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

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

Stanford University

DRAFT: October 4, 2012

PRELIMINARY VERSION: PLEASE DO NOT CIRCULATE WITHOUT PERMISSION

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 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. Ten years have passed since NCLB went into effect. In this paper we investigate whether the Act has been successful at narrowing racial achievement gaps. We do so by testing whether there is an association between the number of years that a cohort 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.

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, though we do find evidence indicating that the effect of NCLB varies across states. Moreover, we find that the effect of NCLB on the white-black gap depends in part on whether a majority of black students are in schools where there are enough black students to meet the state-determined NCLB minimum subgroup size reporting threshold. In states where relatively few black students are in schools held accountable for their black students' performance, NCLB actually appears to have led to a widening of the white-black achievement gap. Nonetheless, the impact of NCLB on achievement gaps—whether positive or negative—is generally very modest in size, on the order of changing gaps by $1/100^{th}$ of a standard deviation per year on average.

Introduction

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. Although racial gaps narrowed substantially in the 1970s and 1980s (Grissmer, Flanagan and Williamson 1998; Hedges and Nowell 1998; Hedges and Nowell 1999; Neal 2006), they narrowed only slightly in the 1990s, and remained very large in 2001 (roughly 0.75-1.0 standard deviations), when the law was passed (Hemphill, Vanneman and Rahman 2011; Reardon and Robinson 2007; Vanneman et al. 2009). Dissatisfied with these large gaps, as well as with overall levels of achievement, Congress passed the NCLB legislation. Title I 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 disaggregated by race and socioeconomic status, and that sanctions at the school level hinge on these results.

Ten years have passed since NCLB went into effect. In this paper we investigate whether the Act has been successful at narrowing racial achievement gaps. We do so using several different analyses. First, we describe the average trends in within-state achievement gaps from 1990 through 2009. Second, we test whether there is an association between the number of years that a cohort 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 these exposure-gap associations are stronger in states where NCLB was implemented in ways that we expect would lead to a greater focus on 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, though we do find evidence indicating

that the effect of NCLB varies across states. Moreover, we find that the effect of NCLB on the white-black gap depends in part on whether a majority of black students are in schools where there are enough black students to meet the state-determined NCLB minimum subgroup size reporting threshold. In states where relatively few black students are in schools held accountable for their black students' performance, NCLB actually appears to have led to a widening of the white-black achievement gap. Nonetheless, the impact of NCLB on achievement gaps—whether positive or negative—is generally very modest in size, on the order of changing gaps by 1/100th of a standard deviation per year on average.

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 and Katz 2008; Murnane, Willett and 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 and Tienda 2007; Posselt et al. 2010). As a result 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 and Tienda 2007; Bollinger 2003; Carneiro, Heckman and Masterov 2003; Neal and 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 both the National Assessment of Educational Progress (NAEP) and state accountability assessments. We know that achievement gaps in both math and reading between white and black students have narrowed substantially over the last forty years (Grissmer, Flanagan and Williamson 1998; Hedges and Nowell 1999; Hemphill, Vanneman and Rahman 2011; Kober, Chudowsky and Chudowsky 2010; Neal 2005; Vanneman et al. 2009). Despite this progress, the gaps remain large, ranging from two-thirds to slightly less than one standard deviation, depending on the cohort and subject. White-Hispanic gaps have continued closing between 2004 and 2009 (Hemphill, Vanneman and Rahman 2011). Importantly, both the size and trends in achievement gaps show marked heterogeneity across states (Hemphill, Vanneman and Rahman 2011; Kober, Chudowsky and Chudowsky 2010; Vanneman et al. 2009).

Just as gaps vary across states, they vary as children progress through school. Data from the ECLS-K show that the white-black and white-Hispanic gaps are similar in magnitude at kindergarten entry; however, white-black gaps increase during the first six years of schooling in both math and reading, while white-Hispanic gaps decrease during this period (Fryer and Levitt 2004; Fryer and Levitt 2005; Reardon and Galindo 2009; Reardon and Robinson 2007). At kindergarten entry the white-black and white-Hispanic gaps in reading and math are 0.5 and 0.75 standard deviations, respectively. By fifth grade the white-black gaps in reading and math widen to 0.75 and 1.0 standard deviations. Over the same period, white-Hispanic gaps in reading and math narrow to 0.33 and 0.75 standard deviations. In the NAEP data, racial gaps appear to grow in math, and modestly decrease in reading, between fourth and eighth grade.

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 the 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 and Raymond 2004)—may compel 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 of each subgroup. Here, threats of government restructuring or loss of funding may pressure schools to improve the academic performance of students who are unable to demonstrate proficiency. To the extent that these students are disproportionately low-income or racial/ethnic minorities, the law may induce gap closure.

