Left Behind?
The Effect of No Child Left Behind on Academic Achievement Gaps

Sean F. Reardon
Erica H. Greenberg
Demetra Kalogrides
Kenneth A. Shores
Rachel A. Valentino

Stanford University

VERSION: August 12, 2013

Direct correspondence to sean f. reardon (sean.reardon@stanford.edu). The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant #R305D110018 to Stanford University. Research support for Greenberg, Shores, and Valentino was also supported by the Institute of Education Sciences, U.S. Department of Education, through Grant #R305B090016 to Stanford University. We particularly thank Ximena Portilla, Natassia Rodriguez, Donna Saadati-Soto, Ricky Vargas, Austin Block, Leah Weiser, Walter Herring, Sarah Quartey, and James Chu for their expert research assistance, and Ross Santy for his assistance providing data from EdFacts. The findings, conclusions, and opinions here are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

Abstract

One of the goals of the No Child Left Behind Act of 2001 (NCLB; 20 U.S.C. § 6301) was to close racial and socioeconomic achievement gaps. Over a decade has passed since NCLB went into effect. In this paper we investigate whether the Act has been successful at narrowing racial achievement gaps. Overall, our analyses provide no support for the hypothesis that No Child Left Behind has led, on average, to a narrowing of racial achievement gaps. We find that within-state achievement gaps were closing slowly, on average, prior to the passage of the NCLB legislation, and that this trend did not change significantly after the introduction of NCLB. However, we do find evidence indicating that the policy's impact varies systematically across states in ways that are consistent with NCLB's subgroup-specific accountability features. In states facing more subgroup-specific accountability pressure, more between-school segregation, and larger gaps prior to the implementation of the policy, NCLB appears to have narrowed white-black and white-Hispanic achievement gaps; in states facing less pressure, less segregation, and smaller pre-existing gaps, NCLB appears to have led to a widening of white-black and white-Hispanic achievement gaps. We conclude with a discussion of potential explanations for these findings.

Introduction

One of the goals of the No Child Left Behind Act of 2001 (NCLB; 20 U.S.C. § 6301) was to close racial achievement gaps. Although racial gaps narrowed substantially in the 1970s and 1980s (Grissmer, Flanagan, & Williamson, 1998; Hedges & Nowell, 1998, 1999; Neal, 2006), they narrowed only slightly in the 1990s, and were still very large in 2001 (roughly 0.75-1.0 standard deviations) (Hemphill, Vanneman, & Rahman, 2011; Reardon & Robinson, 2007; Vanneman, Hamilton, Baldwin Anderson, & Rahman, 2009). Dissatisfied with these large gaps, as well as with overall levels of achievement, Congress passed the NCLB legislation in 2001. Title I of the Act begins:

> The purpose of this title is to ensure that all children have a fair, equal, and significant opportunity to obtain a high-quality education and reach, at a minimum, proficiency on challenging State academic achievement standards and state academic assessments. This purpose can be accomplished by…closing the achievement gap between high- and low-performing children, especially the achievement gaps between minority and nonminority students, and between disadvantaged children and their more advantaged peers (115 Stat. 1439-40).

The Act mandated that test results be publicly reported for each school, disaggregated by race and socioeconomic status (among other factors), and tied to sanctions at the school level.

Over a decade has passed since NCLB went into effect. In this paper we investigate whether the Act has been successful at narrowing racial achievement gaps. To do so, we first describe the average trends in within-state achievement gaps from 1990 through 2011. Second, we test whether there is an association between the number of years that a cohort of students has been exposed to NCLB by a particular grade and the size of that cohort's achievement gap in that grade, net of state-specific cohort and grade trends. Third, we examine whether the relationship between achievement gaps and NCLB exposure is stronger in states where the structure of the NCLB accountability system led to a more specific focus on the achievement of minority students.

Overall, our analyses provide no support for the hypothesis that No Child Left Behind substantially narrowed racial achievement gaps, on average. We find that within-state achievement

gaps were closing slowly, on average, prior to the passage of the NCLB legislation, and that this trend did not change significantly after the introduction of NCLB. However, we do find evidence indicating that the policy's impact varies across states. In particular, NCLB appears to narrow white-black and white-Hispanic achievement gaps in states where majorities of black and Hispanic students are enrolled in schools that are held accountable for these students' performance—that is, schools where there are sufficient numbers of black and Hispanic students to meet NCLB's state-determined minimum subgroup reporting threshold—and where between-school levels of racial segregation are high and achievement gaps prior to the policy were large. In states where relatively few minority students are in such schools, where racial segregation is low, and where pre-NCLB achievement gaps were small, NCLB actually appears to widen racial achievement gaps.

Achievement Gap Trends and Accountability Policy

Achievement Gaps

Achievement gaps are of particular concern because academic achievement in the K-12 grades is a precursor to college access and success in the labor market. Although it was possible in the 1950s and 1960s to earn a middle-class wage in the U.S. without holding a college degree, the modern U.S. economy has few such low-skill, high-wage jobs remaining (Goldin & Katz, 2008; Murnane, Willett, & Levy, 1995); as a result, a college degree has become increasingly important in the labor market, and has become increasingly important for economic mobility. At the same time, access to college, particularly to more selective colleges, has become increasingly dependent on students' test scores and academic achievement (Alon & Tienda, 2007; Bastedo & Jaquette, 2011; Grodsky & Pattison, in progress; Posselt, Jaquette, Bastedo, & Bielby, 2010). Because of the growing importance of academic achievement, the white-black test score gap now explains virtually all of the white-black difference in college enrollment (including enrollment at the most selective colleges and universities) and most or all of the white-black differences in wages (Alon & Tienda, 2007;

Bollinger, 2003; Carneiro, Heckman, & Masterov, 2003; Neal & Johnson, 1996; Posselt et al., 2010). Eliminating racial achievement gaps is therefore essential for reducing broader racial disparities in U.S. society.

Evidence on the national long-term trend in racial achievement gaps is well documented by the National Assessment of Educational Progress (NAEP). Achievement gaps in both math and reading between white and black students have narrowed substantially over the last forty years (Grissmer et al., 1998; Hedges & Nowell, 1999; Hemphill et al., 2011; Kober, Chudowsky, & Chudowsky, 2010; Neal, 2005; Reardon & Robinson, 2007; Vanneman et al., 2009). Gaps in reading between white and Hispanic students follow the same trend, though gaps in math between these two groups have largely stagnated since the late 1970s, the first year in which NAEP data were disaggregated for Hispanic students (Hemphill et al., 2011; Reardon & Robinson, 2007). Despite this progress, gaps remain large, ranging from two-thirds to slightly less than one standard deviation, depending on the grade and subject. Importantly, both the size of achievement gaps and their trends over time vary considerably across states (Hemphill et al., 2011; Kober et al., 2010; National Center for Education Statistics, 2013; Reardon, Kalogrides, Valentino, Shores, & Greenberg, 2013; Vanneman et al., 2009).

How Might the No Child Left Behind Legislation Affect Academic Achievement Gaps?

NCLB may narrow achievement gaps through several mechanisms. First, the law requires assessment of nearly all students in grades three to eight, along with public reporting of results, disaggregated by subgroup. Illuminating the performance of students from historically low-performing backgrounds—the so-called "informational aspects" of the policy (Hanushek & Raymond, 2004)—may motivate schools and teachers to focus their attention on narrowing gaps (Rothstein, 2004). Second, NCLB may reduce achievement gaps by tying accountability sanctions to the Adequate Yearly Progress (AYP) of each subgroup. Here, an escalating series of consequences

may pressure schools to improve the academic performance of student subgroups with low proficiency rates. After two consecutive years "in need of improvement," a school must offer transfer options to families; after four, corrective actions must be taken to change school personnel or academic functions; after six, the school must be restructured by replacing the administration, teaching staff, or governance structure. If these actions, or the threat of these actions, increase achievement among low-performing student subgroups, achievement gaps may narrow.

In addition to shining a bright light on differential achievement and imposing accountability sanctions, NCLB includes other provisions that may affect existing achievement gaps. For example, its Highly Qualified Teacher provision requires that all teachers have a bachelor's degree, full state certification or licensure, and documented knowledge of the relevant subject matter. In many states, lesser-qualified teachers are over-represented in schools serving low-income and minority students (Clotfelter, Ladd, & Vigdor, 2005; Lankford, Loeb, & Wyckoff, 2002). If NCLB equalizes the distribution of qualified teachers among schools to some extent, and if teachers with these qualifications are more effective at raising student achievement than their less qualified peers, then NCLB may reduce achievement gaps. Finally, the law increased federal support for supplemental education services and school choice options for children in underperforming schools. If more low-income and non-white families make use of these provisions than others, and if these services and options systematically increase student achievement, then these features of the No Child Left Behind Act may narrow achievement gaps, as well.

Despite the federal implementation of NCLB, there are reasons to think that its subgroup-specific accountability pressure, and therefore its effect on achievement gaps, may vary among states (Davidson, Reback, Rockoff, & Schwartz, 2013). One reason for this variation is that NCLB does not require states to hold schools accountable for the AYP of subgroups with too few students to yield reliable information on their achievement.[1] In such a school with few black students, for

[1] NCLB, 2001 Sec. 1111 [b][2][C][v][II].

example, NCLB may create little or no incentive for educators to focus attention on the performance of the small number of black students in the school (though black students' scores would still be included in calculations of the school's overall proficiency rate). Indeed, it may create an incentive to focus primarily on the performance of low-achieving white students. As a result, the NCLB incentive structure may lead to no change in, or even a widening of, the white-black achievement gap in that school. In contrast, a school with a large number of black students will be held accountable for the performance of its black students separately, creating a greater incentive to improve their performance and narrow achievement gaps.

One potential consequence of this feature of the law is that NCLB may be more effective at narrowing achievement gaps in states where more minority students attend schools requiring subgroup-specific reporting of test scores. The proportion of black or Hispanic students in such schools will depend on several factors: 1) the overall proportion of black or Hispanic students in the state; 2) the degree of between-school racial segregation (in highly segregated states, more minority students attend schools with large numbers of same-race peers); 3) the average school size (when most schools are small, fewer students will be in schools meeting the minimum subgroup threshold); and 4) the criteria for determining what number of students is sufficient to require subgroup-specific reporting and accountability. These criteria vary substantially across states (Aspen Institute, 2006; Chudowsky & Chudowsky, 2005; Davidson et al., 2013; Fulton, 2006; Sunderman, 2006; Wei, 2012). As we show below, variation among states in these factors produces considerable variation in the proportion of black and Hispanic students whose test scores were relevant for reporting and accountability purposes.

Prior Evidence on the Effect of No Child Left Behind

The literature is mixed regarding the effects of accountability systems generally, and of No Child Left Behind specifically, on student achievement (Carnoy & Loeb, 2002; Dee & Jacob, 2011;

Gaddis & Lauen, 2011; Hanushek & Raymond, 2004, 2005; Lauen & Gaddis, forthcoming; Lee, 2006; Lee & Reeves, 2012; Lee & Wong, 2004; Wei, 2012; Wong, Cook, & Steiner, 2011). Research on the effects of NCLB is challenged by the difficulty of identifying a plausible counterfactual necessary for estimating the causal impact of the policy. Because NCLB was introduced at the federal level, the policy took effect in all states at the same time. This makes it difficult to disentangle non-NCLB-induced trends from NCLB effects. One solution to this challenge is to leverage variation among states—in either their pre-NCLB state accountability systems or the strength of their NCLB standards—to assess the effect of the policy on student achievement. Strategies of this type have been used convincingly by Dee and Jacob (2011) and Lee (2006), who use variation in pre-NCLB state factors to identify policy effects; by Wong, Cook, and Steiner (2011), who rely on variation in post-NCLB implementation approaches for identification; and by Lee and Reeves (2012), who leverage both types of factors.

Dee and Jacob (2011), Lee (2006), and Lee and Reeves (2012) reason that NCLB should have had a larger impact on achievement trends in states that had no NCLB-like system of "consequential accountability"(CA)[2] prior to the NCLB legislation than in states that already had CA systems before the implementation of the federal law. They conduct a set of comparative interrupted time series analyses, using Main NAEP data from 1990 through the mid- to late-2000s to estimate the effect of the policy. These studies reach conflicting conclusions, possibly due to differences in samples, statistical power, and analytic methods. Lee (2006) finds no consistent effect of NCLB on achievement levels or gaps. Dee and Jacob (2011) find that NCLB improved average math performance, particularly in fourth grade, but did not affect reading performance. Their analyses provide inconclusive evidence of the effects of NCLB on achievement gaps, however, because their subgroup analyses are sometimes based on different sets of states for white, black,

[2] The literature (e.g., Hanushek and Raymond 2004, 2005) defines consequential accountability systems as those that issue incentives and levy sanctions based on measurable outcomes, as opposed to report card or other accountability systems that rely on informational mechanisms alone.

and Hispanic students' scores and do not include statistical tests of whether the effects differ significantly by subgroup.[3] Finally, Lee and Reeves (2012) couple comparative interrupted time series analysis with inverse probability of treatment weighting to address issues of selection bias relevant to the CA/no-CA causal contrast. They find that NCLB significantly decreased racial gaps in eighth grade math between 2002 and 2009 by roughly one-twentieth of a standard deviation. They find no other significant effects in fourth grade math or reading.

Wong, Cook, and Steiner (2011) and Lee and Reeves (2012) adopt a similar comparative interrupted time series approach, but compare post-NCLB changes in achievement trends among states with stronger and weaker implementation of NCLB. Wong, Cook, and Steiner (2011) compare states that instituted "high" proficiency and "low" standards in response to the federal NCLB accountability mandate. Their argument is that states with high standards (defined as standards resulting in fewer than 50% of students meeting the proficiency threshold) experienced more NCLB accountability pressure than states with low standards (where more than 75% meet the threshold). Like previous research (Dee and Jacob, 2011; Lee and Reeves, 2012), they find significant effects of NCLB on average fourth and eighth grade math achievement (but no effect on reading achievement). Lee and Reeves (2012) construct several indices of fidelity and rigor in the implementation of No Child Left Behind, including one related to state standards. Using this identification strategy, they find no consistent pattern of effects of any index on average student achievement. Neither paper uses post-NCLB implementation differences to estimate the effects of NCLB on achievement gaps.

One additional paper investigates the association between post-NCLB state-level implementation factors and racial achievement gaps. Wei (2012) examines the accountability plans submitted by states to the federal government in 2003. She identifies six measures of

[3] Their point estimates suggest that NCLB may have led to a narrowing of the white-black gap in fourth grade math, a narrowing of white-Hispanic gaps in fourth and eighth grade math, but widening of the white-Hispanic gap in fourth grade reading.

accountability stringency: (1) strength of annual measurable objectives (AMO), (2) use of confidence intervals in assessing whether schools meet these objectives, (3) performance indexing, (4) allowance of retesting, (5) minimum subgroup size requirements, and (6) difficulty of state proficiency standards. Wei then estimates the cross-sectional associations between each of these measures of accountability strength and both achievement levels and achievement gaps in 2005. Some of the accountability strength measures are associated with larger gaps; some with smaller gaps; others appear unrelated to achievement gaps. Moreover, these associations are not stable across grades, subjects, and groups. Because her analysis relies on a cross-sectional correlation analysis, these associations should not be interpreted causally.

The studies described above generally argue that NCLB might have affected achievement levels by introducing "consequential accountability" systems. Moreover, they argue, the effects of NCLB in a given state might have been moderated by the stringency of the accountability system adopted by that state. Although these are plausible arguments for how NCLB might have affected *achievement levels*, it is less clear that these features of NCLB would have affected *achievement gaps*. Rather, we reason that NCLB was most likely to affect achievement gaps through its *subgroup-specific* reporting and accountability requirements. By drawing attention to racial/ethnic differences in performance, and by holding schools accountable for the performance of each subgroup, NCLB may have led schools and districts to focus explicitly on reducing achievement gaps at the same time as they focused on increasing achievement levels.

Absent the subgroup-specific nature of the NCLB accountability regime, we might not expect it to be particularly effective at narrowing achievement gaps.[4] Moreover, the subgroup-

[4] That is, unless an accountability regime requires states and schools to focus on relative performance, we reason that it will shift the performance of all groups (to the extent that it is effective) without affecting gaps. It is also possible that an accountability regime would bring up the lower tail of the student achievement distribution and, to the extent that more minority than white students are clustered in that tail, narrow the difference in average test scores between the two groups. However, such an effect might not alter the ordinal ranking of students in the test score distribution, and so would not affect rank-order measures of

specific informational and accountability features of NCLB may be more effective in states where most minority students are in schools where their scores are reported and subject to consequences. This implies that overall accountability stringency may be a less important moderator of NCLB's effect on achievement gaps than is the proportion of minority students in schools meeting the minimum subgroup size criterion. Our analytic strategy below allows us to test this hypothesis.

