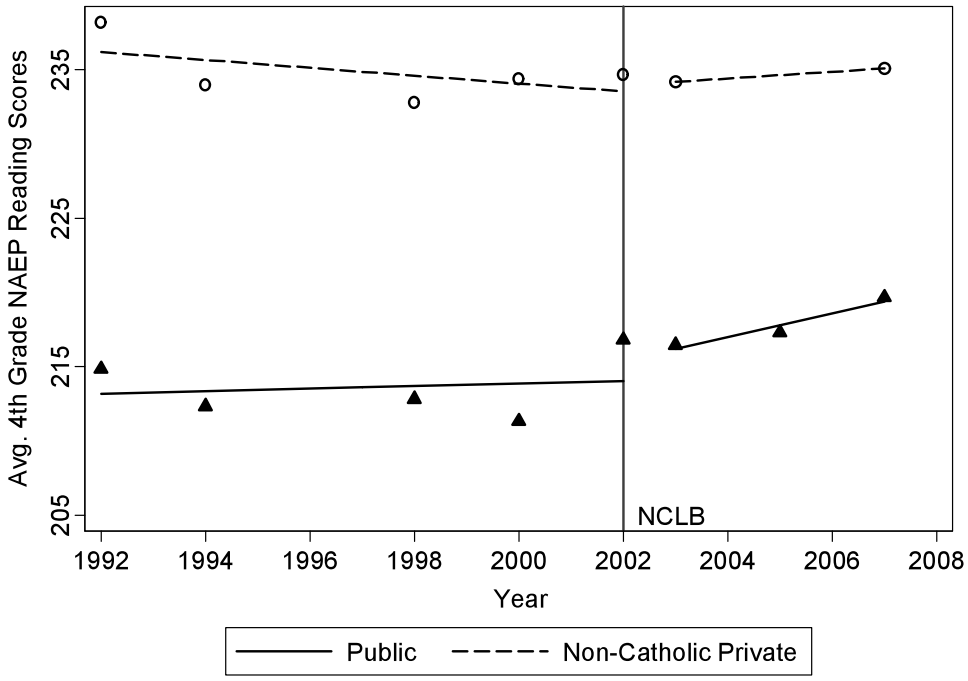

Figure 3a. Main NAEP 4th grade math scores by year:
Public and Catholic schools

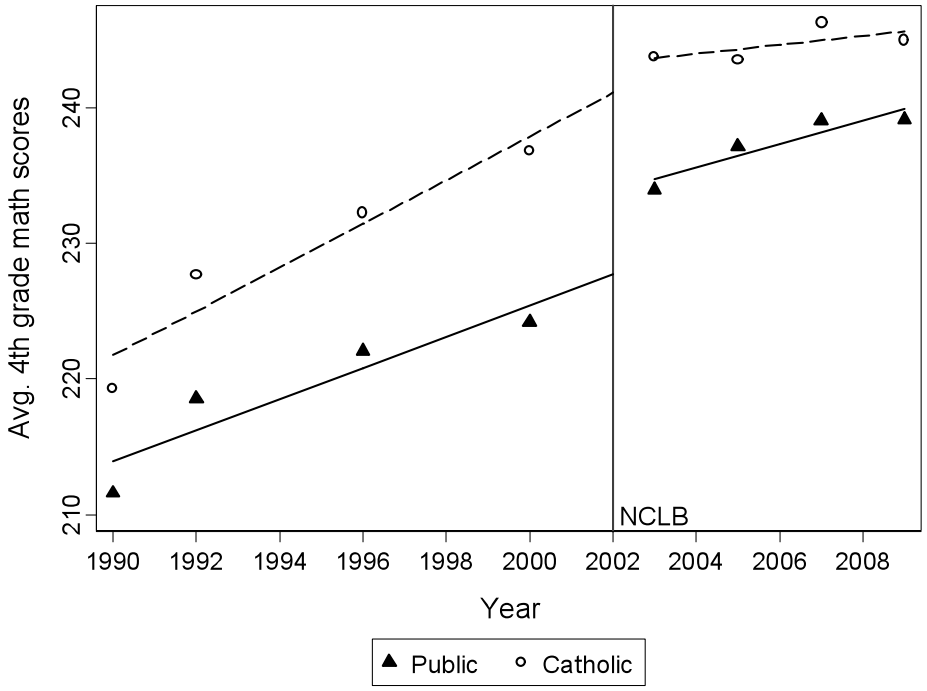

Figure 3b. Main NAEP 4th grade math scores by year:
Public and non-Catholic private schools