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. Given that lesser-qualified teachers are over-represented in schools serving low-income and minority students (Lankford, Loeb and Wyckoff 2002), NCLB may affect achievement gaps by equalizing the distribution of qualified teachers and, therefore, disassociating the relationship between students' background characteristics and the quality of teaching they experience. 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 they systematically increase student achievement, then these facets of No Child Left Behind may close achievement gaps, as well.

There is reason to think that the effect of NCLB on achievement gaps may vary among states. For example, the extent to which NCLB induced subgroup-specific accountability pressure may have varied among states, leading to different effects on achievement gaps. One reason that subgroup-specific accountability may vary among states is that NCLB does not require states to hold schools accountable separately for the performance of subgroups of students when there are

too few students of a subgroup within a school to yield reliable information regarding that subgroup's performance.[1] That is, in a school with few black students in tested grades, the test scores of black students would not be reported separately and the school would not be required to show adequate yearly progress for black students (black students' scores would still be included in calculations of the school's overall proficiency rate, however, though they might matter little given the small number of black students). In such a school, NCLB may create little or no incentive to focus attention on the performance of the small number of black students in the school—indeed, it may create an incentive to focus primarily on the performance of the schools' 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. A school with a large number of black students, in contrast, 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 most minority students are in schools where their group is sufficiently large to require subgroup-specific reporting and accountability than in states where few minority students are in such schools. The proportion of black or Hispanic students who are in such schools depends 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 are in 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 state's criteria for determining what number of students is sufficient to require subgroup-specific reporting and accountability (NCLB gave states some latitude in determining the number of students sufficient to require subgroup-specific reporting). As we show below, states vary considerably in the proportion of black and Hispanic students whose

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

test scores were relevant for accountability purposes. If NCLB affects achievement gaps through incentives induced by the subgroup-specific reporting and accountability requirements, we might expect NCLB to have led to a greater reduction in achievement gaps in states where most minority students were in schools where their scores were reported separately. Moreover, if few minority students in a state are in schools where their scores are reported and determine sanctions, schools in that state may actually be induced to focus more on white students' achievement, possibly widening achievement gaps.

Prior Evidence on the Effect of No Child Left Behind

The research literature is mixed regarding the effects of accountability systems generally, and of No Child Left Behind specifically, on student achievement (Carnoy and Loeb 2002; Dee and Jacob 2011; Gaddis and Lauen 2011; Hanushek and Raymond 2004; Hanushek and Raymond 2005; Lauen and Gaddis forthcoming; Lee 2006; Lee and Wong 2004; Wong, Cook and Steiner forthcoming). In general, these divergent findings may be attributed to differences in the studies' samples, model specifications, and accountability system types they examine.

Research on the effects of NCLB is challenged by the difficulty of identifying a plausible counterfactual necessary for estimating the causal impact of accountability regimes on differential achievement. Because NCLB was introduced at the federal level, the treatment was effectively imposed on all states at the same time, making 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 both Wong, Cook, and Steiner (forthcoming) and Dee and Jacob (2011).

Dee and Jacob (2011) 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. Based on this reasoning, they conduct a set of comparative interrupted time series analyses, using Main NAEP data from 1990-2007, to estimate the effect of NCLB. They find that NCLB improved average math performance, particularly in fourth grade, but did not affect reading performance. Although they do not estimate the effect of NCLB on achievement gaps, they do disaggregate effects by racial/ethnic, gender, and socioeconomic (eligibility for free/reduced-price lunch) subgroup. Their findings suggest the 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. However, their analyses are often based on different sets of states and do not provide statistical tests of the differences in effects between subgroups, making it difficult to determine the size and statistical significance of differences between the effects of NCLB on different subgroups.

Wong, Cook, and Steiner (forthcoming) adopt a similar approach, but compare post-NCLB changes in achievement trends between states that instituted "high" proficiency and "low" standards in response to the federal NCLB accountability mandate. Their argument is that states with high standards (which they define 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 Dee and Jacob (2011) they find significant effects of NCLB on average fourth and eighth grade math achievement (but no effect on reading achievement). They do not estimate the effects of NCLB on achievement gaps, however.

[2] The literature 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.

Analytic Strategy and Hypotheses

We rely on two strategies to identify the effects of NCLB on achievement gaps. First, we reason that any effects of the NCLB accountability regime ought to accumulate as students progress through school. This suggests that we can use differences between cohorts in the number of years they have been exposed to NCLB accountability pressure by a given grade to identify the effects of NCLB. Differences between cohorts in exposure to NCLB may, however, be correlated with other between-cohort differences in factors affecting achievement gaps. To address this threat, our second strategy relies on a difference-in-differences approach, comparing the association between exposure and achievement gaps in states that differ in their implementation of NCLB. Specifically, we posit the following two hypotheses:

Hypothesis 1: Greater exposure to NCLB will lead to smaller gaps.