Analytic Strategy and Hypotheses

Our analysis proceeds in five parts. First, we describe the data and methods used to measure racial achievement gaps in a comparable way across states, grades, years, and test subjects. Second, we describe average within-state trends in white-black and white-Hispanic achievement gaps. Third, we estimate the average effect of NCLB on within-state achievement gaps, using a dose-response model that relies on the assumption that the effects of the NCLB accountability regime accumulate as students progress through school. Fourth, we test whether and how the effect of NCLB varies among states. Our hypothesis is that if NCLB operates via subgroup-specific informational and accountability mechanisms, it should most reduce gaps in states where a large proportion of black and Hispanic students are subject to these mechanisms. That is, states in which most black and Hispanic students attend schools where their group meets the required minimum subgroup reporting size will experience the fastest rates of gap closure. Finally, we investigate whether we can rule out alternative explanations for an association between the magnitude of the NCLB effect and the proportion of minority students with reported scores.

Data Sources and Considerations

Estimating Achievement Gaps

There are three different ways of defining "achievement gaps." First is what we call a

---

achievement gaps. Later, we describe different kinds of gap estimates, some that compare students in an ordinal ranking and others that compare subgroups' mean scores, and estimate the effect of NCLB on each.

"proficiency gap," the between-group difference in the proportions of students scoring above some "proficiency" threshold on a test. Second is an "average score gap," defined as the between-group difference in mean scores on the same test. Third is what we call a "distributional gap," typically described using some summary measure of the relative difference in two groups' test score distributions (such as the difference in means divided by their pooled standard deviation, or the extent to which two distributions overlap). In this paper, we examine all three types of gaps. For reasons explained below, they need not have the same sign, nor trend in the same direction, even when computed from the same data.

The reporting requirements of NCLB make it easy to compute proficiency gaps, but such gaps—and especially their trends—depend heavily on where the proficiency threshold is set relative to the distributions of test scores in the two groups, a point made very clearly by Ho (2008). Indeed, Ho shows that a given trend in test score distributions can lead one to conclude the proficiency gap is widening, remaining constant, or narrowing, depending on where the proficiency threshold is set. This makes proficiency gap trends highly susceptible to where states set their proficiency thresholds, which is an undesirable property for our analysis. Because of the enormous heterogeneity among states in the strictness of their proficiency standards, as well as the heterogeneity in average achievement levels across states, trends in proficiency gaps can be very misleading as indicators of trends in distributional differences. Nonetheless, because NCLB was explicitly designed to narrow proficiency gaps, where "proficiency" was defined individually by each state, it is worth testing whether it does indeed narrow such gaps.

Achievement gaps are more commonly reported using average score differences. One drawback of average score difference measures, however, is that they rely on the assumption that test scores are measured in an interval-scaled metric, meaning that each unit of the score has equal value. Test scores, unlike say, height, have no natural or "cardinal" metric that has equal-interval properties, however. Indeed any monotonic transformation of a test score yields another scale that

may be just as defensible. The interval-scale assumption may therefore be problematic, particularly when comparing trends in achievement gaps, which can be highly sensitive to scale transformations (Reardon, 2008). Another drawback of average score difference measures for our purposes is that state test results are often not reported in ways that allow us to compute subgroup-specific mean scores. Even when mean scores are available, the tests vary across states, grades, and years, making comparisons of mean differences problematic.

Because of the sensitivity of mean or standardized mean difference measures to violations of the interval scale assumption, we rely instead on an alternate distributional gap measure which does not rely on this assumption, the V-statistic (Ho, 2009; Ho & Haertel, 2006; Ho & Reardon, 2012). V is defined as follows: let $P_{a>b}$ be the probability that a randomly chosen individual from group a has a score higher than a randomly chosen individual from group b. Note that this measure depends only on the ordered nature of test scores; it does not depend in any way on the interval-scale properties of the test metric. Now define V as a monotonic transformation of $P_{a>b}$: $V = \sqrt{2}\Phi^{-1}(P_{a>b})$, where Φ^{-1} is the inverse cumulative normal density function. Under this transformation, V can be interpreted as a quasi-effect size. Indeed, if the test score distributions of groups a and b are both normal (regardless of whether they have equal variance), then V will be equal to Cohen's d (the difference in means divided by their pooled standard deviation) .

A useful property of V, however, is that if the test metric is transformed by a non-linear monotonic transformation, Cohen's d will be changed, but V will not. Thus, V can be understood as the value of Cohen's d if the test score metric were transformed into a metric in which both groups' scores were normally distributed. This transformation-invariance property of V is particularly useful when comparing gaps measured using different tests. In order to compare gaps across tests using Cohen's d, we would have to assume that each test measures academic achievement in an interval-scaled metric (so that a score on any test can be written as a linear transformation of a score on any other test). To compare gaps using V, however, we need only to assume that each test

measures achievement in some ordinal-scaled metric, a much more defensible assumption. Reardon and colleagues show that achievement gaps (measured with V) computed from NAEP and from state test data are very highly correlated (the within-state correlation between estimated achievement gap magnitudes is greater than 0.9),[5] indicating that V can be used to compare achievement gaps across a wide range of state and NAEP tests (Shores, Valentino, & Reardon, 2013).

An additional advantage of the V-statistic is that it can be estimated very reliably from either student-level test score data (such as are available for NAEP) or data indicating the number of students of each group in each of several (at least three) proficiency categories (Ho and Reardon 2012). That is, we do not need to know the means and standard deviations of each group's test score distribution; we need only the counts of black, Hispanic, and white students who score "Far Below Basic," "Below Basic," "Basic," "Proficiency," and "Advanced," for example. This makes it possible to easily estimate achievement gaps based on state accountability tests in each state-year-grade-subject for which subgroup-specific proficiency category counts are available.[6]

In sum, this paper employs three different kinds of gap estimates—proficiency, average score, and distributional—the properties of which may yield different estimated treatment effects under different identification strategies. For example, the proficiency measure may overestimate the effects of NCLB on achievement gaps, depending on the state-specific definition of proficiency: if the relevant cut score is set low (and gaps are large prior to 2002), then many white students may appear proficient from the outset, and small changes in minority student achievement may result in large changes in estimated gaps. The average score gap metric, by contrast, may be affected by non-interval scaling of states' test metrics, which may distort gap trends in unknown ways. These differences are important to keep in mind in the interpretation of results, below.

[5] For white-black gaps, the correlation between state-specific gap *trends* estimated from NAEP and state assessments is also very high (greater than 0.9); the gap trend correlation, however, is considerably lower for white-Hispanic achievement gaps (less than 0.25).

[6] See Appendix for details on the estimation of V.

Data

In this paper we use two primary data sources to estimate state-level achievement gaps: NAEP[7] and state assessments. We use NAEP 4th and 8th grade math and reading test score data from 1996 through 2011, and categorical proficiency data (e.g., percentages of students scoring "Below Basic," "Basic," "Proficient," and "Advanced") from state-administered standardized math and reading tests. Most of the state test data comes from tests introduced under the No Child Left Behind Act, but we also use some earlier test score data from states that had state testing programs in place prior to 2002. Typically we have data for grades three through eight, though in some states and years data are available for fewer grades (based on changing federal requirements and increasing state testing capacities). In a small number of states and years, data are available for second grade, as well.[8] From these data we compute estimates of white-black and white-Hispanic gaps in each state-year-grade-subject combination for which we have NAEP and/or state test data. Table 1 shows the number of gap estimates we have from each source for each cohort and grade. Note that the maximum possible number of observations in any state-by-year cell is 200 (50 states x 2 subjects x 2 gap types).[9]

Table 1 here

In total, we have 11,475 state-year-grade-subject-specific achievement gaps estimates, 2,474 from NAEP and 9,001 from state test data. However, the NAEP and state test data we use cover different states, years, and grades. NAEP was administered to representative samples within each state starting in 2003; prior to that, participation in the state NAEP assessment was voluntary,

[7] We use "State NAEP" data, based on math and reading assessments administered to representative samples of fourth- and eighth-graders roughly every two years in each of the 50 states. State NAEP sample sizes are roughly 2,500 students, from approximately 100 schools, in each state-grade-subject.

[8] We do not analyze data from high school because, in many states, not all students take the same tests in high school and because states vary widely in the specific content covered in high school tests and the grades when they are administered.

[9] We exclude the District of Columbia (DC) from our analyses because the white-black achievement gap is substantially higher in DC than in any other state, making it a high-leverage outlier in our descriptive analyses. Its exclusion from our models does not significantly alter our conclusions.

so coverage is incomplete. Likewise, we are able to compute achievement gaps from state test data in all 50 states beginning in 2006, with incomplete coverage in the preceding years. See Appendix B for further details on the sources of state data used for our analyses, and the methods used to reconcile differences among the sources in cases where we had estimates from multiple sources.

Considerations Regarding the Use of NAEP and State Test Data

Each of the data sources employed in this paper has its own distinct set of advantages and disadvantages. First, NAEP and state tests cover different combinations of years and grades, as noted in Table 1 above. Note that cohorts of students entering kindergarten prior to Fall 1994 would have been in high school before NCLB was enacted, while cohorts entering in Fall 2002 or later would have experienced their entire elementary school career under NCLB; cohorts entering kindergarten from 1994 to 2001 experienced NCLB in some grades (their later grades) but not all grades. The dashed lines in Table 1 distinguish the year-grade observations that correspond to pre-1994, 1994-2001, and post-2001 cohorts. These lines make clear that the NAEP data primarily include cohorts of students who entered kindergarten prior to the implementation of NCLB (prior to the fall of 2002). The state data not only include far more observations, but also include much more data from post-2001 cohorts. These differences in the coverage of the NAEP and state data provide complementary identification opportunities, as we discuss below.

The NAEP and state test data differ in a number of other ways, as well. NAEP tests are the same across states and change little over time; state tests are more closely aligned with specific state standards; NAEP tests are low-stakes for both students and schools; state tests have high-stakes for schools; and the state estimates are based on much larger numbers of students than the NAEP data. We explore these differences, and their consequences for achievement gap estimates, elsewhere (Shores et al., 2013). Importantly for the current study, we find that these data sources yield statistically indistinguishable estimates of average achievement gap trends over time, on

average. As a result, we pool data from both sources in our primary analyses below.

State Accountability Measures

We characterize states by the extent to which their implementation of NCLB was likely to focus educators' attention on the performance of black and Hispanic students. As noted above, because each state could set its own minimum subgroup size—the number of students of a subgroup in a school below which scores for that subgroup were not required to be reported or used in determining sanctions—and because states vary in the size of their black and Hispanic student bodies, their levels of between-school racial segregation, and their average school size, states vary in the proportion of black and Hispanic students whose test scores were relevant for accountability purposes. We compute, for each state, the proportion of black and Hispanic students who were in schools in Spring 2002 (prior to the first year of NCLB implementation) where their group met the minimum subgroup size threshold as defined under NCLB. As Figure 1 shows, there is a great deal of variation among states in the proportions of students of different subgroups in schools where their scores are reported for subgroup-specific accountability purposes.

Figure 1 here

Covariates

In many of our models, we include a variety of state-level time-varying and time-invariant covariates, both to reduce possible bias and to improve the precision of our estimates. We construct these covariates using data from two main sources: the Current Population Survey (CPS) and the Common Core of Data (CCD). From the CPS, we compute the white-black and white-Hispanic average household income ratio, poverty ratio, and unemployment rate ratio for each state and year. From the CCD, we compute the levels of white-black and white-Hispanic school segregation and the proportion of public school students who are black and Hispanic, for each state

and year. The method and rationale for constructing both time-varying and time-invariant covariates is explained in Appendix A.

Recent Within-State Trends in Racial Achievement Gaps

To begin, we examine the average within-state trend in the white-black and white-Hispanic achievement gaps in math and reading from 1996 through 2011. Figures 2 and 3 show these trends, as estimated from both NAEP and state accountability test data.[10] Two features of the figures are notable. First, both white-black and white-Hispanic achievement gaps have been narrowing, albeit slowly and unevenly, over the last 15 years; this trend is evident in both the NAEP and state test data. Second, we find no evidence to suggest that racial achievement gaps began closing faster after the introduction of NCLB in 2002. In fact, NAEP and state data suggest that the rate of narrowing of white-black gaps in math and reading has slowed significantly over the last decade. With respect to white-Hispanic gaps, we observe a similar trend until 2007, followed by an increase in rates of gap closure in more recent years. At these rates, it will take more than 50 years for white-Hispanic gaps to be eliminated. For white-black gaps, it will take more than a century.

Figures 2 and 3 here

Figures 2 and 3 align with the findings in Shores et al. (2013), who estimate trends in achievement gaps using only NAEP and state test data from the same state, year, grade, and subject.

---

[10] The trends displayed in Figures 2 and 3 indicate the trend in the estimated year fixed effects (the $\hat{\Gamma}_y$'s) from the precision-weighted random-effects model

$$\hat{G}_{ysg} = \lambda_s + \Gamma_y + u_{\gamma 1s}(year_y - 2002) + u_{\gamma 2s}(year_y - 2002)^2 + \alpha_s(grade_g - 4) + e_{ysg} + \epsilon_{ysg},$$

$$e_{ysg} \sim N[0, \sigma^2]$$

$$\epsilon_{ysg} \sim N[0, \omega_{ysg}^2] = N[0, var(\hat{G}_{ysg})]$$

$$\begin{bmatrix} \lambda_s \\ u_{\gamma 1s} \\ u_{\gamma 2s} \\ \alpha_s \end{bmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \\ 0 \\ \alpha \end{pmatrix}, \begin{pmatrix} \tau_\lambda & \tau_{\lambda u_1} & \tau_{\lambda u_2} & \tau_{\lambda\alpha} \\ \tau_{u_1\lambda} & \tau_{u_1} & \tau_{u_1u_2} & \tau_{u_1\alpha} \\ \tau_{u_2\lambda} & \tau_{u_2u_1} & \tau_{u_2} & \tau_{u_2\alpha} \\ \tau_{\alpha\lambda} & \tau_{\alpha u_1} & \tau_{\alpha u_2} & \tau_\alpha \end{pmatrix}\right].$$

The year fixed effects describe the average trend in the size of within-state achievement gaps. The model allows each state to have a state-specific intercept (λ_s), a state-specific linear grade trend (α_s), and state-specific linear and quadratic deviations (u_{y1s} and u_{y2s}) from the average temporal trend described by the year fixed effects.

While these figures reveal clear trends at the national level, they obscure substantial variation across states. White-black gaps show greater variability than white-Hispanic gaps, in general, and both gap types differ by data source and academic subject. The standard deviations of the annual change in white-black gaps estimated from NAEP data are 0.009 (math) and 0.008 (reading), and are 0.018 (math) and 0.015 (reading) when based on state test data. The same estimates for white-Hispanic gaps are 0.005 (math) and 0.007 (reading) in NAEP, and 0.012 (math) and 0.013 (reading) in state test data. These standard deviations are large relative to the average trends over time, indicating that gaps are growing in some states and narrowing in others.