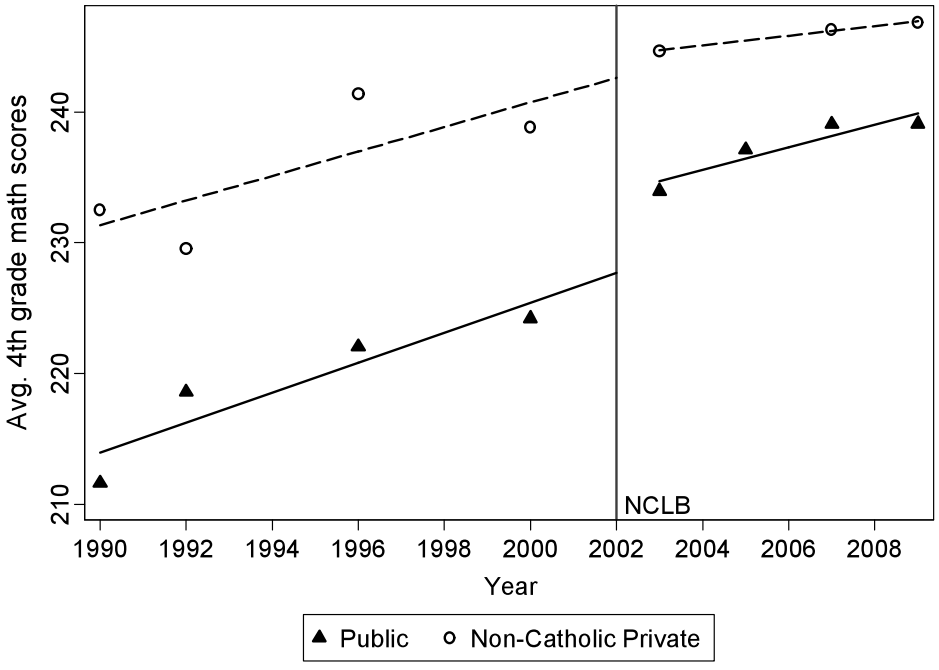

Figure 4a. Main NAEP 8th grade math scores by year:
Public and Catholic schools

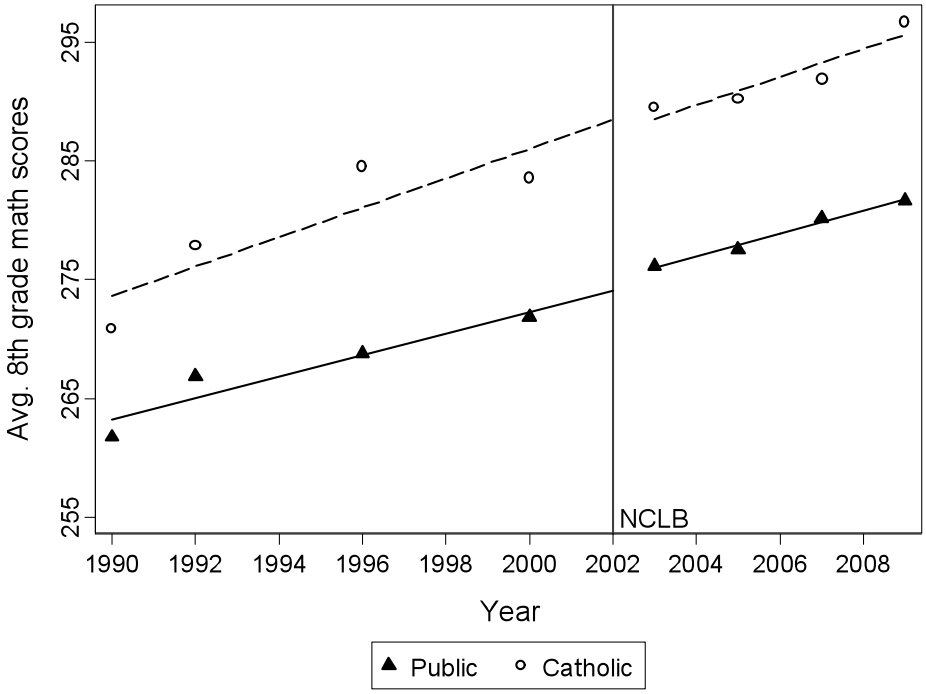

Figure 4b. Main NAEP 8th grade math scores by year:
Public and non-Catholic private schools

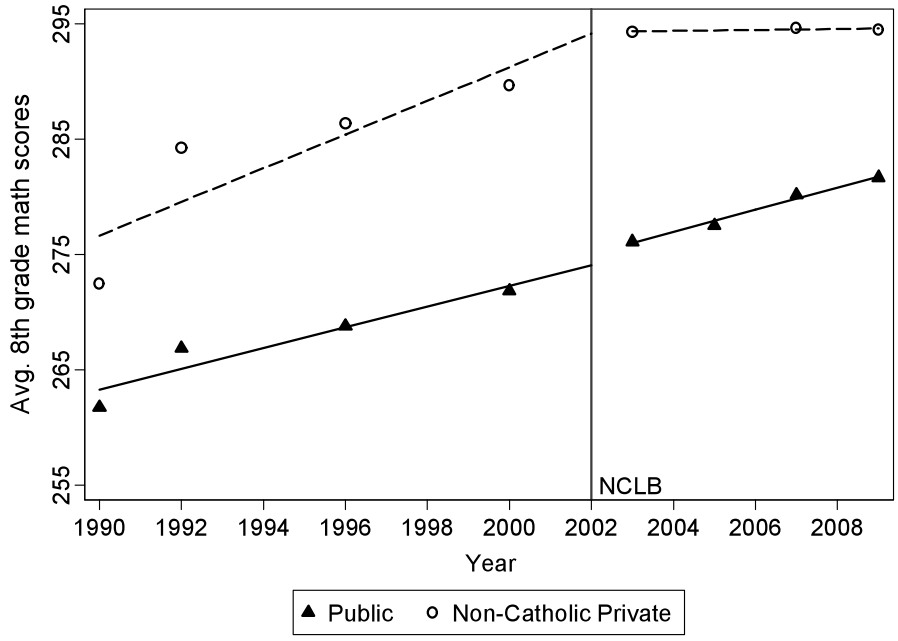

Figure 5a. Main NAEP 4th grade reading scores by year:
High vs. med. vs. low proficiency standard states

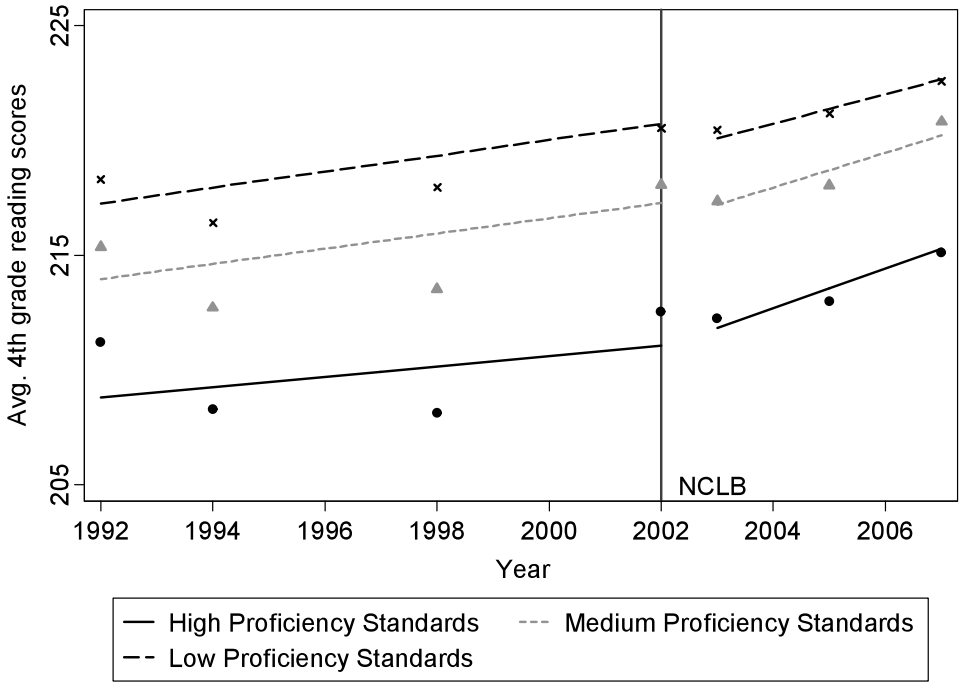

Figure 5b. Main NAEP 4th grade math scores by year:
High vs. med. vs. low proficiency standard states