That is, in a given grade, the achievement gap will be smaller, on average, for cohorts that have spent more years under the NCLB regime than for cohorts that have spent fewer years under the regime. Put differently, within a given cohort, the achievement gap will narrow faster (or widen less rapidly) as the cohort progresses through school during the NCLB regime than prior to it.

Hypothesis 2: The association between the estimated NCLB effect and the proportion of minority students in schools that meet the state's minimum subgroup size reporting threshold will be negative.

As we argued above, we expect that the policy will exert more pressure to narrow gaps on states in which larger proportions of minority students are in schools where the minimum subgroup threshold is met. That is, exposure to NCLB will narrow gaps faster (or widen them less rapidly) in states where more minority students are in schools meeting the subgroup reporting threshold.

Data and Methods

Estimating Achievement Gaps

There are two different ways of defining "achievement gaps." First is what we call a "proficiency gap," the between-group difference in the proportions of students scoring above some "proficiency" threshold on a test. Second is what we call a "distributional gap," typically described using some summary measure of the difference between the test score distributions in two groups (such as the difference in means, or the difference in means divided by their pooled standard deviation). These two types of gaps, computed from the same data, need not have the same sign, nor trend in the same direction.

The data 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, in some sense, NCLB is explicitly designed to narrow proficiency gaps, as defined by where states set their proficiency threshold, so it is worth testing whether it does indeed narrow such gaps.

Achievement gaps are more commonly reported using distributional gap measures, such as mean differences or standardized mean differences. One drawback of mean and standard deviation-based measures, however, is that they rely on the assumption that test scores are measured in an interval-scaled (or cardinal scale) metric, meaning that each unit of the score has

equal value. This assumption that may be problematic, particularly when comparing trends in achievement gaps, which can be highly sensitive to the interval-scale assumption (Reardon 2008).

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 and Reardon 2012; Ho 2009; Ho and Haertel 2006). 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 Ho and colleagues 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) (Ho and Reardon 2012).

A nice 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 assume that each test measures achievement in some ordinal-scaled metric, a much more defensible assumption.

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, under an NCES restricted data

use license) or data on the counts of students of each group in each of several (at least three) proficiency categories. 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.[3]

Data

In this paper we use two primary data sources to estimate state-level achievement gaps: NAEP[4] and state assessment data. We use state NAEP test score data from 4th- and 8th-graders between 1990 and 2009 in math and reading, and categorical proficiency data (e.g., percentages of students scoring "Below Basic," "Basic," "Proficient," and "Advanced") from state-level accountability tests. Most of the state accountability test data comes from tests introduced beginning in 2002 under the No Child Left Behind Act, but we also use some earlier test score data from states that had accountability testing in place prior to that. These data have been collected by federal and state departments of education and are disaggregated by subgroup, subject, grade, and year. Typically we have data for grades three through eight, though in some states/years data are available for fewer years (because tests were not given in each of these grades); in a small number of states/years, data are available for second grade as well. We do not analyze data from secondary grades, as states vary in the specific content covered in such tests and the ages of students tested. We use only data from math and reading, as these two subjects are those most consistently reported and align with those tested in NAEP. From these data we compute estimates of white-black and white-Hispanic gaps in each state-by-year-by-grade-by subject for which we have NAEP

[3] See Appendix for details on the estimation of V. [TO BE ADDED IN A FUTURE DRAFT]
[4] 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 and Washington, DC. State NAEP sample sizes are roughly 2,500 students, from approximately 100 schools, in each state-grade-subject.

and/or state test data.

With respect to state standardized test data, our dataset includes outcomes from as far back as 1997 for some states, and for a substantial number of states (26) starting in 2002. These states do not appear to be clustered in any specific region. We have data for virtually all states beginning in 2006. 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 retrieved these data from three data sources: state Department of Education websites, the Center on Education Policy (CEP) website, and directly from the U.S. Department of Education. To maximize the amount of years we could analyze, we combined these data sources to fill in gaps when one source was missing an observation for any year, grade, and/or subject combination. See Appendix B for further details on the sources of state data used for our analyses, and the methods used to determine which data sources were the most valid.

State Accountability Measures

We characterize states by the extent to which their implementation of NCLB was likely to focus attention on 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 and were not 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. Figure 1 describes the variation among states in the proportions of students of different subgroups in schools where their scores are reported and

consequential. There is a great deal of variation in the proportions of students subject to accountability reporting among states.

Figure 1 here

Covariates

We include a variety of state-level time-varying and time-invariant covariates in our models, 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 income ratio, poverty ratio, and unemployment rate ratio for each state and year. For each state-cohort-grade combination, we then average these ratios over the years from a cohort's birth year to the year before that cohort entered kindergarten to construct a measure of the average ratio experienced by a cohort during preschool. We construct similar measures of the cumulative exposure to each of these ratio variables from kindergarten through each grade in which we observe a cohort's achievement gap. 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. For each state-cohort-grade combination, we compute the cumulative exposure of a cohort to the variable through a given grade. The rationale for this method of constructing the covariates is explained in Appendix A.