The Average Effect of NCLB on Racial Achievement Gaps

Methodological Approach

In Figures 2 and 3, above, we describe the trends in achievement gaps across calendar years (pooling across grades). For the remainder of the analyses, however, we describe trends across student cohorts rather than calendar years. This change in perspective allows us to examine how the achievement gap changes across cohorts experiencing different numbers of years of schooling prior to and during the NCLB regime. In Appendix A, we derive a model for the achievement gap for a given cohort c in grade g in state s. This model describes the achievement gap as a state-specific function of grade (denoted gr_g), cohort (denoted coh_c) and test-subject, as well as the number of years the cohort has been exposed to NCLB by the end of grade g (exp_{cg}). In some specifications, we include vectors of cohort-by-state and cohort-by-state-by grade covariates ($\mathbf{X}_{cs}$ and $\mathbf{W}_{csg}$). We estimate the relevant parameters using precision-weighted random coefficients models of the following form:

$$\hat{G}_{csgt} = \lambda_s + \alpha_s(gr_g) + \eta(E_g) + \beta(gr_g \cdot coh_c^*) + \gamma_{1s}(coh_c^*) + \gamma_{2s}(coh_c^*)^2 + \zeta(sub_t) + \delta_s(exp_{cg}) + \mathbf{X}_{cs}\mathbf{A} + \mathbf{W}_{csg}\mathbf{B} + e_{csgt} + \epsilon_{csgt},$$

$$e_{csgt} \sim N[0, \sigma^2]$$

$$\epsilon_{vsgt} \sim N\left[0, \omega^2_{csgt}\right] = N\left[0, var\left(\hat{G}_{csgt}\right)\right]$$

$$\begin{bmatrix} \lambda_s \\ \gamma_{1s} \\ \gamma_{2s} \\ \alpha_s \\ \delta_s \end{bmatrix} \sim N\left[\begin{pmatrix} \lambda \\ \gamma_1 \\ \gamma_2 \\ \alpha \\ \delta \end{pmatrix}, \begin{pmatrix} \tau_\lambda & \tau_{\lambda\gamma_1} & \tau_{\lambda\gamma_2} & \tau_{\lambda\alpha} & \tau_{\lambda\delta} \\ \tau_{\gamma_1\lambda} & \tau_{\gamma_1} & \tau_{\gamma_1\gamma_2} & \tau_{\gamma_1\alpha} & \tau_{\gamma_1\delta} \\ \tau_{\gamma_2\lambda} & \tau_{\gamma_2\gamma_1} & \tau_{\gamma_2} & \tau_{\gamma_2\alpha} & \tau_{\gamma_2\delta} \\ \tau_{\alpha\lambda} & \tau_{\alpha\gamma_1} & \tau_{\alpha\gamma_2} & \tau_\alpha & \tau_{\alpha\delta} \\ \tau_{\delta\lambda} & \tau_{\delta\gamma_1} & \tau_{\delta\gamma_2} & \tau_{\delta\alpha} & \tau_\delta \end{pmatrix}\right]$$

[1]

Here $\hat{G}_{csgt}$ is the estimated achievement gap in state s in subject t for cohort c in grade g; coh^*_c is a continuous variable indicating the calendar year in which the cohort entered kindergarten, centered at 2002; gr_g is a continuous variable indicating the grade in which $\hat{G}_{csgt}$ is measured (gr_g is centered at -1, so that it measures the number of years of schooling students have had by the spring of grade g); E_g is a variable defined as $\frac{1}{2}\left(gr_g^2 - gr_g\right)$ (see Appendix A); sub_t is a dummy variable indicating whether $\hat{G}_{csgt}$ is a math or reading gap; and exp_{cg} is the number of years that cohort c has been exposed to NCLB by the spring of grade g.

The key parameter of interest is δ, the average annual effect of NCLB on the achievement gap within a cohort. The error term ϵ_{csgt} is the sampling error of $\hat{G}_{csgt}$; we set its variance ω^2_{csgt} to be equal to the square of the standard error of $\hat{G}_{csgt}$. We estimate the parameters of this model, as well as σ^2 and the $\boldsymbol{\tau}$ matrix, using maximum likelihood methods as implemented in the HLM v7 software (Raudenbush, Bryk, & Congdon, 2012).

The identification of δ in Model [1] comes from two sources of variation in exp_{cg}. First, for cohorts who entered kindergarten in Fall 2002 or earlier, $exp_{cg} = 0$ prior to the 2002-03 school year, and then increases across grades (within a cohort) or across cohorts (within a grade) after the 2001-02 year. Thus, for pre-2003 cohorts, δ is the average within-state difference in the trend in the achievement gap across grades within a cohort before and after Spring 2002; equivalently, δ is the average within-state difference in the trend in the achievement gap across cohorts within a grade before and after Spring 2002. Second, for years after 2002, $exp_{cg} = coh^*_c + gr_g$ for cohorts

entering kindergarten prior to 2003, but $exp_{cg} = gr_g$ for later cohorts. Thus, after 2002, δ is the average within-state difference in the trend in the achievement gap across cohorts within a grade between pre-2003 cohorts and later cohorts.

Figure A1 helps to clarify these different sources of variation in exp_{cg} (see also the discussion in Appendix A). The first source of variation is represented by the transition from yellow to green shading; the second source of variation is represented by the transition from green to blue shading. Because almost all of the available NAEP data fall in the yellow and green regions (the 2009 and 2011 4th grade NAEP data, corresponding to the 2004 and 2006 cohorts, are an exception), the estimates based on NAEP data rely on the first source of variation in exp_{cg}. The estimates based on state test data rely on both, but more heavily on the second source of variation, as most of the state data are collected after 2002.

We fit several versions of Model [1], each using different subsets of our data. Our most comprehensive models pool all the data—both NAEP and state data, math and reading gaps, and data from all available cohorts and years—for both white-black and white-Hispanic gap estimates. We then fit models that use different subsets of the data, defined by data source, gap measure, and subject. We focus on estimating models that use the V-statistic as the outcome, as we are most interested in the effects of NCLB on the distributional gap. However, we also fit models using the mean difference and proficiency gap measures for comparison.

Because NCLB applied to all states beginning in Fall 2002, there is no variation among states in the exposure variable within a given cohort and grade. Thus, the identification of δ in Model [1] depends on the assumption that there is no other factor that affected all states' achievement gap trends in a similar way following 2002. In other words, to interpret the coefficient δ as the effect of NCLB, we must assume that no other policy or demographic change in 2002, net of the demographic trends captured by our control variables, had a cumulative effect on achievement gaps in the years following 2002. We discuss the plausibility of this assumption below.

Our model also assumes the annual impact of NCLB on achievement gaps is constant over time. One might imagine that the slow and uneven pace of implementation of NCLB, as well as the fact that its sanctions took some years to take effect, might lead the effects of NCLB to be larger in later years than in initial years. Alternately, one might imagine that the effects of NCLB on a cohort of students are evident quickly after that cohort encounters NCLB, but do not accumulate over time, a pattern similar to that found in prior studies of NCLB. These earlier studies rely on comparative interrupted time series models that allow the policy to have an immediate effect, as well as effects that develop over time (Dee and Jacob, 2011; Lee, 2006; Lee and Reeves, 2012; Wong, Cook, and Steiner, 2011). When we include additional parameters in our models to allow for temporal variation in the effectiveness of the policy, however, we find no evidence that the effects change systematically over time, and no evidence of any immediate effect at the policy's initial implementation. We therefore report estimates from models that assume the annual additive effect of NCLB on achievement gaps is constant across years and grades.

Results

Table 2 reports the estimates of δ, the average effect of each additional year of exposure to NCLB on a cohort's achievement gap, obtained from fitting Model [1] to different combinations of data sources, measures, and test subjects. We report the estimates from models with and without the vectors of covariates $\mathbf{X}_{cs}$ and $\mathbf{W}_{csg}$. These include black/white (or Hispanic/white, as appropriate) income ratio, poverty ratio, and unemployment ratio, as well as the proportion of black (or Hispanic) students in public schools and the level of black/white (or Hispanic/white) school segregation. In general, the estimates change little when we add the covariates, suggesting that there is little systematic confounding of exposure to NCLB and our cohort- and time-varying covariates.

Table 2 here

The top panel of Table 2 contains the estimated effect of NCLB using the "distributional gap" measure V, computed from pooled NAEP and state accountability data. When using all observations (all available cohorts, years, and subjects), the estimated effect of exposure to NCLB is statistically indistinguishable from 0 for white-black gaps ($\hat{\delta} = -0.005, se = 0.003$ without covariates, $\hat{\delta} = -0.004, se = 0.004$ with covariates). Likewise, for white-Hispanic gaps, this effect is insignificant in both the base model and after including covariates ($\hat{\delta} = +0.005, se = 0.004$). The second and third panels of the Table report the same parameters as estimated separately from NAEP and state data. The NAEP estimates show a marginally significant effect of NCLB on both white-black and white-Hispanic achievement gaps. Here, one year of exposure to the policy induces an increase in achievement gaps of 0.007 ($se = 0.004, p < .10$) standard deviations per year. These point estimates increase slightly after including covariates, attaining clearer statistical significance in the case of white-black gaps ($\hat{\delta} = +0.009, se = 0.004, p < .05$ for white-black gaps, $\hat{\delta} = +0.008, se = 0.004, p < .10$ for white-Hispanic gaps,). The third panel shows no significant effects on distributional gaps computed from state test data. One reason for the difference between the NAEP and the state data is that they rely on data from different years, cohorts, and grades. When we restrict the models to include only the pre-2003 cohorts or only data from 2003 and later (to restrict the analyses to only one of the two sources of variation in exposure to NCLB described above), we find virtually no evidence of significant effects of NCLB and no significant differences between the NAEP and state data estimates (results available upon request).

The fourth panel of Table 2 shows the estimated effects of NCLB on achievement gaps measured as differences in average test scores. Here there is no evidence of any significant or sizeable effect on achievement gaps. The bottom panel of Table 2, which shows estimated effects on proficiency gaps, suggests that NCLB narrowed white-black differences in state-defined proficiency gaps by roughly one-half ($se = 0.24, p < .05$) of a percentage point per year. The comparable estimates for white-Hispanic gaps are not as large and not statistically significant.

How are we to interpret the significant and negative coefficients on the white-black proficiency gap in light of the non-significant (or sometimes positive) coefficients on the distributional and mean difference gap measures? We might conclude that the bottom panel implies that NCLB has been effective at narrowing white-black gaps when gaps are measured by the metric NCLB privileges. As Ho (2008) notes, however, proficiency gaps and their trends are highly sensitive to the location of the proficiency threshold in the test score distribution. Indeed, proficiency gaps can narrow or widen as a result of changes in the location of the proficiency threshold and/or changes in average test scores, *even if all students retain their same position in the distribution*. This suggests the need for caution in interpreting the proficiency gap coefficients as evidence that achievement gaps systematically narrowed as a result of exposure to NCLB.

Between-State Variation in the Average Effect of NCLB on Racial Achievement Gaps

Although Table 2 suggests that NCLB has not narrowed achievement gaps, on average, these averages may mask considerable heterogeneity among states in the effect of NCLB. Indeed the estimated standard deviation of the effect of NCLB on the white-black and white-Hispanic gaps is 0.013 and 0.018, respectively (deviance tests reject the null hypothesis that these standard deviations are zero; $p < .001$). These standard deviations are larger than the estimated average effects, indicating that there are some states where the effect of NCLB on gaps is positive and others where it is negative. Figure 4 shows the estimated state-specific effects of exposure to NCLB. Specifically, the figure shows the Empirical Bayes estimate of δ_s from models that pool math and reading and NAEP and state test data, and that include the vectors of covariates.

Figure 4 here

Methodological Approach

Having found that the effect of NCLB varies significantly across states, we add to Model [1] a

term representing the interaction of exp_{cg} and a variable indicating the proportion of black (or Hispanic) students in Spring 2002 who were in schools that would meet the state's minimum subgroup reporting size threshold. A negative coefficient on this interaction term would be consistent with our theoretical expectations; that is, it would imply that achievement gaps narrowed more after the start of NCLB in states where a larger proportion of black or Hispanic students are in schools where their groups' test scores are consequential for determining whether the school is meeting NCLB's Adequate Yearly Progress benchmarks.

Results

The estimated coefficients on exp_{cg} and the interaction term are shown in Table 3. Several patterns are evident in the coefficients reported in Table 3. First, the coefficient on the exposure variable is often positive and significant in these models. This implies that in states where no or very few black and Hispanic students were in schools meeting the minimum subgroup size (as was true in Vermont, Idaho, and Montana for white-black gaps, and in Maine, Montana, and West Virginia for white-Hispanic gaps), achievement gaps actually grow with increased exposure of cohorts to NCLB. The negative (and significant) coefficients on the interaction term, however, indicate that gaps widened less, or narrowed, in states where the proportion of minority students subject to test score reporting was larger. In a hypothetical state with 100 percent of black students in schools held accountable for their performance, these results suggest that NCLB would have a total effect of narrowing white-black distributional gaps by 0.017 standard deviations per year. The estimates are consistent in sign and general magnitude across model specifications, though the standard errors are smaller when using state data compared to NAEP. They are also consistent with previous work (Dee and Jacob 2011), in which the effect of NCLB on black students' achievement is larger in states with larger black populations. Figure 5 illustrates these estimates by plotting the Empirical Bayes estimates of the state-specific NCLB effects against the proportion of

black students in schools where they met the states minimum subgroup size reporting threshold.

Table 3 here

Figure 5 here

Table 3 also indicates that the impact of NCLB on the white-Hispanic achievement gap varies systematically with the proportion of minority students in schools meeting the state's minimum subgroup size. However, these estimates are statistically significant only in the models that employ state data; the NAEP coefficients are much smaller in magnitude and have larger standard errors. Figure 6 shows that the relationship between NCLB-induced changes in white-Hispanic gaps and the proportion of Hispanic students in schools subject to accountability is considerably more noisy than the equivalent relationship for white-black gaps. Nevertheless, across both racial achievement gaps, there does appear to be a relationship between the proportion of minority students in schools where their scores are reported and the magnitude of the effect of NCLB on the achievement gap, though this relationship is stronger and more robust for white-black gaps than for white-Hispanic gaps.

Figure 6 here

Additional Analyses

Although the analyses above show that the effects of NCLB on achievement gaps vary systematically with respect to our measure of subgroup-specific accountability pressure, it is not clear that this association should be interpreted causally. Prior research on NCLB has identified a number of aspects of both pre-NCLB policy context and post-NCLB policy implementation that vary among states and that have been hypothesized to moderate the impact of NCLB on academic achievement. Dee and Jacob (2011), for example, argue that NCLB represented a larger change in states that had no pre-NCLB consequential accountability system than in states that already had NCLB-type systems in place. Others have argued that between-state differences in the

implementation of NCLB—in the rigor of standards set by states, the use of performance indexing or annual measureable objectives to rate school performance, and the availability of funds for schools in need of improvement, for example—may lead to differential effects of NCLB across states (see Lee 2006; Lee and Reeves 2012; Wei 2012; Wong, Cook, and Steiner forthcoming).

To test whether the effects of NCLB vary in relation to these features of pre-NCLB context and post-NCLB implementation, we fit a set of models in which we include the interaction of exp_{cg} with each of the NCLB policy variables used in prior research on NCLB. These measures were provided to us by the authors of those papers, and are described in detail in the relevant papers (Dee and Jacob 2011; Lee 2006; Lee and Reeves 2012; Wei 2012; Wong, Cook, and Steiner 2011). The results of these analyses, reported separately by subject and gap type, appear in Table 4. (We report them separately by subject because some of the accountability measures are subject-specific.) None of these accountability implementation measures show any sizeable or consistent relationship to racial achievement gaps.

Table 4 here

Looking across Table 4, however, our proportion accountable variable has the largest and most consistently significant relationship with both white-black and white-Hispanic gaps. But we remain uncertain about the extent to which this relationship is determined by features of the law—i.e., the minimum subgroup reporting threshold left to states' discretion in the legislation—as compared with state-specific characteristics like demographic composition and within-school levels of segregation. In fact, Wei's measure of minimum subgroup size in 2003 (in the fifth panel) has no significant relationship with math or reading gaps. We probe other determinants of proportion accountable, and of NCLB's effect on achievement gaps, below.

Robustness Checks

The above analysis suggests that between-state variation in the effects of NCLB on

achievement gaps is not driven by general accountability pressure, but rather by something associated with subgroup-specific accountability pressure. That is, the degree to which NCLB focused accountability pressure on all students and schools in a state does not appear to be associated with changes in achievement gaps in that state, all other things equal. Whether subgroup-specific accountability pressure *per se* is driving the variation in observed effects is less clear, however. It could be that other factors correlated with this pressure—such as school segregation levels or state racial composition—play a role in moderating the effects of NCLB, independent of their impact on the proportion of students whose scores are relevant for subgroup-specific accountability purposes. To investigate this hypothesis, we conduct three partial tests of whether the proportion of minority students in schools where their scores are reported has a causal effect on the size of the NCLB effect.