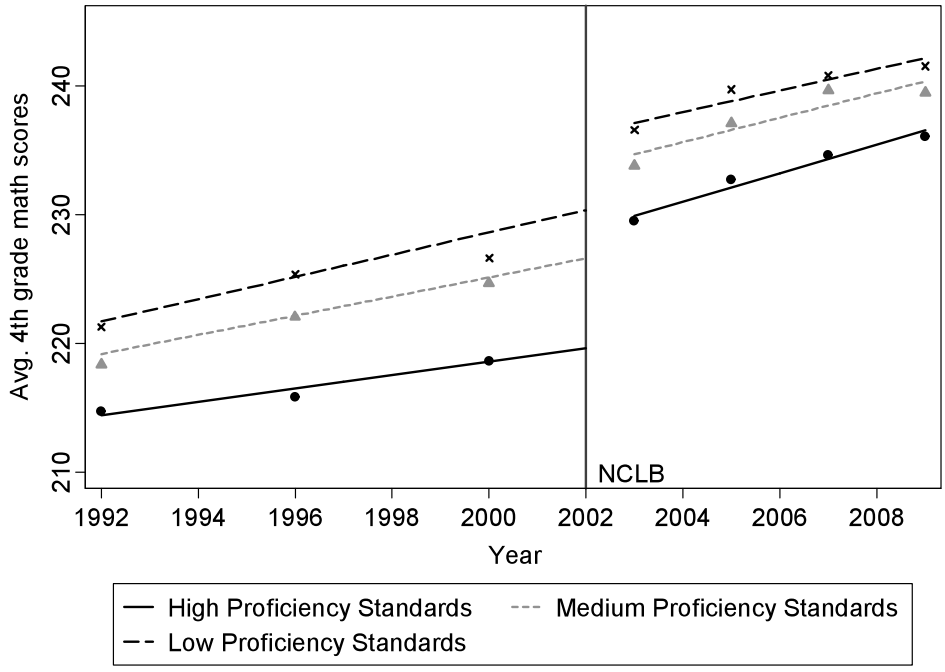

Figure 5c. Main NAEP 8th grade math scores by year:
High vs. med. vs. low proficiency standard states

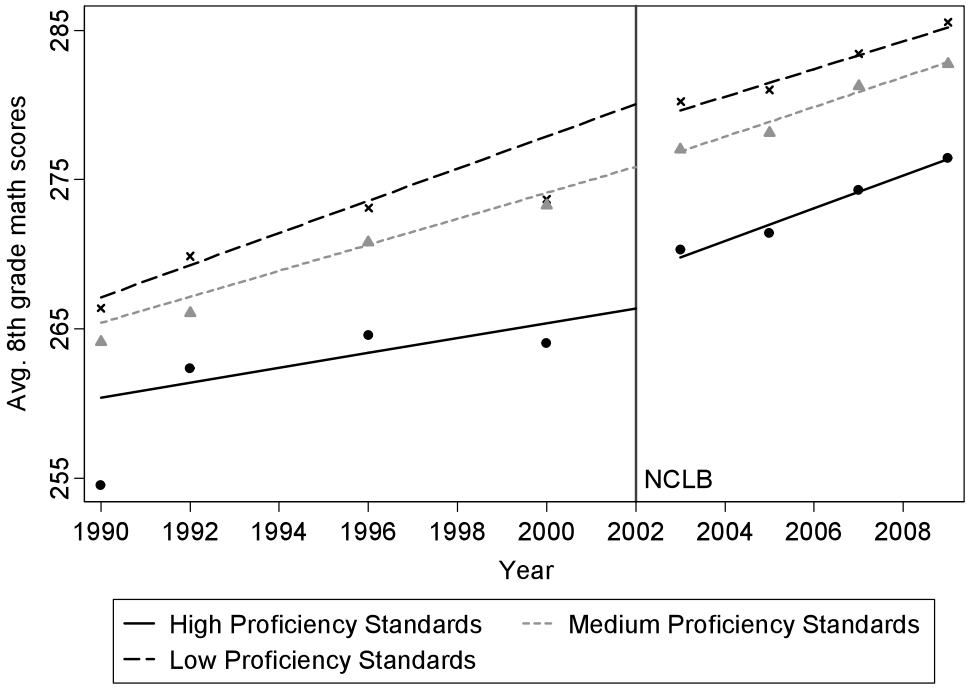

Figure 6a. Main NAEP 4th grade reading scores by year:
Standards and consequential accountability combined contrasts