Considerations Regarding the Use of NAEP and State Test Data

We use both NAEP and state accountability test data here. Each has a distinct set of advantages and disadvantages. First, they cover different combinations of cohorts and grades. Table 1 describes the number of state-by-subject-by-subgroup achievement gap observations we have for each cohort and grade. The maximum possible in any cell here is 204 (51 states x 2

subjects x 2 subgroups). Note that cohorts of students entering kindergarten prior to 1994 would have been in high school before NCLB was enacted, while cohorts entering in 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. Table 1 shows 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 include much more data from post-NCLB cohorts. These differences in the coverage of the NAEP and state data will be of interest to us below.

Table 1 here

The NAEP and state test data differ in a number of other ways as well. NAEP assessments have the advantage of being based on a set of tests that has remained relatively unchanged over the last two decades, making comparisons among states and years relatively straightforward. Moreover, the NAEP assessments are low-stakes tests, and are reported only at the state level, meaning there is little inventive for schools or teachers to teach to the test or to otherwise influence their students' scores (on the other hand, the low-stakes nature of the NAEP assessments may mean students have little incentive to perform as well as they are capable). However, because the NAEP assessments are administered to relatively small samples in only two grades and only every other year, they may provide less reliable estimates of achievement gaps and their trends than the state tests, which are administered to virtually all students in grades 2-8 each year. The state tests, in contrast, are high-stakes tests (meaning that schools are held accountable for their results, which may distort scores in unpredictable ways); they are aligned with state standards (which may increase the validity of the gap estimates relative to the standards espoused by each state, but which also means that gaps may not be comparable across states); and they have changed over time, potentially complicating trend analyses. Many of the concerns about the state tests, however, are much less problematic for the analysis of achievement gaps than the analysis of achievement

levels. When we compute the V-statistic, we rely only on the assumption that the test measures achievement in a given domain in some ordinal metric. Differences among tests that measure the same content domain with comparable reliability will therefore yield similar gap estimates, even if the test metrics differ substantially. Below we present some evidence that the state test data and the NAEP data yield similar estimates of the size and trends in achievement gaps across states. Moreover, our primary results here do not differ when based on NAEP or state data.

Methods

In Appendix A, we derive a model for the relationship between the size of the achievement gap for a given cohort c in grade g in state s. This model expresses the achievement gap as a state-specific function of grade (denoted gr_g), cohort (denoted coh_c), a variable $E_g = \frac{1}{2}\left(gr_g^2 - gr_g\right)$, vectors of cohort-by-state and cohort-by-state-by grade covariates ($\mathbf{X}_{cs}$ and $\mathbf{W}_{csg}$), and the number of years the cohort has been exposed to NCLB by the end of grade g (exp_{cg}):

$$G_{csg} = \lambda_s + \alpha_s\left(gr_g\right) + \eta\left(E_g\right) + \beta\left(gr_g \cdot coh_c\right) + \gamma_s(coh_c) + \delta_s\left(exp_{cg}\right) + \mathbf{X}_{cs}\mathbf{A} + \mathbf{W}_{csg}\mathbf{B} + e_{csg}. \quad [1]$$

To fit this model, we pool the math and reading gap estimates into a single data set and estimate the effects of NCLB on achievement gaps using a set of precision-weighted random coefficients models of the form:

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

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

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

$$\begin{bmatrix} \lambda_s \\ \gamma_s \\ \alpha_s \\ \delta_s \end{bmatrix} \sim N\left[\begin{pmatrix} \lambda \\ \gamma \\ \alpha \\ \delta \end{pmatrix}, \begin{pmatrix} \tau_\lambda & \tau_{\lambda\gamma} & \tau_{\lambda\alpha} & \tau_{\lambda\delta} \\ \tau_{\gamma\lambda} & \tau_\gamma & \tau_{\gamma\alpha} & \tau_{\gamma\delta} \\ \tau_{\alpha\lambda} & \tau_{\alpha\gamma} & \tau_\alpha & \tau_{\alpha\delta} \\ \tau_{\delta\lambda} & \tau_{\delta\gamma} & \tau_{\delta\alpha} & \tau_\delta \end{pmatrix}\right]$$

[2]

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); sub_t is a dummy variable indicating whether $\hat{G}_{csgt}$ is a math or reading gap; $\mathbf{X}_{cs}$ and $\mathbf{W}_{csg}$ are vectors of cohort-by-state and cohort-by-state-by grade covariates, respectively (the specific form of these covariates is described in Appendix A); 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_{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 $\boldsymbol{\tau}$ matrix, using the HLM v7 software.

The identification of δ in model (2) 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 linearly 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 2 helps to clarify these different sources of variation in exp_{cg} (see also the discussion

in Appendix A). In Figure 2, 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. Because almost all of the available NAEP data fall in the yellow and green regions of Figure 2 (the 2009 4th grade NAEP data, corresponding to the 2004 cohort, are an exception), our models using NAEP data rely on the first source of variation in exp_{cg}. Our models using 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.