First, we include a set of control variables as interactions in Model [1]. Specifically, we include interactions of the exposure variable (exp_{cg}) with racial composition, school segregation, and average school size (the three state demographic factors that influence the proportion accountable). We also include an interaction of the exposure variable with a variable indicating the size of the state's pre-NCLB achievement gap. Second, we instrument for proportion accountable with state minimum subgroup threshold. Under the assumption that states set their minimum subgroup threshold exogenously (that is, minimum subgroup thresholds were not set in ways associated with states' potential NCLB effect outcomes, net of school segregation levels, racial composition, average school size, and pre-NCLB achievement gaps), this IV model will yield causal estimates of the within-state effect of proportion accountable on racial achievement gaps. Unfortunately, these IV analyses are underpowered (the F-statistics from these models are 6.4 and 23.8 for the white-black and white-Hispanic models, respectively, and the standard errors are 2.5 to

8 times larger than the estimates shown in Table 3), so we do not report the findings here.[11] Third, we pool the white-black and white-Hispanic NCLB effect estimates and test whether the NCLB effect within a state is larger for the subgroup that has more students in schools where their scores are reported, controlling for differences between the subgroups in population size, segregation, and pre-NCLB gap magnitudes. The results of these analyses are described below.

We begin by estimating the pairwise correlations between the proportion of black and Hispanic students subject to accountability in each state and several other relevant factors. These factors include: the minimum subgroup reporting threshold, the demographic make-up of public school students in the state, average school size for schools serving grades 2 through 8, subgroup-specific racial segregation, achievement gaps in 2003 (prior to full implementation of the policy), and the state's mean NAEP score in 2003. Table 5 reports these correlations.

Table 5 here

In the first row, we find that the proportion of black and Hispanic students subject to accountability is not highly correlated with state subgroup reporting thresholds. These correlations (-0.078 for proportion black students accountable and -0.136 for Hispanics accountable) are in the expected direction, but they suggest that state demographic factors may play a larger role than the minimum subgroup size in determining the population subject to reporting and accountability requirements under NCLB. Instead, the proportion of both black and Hispanic students subject to accountability is highly correlated with average school enrollment in schools with grades 2 through 8, school-level segregation, and the relevant NAEP gap in 2003 (0.532, 0.580, and 0.750 for black students; and 0.446, 0.534, and 0.688 for Hispanic students, respectively). These high correlations suggest the possibility of multicollinearity in subsequent analyses. Although we proceed with tests of the causal relationship between state-level proportion accountable and NCLB effect size, we do so with caution based on the results in Table 5.

[11] The point estimates from these models are positive (unlike the estimates in Table 3), but have very large standard errors and are not distinguishable from 0. Results available from authors upon request.

Our first test includes additional interaction terms as control variables in Model [1]. We construct interactions of the exposure measure with four covariates that are moderately correlated with the proportion accountable: average school size, percent black (or Hispanic) public school enrollment, average white-black (or white-Hispanic) segregation level, and pre-NCLB achievement gap size. The estimates from the models including these interactions are shown in Table 6. After controlling for these state-level characteristics, the interaction of exposure to NCLB and the proportion of students in schools where their scores are reported remains negative and statistically significant, suggesting that the effects observed in Table 3 above are robust to the inclusion of additional control variables. However, these patterns are not consistent across test subjects or across data sources—the coefficients are negative and statistically significant only when we use the state data, but not in the NAEP data. Moreover, the same terms in the white-Hispanic models are generally near zero and not statistically significant. Although these findings imply that subgroup-specific accountability, over and above state-level demographic or pre-NCLB educational factors, may be associated with post-NCLB trends in the white-black achievement gap, it is not clear why this is not also evident in the trends in white-Hispanic gaps.

Table 6 here

We next look within states to examine the association between NCLB effects and proportion of minority students in schools held accountable for their performance. Specifically, we examine the relationship between the estimated annual effect of NCLB on the achievement gap between whites and students of group m ($\hat{\delta}_{ms}$) and characteristics of group m's distribution among schools in that state, including the proportion of group m students subject to accountability, segregation levels of group m, the proportion of students of group m in the state's public schools., and the size of the gap between group m and white students in 2003. We fit the following precision-weighted state random-effects model with within-state centered covariates:

$$\hat{\delta}_{ms} = \mathbf{A} + (\mathbf{X}_{ms} - \overline{\mathbf{X}}_s)\mathbf{B} + \left(\overline{\mathbf{X}}_s - \overline{\overline{\mathbf{X}}}\right)\mathbf{\Gamma} + u_s + v_{ms} + \omega_{ms}$$

$$u_s \sim N[0, \tau]$$
$$v_{ms} \sim N[0, \sigma^2]$$
$$\omega_{ms} \sim N\left[0, var\left(\hat{\delta}_{ms}\right)\right]$$

[2]

where s indexes states and m indexes minority groups (black or Hispanic). The vector $\mathbf{X}_{ms}$ includes the proportion of group m in schools where they met the minimum subgroup threshold in state s, the segregation of group m from whites in state s, the proportion of group m in the total enrollment in state s, and the size of the gap in 2003 (and a dummy variable indicating whether the observation pertains to black or Hispanic effects). **B** is the parameter of interest; it represents the average within-state association between δ_{ms} and $\mathbf{X}_{ms}$.

Table 7 here

The results of these within-state analyses appear in Table 7. Here, after controlling for racial segregation, the proportion of minority public school enrollment, and 2003 achievement gaps, we find that the proportion of black and Hispanic students in schools held accountable for their performance has no statistically significant independent association with the effect of NCLB on racial achievement gaps. This suggests that it may not be the extent of subgroup-specific accountability pressure *per se* that drives the effects of NCLB on achievement gaps, but other state-level factors. Other factors correlated with state demographic composition—for example, the distribution of teaching talent—may be the "active ingredients" through which NCLB affected racial achievement gaps.

In fact, one provision of NCLB, separate from its accountability regime, is the Highly Qualified Teacher (HQT) provision. This provision requires that all teachers have a bachelor's degree, full state certification or licensure, and documented knowledge of the relevant subject matter. NCLB may affect achievement gaps by equalizing the distribution of qualified teachers and, therefore, weakening the relationship between students' background characteristics and the quality

of teaching they experience. We know of no papers that examine this question directly. In additional analyses not shown here, we estimated the effect of NCLB on changes in the distribution of teachers with respect to students of different races, using the Schools and Staffing Survey (SASS). We find no meaningful change in any of several measures of observable teaching quality that might confound the findings, above.[12] However, we do find evidence in the extant literature to suggest that increases in teacher compensation due to NCLB were largely focused in high-poverty schools (Dee, Jacob, & Schwartz, 2013). To the extent that high-poverty and high-minority schools overlap, then, differential increases in compensation may have induced unobserved increases in teacher quality for black and Hispanic students.

Finally, the estimated effect of NCLB on achievement gaps may be explained by changes in retention policies, special education classification, or definitions of "continuous student enrollment" that determine which students are subject to the law's testing and reporting requirements (cf. Davidson et al., 2013). Any of these processes may exempt students from the requirements of regular testing; to the extent that the students removed are disproportionately black or Hispanic, these processes may confound the estimated effects of the policy. A rigorous test of this hypothesis is beyond the scope of the current study. However, it remains a potential alternative explanation for the findings observed here.

[12] Using the 1993-1994, 1999-2000, 2003-2004, and 2007-2008 waves of the Schools and Staffing Survey, we examined changes in the white-minority gap in exposure to high quality teachers over time across states. We considered four measures of high quality teachers to approximate NCLB's definition of "high quality." We measured differences in exposure to teachers with Master's degrees (there was no variation in exposure to teachers with Bachelor's degrees), teachers with regular or standard certification, and teachers with any certification (relative to none). We also computed the white/minority ratio of exposure to teacher experience. We used interrupted time series models to examine differences in exposure pre- and post-NCLB among states with varying segregation levels, minority population sizes, and proportions of students being held accountable. Findings are available from the authors upon request. Overall, we find no significant post-NCLB changes in the gap in any measure of exposure to high quality teachers; nor do we find evidence that that the gaps changed more in states where more minority students were in schools where their scores were used for subgroup-specific accountability purposes.

Conclusion

Overall, we find that racial achievement gaps have been closing slowly since 1990. This is true for both white-black and white-Hispanic gaps. Based on this trend, we turn to the period of No Child Left Behind and ask whether this federal policy, which explicitly aimed to narrow gaps between minority and nonminority students, was successful at achieving its goal. We find no consistent evidence that NCLB has narrowed achievement gaps, on average. Our estimates are very precise, and we can rule out the possibility that NCLB had, on average, meaningfully large effects (effects larger than 0.01 standard deviations change per year) on achievement gaps.

Despite the fact that NCLB appears to have had no average effect on achievement gaps, its effect does appear to vary among states. Moreover, the effects of NCLB vary with the proportion of minority students in schools where they are subject to accountability pressure. This is consistent with the hypothesis we framed at the start of the paper—that is, that greater information about achievement gaps and greater subgroup-specific accountability pressure on schools should lead to more rapid narrowing of these gaps. Our supplemental analyses suggest, however, that we cannot rule out the possibility that other processes may be at work instead of (or in addition to) those hypothesized. Our analyses suggest that NCLB has been most effective at narrowing achievement gaps in states with segregated minority student populations and in states where achievement gaps were largest prior to NCLB. These characteristics are highly correlated with the proportion of minority students in schools where their scores are reported, but we cannot isolate the latter as the cause of NCLB's greater effectiveness.

As Congress considers reauthorization of the Elementary and Secondary Education Act, our findings suggest the need for prudence in the revision of its accountability regime. If the proportion of students subject to testing and reporting requirements does, in fact, influence racial achievement gaps—either through subgroup-specific accountability pressure, or through associated factors like state demographic composition or existing educational inequality—then related provisions should

be subject to special scrutiny. As Davidson and colleagues (2013) argue, "Complex and off the radar of all but the most embedded policymakers and researchers, these esoteric rules have substantive impacts on schools" (p. 3). These authors conclude that accountability policies should be "sensibly standardized" (p. 23) in order to achieve uniform goals, but the findings in this study suggest otherwise, at least with respect to racial achievement gaps. Instead, state context, particularly student body racial composition, segregation, and other determinants of the proportion of minority students included in school-level AYP calculations, might be factored into the overall accountability scheme. In addition, distributional achievement gaps might be placed alongside subgroup-specific proficiency rates as an accountability metric of interest.

Despite its intentions, there is no evidence that NCLB-style accountability has led to any substantial narrowing of achievement gaps. Although there is variation among states in the effects of NCLB, comparing the magnitude of these effects is akin to comparing the speed of different glaciers: some are retreating, some advancing, but none so fast that one would notice a meaningful difference except over a span of decades (or centuries). Even in those states where NCLB's effects on achievement gaps have been greatest, our estimates suggest that NCLB has narrowed achievement gaps at a rate of only two-one-hundredths of a standard deviation per year. Over a student's K-8 career, this would still only narrow the achievement gap by less than one-fifth of a standard deviation. NCLB's framers aimed to "ensure that all children have a fair, equal, and significant opportunity to obtain a high-quality education." With respect to racial achievement gaps, our findings suggest that NCLB has not been successful at this goal. For future education policies to be more successful, we will likely have to adopt a different, perhaps more deliberate, set of strategies.

References

Alon, S., & Tienda, M. (2007). Diversity, opportunity, and the shifting meritocracy in higher education. *American Sociological Review, 72*(4), 487-511.

Aspen Institute. (2006). Commission staff research report: Children with disabilities and LEP students: Their impact on the AYP determination of schools *Commission on No Child Left Behind*. Aspen, CO: The Aspen Institute.

Bastedo, M. N., & Jaquette, O. (2011). Running in place: Low-income students and the dynamics of higher education stratification. *Educational Evaluation and Policy Analysis, 33*(3), 318-339.

Bollinger, C. (2003). Measurement error in human capital and the black-white wage gap. *The Review of Economics and Statistics, 85*(3), 578-585.

Carneiro, P., Heckman, J. J., & Masterov, D. V. (2003). Labor market discrimination and racial differences in premarket factors *NBER Working Paper*. Cambridge, MA: National Bureau of Economic Research.

Carnoy, M., & Loeb, S. (2002). Does external accountability affect student outcomes? A cross-state analysis. *Educational Evaluation and Policy Analysis, 24*(4), 305-331.

Chudowsky, N., & Chudowsky, V. (2005). States test limits of federal AYP flexibility. Washington, DC: Center on Education Policy.

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. (2005). Who teaches whom? Race and the distribution of novice teachers. *Economics of Education Review, 24*(4), 377-392.

Davidson, E., Reback, R., Rockoff, J. E., & Schwartz, H. L. (2013). Fifty ways to leave a child behind: Idiosyncrasies and discrepancies in states' implementation of NCLB *Working Paper Series. Working Paper 18988*. Cambridge, MA: National Bureau of Economic Research.

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

Dee, T. S., Jacob, B., & Schwartz, N. L. (2013). The effects of NCLB on school resources and practices. *Educational Evaluation and Policy Analysis*.

Fulton, M. (2006). Minimum subgroup size for Adequate Yearly Progress (AYP): State trends and highlights *StateNotes*. Denver, CO: Education Commission of the States.

Gaddis, S. M., & Lauen, D. L. (2011). *Has NCLB accountability narrowed the black-white test score gap?*

Goldin, C., & Katz, L. F. (2008). *The Race Between Education and Technology*. Cambridge, MA: Harvard University Press.

Grissmer, D. W., Flanagan, A., & Williamson, S. (1998). Why did the Black-White score gap narrow in the 1970s and 1980s? In C. Jencks & M. Phillips (Eds.), *The Black-White Test Score Gap* (pp. 182-228). Washington, D.C.: Brookings Institution Press.

Grodsky, E., & Pattison, E. (in progress). Changes in ascribed and achieved advantage in American higher education.

Hanushek, E. A., & Raymond, M. E. (2004). The effect of school accountability systems on the level and distribution of student achievement. *Journal of the European Economic Association, 2*(2-3), 406-415.

Hanushek, E. A., & Raymond, M. E. (2005). Does school accountability lead to improved student performance? *Journal of Policy Analysis and Management, 24*(2), 297-327.

Hedges, L. V., & Nowell, A. (1998). Black-White Test Score Convergence Since 1965. In C. Jencks & M. Phillips (Eds.), *The Black-White Test Score Gap* (pp. 149-181). Washington, D.C.: Brookings Institution Press.

Hedges, L. V., & Nowell, A. (1999). Changes in the black-white gap in achievement test scores. *Sociology of Education, 72*(2), 111-135.

Hemphill, F. C., Vanneman, A., & Rahman, T. (2011). Achievement Gaps: How Hispanic and White Students in Public Schools Perform in Mathematics and Reading on the National Assessment

of Educational Progress. Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.

Ho, A. D. (2008). The problem with "proficiency": Limitations of statistics and policy under No Child Left Behind. *Educational Researcher, 37*(6), 351-360.

Ho, A. D. (2009). A nonparametric framework for comparing trends and gaps across tests. *Journal of Educational and Behavioral Statistics, 34*, 201-228.

Ho, A. D., & Haertel, E. H. (2006). Metric-free measures of test score trends and gaps with policy-relevant examples. Los Angeles, CA: Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing, Graduate School of Education & Information Studies.

Ho, A. D., & Reardon, S. F. (2012). Estimating achievement gaps from test scores reported in ordinal 'proficiency' categories. *Journal of Educational and Behavioral Statistics, 37*(4), 489-517.

Kober, N., Chudowsky, N., & Chudowsky, V. (2010). State Test Score Trends through 2008-09, Part 2: Slow and Uneven Progress in Narrowing Gaps. *Center on Education Policy*, 79.

Lankford, H., Loeb, S., & Wyckoff, J. (2002). Teacher sorting and the plight of urban schools: A descriptive analysis. *Educational Evaluation and Policy Analysis, 24*(1), 37-62.

Lauen, D. L., & Gaddis, S. M. (forthcoming). Shining a light or fumbling in the dark? The effects of NCLB's subgroup-specific accountability on student achievement. *Educational Evaluation and Policy Analysis*.

Lee, J. (2006). Tracking achievement gaps and assessing the impact of NCLB on the gaps: An in-depth look into national and state reading and math outcome trends. Cambridge, MA: The Civil Rights Project at Harvard University.

Lee, J., & Reeves, T. (2012). Revisiting the impact of NCLB high-stakes school accountability, capacity, and resources: State NAEP 1990-2009 reading and math achievement gaps and trends. *Educational Evaluation and Policy Analysis, 34*(2), 209-231.

Lee, J., & Wong, K. K. (2004). The impact of accountability on racial and socioeconomic equity: Considering both school resources and achievement outcomes. *American Educational Research Journal, 41*(4), 797-832.