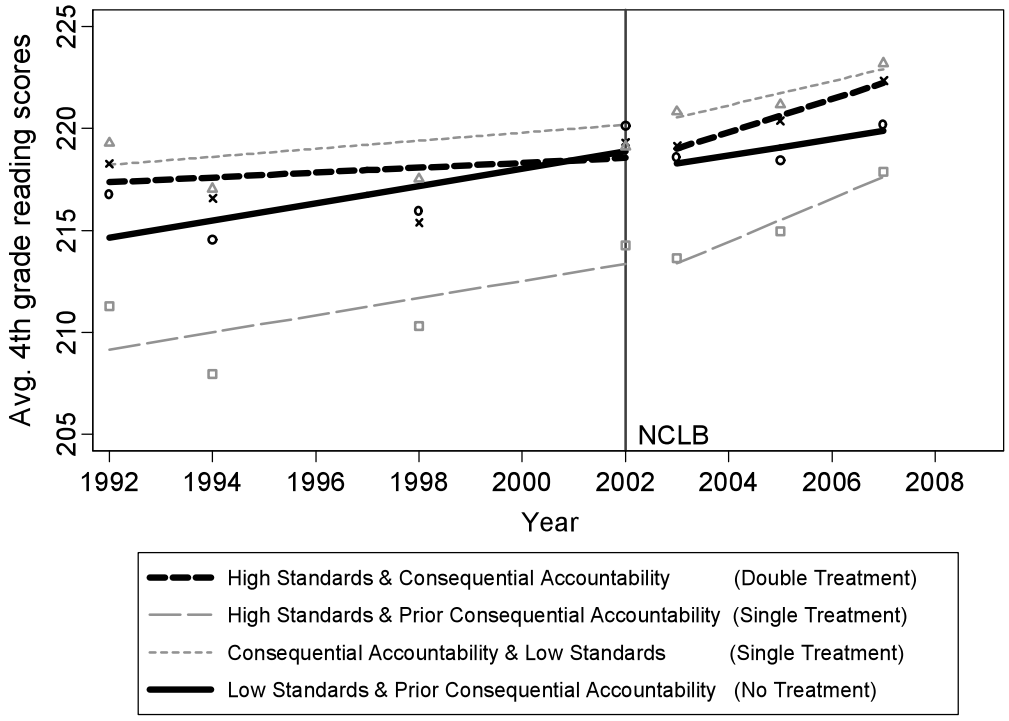

Figure 6b. Main NAEP 4th grade math scores by year:
Standards and consequential accountability combined contrasts

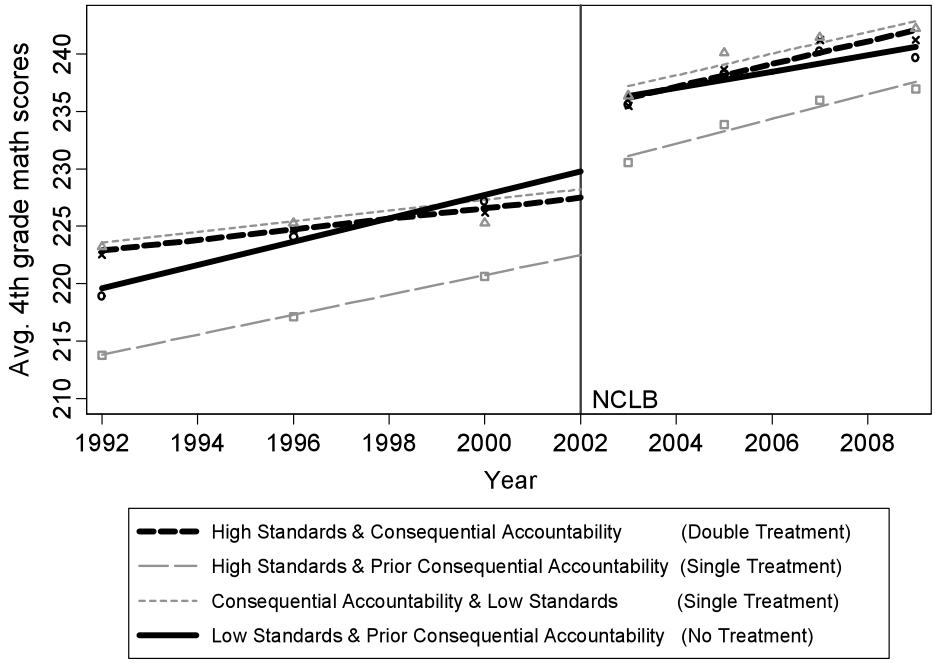

Figure 6c. Main NAEP 8th grade math scores by year:
Standards and consequential accountability combined contrasts

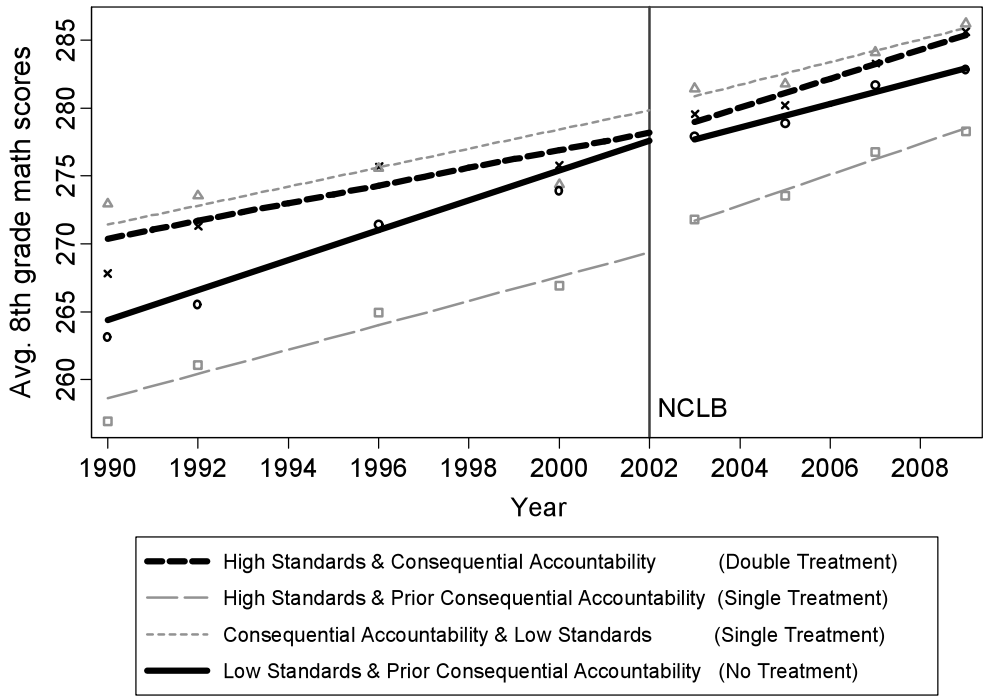

Table 1. Distribution of student enrollment and school composition profiles from 1994 to 2006 for Catholic, other private and public schools

	Student Enrollment				Pubil-to-Teacher Ratio		
	Catholic	Other Private	Public		Catholic	Other Private	Public
1994	5.73	4.72	89.55	1994	18.12	11.69	17.54
1996	5.67	4.74	89.60	1996	17.89	11.22	17.41
1998	5.58	4.87	89.56	1998	17.33	10.85	16.94
2000	5.38	4.81	89.81	2000	16.76	16.76	16.25
2002	5.26	5.13	89.61	2002	16.18	10.35	16.07
2004	4.88	4.93	90.18	2004	15.47	9.93	16.14
2006	4.56	5.07	90.37	2006	15.05	9.71	15.85