Figure 2 here

We fit several versions of model (2), each using different subsets of our data. Our most comprehensive models pool all our data—both NAEP and state data, math and reading gaps, and data from all available cohorts and years (we do not pool white-black and White-Hispanic gap estimates, however). We then fit models that use different subsets of the data: using only NAEP or only state data; limiting the data to pre-2003 cohorts or post-2002 years; and fitting the models separately for math and reading outcomes. We focus on estimating models that use the V-statistic as the outcome, as the distributional gap measure is of most interest. However, we also fit a set of models using the difference in proficiency rates as the outcome (these models necessarily use only the state test data), to assess whether NCLB affected proficiency gaps. In all the models, we include a set of cohort- and time-varying covariates that might be correlated with exp_{csgt} and that might impact trends in achievement gaps. These include cohort- and state-specific measures of the black/white (or Hispanic/white, as appropriate) income ratio, poverty ratio, and unemployment ratio, as well as measures of the proportion black (or Hispanic) in public schools and the level of black/white (or Hispanic/white) school segregation. These measures and their construction are described in more detail in Appendix A. In general, we find that the inclusion of these covariates

has little effect on either the coefficients of interest or their standard errors.

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 (2) 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, there was no other policy or demographic change in 2002, net of the demographic trends captured by our control variables, that had a cumulative effect on achievement gaps in the years following 2002.

One of the advantages of fitting Model (2) using a random coefficients model is that it allows each state to have a different intercept (a different-size pre-K achievement gap in 2002), a different pre-NCLB linear time trend in the achievement gap, a different pre-NCLB grade slope in the achievement gap, and a different effect of NCLB. We use a deviance test to test the null hypothesis that the NCLB effect (δ_s) is constant across states. In general, our results indicate that we can reject this hypothesis: the effect of NCLB varies among states. Finding that the estimated effect varies across states, we then estimate a second set of models to test the hypothesis that the state-specific effect of NCLB is negatively associated with the proportion of minority students in schools where their group's scores were required to be reported. Specifically, we interact exp_{cg} with 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. Negative coefficients on these interaction terms would support the hypothesis that NCLB narrowed achievement gaps more in states where more minority students were subject to accountability pressure, consistent with our theoretical expectations.

Results

Trends in Achievement Gaps

We begin by describing the trends in white-black and white-Hispanic achievement gaps in

math and reading for cohorts of students who entered kindergarten from 1991 through 2006. Figures 1 and 2 shows these trends, estimated using three different sources/measures: 1) Cohen's d based on 4th and 8th grade NAEP data from 1995 through 2009; 2) V based on 4th and 8th grade NAEP data from 1995 through 2009; 3) V based on state accountability test data from grades 2-8 from 1997 through 2010.[5] Three features of the figures are notable. First, the magnitude and trend based on NAEP data are virtually identical for the Cohen's d and V measures. This suggests that using V in our analyses will yield similar results as if we had used Cohen's d. Second, the magnitude of V based on state data is generally smaller than V based on NAEP (though this is not true for the white-Hispanic reading gaps, perhaps because of different exclusion criteria in NAEP and state tests). One reason for this is that NAEP scores are corrected to account for measurement error, while the state test score data are not; this tends to attenuate the state gap estimates relative to the NAEP gaps, as we see here.[6] Third, both white-black and White-Hispanic achievement gaps have been narrowing, albeit slowly, over the last two decades; this pattern is consistent across each of the three different gap measures. Both the NAEP and state data suggest that the rate of narrowing of the white-black gap in math has slowed in the most recent cohorts; but this trend is less evident in reading and not evident in the White-Hispanic gap trends.

Figures 3 and 4 here

Figures 3 and 4 are based on a set of non-parametric trend models, and show only average trends across states and grades. In order to examine the variation in trends across states, we fit a

[5] The trends displayed in Figures 3 and 4 indicate the trend in the estimated cohort fixed effects (the $\hat{\Gamma}_c$'s) from the model

$$\hat{G}_{csg} = \lambda_s + \Gamma_c + u_{\gamma s}(coh_c^*) + \alpha_s(grade_g - 4) + e_{csg} + \epsilon_{csg},$$
$$e_{csg} \sim N[0, \sigma^2]$$
$$\epsilon_{vsg} \sim N[0, \omega_{csg}^2] = N[0, var(\hat{G}_{csg})]$$
$$\begin{bmatrix} \lambda_s \\ u_{\gamma s} \\ \alpha_s \end{bmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \\ \alpha \end{pmatrix}, \begin{pmatrix} \tau_\lambda & \tau_{\lambda\gamma} & \tau_{\lambda\alpha} \\ \tau_{\gamma\lambda} & \tau_\gamma & \tau_{\gamma\alpha} \\ \tau_{\alpha\lambda} & \tau_{\alpha\gamma} & \tau_\alpha \end{pmatrix}\right].$$