Murnane, R. J., Willett, J. B., & Levy, F. (1995). The Growing Importance of Cognitive Skills in Wage Determination. *Review of Economics and Statistics, 78*(2), 251-266.

National Center for Education Statistics. (2013). The Nation's Report Card: Mega states: An analysis of student performance in the five most heavily populated states in the nation (NCES 2013 450). Washington, DC: Institute of Education Sciences, U.S. Department of Education.

Neal, D. A. (2005). *Why has Black-White skill convergence stopped?* manuscript. University of Chicago.

Neal, D. A. (2006). Why has Black-White skill convergence stopped? In E. A. Hanushek & F. Welch (Eds.), *Handbook of the Economics of Education* (Vol. 1, pp. 511-576). New York: Elsevier.

Neal, D. A., & Johnson, W. R. (1996). The role of premarket factors in black-white wage differences. *The Journal of Political Economy, 104*(5), 869-895.

Posselt, J., Jaquette, O., Bastedo, M., & Bielby, R. (2010). *Access without equity: Longitudinal analyses of institutional stratification by race and ethnicity, 1972-2004*. Paper presented at the Annual meeting of the Association for the Study of Higher Education, Indianapolis, IN.

Raudenbush, S. W., Bryk, A. S., & Congdon, R. (2012). HLM 7 for Windows. Skokie, IL: Scientific Software International, Inc.

Reardon, S. F. (2008). Thirteen Ways of Looking at the Black-White Test Score Gap *IREPP Working Paper*. Stanford, CA: Working Paper Series, Institute for Research on Educational Policy and Practice, Stanford University.

Reardon, S. F., & Ho, A. D. (2013). *Addressing measurement error and sampling variability in nonparametric gap estimation*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, CA.

Reardon, S. F., Kalogrides, D., Valentino, R. A., Shores, K. A., & Greenberg, E. (2013). *Patterns, trends, and between-state variation in racial academic achievement gaps* Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA.

Reardon, S. F., & Robinson, J. P. (2007). Patterns and trends in racial/ethnic and socioeconomic academic achievement gaps. In H. F. Ladd & E. B. Fiske (Eds.), *Handbook of Research in Education Finance and Policy* (pp. 497-516). New York, NY: Routledge.

Rothstein, R. (2004). A wider lens on the black-white achievement gap. *Phi Delta Kappan, October*, 104-110.

Shores, K. A., Valentino, R. A., & Reardon, S. F. (2013). *Trends in nonparametric achievement gaps in the NCLB era*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, CA.

Sunderman, G. L. (2006). The unraveling of No Child Left Behind: How negotiated changes transform the law. Cambridge, MA: The Civil Rights Project at Harvard University.

Vanneman, A., Hamilton, L., Baldwin Anderson, J., & Rahman, T. (2009). Achievement Gaps: How Black and White Students in Public Schools Perform in Mathematics and Reading on the National Assessment of Educational Progress. Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.

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

Wong, M., Cook, T. D., & Steiner, P. M. (2011). 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 (WP-09-11) *Institute for Policy Research Working Paper Series*. Evanston, IL: Northwestern University.

Table 1: Number of Achievement Gap Estimates, by Year, Grade, and Data Source

	Year (Spring of Grade)																
Grade	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	Total
State Data																	
2								4	8	8	12	12	12	12	8		76
3				3	8	21	46	56	79	115	170	186	197	199	198	192	1470
4		2	2	28	39	45	54	62	100	123	171	186	194	199	124	192	1521
5				3	10	25	49	58	79	117	174	189	197	200	198	192	1491
6				8	12	28	36	42	65	86	172	188	198	199	198	192	1424
7				6	6	14	26	28	60	102	176	189	198	198	198	192	1393
8					37	42	68	81	112	139	172	188	198	199	198	192	1626
Total		2	2	48	112	175	279	331	503	690	1047	1138	1194	1206	1122	1152	9001
NAEP Data																	
4	86		80		82		93	200		200		200		200		183	1324
8					80		93	200		200		200		200		177	1150
Total	86		80		162		186	400		400		400		400		360	2474

Note: Cell counts indicate the total number of state achievement gap estimates in the analytic sample. Counts include gaps in up to two subjects (math and reading) and for up to two groups (white-black and white-Hispanic gaps) for each state. Dashed lines distinguish cohorts who entered kindergarten before Fall 1994, between Fall 1994 and Fall 2001, and Fall 2002 or later.

Table 2: Estimated Assocation of White-Black and White-Hispanic Achievement Gaps With Years of Exposure to NCLB

	White-Black Gaps			White-Hispanic Gaps		
	Pooled Subjects	Math Only	Reading Only	Pooled Subjects	Math Only	Reading Only
All Data (Distributional Gap)						
Base Model	-0.005	-0.006	-0.004	0.005	0.006	0.003
	(0.003)	(0.004)	(0.003)	(0.004)	(0.004)	(0.004)
With Covariates	-0.004	-0.005	-0.002	0.005	0.006	0.003
	(0.004)	(0.004)	(0.003)	(0.004)	(0.004)	(0.004)
NAEP Data (Distributional Gap)						
Base Model	0.007 +	0.003	0.014 *	0.007 +	0.004	0.01 +
	(0.004)	(0.005)	(0.006)	(0.004)	(0.006)	(0.006)
With Covariates	0.009 *	0.003	0.014 *	0.008 +	0.005	0.011 *
	(0.004)	(0.005)	(0.007)	(0.004)	(0.006)	(0.005)
State Data (Distributional Gap)						
Base Model	-0.002	-0.004	-0.004	0.005	0.008 *	0.001
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
With Covariates	-0.002	-0.003	-0.002	0.005	0.008 +	0.002
	(0.004)	(0.004)	(0.003)	(0.004)	(0.004)	(0.004)
NAEP Data (Average Score Gap)						
Base Model	0.001	-0.003	0.003	0.003	0	0.007
	(0.004)	(0.005)	(0.006)	(0.005)	(0.006)	(0.007)
With Covariates	0.002	-0.003	0.003	0.005	0.003	0.008
	(0.004)	(0.005)	(0.006)	(0.004)	(0.006)	(0.007)
State Data (Proficiency Gap)						
Base Model	-0.526 *	-0.659 *	-0.541 *	-0.283	-0.267	-0.377
	(0.218)	(0.281)	(0.229)	(0.223)	(0.259)	(0.230)
With Covariates	-0.513 *	-0.735 *	-0.556 *	-0.336	-0.386	-0.417 +
	(0.236)	(0.289)	(0.238)	(0.228)	(0.263)	(0.229)

Note: Each cell indicates the estimated annual effect of exposure to NCLB (the coefficient on the variable indicating the number of years of exposure to NCLB). Each coefficient is from a separate model. Robust standard errors are in parentheses. + p<.10; * p<.05; ** p<.01; *** p<.001.

Table 3: Estimated Association of White-Black and White-Hispanic Achievement Gaps With Years of NCLB Exposure and Its Interaction with Proportion of Black or Hispanic Students in Schools Subject to Accountability

	White-Black Gaps						White-Hispanic Gaps					
	Pooled Subjects		Math Only		Reading Only		Pooled Subjects		Math Only		Reading Only	
All Data (Distributional Gap)												
Exposure	0.014	*	0.02	**	0.013	*	0.014	**	0.011	+	0.012	*
	(0.007)		(0.007)		(0.006)		(0.005)		(0.006)		(0.005)	
Exposure*Proportion Accountable	-0.031	**	-0.043	***	-0.026	**	-0.018	**	-0.011		-0.017	**
	(0.010)		(0.009)		(0.008)		(0.006)		(0.009)		(0.006)	
NAEP data (Distributional Gap)												
Exposure	0.024	**	0.02	*	0.03	*	0.008		0.006		0.011	
	(0.009)		(0.009)		(0.013)		(0.006)		(0.009)		(0.008)	
Exposure*Proportion Accountable	-0.026	+	-0.028	+	-0.027		-0.001		-0.001		0.001	
	(0.013)		(0.014)		(0.018)		(0.008)		(0.011)		(0.010)	
State Data (Distributional Gap)												
Exposure	0.014	*	0.016	+	0.015	*	0.012	*	0.014	**	0.006	
	(0.006)		(0.008)		(0.006)		(0.005)		(0.005)		(0.006)	
Exposure*Proportion Accountable	-0.026	**	-0.032	**	-0.028	**	-0.013	*	-0.012	+	-0.009	
	(0.009)		(0.010)		(0.009)		(0.006)		(0.007)		(0.007)	
NAEP Data (Average Score Gap)												
Exposure	0.017	*	0.017	*	0.018		0.006		0.001		0.012	
	(0.008)		(0.008)		(0.012)		(0.006)		(0.008)		(0.009)	
Exposure*Proportion Accountable	-0.025	*	-0.031	**	-0.024		-0.001		0.004		-0.006	
	(0.012)		(0.012)		(0.015)		(0.008)		(0.011)		(0.010)	
State Data (Proficiency Gap)												
Exposure	0.552		0.998	*	0.026		0.364		0.226		0.612	*
	(0.419)		(0.424)		(0.532)		(0.287)		(0.380)		(0.284)	
Exposure*Proportion Accountable	-1.858	**	-2.845	***	-1.045		-1.573	***	-1.346	*	-2.296	***
	(0.657)		(0.632)		(0.773)		(0.463)		(0.613)		(0.451)	

Note: All models include controls for grade, cohort, and time-varying economic and school composition and segregation covariates. Robust standard errors are in parentheses. + $p<.10$; * $p<.05$; ** $p<.01$; *** $p<.001$.

Table 4: Estimated Association of Achievement Gaps with Years of NCLB Exposure and Its Interaction with Various Treatment Typologies

	Math				Reading			
	White-Black		White-Hispanic		White-Black		White-Hispanic	
Exposure*Proportion Accountable	-0.043	***	-0.011		-0.026	**	-0.017	**
	(0.009)		(0.009)		(0.008)		(0.006)	
Exposure*Consequential Accountability After NCLB	0.0019		0.0053	*	-0.0023		0.0014	
(Dee and Jacob)	(0.004)		(0.003)		(0.003)		(0.003)	
Exposure * Low Standards Under NCLB	-0.0007		-0.0017		0.0023		-0.0059	
(Wong, Cook, Steiner)	(0.004)		(0.004)		(0.004)		(0.004)	
Exposure * High Standards Under NCLB	0.0034		0.0046		-0.0001		-0.0018	
(Wong, Cook, Steiner)	(0.004)		(0.005)		(0.004)		(0.004)	
Exposure*Performance Indexing	-0.0053		-0.0014		-0.0058		0.0025	
(Wei)	(0.004)		(0.005)		(0.004)		(0.005)	
Exposure*Minimum Subgroup Size	-0.0001		-0.0001		0.0001		-0.0001	
(Wei)	(0.000)		(0.000)		(0.000)		(0.000)	
Exposure*Annual Measureable Objectives	0.0002		0.0007	*	-0.0004		0.0002	
(Wei)	(0.000)		(0.000)		(0.000)		(0.000)	
Exposure*NCLB Implementation (z-score)	-0.0023		-0.0019		-0.0005		-0.0036	*
(Lee and Reeves)	(0.003)		(0.002)		(0.002)		(0.002)	
Exposure*Rigor of Standards	0.0031		0.0043		0.0001		0.0011	
(Lee and Reeves)	(0.002)		(0.003)		(0.002)		(0.003)	
Exposure*Data Tracking Capacity	-0.001	+	-0.0003		-0.0012	+	-0.0006	
(Lee and Reeves)	(0.001)		(0.001)		(0.001)		(0.000)	
Exposure*Funding Capacity for SINI (in 1000s)	0		0		0		0	
(Lee and Reeves)	(0.000)		(0.000)		(0.000)		(0.000)	

Note: All models include controls for grade, cohort, and time-varying economic and school composition and segregation covariates. Robust standard errors are in parentheses. + p<.10; * p<.05; ** p<.01; *** p<.001.

Table 5: Pairwise Correlations, Proportion Accountable with Other Measures

		Proportion Black Students Accountable	Proportion Hispanic Students Accountable
(1)	Reporting Threshhold	-0.08	-0.14
(2)	Proportion White in the State	-0.40	-0.33
(3)	Proportion Black/Hispanic in the State	0.73	0.65
(4)	Proportion Whites Accountable	0.31	-0.10
(6)	Average School Enrollment in Grades 2-8	0.53	0.45
(7)	White-Black/-Hispanic Segregation	0.58	0.53
(8)	White-Black/-Hispanic NAEP Gap in 2003	0.75	0.69
(9)	State Mean NAEP Score in 2003	-0.23	0.06
(10)	NCLB Exposure Effect	-0.66	0.02

Table 6: Estimated Association of White-Black and White-Hispanic Achievement Gaps With Years of NCLB Exposure and Its Interaction with Proportion of Black or Hispanic Students in Schools Subject to Accountability

	White-Black Gaps						White-Hispanic Gaps					
	Pooled Subjects		Math Only		Reading Only		Pooled Subjects		Math Only		Reading Only	
All Data (Distributional Gap)												
Exposure	0.013	+	0.013		0.016	*	0.003		-0.001		0.005	
	(0.007)		(0.011)		(0.007)		(0.005)		(0.006)		(0.007)	
Exposure*Proportion Accountable	-0.028	*	-0.03		-0.031	*	0.006		0.015		-0.001	
	(0.013)		(0.019)		(0.012)		(0.009)		(0.010)		(0.013)	
Exposure*Average School Size	0		0		0	+	0	+	0		0	*
	(0.000)		(0.000)		(0.000)		(0.000)		(0.000)		(0.000)	
Exposure*Percent Black	-0.011		-0.007		-0.007		-0.006		-0.013		0.004	
	(0.012)		(0.017)		(0.012)		(0.013)		(0.013)		(0.016)	
Exposure*BW Segregation (H)	-0.026	*	-0.034	*	-0.023	+	-0.006		-0.013		-0.002	
	(0.013)		(0.016)		(0.014)		(0.011)		(0.013)		(0.013)	
Exposure*Size of Gap in 2003	-0.014		0.004		-0.034	**	0.019	+	0.023	**	0.012	
	(0.012)		(0.019)		(0.013)		(0.010)		(0.007)		(0.013)	
NAEP data (Distributional Gap)												
Exposure	0.005		-0.004		0.004		-0.003		-0.005		-0.001	
	(0.011)		(0.014)		(0.015)		(0.008)		(0.012)		(0.009)	
Exposure*Proportion Accountable	0.009		0.015		0.019		0.025	+	0.021		0.026	+
	(0.018)		(0.026)		(0.023)		(0.013)		(0.019)		(0.014)	
Exposure*Average School Size	0		0		0		0		0		0	
	(0.000)		(0.000)		(0.000)		(0.000)		(0.000)		(0.000)	
Exposure*Percent Black	-0.027		-0.02		-0.046	*	0.005		0.014		0.005	
	(0.017)		(0.027)		(0.020)		(0.013)		(0.021)		(0.013)	
Exposure*BW Segregation (H)	0.001		-0.002		0.003		-0.016		-0.039	*	-0.002	
	(0.016)		(0.018)		(0.018)		(0.015)		(0.017)		(0.020)	
Exposure*Size of Gap in 2003	0.029	+	0.044	+	0.03		0.033	***	0.023	+	0.037	*
	(0.017)		(0.023)		(0.019)		(0.010)		(0.013)		(0.015)	
State Data (Distributional Gap)												
Exposure	0.019	**	0.016		0.022	**	0.006		0.004		0.003	
	(0.007)		(0.010)		(0.008)		(0.006)		(0.006)		(0.008)	
Exposure*Proportion Accountable	-0.035	**	-0.032	*	-0.042	***	-0.001		0.01		-0.002	
	(0.011)		(0.016)		(0.012)		(0.010)		(0.008)		(0.015)	
Exposure*Average School Size	0		0		0	*	0	*	0	+	0	*
	(0.000)		(0.000)		(0.000)		(0.000)		(0.000)		(0.000)	
Exposure*Percent Black	-0.001		-0.003		0.004		0.009		-0.01		0.017	
	(0.010)		(0.015)		(0.013)		(0.017)		(0.016)		(0.022)	
Exposure*BW Segregation (H)	-0.026	*	-0.041	**	-0.023	+	-0.002		-0.016		0.006	
	(0.011)		(0.016)		(0.014)		(0.014)		(0.015)		(0.017)	
Exposure*Size of Gap in 2003	-0.034	***	-0.021		-0.045	***	0.016		0.015	+	0.011	
	(0.009)		(0.015)		(0.013)		(0.010)		(0.008)		(0.013)	
NAEP Data (Average Score Gap)												
Exposure	0.005		0.002		0.001		-0.002		-0.006		0.004	
	(0.010)		(0.013)		(0.014)		(0.006)		(0.011)		(0.009)	
Exposure*Proportion Accountable	-0.003		-0.004		0.007		0.017	+	0.02		0.009	
	(0.015)		(0.023)		(0.021)		(0.010)		(0.018)		(0.013)	
Exposure*Average School Size	0		0		0		0		0		0	
	(0.000)		(0.000)		(0.000)		(0.000)		(0.000)		(0.000)	
Exposure*Percent Black	-0.021		-0.009		-0.031	+	0.014		0.031	+	0.006	
	(0.015)		(0.025)		(0.019)		(0.010)		(0.018)		(0.013)	
Exposure*BW Segregation (H)	0.009		0.001		0.004		-0.016		-0.028	+	-0.009	
	(0.014)		(0.018)		(0.015)		(0.014)		(0.017)		(0.020)	
Exposure*Size of Gap in 2003	0.021		0.032		0.022		0.032	***	0.034	*	0.021	
	(0.016)		(0.021)		(0.017)		(0.010)		(0.014)		(0.013)	
State Data (Proficiency Gap)												
Exposure	-0.491		-0.755		-0.396		-0.534		-0.985	+	-0.066	
	(0.493)		(0.533)		(0.628)		(0.413)		(0.505)		(0.438)	
Exposure*Proportion Accountable	-0.063		0.208		-0.303		0.2		1.062		-0.887	
	(0.871)		(0.977)		(0.992)		(0.685)		(0.792)		(0.859)	
Exposure*Average School Size	-0.002	*	-0.001		-0.002		-0.001		-0.001		-0.001	
	(0.001)		(0.001)		(0.001)		(0.001)		(0.001)		(0.001)	
Exposure*Percent Black	-1.602		-2.41	*	-0.909		0.073		0.047		-0.047	
	(1.207)		(1.223)		(1.160)		(1.340)		(1.436)		(1.580)	
Exposure*BW Segregation (H)	-3.93	**	-3.103	*	-3.493	**	-1.933	*	-2.997	**	-0.774	
	(1.314)		(1.403)		(1.309)		(0.822)		(1.091)		(0.915)	
Exposure*Size of Gap in 2003	-0.914		0.735		-1.913	*	1.773	***	2.353	***	1.44	*
	(0.943)		(1.037)		(0.848)		(0.413)		(0.456)		(0.643)	