	Percent Hispanic				Percent Black		
	Catholic	Other Private	Public		Catholic	Other Private	Public
1994	10.66	5.25	12.39	1994	8.81	9.70	16.15
1996	10.07	4.32	13.26	1996	7.91	8.20	16.45
1998	10.04	4.21	14.22	1998	7.65	8.12	16.64
2000	10.59	10.59	15.40	2000	7.82	7.82	16.82
2002	11.21	4.63	16.94	2002	7.82	8.73	16.84
2004	11.21	5.00	18.34	2004	7.60	8.73	16.88
2006	11.86	5.28	19.66	2006	7.43	8.81	16.81

Source: Common Core Data and Private School Universe Survey

Table 2. Mean percentage of students meeting proficiency level on state and NAEP Tests[1]

High Proficiency Standard States			Low Proficiency Standard States		
State	State Test	NAEP Test	State	State Test	NAEP Test
Arizona	46	24	Colorado	83	35
Arkansas	46	25	Connecticut	76	39
California	36	22	Georgia	76	25
District of Columbia	48	8	Minnesota	75	40
Hawaii	31	21	Nebraska	80	33
Kentucky	47	27	New Hampshire	76	39
Maine	35	34	North Carolina	85	34
Massachusetts	50	41	Tennessee	80	24
Missouri	29	32	Texas	84	28
Rhode Island	45	28	Virginia	75	35
South Carolina	26	27	Wisconsin	77	35
Washington	46	34			
Wyoming	39	35			
Mean	40	27		79	33

[1]Results are averaged across grades (4th and 8th grade), subjects (math and reading) in year 2003 for state and NAEP assessment.

Note: When state assessment data are missing in the grade examined, data from the next lower grade are used and if not available then data are from the next higher grade.

Source: Consolidated State Performance Report and Institute of Education Science

Table 3. State student proficiency rates over time

State	Group	SA01	SA02	SA03	SA04	SA05
Alabama	Med	67	.	61	69	72
Alaska	Med	77	67	70	70	73
Arizona	High	48	.	46	48	64
Arkansas	High	37	.	46	54	48
California	High	33	31	36	37	42
Colorado	**Low**	**54**	**55**	**83**	**84**	**85**
Connecticut	Low	76	76	76	76	74
Delaware	Med	65	68	68	70	72
District of Columbia	**High**	**24**	**21**	**48**	**46**	**45**
Florida	**Med**	**58**	**21**	**56**	**59**	**60**
Georgia	Low	69	72	76	78	78
Hawaii	**High**	**.**	**58**	**31**	**33**	**35**
Idaho	Med	.	.	70	79	83
Illinois	Med	63	64	62	66	67
Indiana	Med	69	68	70	70	72
Iowa	Med	71	71	73	74	77
Kansas	Med	63	63	69	72	76
Kentucky	High	43	45	47	52	53
Louisiana	Med	53	49	57	58	60
Maine	High	34	34	35	35	41
Maryland	**Med**	**38**	**31**	**56**	**63**	**68**
Massachusetts	High	47	48	50	52	49
Michigan	Med	63	57	58	62	68
Minnesota	**Low**	**51**	**49**	**75**	**70**	**76**
Mississippi	Med	58	62	67	73	70
Missouri	High	29	30	29	30	32
Montana	Med	74	70	73	58	64
Nebraska	Low	75	68	80	84	88
Nevada	Med	53	51	53	47	49
New Hampshire	**Low**	**33**	**34**	**76**	**76**	**.**
New Jersey	Med	70	70	69	72	74
New Mexico	Med	38	.	67	52	42
North Carolina	Low	81	84	85	87	88
North Dakota	Med	74	60	61	66	73
Ohio	Med	59	61	61	67	69
Oklahoma	Med	66	65	67	70	71
Oregon	Med	69	71	70	70	75
Pennsylvania	Med	55	55	57	63	65
Rhode Island	High	55	.	45	52	.
South Carolina	High	26	29	26	31	32
South Dakota	Med	57	57	72	77	80
Tennessee	Low	.	.	80	82	87
Texas	Low	92	93	84	81	76
Utah	Med	64	68	74	75	76
Virginia	Low	71	73	75	78	81
Washington	High	44	48	46	60	65
West Virginia	Med	58	.	65	75	77
Wisconsin	Low	64	67	77	75	77
Wyoming	High	.	.	39	44	41

Note: Bolded text highlight states where proficiency rates increased by over 20 percentage points
Source: U.S. Department of Education

Table 4. Correlation of state assessment proficiency rates between year 2001 and 2005

	Year 2001	Year 2002	Year 2003	Year 2004
Year 2001				
Year 2002	0.91			
Year 2003	0.85	0.82		
Year 2004	0.81	0.79	0.95	
Year 2005	0.80	0.77	0.94	0.98

Table 5: Percentage of schools and teachers meeting NCLB Requirements in 2007: By level of proficiency standards

	High	Med	Low	H vs. L Diff
AYP Determination				
Schools that Failed to Make AYP	35.88	27.44	21.25	14.63 *
Title I Schools that Failed to Make AYP	37.15	26.78	21.12	16.03 *
Schools that were Unsuccessful in Appealing their Failed AYP Status	70.43	50.54	61.95	8.47 *
Teacher Certif ication				
Deemed Not Highly Qualifed Teachers	5.95	6.93	2.41	3.54 *
Schools in Need of Improvement Status (Based on All Schools)				
Schools in Early Improvement Status	10.43	8.58	4.97	5.46 *
Schools in Corrective Action	2.11	1.45	1.02	1.09 *
Schools in Restructuring	2.16	2.12	0.71	1.45 *
Schools in Need of Improvement	14.69	12.15	6.70	8.00 *
Schools in Need of Improvement Status (Based on Title I Schools Only)				
Schools in Early Improvement Status	18.05	15.39	8.91	9.14 *
Schools in Corrective Action	3.65	2.60	1.82	1.82 *
Schools in Restructuring	3.73	3.79	1.27	2.46 *
Schools in Need of Improvement	25.43	21.78	12.00	13.43 *
Based on All Title I Schools that Did Not Make AYP				
In Need of Improvement Status in Year 1 and 2: Initial Actions				
Students Eligible for School Choice	21.43	10.19	5.04	16.39 *
Students Eligible for Supplemental Services	11.40	7.13	2.92	8.49 *
In Need of Improvement Status in Year 3: Corrective Actions				
Institute New Curriculum (CA)	7.16	5.06	2.76	4.40 *
Appointed Outside Expert Advice (CA)	5.99	2.54	4.63	1.36 *
Decreased Management Authority (CA)	3.12	1.75	0.78	2.33 *
Replaced School Staff (CA)	1.70	1.38	1.19	0.51 *
Extended School Day (CA)	1.44	1.64	0.75	0.69 *
Restructure their Internal Organization (CA)	3.22	2.43	2.72	0.49 *
Replaced Principal (CA)	0.04	2.71	0.48	-0.44 *
Total Corrective Actions (assuming no overlap)	22.66	17.51	13.32	9.34
Total Corrective Actions (assuming full overlap)	8.75	4.41	4.23	4.52
In Need of Improvement Status in Year 4: Restructuring Actions				
Taken Over By the State (RA)	0.99	0.63	0.00	0.99 *
Replaced School Staff (RA)	1.13	1.59	0.19	0.95 *
Contract Private Company to Run School (RA)	0.71	0.56	0.00	0.71 *
Reopen as a Charter Schools (RA)	0.06	0.04	0.11	-0.05
Took Other RA Actions (RA)	7.79	4.44	1.72	6.07 *
Total Restructuring Actions (assuming no overlap)	10.68	7.26	2.01	8.67
Total Restructuring Actions (assuming full overlap)	9.80	3.84	1.69	8.11