[6] Figures A1-A2 in the Appendix are similar to Figures 3-4 here, but limit the NAEP and state test data samples to state-cohort-grade-subject cases where we have gap estimates from both sources. These figures show very similar patterns as in Figures 3-4, indicating that the difference in the magnitude of the trends is not an artifact of the different samples of states/cohorts/grades/subjects used in Figures 3-4.

set of random coefficient trend models that allow us to estimate both the average linear trend across states and the extent to which the linear trends vary among states. Table 2 reports the results of these models. The top three panels of the table shows estimated parameters from 18 models describing the trends in the white-black and White-Hispanic gaps, measured by the V-statistic, using different combinations of data sources and test subjects. Across all 18 models, however, the estimated cohort slope is always negative and statistically significant, ranging from estimates of -0.005 standard deviations per year to -0.010 standard deviations per year.[7] In the models that pool both state and NAEP V gap estimates and that pool both math and reading gap estimates, the estimated trend in the white-black and white-Hispanic gaps are -0.008 (se = 0.002, $p < .001$) and -0.009 (se = 0.002, $p < .001$) standard deviations per year, respectively. At this rate, the white-black and white-Hispanic gaps will be eliminated in roughly 70-100 years.

Table 2 here

The fourth panel of Table 2 reports estimates of the gap trend model using Cohen's d as a measure of the gap, rather than V. The results of these models are very similar to the models using the V-statistic. Finally, the bottom panel of Table 2 presents the results of models using the proficiency gap measure as the outcome. In general, the proficiency gaps declined across cohorts born in the 1990s and 2000s at a rate of 0.3-0.45 percentage points per year for white-black gaps and at a rate nearly double that for white-Hispanic gaps.

To test whether the achievement gaps narrow faster after the start of NCLB, we fit a set of models like those described in Equation [2] above. The key parameter of interest here is the coefficient δ, the average effect of each additional year of exposure to NCLB on a cohort's achievement gap. Tables 2 and 3 report these estimated coefficients from a set of models using different combinations of data sources, measures, and samples of observations. We report the

[7] Table A1 in the Appendix reports models like those in Table 2, but based on the overlapping NAEP and state test data samples. In these models, the trend in achievement gaps appears to be declining somewhat faster in the models based on the state data than in the NAEP models, but the difference is not statistically significant.

estimates from models with and without the vectors of covariates $\mathbf{X}_{cs}$ and $\mathbf{W}_{csg}$. 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.

Tables 3 and 4 here

The top left panels of Tables 3 and 4 contain the estimated effect of NCLB using pooled math and reading gap data and pooled NAEP and state accountability data. When using all observations (all available cohorts and years), the estimated effect of exposure to NCLB is not statistically different than 0 for white-black gaps or white-Hispanic gaps ($\hat{\delta} = -0.005, se = 0.003$ for white-black gap; and $\hat{\delta} = 0.003, se = 0.005$ for white-Hispanic gap). Given the small standard errors, we can rule out meaningfully large effects. Although there is some variation across the two data sources, test subjects, and samples of observations, there is little evident pattern to the result. In Table 3, there is some evidence that NCLB *widened* white-black achievement gaps (based on the generally positive, and sometimes significant, coefficients in the "Pre-2003 Cohorts" columns), but this does not hold across most of the models (particularly the "Post-2002 Data" estimates, where there is no evidence of a significant effect). In Table 4, there is likewise some evidence that NCLB widened white-Hispanic achievement gaps (here based on the estimates in the "Post-2002 Data" column), though again the estimates are inconsistent. In neither Table 3 nor 4, however, is there any evidence to suggest a substantial or statistically significant association between the number of years of exposure to NCLB and the size of achievement gaps.

In additional analyses not shown here, we added a term to the models for Tables 3 and 4 to test whether the effect of NCLB changes across grade levels (see Appendix A for details on this model). We found no evidence to suggest any trend in the magnitude of the effect of NCLB across grades.

Although Tables 3 and 4 suggest 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 0.007 and 0.017, respectively. 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 5 shows the estimated state-specific effects of exposure to NCLB. 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 5 here

In Tables 5 and 6 we report the results from models testing whether the association between exposure to NCLB and achievement gaps is larger (more negative) in states where a larger proportion of black or Hispanic students are in schools meeting the state's minimum subgroup size threshold—in schools where their test scores are consequential for accountability.