Note: All models include controls for grade, cohort, and time-varying economic and school composition and segregation covariates. Robust standard errors are in parentheses. + $p<.10$; * $p<.05$; ** $p<.01$; *** $p<.001$.

Table 7: Within-State Association Between Estimated Annual NCLB Effect and Proportion of Minority Students in Schools Meeting Minimum Subgroup Size

	Model 1	Model 2	Model 3	Model 4	Model 5
Intercept	-0.001	-0.001	0	0	0.001
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Within-State Variables					
Proportion Accountable		-0.012	-0.002	-0.001	-0.003
		(0.008)	(0.007)	(0.014)	(0.014)
Black Effect			-0.009 ***	-0.007 *	-0.007 **
			(0.002)	(0.003)	(0.003)
Segregation				-0.005	-0.004
				(0.023)	(0.023)
Proportion Enrollment				0.011	0.013
				(0.016)	(0.016)
Gap in 2003 (Averaged Across Grades)				0.013	0.012
				(0.012)	(0.012)
Between-State Variables					
Mean Proportion Accountable					-0.003
					(0.014)
Mean Segregation					-0.008
					(0.013)
Mean Proportion Enrollment					-0.055 **
					(0.019)
Mean Gap in 2003					-0.002
					(0.011)
Within-State Residual SD	0.0078	0.0076	0.0046	0.0044	0.0039
Between-State Residual SD	0.0053	0.0051	0.0068	0.0069	0.0054

Note: All within-state variables are centered around their state means; all between-state variables are centered around their population mean. Sample includes 49 states (ND omitted).

Figure 1: Distribution of Proportions of Black and Hispanic Students in Schools Meeting Minimum Subgroup Reporting Size

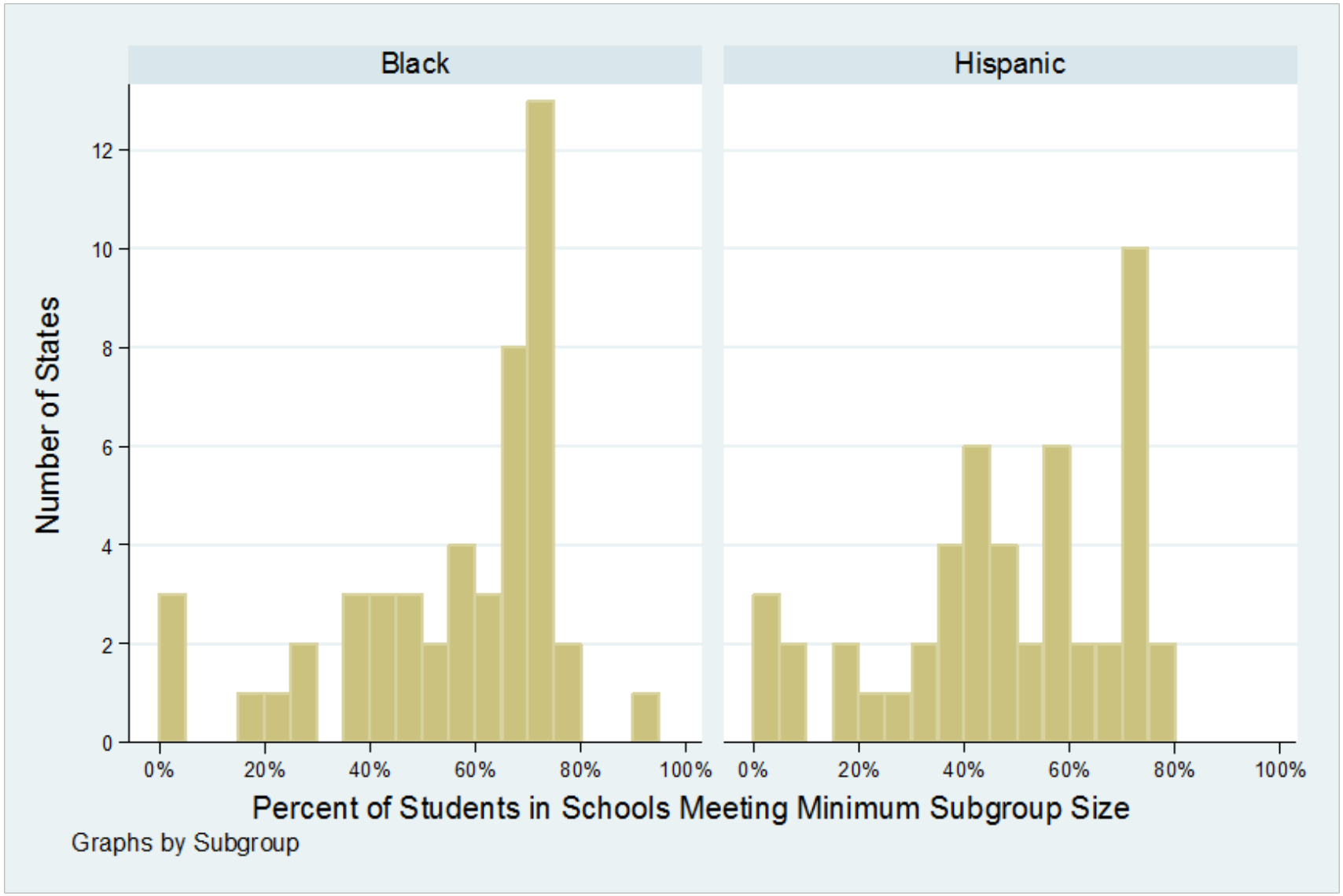


Figure 2: White-Black Achievement Gap Trends, Math and Reading, 1991-2006 Cohorts

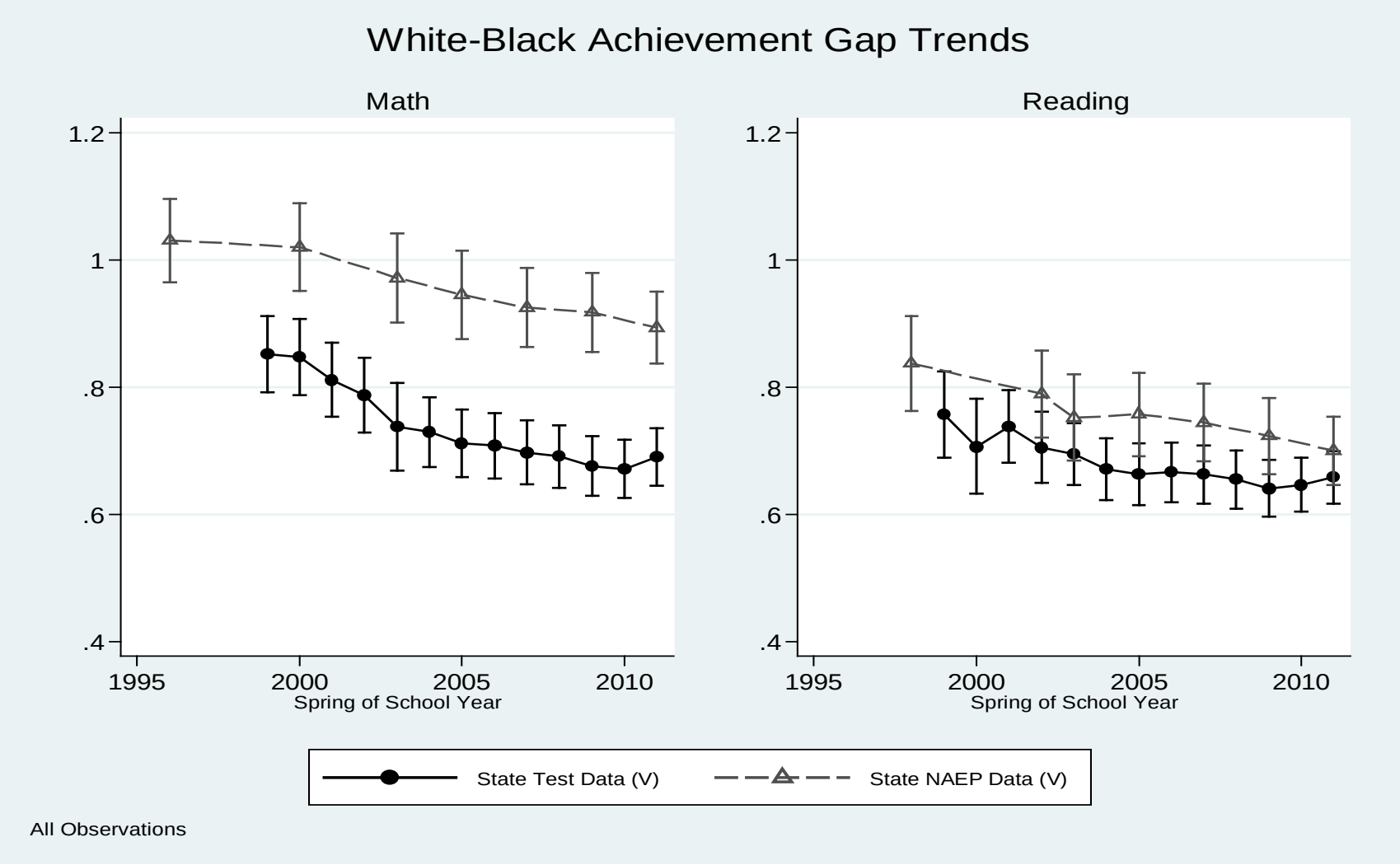


Figure 3: White-Hispanic Achievement Gap Trends, Math and Reading, 1991-2006 Cohorts

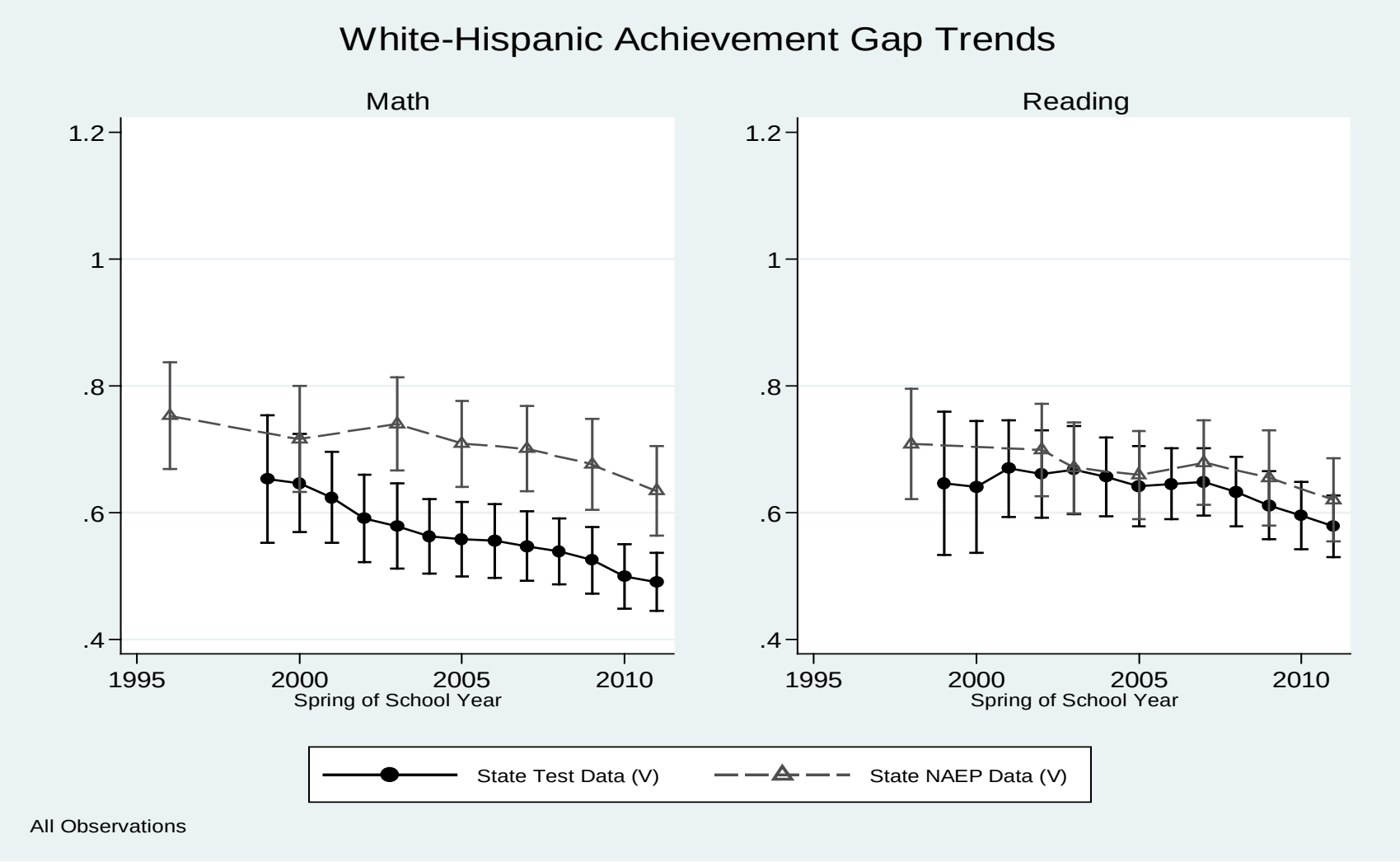


Figure 4: Estimated Annual Effect of NCLB on Achievement Gaps, by State (Empirical Bayes Estimates)