* Hypothesis test is conducted at the school level between high and low performance standard states and is statistically significant at $\alpha = .05$

Source: 2007 Consolidated State Performance Report

Table 6. Difference in differences in mean, slope and total change post-NCLB for public vs. private school contrasts: Analyses based on Trend and Main NAEP

	4th Grade Reading			4th Grade Math			8th Grade Math		
	Coef.	S.E.	t	Coef.	S.E.	t	Coef.	S.E.	t
Public vs. Catholic (Trend NAEP)									
Diff. in Mean Δ (2004)	3.92	3.28	1.20	10.93	5.53	1.97*	7.26	2.03	3.58*
Public vs. Catholic (Main NAEP)									
Diff. in Mean Δ	-2.06	3.94	-0.52	3.92	4.38	0.89	2.06	4.69	0.44
Diff. in Slope Δ	1.73	1.04	1.66	1.00	0.76	1.33	0.12	0.81	0.15
Diff in Total Δ (2007 or 2009)[1]	6.60	4.35	1.52	10.96	6.20	1.77+	2.91	6.64	0.44
Diff in Total Δ (Exclude 1990)	-	-	-	10.73	3.43	3.13*	0.26	5.57	0.05
Public vs. Non-Catholic (Main NAEP)									
Diff. in Mean Δ	0.95	4.25	0.22	4.41	6.11	0.72	0.77	5.79	0.13
Diff. in Slope Δ	0.22	1.10	0.20	0.29	1.03	0.28	1.48	0.98	1.51
Diff in Total Δ (2007 or 2009)	2.06	4.68	0.44	6.46	8.39	0.77	11.16	7.95	1.40
Diff in Total Δ (Exclude 1990)	-	-	-	13.90	9.72	1.43	5.57	1.39	4.00*

+ $p<0.1$, * $p<0.05$

[1] Reading estimates for 2007, math estimates for 2009

Table 7. Effect sizes in percentile (Pct), standard deviation (SD), and months of learning (Months): Public vs. private school contrast

	4th Grade Reading			4th Grade Math			8th Grade Math		
	SD[1]	Months[2]	Pct.[3]	SD	Months	Pct.	SD	Months	Pct.
Public vs. Catholic (Trend NAEP)[4]									
Diff. in Total Δ (2004)	0.11	3.22	-	0.34	7.20	-	0.24	12.99	-
Public vs. Catholic (Main NAEP)									
Diff. in Mean Δ	-0.06	-1.79	-0.03	0.15	3.17	0.15	0.06	3.31	0.07
Diff. in Slope Δ	0.05	1.50	0.05	0.04	0.81	0.09	0.00	0.19	0.01
Diff. in Total Δ (2007 or 2009)[5]	0.19	5.74	0.43	0.41	8.86	0.57	0.09	4.67	0.12
Public vs. Non-Catholic (Main NAEP)									
Diff. in Mean Δ	0.03	0.83	0.04	0.17	3.56	0.17	0.02	1.24	0.02
Diff. in Slope Δ	0.01	0.19	0.01	0.01	0.24	0.04	0.04	2.38	0.04
Diff. in Total Δ (2007 or 2009)	0.06	1.79	0.08	0.24	5.22	0.38	0.33	17.90	0.49

[1] Effects sizes are computed using group averaged grade- and subject-specific standard deviations of student test scores provided by NAEP.
For Main NAEP Catholic vs. public analyses, SD=35 for 4th grade reading, SD=27 for 4th grade math, SD=34 for 8th grade math
For Main NAEP other private vs. public analyses, SD=35 for 4th grade reading, SD=27 for 4th grade math, SD=35 for 8th grade math
For Trend NAEP Catholic vs. public analyses, SD=37 for 4th grade reading, SD=33 for 4th grade math, SD=31 for 8th grade math

[2] Gains in months are based on the average grade- and subject-specific effect size in moving from one grade to next on nationally normed tests

[3] Gains in percentile rank are calculated based on the distribution of state ranking observed in 2002.

[4] Pecentile cannot be calculated because no distributional data on Trend NAEP are available at the state level to calculate changes in rank.