Tables 5 and 6 here

Several patterns are evident in the coefficients reported in Table 5. First, the coefficient on the exposure variable is often positive and significant in these models, particularly when the models are fit using the pre-2003 cohorts. This implies that in states where no black students were in schools meeting the minimum subgroup size (as was true in VT, ID, MT, and was nearly true in WY), the white-black gap actually grew with increased exposure of cohorts to NCLB. The negative (and significant) coefficients on the interaction term, however, indicate that the white-black gap widened less, or narrowed, in states where the proportion of black students subject to test score reporting was larger. One of the key patterns evident in Table 5 is that the estimates from the models using NAEP and state test data differ somewhat when all observations are used, but this appears to be driven entirely by the fact that the NAEP models rely almost entirely on pre-2003 cohorts while the state data models rely much more on the post-2002 data years. A comparison of the NAEP and state models that rely only on pre-2003 cohorts shows the two data sources yield very similar results: NCLB appears to have increased achievement gaps in states where few black

students' scores were reported at the school level, but had a much smaller or zero effect in states where most black students' scores were reported. In the years following the introduction of NCLB (the post-2001-02 school years), however, the estimates from the state data suggest that NCLB had no impact on achievement gaps for states where few black students were subject to accountability, and reduced gaps modestly where most black students were subject to accountability. Figures 6-8 illustrate these figures 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.

Figures 6-8 here

Table 6 presents the corresponding estimates from the White-Hispanic gap models. In these models, however, there is no evidence that the impact of NCLB varies systematically with the proportion of Hispanic students in schools meeting the state's minimum subgroup size. As in Table 4, there is some evidence that NCLB widened achievement gaps, on average, but this is only evident in the models using post-2002 data. In general, exposure to NCLB does not seem strongly associated with the white-Hispanic gap, even in states where most Hispanic students are in schools where their scores are reported.

Discussion

One way in which NCLB may affect academic achievement gaps is by holding schools accountable for the average test scores of both each subgroup (black, Hispanic, white) separately. However, NCLB does not, in fact, require all schools to be held accountable for the test scores of each subgroup. If a school enrolls fewer than some minimum number of students (a number set by each state), the scores for that subgroup are not reported separately, and the school is not held accountable for that specific subgroup's performance (though of course, the students in that subgroup still contribute to the overall scores of the school). This suggests that NCLB may create

more pressure for schools to improve minority students' scores in some schools than it does in others, which may lead to differential effectiveness of NCLB across states in closing achievement gaps. This in turn suggests that states where most minority students are in schools where their scores are reported—states with large minority populations, high levels of segregation, and/or low minimum subgroup size thresholds—will be the states where NCLB will tend to have the largest impact on reducing achievement gaps.

Our results in this paper are somewhat consistent with this story, as least with respect to white-black gaps. Although we find little evidence that the average within-state effect of exposure to the NCLB regime affected achievement gaps (and what evidence we do find tilts more toward an effect of increasing gaps rather than narrowing them), we do find evidence that the effect varies moderately across states. Moreover, the effect of NCLB on the white-black achievement gap is positive (it widens the gap) in states where few black students are in schools where black students' scores are reported; however, the effect is less positive or even negative (it narrows gaps) in states where most black students are in schools where their scores are reported. This pattern does not hold, however, for the white-Hispanic gap.

One of the puzzles in our findings, however, is that if we estimate the NCLB effects using data only from cohorts who entered kindergarten prior to Fall 2003, we observe somewhat different patterns of results than if we estimate the models using data only from the school year 2002-03 and later. Moreover, the pattern of these differences is opposite in the white-black and White-Hispanic gap models. Generally speaking, the estimated effects of NCLB on white-black gaps based on the pre-2003 cohorts are larger (i.e., more positive, implying NCLB widened achievement gaps) than those based on post-2002 data. The opposite is true for the White-Hispanic models. There are several possible explanations for these patterns. First, the models based on pre-2003 cohorts and on post-2002 data rely on different sources of variation in exposure to NCLB for identification of their effects. The pre-2003 cohort models rely largely on variation between

cohorts in their exposure to NCLB in the later grades; while the pre-2002 data models rely largely on variation in exposure in the early grades. If the effects of NCLB differ systematically across grades, this could account for the differences in estimates. However, we did fit a set of models that explicitly test for variation in the NCLB effect across grads (not shown), and found no evidence of such variation.

Second, NCLB took some time to be implemented and the incentives of NCLB changed over time, so the effect of NCLB may have varied across time. Schools (and states) had much more information about the performance of black and Hispanic students in the later years of NCLB than in the first few years, and the sanctions associated with failure to make adequate yearly progress grew over time. These factors may have increased the likelihood that schools and states would focus efforts on improving minority students' achievement, and would have led to a narrowing of achievement gaps. This narrative might account for the improvement in the effects of NCLB on the white-black gaps (from the pre-2003 cohort models to the post-2002 data models), but do not explain why we see the opposite the pattern in the White-Hispanic effects.

A third possibility is that the model is misspecified. The model assumes that the trend in the achievement gap within a given state and grade would have been linear in the absence of NCLB. A violation of this assumption might lead to differences in the estimates based on different subsets of the data. Put differently, the exposure to NCLB may not be exogenous, and correlations of unobserved time-varying factors may differ in different subsets of the data, leading to different degrees of bias. If this is the case, then it is not clear whether we should give more credence to the pre-2003 cohort models or the post-2002 data models.