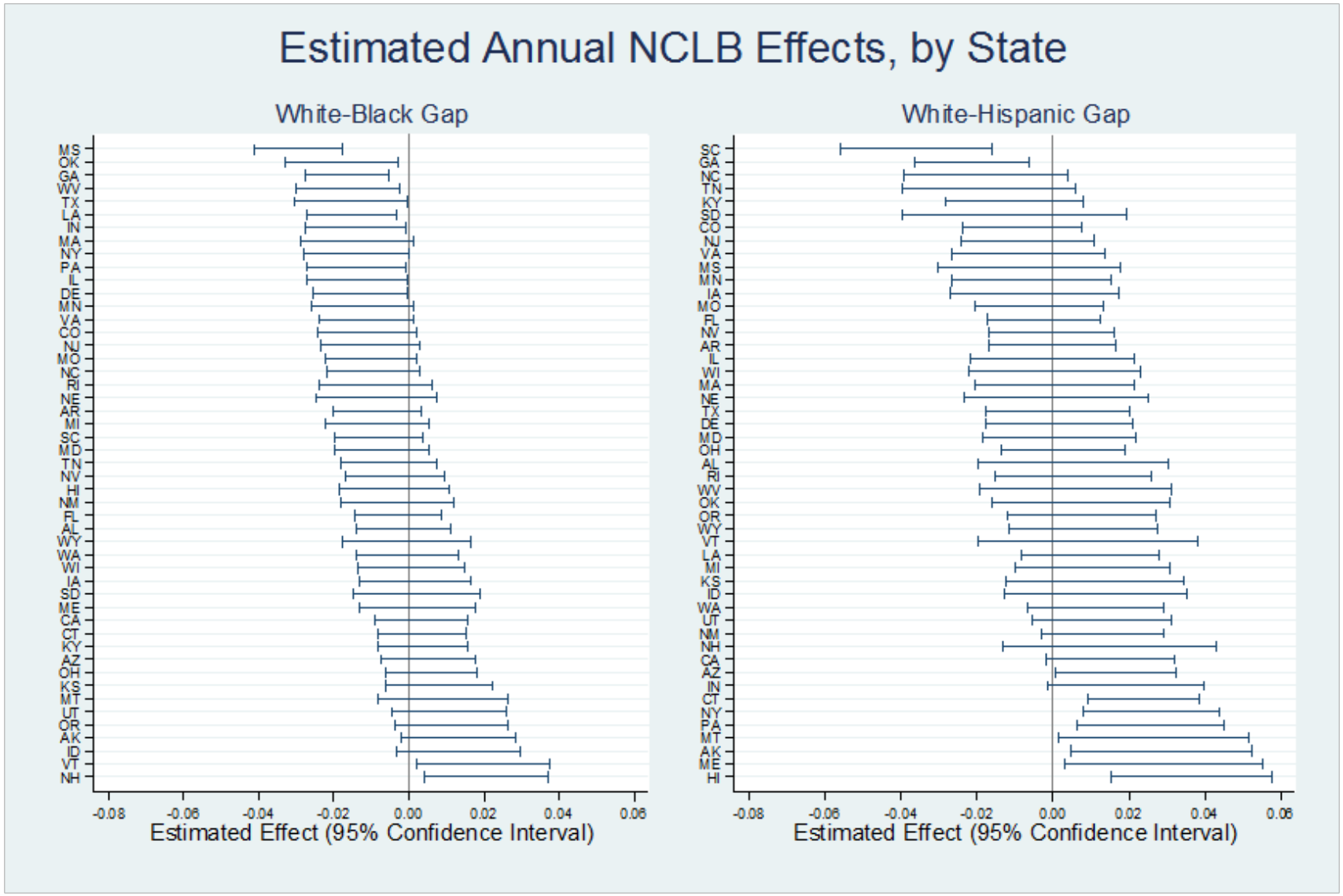


Figure 5: Estimated State-Specific NCLB Annual Effect on White-Black Achievement Gap, by Proportion of Black Students in Schools Meeting State Minimum Subgroup Size Threshold
Estimates from data pooled across test subjects, data sources, and all cohorts/years

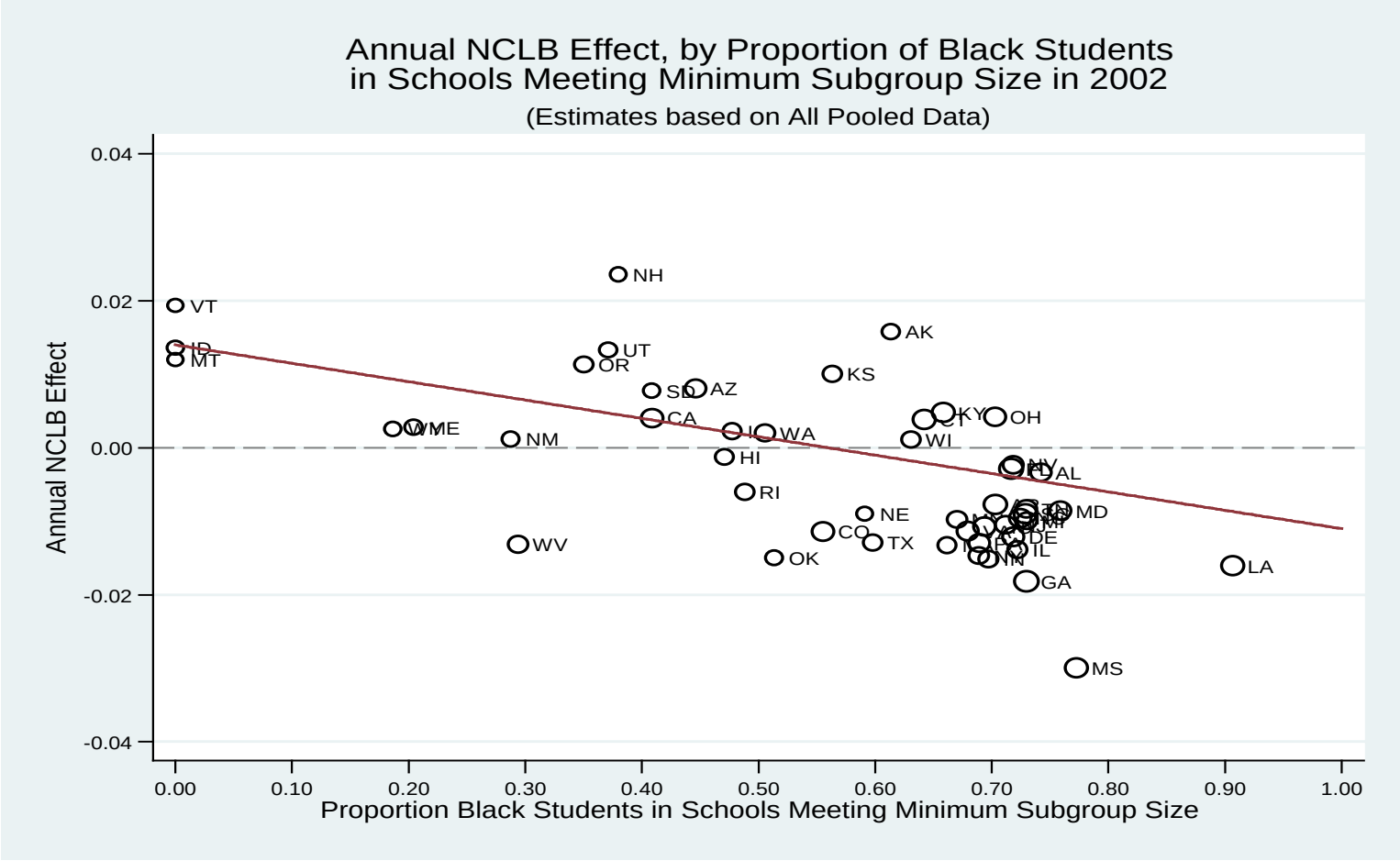
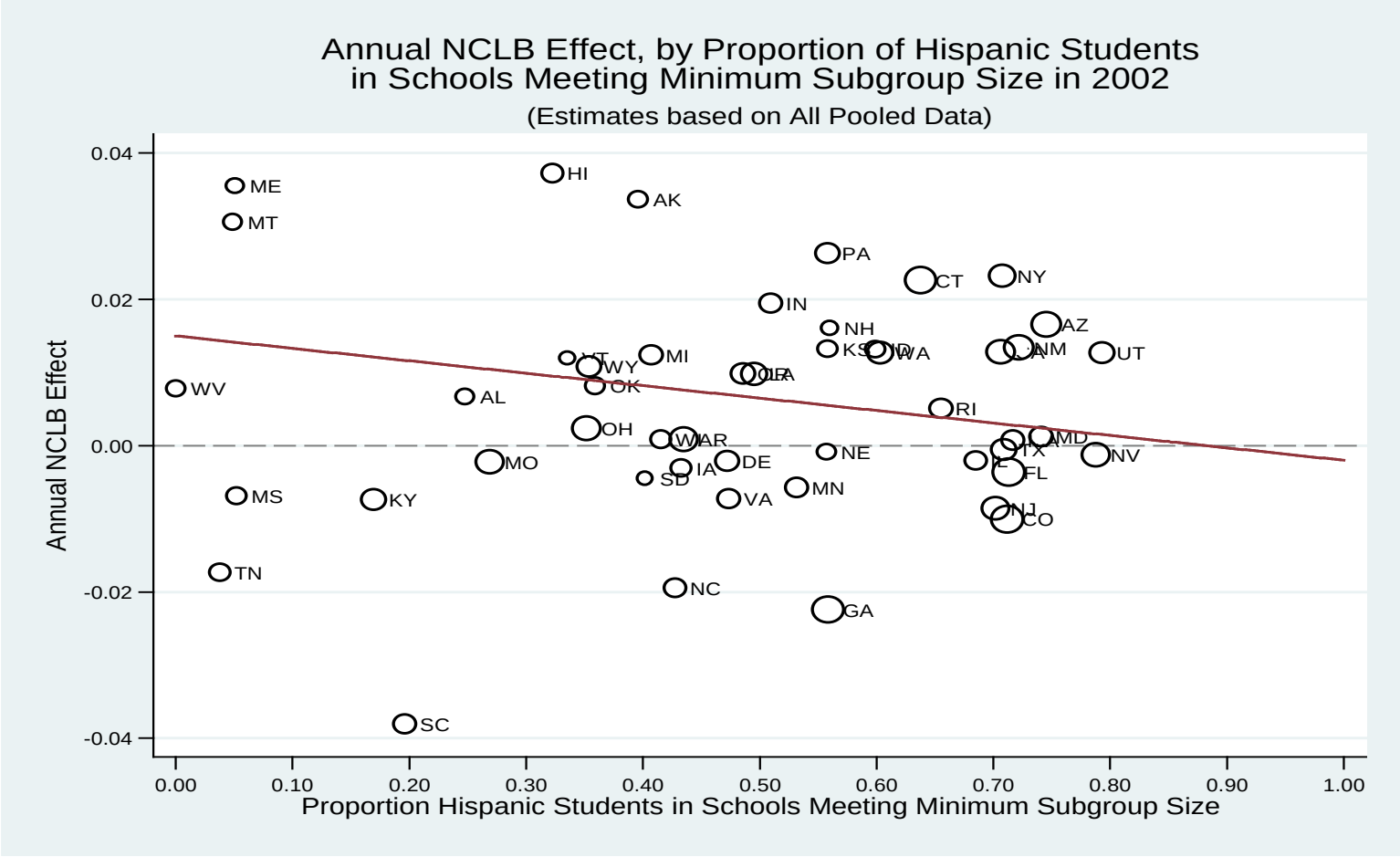


Figure 6: Estimated State-Specific NCLB Annual Effect on White-Hispanic Achievement Gap, by Proportion of Hispanic Students in Schools Meeting State Minimum Subgroup Size Threshold
Estimates from data pooled across test subjects, data sources, and all cohorts/years



Appendix A: Modeling the Effect of NCLB

Notation

We begin by defining some notation. Each of our observations pertains to an achievement gap in a particular grade (indexed by g, where $g = 0$ for kindergarten; $g = 1$ for first grade, and so on) and state (indexed by s) for a particular cohort of students (indexed by c). We denote cohorts of students by the calendar year in which they entered kindergarten; for example, a 6th grade observation in Spring 2008 pertains to the 2001 cohort of students (students who entered kindergarten in Fall 2001). Let coh_c, gr_g, and yr_{cg} denote the cohort, grade, and spring calendar year, respectively, of an observation in cohort c and grade g. We center yr_{cg} and coh_c at 2002 in all our models, defining $yr^*_{cg} = yr_{cg} - 2002$ and $coh^*_c = coh_c - 2002$ (so $yr^*_{cg} > 0$ for observations made during the NCLB era—in Spring 2003 or later; and $coh^*_c = 0$ for the first cohort who entered kindergarten during the NCLB era). We define $gr_g = g + 1$, so that $gr_0 = 1$ (i.e., gr_g indicates the number of years a cohort has been in school by the spring of grade g). Note that

$$yr^*_{cg} = coh^*_c + gr_g. \qquad [A1]$$

A Model for the Development of Achievement Gaps

Now let G_{csg} be the achievement gap in the spring of grade g for students in cohort c in state s (in this notation, G_{cs0} is the gap for cohort c in the spring of their kindergarten year, and $G_{cs(-1)}$ is the gap when these children entered kindergarten). We can express the initial achievement gap at kindergarten entry (more specifically, in the spring before they enter kindergarten) in state s for cohort c as a state-specific function f_s of the cohort, plus some linear function of a vector cohort-by-state covariates ($\mathbf{X}_{cs}$, which includes, in our models, the average white-black [or white-Hispanic] income, poverty, and unemployment ratios in state s during the

pre-kindergarten years of cohort c), plus some mean-zero error term, ν_{cs}:

$$G_{cs(-1)} = \lambda_s + f_s(coh_c^*) + \mathbf{X}_{cs}\mathbf{A} + \nu_{cs}.$$

[A2]

Here λ_s is the size of the achievement gap prior to kindergarten entry (after adjusting for $\mathbf{X}_{cs}$) for the cohort that entered kindergarten in Fall 2002 (the first cohort who entered school when NCLB was in effect) in state s, and f_s describes the shape of the trend in the size of this pre-kindergarten gap in state s. Note that we do not include an NCLB-effect parameter in Equation (2) because we do not expect NCLB to affect pre-kindergarten academic achievement gaps.

We can express the gap in later grades as the sum of the same cohort's gap in the prior grade/year plus some cohort-state-grade-specific change, δ_{csg}:

$$G_{csg} = G_{cs(g-1)} + \delta_{csg}.$$

[A3]

Now we can write the change in the gap during grade g for cohort c as a function of a state fixed effect (v_s), a linear cohort effect (β), a linear grade effect (η), an effect of some vector of covariates $\mathbf{w}_{csg}$, a state-specific effect of the presence of NCLB (δ_s), and a mean-zero error term (e_{csg}):

$$\delta_{csg} = \alpha + v_s + \beta(coh_c^*) + \eta(g) + \delta_s T_{cg} + \mathbf{w}_{csg}\mathbf{B} + e_{csg},$$

[A4]

where T_{cg} indicates the presence of NCLB in the year in which cohort c completed grade g; that is $T_{cg} = 1$ if $yr_{cg}^* > 0$ and $T_{cg} = 0$ otherwise. Note that this model assumes that the effect of NCLB on achievement gaps is constant across years and grades (but not necessarily across states). A more flexible model that lets the effect of NCLB vary across years and grades would be

$$\delta_{csg} = \alpha + v_s + \beta(coh_c^*) + \eta(g) + \delta_{0s} T_{cg} + \delta_{1s}\left(T_{cg} \cdot (yr_{cg} - 2006)\right) + \delta_{2s}\left(T_{cg} \cdot (g-4)\right) + \mathbf{w}_{csg}\mathbf{B} + e_{csg}.$$

[A5]

Here δ_{0s} is the NCLB effect on the gap during 4th grade in 2006 in state s, δ_{1s} is the linear trend in the effect of NCLB across years in state s, and δ_2 is the linear trend in the effect of NCLB across grades in state s.[13]

Now it is useful to define several cumulative variables. First, we define exp_{cg} as the number of years a cohort c has been exposed to NCLB by the time it reaches spring of grade g. That is, $exp_{cg} = \sum_{k=0}^{g} T_{ck}$. Second, we define $E_g = \sum_{k=0}^{g} k = \frac{1}{2}(g^2 + g) = \frac{1}{2}\left(gr_g^2 - gr_g\right)$. Third, we define $expgr_{cg} = \sum_{k=0}^{g}\left(T_{ck} \cdot (k-4)\right)$. And fourth, we define $expyr_{cg} = \sum_{k=0}^{g}\left(T_{ck} \cdot (yr_{ck} - 2006)\right)$. And fifth, we define $\mathbf{W}_{csg}$ as the cumulative exposure vector of cohort c in state s to the covariate vector $\mathbf{w}$ from kindergarten through grade g. That is, $\mathbf{W}_{csg} = \sum_{k=0}^{g} \mathbf{w}_{csk}$. These cumulative variables will play a role in our model below.