[5] Reading estimates for 2007, math estimates for 2009

Table 8. Difference in differences in mean, slope and total compositional change post-NCLB: Public vs. private school contrast

	Pct. Black			Pct. Hispanic			Pct. White		
Public vs. Catholic	Coef.	Std. Err.	t	Coef.	Std. Err.	t	Coef.	Std. Err.	t
Diff. in Mean Δ	0.18	0.75	0.24	-0.29	0.65	-0.45	1.10	1.88	0.59
Diff. in Slope Δ	0.02	0.23	0.08	0.25	0.20	1.25	0.13	0.58	0.23
Diff in Total Δ (2007)	0.28	0.69	0.40	0.94	0.59	1.59	1.75	1.73	1.02
Public vs. Other Private									
Diff. in Mean Δ	-5.04	12.16	-0.41	2.07	39.10	0.05	-41.01	84.11	-0.49
Diff. in Slope Δ	0.16	0.31	0.50	-0.12	1.01	-0.11	1.11	2.17	0.51
Diff in Total Δ (2007)	-4.26	10.59	-0.40	1.50	34.05	0.04	-35.48	73.26	-0.48

	Pct. Student			Pct. 4th Graders			Pct. 8th Graders		
Public vs. Catholic	Coef.	Std. Err.	t	Coef.	Std. Err.	t	Coef.	Std. Err.	t
Diff. in Mean Δ	-0.06	0.92	-0.07	-0.02	0.32	-0.05	0.05	0.55	0.09
Diff. in Slope Δ	-0.06	0.28	-0.20	-0.08	0.10	-0.85	-0.03	0.17	-0.18
Diff in Total Δ (2007)	-0.35	0.79	-0.44	-0.43	0.28	-1.57	-0.10	0.48	-0.21
Public vs. Other Private									
Diff. in Mean Δ	0.39	10.89	0.04	-0.48	4.93	-0.10	-0.07	6.45	-0.01
Diff. in Slope Δ	-0.02	0.28	-0.07	0.00	0.13	0.01	-0.01	0.17	-0.05
Diff in Total Δ (2007)	0.30	9.49	0.03	-0.47	4.30	-0.11	-0.11	5.62	-0.02

	Student-to-Teacher Ratio		
Public vs. Catholic	Coef.	Std. Err.	t
Diff. in Mean Δ	-0.49	0.58	-0.85
Diff. in Slope Δ	0.00	0.18	-0.01
Diff in Total Δ (2007)	-0.50	0.50	-0.99
Public vs. Other Private			
Diff. in Mean Δ	5.50	34.32	0.16
Diff. in Slope Δ	-0.22	0.88	-0.25
Diff in Total Δ (2007)	4.39	29.90	0.15

Source: Common Core Data and Private School Universe Survey

Table 9. Difference in differences in mean, slope and total change post-NCLB for state contrasts: Analyses based on Main NAEP

	4th Grade Reading			4th Grade Math			8th Grade Math		
	Coef.	S.E.	t	Coef.	S.E.	t	Coef.	S.E.	t
High vs. Low Proficiency Standards									
Diff in Mean Δ	1.43	0.97	1.47	3.77	1.76	2.14*	2.63	1.90	1.38
Diff. in Slope Δ	0.33	0.36	0.93	0.52	0.39	1.33	0.62	0.30	2.04*
Diff in Total Δ (2007 or 2009)[1]	3.10	1.81	1.71+	7.38	3.33	2.22*	6.97	3.38	2.06*
State Contrast - Continuous									
Diff. in Mean Δ	1.60	0.80	2.00*	3.34	1.67	2.00*	3.12	2.13	1.46
Diff. in Slope Δ	0.19	0.38	0.50	0.38	0.38	1.00	0.53	0.27	2.00*
Diff. in Total Δ (2007 or 2009)	2.45	1.54	1.59	5.90	2.70	2.18*	6.82	3.06	2.23*

+ $p<0.1$, * $p<0.05$

[1] Reading estimates for 2007, math estimates for 2009

Note: The calcuation of the total change for the continuous measure is based on a 38 percent contrast as oppose to a 100 percent contrast found in the traditional treatment vs. no treatment analyses. This is done to provide a more meaningful comparison between groups. The 38 percent is based on the difference in the average state assessment rate between high and low proficiency standard states.

Table 10. Effect sizes in percentile (Pct), standard deviation (SD), and months of learning (Months): State contrast

	4th Grade Reading			4th Grade Math			8th Grade Math		
	SD[1]	Months[2]	Pct.[3]	SD	Months	Pct.	SD	Months	Pct.
High vs. Low Proficiency Standards									
Diff. in Mean Δ	0.04	1.16	0.02	0.13	2.89	0.17	0.07	3.98	0.07
Diff. in Slope Δ	0.01	0.27	0.01	0.02	0.39	0.00	0.02	0.94	0.01
Diff in Total Δ (2007 or 2009)[4]	0.08	2.52	0.05	0.26	5.65	0.21	0.19	10.56	0.21
State Contrast - Continuous									
Diff. in Mean Δ	0.04	1.29	0.02	0.09	1.94	0.17	0.08	4.59	0.12
Diff. in Slope Δ	0.01	0.15	0.01	0.01	0.22	0.00	0.01	0.78	0.01
Diff. in Total Δ (2007 or 2009)	0.07	1.99	0.06	0.16	3.42	0.22	0.18	10.05	0.24

[1] Effects sizes are computed using grade- and subject-specific standard deviations of individual student test score data provided by NAEP. SD=37 for 4th grade reading, SD=28 for 4th grade math, SD=36 for 8th grade math

[2] Gains in months are based on the average grade- and subject-specific effect size in moving from one grade to next on nationally normed tests

[3] Gains in percentile rank are calculated based on the distribution of state ranking observed in 2002.

[4] Reading estimates for 2007, math estimates for 2009

Table 11. Percentage of students identified with a disability and limited English proficient:
By contrast group, years immediately around NCLB's implementation, and averaged across all the years pre- or post-NCLB

	High Proficiency Standard States			Low Proficiency Standard States		
Year immediately before and after NCLB	*2000 or 2002* [1]	*2003*	*Diff*	*2000 or 2002*	*2003*	*Diff*
4th Grade Math	18.92	20.77	1.85	16.09	19.08	2.99
8th Grade Math	17.00	18.77	1.77	15.36	17.15	1.79
4th Grade Reading	19.23	21.00	1.77	18.09	18.69	0.60
Years averaged before and after NCLB	*1990-2002*	*2003-2009*	*Diff*	*1990-2002*	*2003-2009*	*Diff*
4th Grade Math	14.84	21.28	6.44	13.42	19.90	6.48
8th Grade Math	12.27	18.44	6.16	11.19	17.03	5.83
4th Grade Reading	15.37	21.51	6.14	14.17	19.79	5.62

	Public			Catholic			Non-Catholic Private		
Year immediately before and after NCLB	*2000 or 2002*	*2003*	*Diff*	*2000 or 2002*	*2003*	*Diff*	*2000 or 2002*	*2003*	*Diff*
4th Grade Math	19.16	22.28	3.12	2.89	3.94	1.05	2.16	3.89	1.72
8th Grade Math	14.36	18.54	4.19	1.91	3.17	1.27	4.42	3.37	-1.05
4th Grade Reading	20.56	21.91	1.35	1.52	3.20	1.68	2.42	4.40	1.98
Years averaged before and after NCLB	*1990-2002*	*2003-2009*	*Diff*	*1990-2002*	*2003-2009*	*Diff*	*1990-2002*	*2003-2009*	*Diff*
4th Grade Math	11.25	22.67	11.42	2.89	4.33	1.45	2.16	4.67	2.50
8th Grade Math	11.67	18.67	7.00	1.91	3.67	1.76	4.42	4.00	-0.42
4th Grade Reading	16.00	22.67	6.67	2.36	4.33	1.97	2.42	5.00	2.58

[1]Closest year prior to NCLB that data is available for math (2000) and reading (2002).
Source: National Center for Educational Statistics and the Educational Testing Service.