Now, substituting [A5] and [A2] into [A3], we have

$$
\begin{aligned}
G_{csg} &= G_{cs(-1)} + \sum_{k=0}^{g} \delta_{csk} \\
&= [\lambda_s + f_s(coh_c^*) + \mathbf{X}_{cs}\mathbf{A} + \nu_{cs}] \\
&\qquad + \sum_{k=0}^{g} \left[\alpha + v_s + \beta(coh_c^*) + \eta(k) + \delta_{0s} T_{ck} + \delta_{1s}\left(T_{ck} \cdot (yr_{ck} - 2006)\right)\right. \\
&\qquad \left. + \delta_{2s}\left(T_{ck} \cdot (k-4)\right) + \mathbf{w}_{csk}\mathbf{B} + e_{csk}\right] \\
&= [\lambda_s + f_s(coh_c^*) + \mathbf{X}_{cs}\mathbf{A} + \nu_{cs}] + (g+1)\left(\alpha + v_s + \beta(coh_c^*)\right) + \eta\left(E_g\right) + \delta_{0s}\left(exp_{cg}\right) \\
&\qquad + \delta_{1s}\left(expyr_{cg}\right) + \delta_{2s}\left(expgr_{cg}\right) + \mathbf{W}_{csg}\mathbf{B} + \sum_{k=0}^{g} e_{csk} \\
&= \lambda_s + f_s(coh_c^*) + \alpha_s\left(gr_g\right) + \beta\left(gr_g \cdot coh_c^*\right) + \eta\left(E_g\right) + \delta_{0s}\left(exp_{cg}\right) + \delta_{1s}\left(expyr_{cg}\right) \\
&\qquad + \delta_{2s}\left(expgr_{cg}\right) + \mathbf{X}_{cs}\mathbf{A} + \mathbf{W}_{csg}\mathbf{B} + e'_{csg}
\end{aligned}
$$

[A6]

[13] We center yr_{cg} on 2006 and g on 4th grade so that the coefficient δ_{0s} is the annual effect of NCLB at a time and grade that corresponds to meaningful time/grade. In addition, centering yr and g improves the precision with which δ_0 can be estimated.

where $\alpha_s = \alpha + v_s$; and $e'_{csg} = \nu_{cs} + \sum_{k=0}^{g} e_{csk}$. Equation [A6] implies that we can estimate δ_{0s}, δ_{1s}, and δ_{2s} by using a random coefficients model to regress G_{csg} on coh^*, gr, $gr \cdot coh^*$, E, $\mathbf{X}$, $\mathbf{W}$, exp, $expyr$ and $expgr$:

$$\hat{G}_{csg} = (\lambda + u_{\lambda s}) + (\gamma + u_{\gamma s})(coh_c^*) + (\alpha + u_{\alpha s})(gr_g) + \beta(gr_g \cdot coh_c^*) + \eta(E_g) + \mathbf{X}_{cs}\mathbf{A} + \mathbf{W}_{csg}\mathbf{B}$$
$$+ (\delta_0 + u_{\delta 0s})(exp_{cg}) + (\delta_1 + u_{\delta 1s})(expyr_{cg}) + (\delta_2 + u_{\delta 2s})(expgr_{cg}) + e'_{csg} + \epsilon_{csg}$$
$$e'_{csg} \sim N[0, \sigma^2]$$
$$\epsilon_{csg} \sim N[0, \omega_{csg}^2] = N[0, var(\hat{G}_{csg})]$$
$$\begin{bmatrix} u_{\lambda s} \\ u_{\gamma s} \\ u_{\alpha s} \\ u_{\delta 0s} \\ u_{\delta 1s} \\ u_{\delta 2s} \end{bmatrix} \sim N[\mathbf{0}, \boldsymbol{\tau}] .$$

[A7]

Here $\hat{G}_{csgt}$ is the estimated achievement gap in state s in subject t for cohort c in grade g; sub_{csgt} is a dummy variable indicating whether $\hat{G}_{csgt}$ is a math or reading gap; λ is the average pre-kindergarten achievement gap across states for the cohort entering kindergarten in 2002; γ is the average cohort trend in pre-kindergarten achievement gaps across states, α' is the average grade-to-grade change in the achievement gap across states in the absence of NCLB, ζ is the average difference between achievement gaps in math and reading; and δ_1, δ_2, and δ_3 are the key parameters of interest, describing the average annual effect of NCLB on the achievement gap and its relationship to time and grade. The error term ϵ_{csgt} is the sampling error of $\hat{G}_{csgt}$; we set its variance ω_{csgt}^2 to be equal to the square of the standard error of $\hat{G}_{csgt}$. We estimate the parameters of this model, as well as σ^2 and the unconstrained variance-covariance matrix $\boldsymbol{\tau}$, using the HLM v7 software (Raudenbush et al., 2012).

Understanding the Source of Identification of the NCLB Effect

The estimated coefficient δ indicates the average annual effect of NCLB on the achievement gap within a cohort. To understand the variation in the data that identifies this parameter, it is useful to note that, if we define a variable N_c such that $N_c = 1$ if $coh_c^* > 0$ and $N_c = 0$ otherwise, then we can write exp_{csg} as:

$$\begin{aligned} exp_{cg} &= \sum_{k=0}^{g} T_{ck} \\ &= T_{cg} \cdot yr_{cg}^* - N_c \cdot coh_c^* \\ &= (T_{cg} - N_c) coh_c^* + T_{cg} \cdot gr_g. \end{aligned}$$

[A8]

Figure A1 below helps to visualize the relationship between cohort, grade, and exposure:

Figure A1: Exposure to NCLB, by cohort and grade

		Cohort (Fall of Kindergarten Entry Year)																				
Grade	....	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008
K	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	2	2	2	2	2	2	2
2	0	0	0	0	0	0	0	0	0	0	0	0	0	1	2	3	3	3	3	3	3	3
3	0	0	0	0	0	0	0	0	0	0	0	0	1	2	3	4	4	4	4	4	4	4
4	0	0	0	0	0	0	0	0	0	0	0	1	2	3	4	5	5	5	5	5	5	5
5	0	0	0	0	0	0	0	0	0	0	1	2	3	4	5	6	6	6	6	6	6	6
6	0	0	0	0	0	0	0	0	0	1	2	3	4	5	6	7	7	7	7	7	7	7
7	0	0	0	0	0	0	0	0	1	2	3	4	5	6	7	8	8	8	8	8	8	8
8	0	0	0	0	0	0	0	1	2	3	4	5	6	7	8	9	9	9	9	9	9	9

Pre-2003 kindergarten cohort; not subject to NCLB in current year
Pre-2003 kindergarten cohort; subject to NCLB in current year
Post-2002 kindergarten cohort; subject to NCLB in current year

Now, to understand the variation in exp_{cg} that is used to identify δ, it is useful to take the partial derivative of Equation [A7] with respect to coh^* (holding grade constant):

$$\frac{\partial G}{\partial coh^*} = \begin{cases} \gamma_s + \beta \cdot gr & if \quad T = 0, N = 0 \\ \gamma_s + \beta \cdot gr + \delta(1 + 2coh^*) & if \quad T = 1, N = 0 \\ \gamma_s + \beta \cdot gr & if \quad T = 1, N = 1 \end{cases}$$

[A9]

Similarly, the partial derivative with respect to gr (holding cohort constant) is

$$\frac{\partial G}{\partial gr} = \begin{cases} \alpha_s + \beta coh^* + \eta(2gr + 1) & if \quad T = 0 \\ \alpha_s + \beta coh^* + \eta(2gr + 1) + \delta & if \quad T = 1 \end{cases}$$

[A10]

These expressions make clear that the model relies on two distinct sources of variation in exp_{cg} to identify the NCLB effect δ. First, for cohorts entering kindergarten prior to 2003 (for whom $N = 0$), $exp_{cg} = 0$ prior to 2003, and then increases linearly across grades (within a cohort) or across cohorts (within a grade) after 2002. Using this variation, δ is the difference in the grade slope ($\partial G/\partial gr$) within a cohort before and after 2002; equivalently, δ is the difference in the cohort slope ($\partial G/\partial coh^*$) within a grade before and after 2002. Note that if we limit the sample to observations from the pre-2003 cohorts, Model [A7] is very similar to an interrupted time series model. If we drop the E_g and $gr_g \cdot coh_c^*$ variables, [A7] is mathematically identical to an interrupted time series model.

Second, for years after 2002 (when $T = 1$), $exp_{cg} = coh_c^* + gr_g$ for cohorts entering kindergarten prior to 2003 (for whom $N = 0$), but $exp_{cg} = gr_g$ for later cohorts (for whom $N = 1$). Using this variation, δ is the difference in the cohort slope ($\partial G/\partial coh^*$) within a grade between pre-2003 cohorts and later cohorts.

In Figure A1, the first source of variation is represented by the transition from yellow to green shading; the second source of variation is represented by the transition from green to blue shading. To the extent that we have observations in the yellow and green regions, we can use the first source of variation to estimate δ; if we have observations in the green and blue regions, we can use the second source of variation.

A Note on Covariates

As noted above, we construct these covariates using data from two main sources: the Current Population Survey and the Common Core of Data. The CPS gives us information about the relative economic position of minorities and whites, while the CCD gives us information about the racial composition and segregation levels of schools. The CPS measures are constructed as follows:

we restrict the CPS data to records pertaining to children ages 0-14 years old. We then collapse the person-level data by state, cohort, age, and race, using the sampling weights to make the figures representative of the average child's household in each cell. We then construct the ratio of black/Hispanic household income to white household income, the ratio of black/Hispanic household poverty rates to white household poverty rates, and the ratio of black/Hispanic unemployment rates to white unemployment rates. Exposure to unemployment is measured by whether any adult in the child's household is unemployed. We also include a measure of the difference in years of schooling completed among blacks/Hispanics and whites, using the highest level of schooling completed by adults in the household. We use these data to construct two sets of measures. First, we construct measures denoted by $\mathbf{X}_{cs}$ in the equations above. This vector is constructed by taking the average of the CPS measures from birth to age 5 within each state and cohort. It varies between states and cohorts but is constant within states and cohorts across grades. $\mathbf{X}_{cs}$ reflects the amount of racial economic inequality experienced by students in a given a given state and cohort in early childhood. Second, we construct measures denoted by $\mathbf{W}_{csg}$. This vector is constructed by taking the sum of the CPS measures, starting from age 6; it reflects the running sum of the ratios by state and cohort in each year. $\mathbf{W}_{csg}$ varies across grades for each state and cohort, reflecting changes in racial economic inequality experienced by students as they progress through school.

We use the CCD data to measure the proportion of public school students that are black or Hispanic as well as the between school racial segregation within states. The measure of segregation we use is the information theory index (H), which ranges from 0 (no segregation) to 1 (complete segregation). The $\mathbf{X}_{cs}$ vector described above is not relevant for the CCD measures since students are only exposed to the racial composition or segregation levels of their schools after age 5, once they enter school. Therefore, we only use the $\mathbf{W}_{csg}$ version of the measures for the racial composition of public schools and segregation levels, constructed as described above.

Appendix B: State Test Score Data Sources and Cleaning Procedures

State-level categorical proficiency data were collected from three different sources. The first source is from state departments of education websites. Many state departments of education make state-level data, disaggregated by subject, subgroup, and year, publically available in excel files online. We were able to collect data for 18 states through this method. These data included observations for at least one state (Colorado) as far back as 1997, and for about half of the states as early as 2004. After collecting these data, we were able to retrieve four years of data, spanning 2007 to 2010 for 49 of the 50 states from EDFacts. EDFacts is an initiative within the federal Department of Education designed to centralize proficiency data supplied from state education agencies (SEAs). Finally, we were able to retrieve data for all 50 states from the Center on Education Policy (CEP) website (http://www.cep-dc.org/). These data included observations for 6 states as far back as 1999, for 25 states as far back as 2002, and for the majority of states dating back to 2005. As necessary, we rely on the Common Core of Data (CCD) for accurate sample sizes in each state, grade, subgroup, and year when sample sizes were not reported.

We merged these three data sets to generate a master data set consisting of the maximal number of state by year by subgroup by subject observation points. We created a data-quality checking method to determine which data set would be the default if we had duplicate observations across the three sources. See Table 1 for the number of observations we have for each state by year.

Our rules for determining the default data set were as follows. First, for observations with just one data set, we conducted an internal quality check by summing percentages across categories. If the categories summed to an amount between 98% and 102% (to account for rounding errors), we considered these data to be good quality. We dropped observations that did not fit this criterion. When we had observations from more than one data source, we first did the

above check across each of the sources, and if one source summed to a percent between 98 and 102, but the other(s) did not, we retained the observation from the data source that met this criterion and dropped the observation(s) that did not.

When both (or perhaps all three) data sets had categories that summed to this acceptable range, and when all contained the same number of proficiency categories, we generated difference scores in the percent of students scoring proficient within a given category across data sets. When the absolute difference across the categories was less than 4%, we considered both data sources to have consistent and good quality data. This allowed for, on average, a 1% difference between two data sources in a given category, as most states provide data from four proficiency categories. When data did not meet this criterion across any two data set combinations, we computed *V* gap estimates for both data sources, and conducted *t*-tests to determine whether the generated gaps were significantly different across the two sources. If we failed to reject the null that there was no difference between the two computed gaps, we kept the observation for both data sets. Also, as a robustness check, we conducted the same *t*-test check even for those data sources that were off by no more than 4% across the categories. Finally, if data sets both had categories that summed to a range between 98% and 102%, but one data set had more categories available than the other, we kept the observation from the data set with more categories.

If data sources did not match (within an acceptable range of 4% across categories) and did not meet any of the other above mentioned quality checks, observations were dropped. In the end, we dropped a total of 5.4% of the total possible unique state by grade by year by subject observations. One percent of these observations were dropped because the data failed the t-test check, while the majority (4.4%) of the drops occurred because the proficiency categories did not sum to a reasonable range of 98% to 102% across all data sets available for the unique observation.

Our master data set, which was used for the analysis conducted for this study drew 78.9% of its data from CEP, 14.5% of its data from EDFacts, and 5.3% of its data from the data collected

from state department of education websites. In cases where we deemed CEP and at least one of the other two data sets to be accurate we used CEP data as our default for analysis purposes. When we had determined that EdFacts and state website data were both accurate, we used EDFacts data as our default source. The fact that such a large portion of our final data set was constructed from CEP data rather than one of the other sources is partially due to the fact that we chose it as a default when CEP and at least one other data set were found to provide valid data. We could just have easily selected one of the other data sets as our default.

Appendix C: Computation of the *V*-statistic

The estimation procedure used to compute V is described in detail by Ho and Reardon (2012). In particular, we employ the maximum likelihood method they describe (executed through the -rocfit- command in STATA). This method yields standard errors on each gap estimate, which are included as weights ($\frac{1}{se^2}$) in the precision-weighted coefficients models described above. Following Reardon and Ho (2013), we also ensure that each gap estimate is disattenuated for measurement error. The plausible value NAEP scores we use eliminate attenuation bias due to item-level measurement-error in the NAEP assessments (measurement error that arises from item-level unreliability). State test score data are not corrected for this kind of measurement error, however, so our gap estimates based on state data will be biased toward 0. We correct this bias by multiplying state data-derived achievement gaps by a factor of $1/\sqrt{r}$, where r is the reliability of the state test. We assume a reliability of 0.9 for all the state tests,[14] and multiply the estimated state test score gaps by a factor of $\frac{1}{\sqrt{0.9}} = 1.054$. We use these reliability-adjusted gap estimates in all of our analyses.

[14] Reardon and Ho (2013) gathered information on the reliability of state tests reading and math from 46 states. The reported reliabilities typically range from 0.85 to 0.95. In practice, the choice to adjust by 0.9 has little effect on our results: none of our findings differ substantively if we instead assume reliabilities of 0.8 or 1.0.

Finally, note that our estimates do not correct for attenuation bias due to test-retest unreliability in both tests. That is, both tests contain measurement error due to the fact that the same student will not perform identically well on the same test on two different days. This type of measurement error is common to both NAEP and state tests, and so does not affect our estimates based on NAEP and state data differently.