Table 12. Percentage of student with disability and limited English proficiency excluded from NAEP testing:
By contrast group, years immediately around NCLB's implementation, and averaged across all the years pre- or post-NCLB

	High Proficiency Standard States			Low Proficiency Standard States		
Year immediately before and after NCLB	*2000 or 2002* [1]	*2003*	*Diff*	*2000 or 2002*	*2003*	*Diff*
4th Grade Math	4.33	3.31	-1.03	4.09	3.85	-0.24
8th Grade Math	3.42	3.69	0.28	4.45	3.77	-0.69
4th Grade Reading	6.15	5.85	-0.31	6.91	5.85	-1.06
Years averaged before and after NCLB	*1990-2002*	*2003-2009*	*Diff*	*1990-2002*	*2003-2009*	*Diff*
4th Grade Math	6.92	3.26	-3.66	6.04	3.21	-2.84
8th Grade Math	6.06	4.03	-2.03	5.00	3.67	-1.33
4th Grade Reading	6.79	5.90	-0.89	6.17	5.92	-0.25

	Public			Catholic			Non-Catholic Private		
Year immediately before and after NCLB	*2000 or 2002*	*2003*	*Diff*	*2000 or 2002*	*2003*	*Diff*	*2000 or 2002*	*2003*	*Diff*
4th Grade Math	4.24	3.92	-0.33	0.00	0.11	0.11	0.37	0.33	-0.04
8th Grade Math	3.94	3.76	-0.18	0.00	0.04	0.04	0.30	0.68	0.37
4th Grade Reading	6.80	6.31	-0.49	0.64	0.85	0.21	1.03	0.62	-0.40
Years averaged before and after NCLB	*1990-2002*	*2003-2009*	*Diff*	*1990-2002*	*2003-2009*	*Diff*	*1990-2002*	*2003-2009*	*Diff*
4th Grade Math	5.67	2.50	-3.17	0.00	0.15	0.15	0.00	0.31	0.31
8th Grade Math	5.00	4.00	-1.00	0.00	0.18	0.18	0.00	0.55	0.55
4th Grade Reading	6.50	6.33	-0.17	0.64	0.64	0.00	0.64	0.98	0.34

[1] Closest year prior to NCLB that data is available for math (2000) and reading (2002).

Source: National Center for Educational Statistics and the Educational Testing Service.

Table 13. Difference in differences in mean, slope and total change post-NCLB for three types of treatment groups
Analyses based on Main NAEP first using a subset of states, then all states

		4th Grade Reading				4th Grade Math				8th Grade Math			
		Coef.	S.E.	t		Coef.	S.E.	t		Coef.	S.E.	t	
Wong et.al. vs Dee &Jacob (sub)													
Diff in Total Δ (2007 or 2009)[1]	HS+CA	5.60	2.50	2.24	*	10.10	4.28	2.36	*	3.84	4.88	0.79	
Diff in Total Δ (2007 or 2009)	HS	1.34	2.67	0.50		9.82	4.39	2.24	*	9.33	5.28	1.77	+
Diff in Total Δ (2007 or 2009)	CA	1.98	2.56	0.78		9.45	3.76	2.51	*	-1.34	3.93	-0.34	
Interaction Effect[2]		2.28	2.97	0.77		-9.17	4.86	-1.89	+	-4.15	5.96	-0.70	
		Coef.	S.E.	t		Coef.	S.E.	t		Coef.	S.E.	t	
Wong et.al.vs Dee & Jacob(all)													
Diff in Total Δ (2007 or 2009)	HS+CA	4.19	1.68	2.49	*	5.81	3.41	1.70	+	3.45	3.76	0.92	
Diff in Total Δ (2007 or 2009)	HS	1.96	1.80	1.09		2.87	3.41	0.84		3.89	3.49	1.11	
Diff in Total Δ (2007 or 2009)	CA	2.05	1.66	1.24		7.07	2.91	2.43	*	1.97	2.79	0.71	
Interaction Effect		0.18	2.18	0.08		-4.12	4.00	-1.03		-2.41	4.69	-0.51	

+ p<0.1, * p<0.05

Note: HS = High Standards, CA= Consequential Accountability after NCLB

Coefficient are based on comparisons with the reference group of low standards and prior CA

[1] Reading estimates for 2007, math estimates for 2009

[2] Interaction effect is calculated by subtracting the coefficients HS and CA from HS+CA.

Table 14. Effect sizes in percentile (Pct), standard deviation (SD), and months of learning (Months): Analyses based on Main NAEP using all states

	SD[1]	Months[2]	Pct.[3]	SD	Months	Pct.	SD	Months	Pct.
Wong et.al.(median) vs Dee & Jacob(all)									
Diff. in Total Δ (2007 or 2009) HS+CA	0.11	3.40	0.14	0.21	4.45	0.23	0.10	5.22	0.12
Diff. in Total Δ (2007 or 2009) HS	0.05	1.59	0.03	0.10	2.19	0.16	0.11	5.89	0.13
Diff. in Total Δ (2007 or 2009) CA	0.06	1.66	0.04	0.25	5.41	0.30	0.05	2.99	0.05

[1] Effects sizes are computed using grade- and subject-specific standard deviations of individual student test score data provided by NAEP. SD=37 for 4th grade reading, SD=28 for 4th grade math, SD=36 for 8th grade math

[2] Gains in months are based on the average grade- and subject-specific effect size in moving from one grade to next on nationally normed tests

[3] Gains in percentile rank are calculated based on the distribution of state ranking observed in 2002.