Despite the differences in the estimates of the average NCLB effect between the two set of models, however, one pattern is relatively consistent in the white-black models. The coefficient on the interaction term between exposure to NCLB and the proportion of students in schools meeting the minimum subgroup threshold is negative and significant, implying that NCLB is more effective

at narrowing the white-black achievement gaps in states where most black students are in schools meeting the minimum subgroup size. This is not true for the White-Hispanic gaps, for reasons that are not clear. There is much more variation among states in the estimated effects of NCLB on White-Hispanic achievement gaps, however, which makes it much less likely that we would detect a significant coefficient on the interaction term (the standard errors suggest that the data have 80% power to detect a coefficient with absolute value of roughly 0.030-0.045).

Although the white-black gap models are consistent with theoretical predictions that NCLB should be most effective at narrowing achievement gaps in states where most black students are in schools where they meet the minimum subgroup size, it is not obvious that we can interpret these as effects of the minimum subgroup size threshold. That is, it is not clear that requiring states to lower their minimum subgroup size thresholds would lead to more reduction of achievement gaps. Much of the variation in the proportion of students in schools meeting minimum subgroup sizes is due to variation among states in racial composition and racial segregation. It may be that NCLB has been most effective at narrowing achievement gaps (or keeping them from widening) in states with high proportions of students in schools meeting the minimum subgroup size because NCLB is most effective in more segregated school systems, or in states where minority students make up a larger proportion of all students (because perhaps there is more political pressure to improve minority students' scores when they are a larger constituency), or in states where pre-NCLB performance levels were particularly low (which is correlated with racial composition and segregation).

Conclusion

Overall, we find no consistent evidence that NCLB has been effective, on average, at narrowing achievement gaps. Indeed there is some evidence that NCLB may have widened achievement gaps, though the estimates are inconsistent across models and subsamples of our data. Our estimates are very precise, however, so we can rule out the possibility that NCLB had, on

average meaningfully large effects (effects larger than 0.010-0.015 standard deviations change per year) on achievement gaps.

Despite the fact that NCLB appears to have had, at best, no average effect on achievement gaps, the effects appear to vary among states. Moreover, the effects of NCLB vary inversely with the proportion of minority students in schools where they are subject to accountability pressure. While these results do not prove that NCLB had larger effects because of the increased awareness and incentives to improve minority students' performance that may have resulting from the larger share of minority students subject to NCLB reporting, our finding is certainly consistent with a theoretically defensible model of how NCLB may operate.

References

Alon, Sigal, and Marta Tienda. 2007. "Diversity, opportunity, and the shifting meritocracy in higher education." *American Sociological Review* 72(4):487-511.

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

Carneiro, Pedro, James J. Heckman, and Dimitry V. Masterov. 2003. "Labor market discrimination and racial differences in premarket factors " in *NBER working paper*. Cambridge, MA: National Bureau of Economic Research.

Carnoy, Martin , and Susanna Loeb. 2002. "Does external accountability affect student outcomes? A cross-state analysis." *Educational Evaluation and Policy Analysis* 24(4):305-31.

Dee, Thomas S., and Brian Jacob. 2011. "The impact of No Child Left Behind on student achievement." *Journal of Policy Analysis and Management* 30(3):418-46.

Fryer, R.G., and S.D. Levitt. 2004. "Understanding the black-white test score gap in the first two years of school." *Review of Economics and Statistics* 86(2):447-64.

—. 2005. "The black-white test score gap through third grade." in *Working Paper Series. Working Paper 11049*. Cambridge, MA: National Bureau of Economic Research.

Gaddis, S. Michael, and Douglas Lee Lauen. 2011. "Has NCLB accountability narrowed the black-white test score gap?".

Goldin, Claudia, and Lawrence F. Katz. 2008. *The Race Between Education and Technology*. Cambridge, MA: Harvard University Press.

Grissmer, David W., Ann Flanagan, and Stephanie Williamson. 1998. "Why did the Black-White score gap narrow in the 1970s and 1980s?" Pp. 182-228 in *The Black-White Test Score Gap*, edited by Christopher Jencks and Meredith Phillips. Washington, D.C.: Brookings Institution Press.

Hanushek, Eric A., and Margaret E. Raymond. 2004. "The effect of school accountability systems on the level and distribution of student achievement." *Journal of the European Economic Association* 2(2-3):406-15.

—. 2005. "Does school accountability lead to improved student performance?" *Journal of Policy Analysis and Management* 24(2):297-327.

Hedges, Larry V., and Amy Nowell. 1998. "Black-White Test Score Convergence Since 1965." Pp. 149-81 in *The Black-White Test Score Gap*, edited by Christopher Jencks and Meredith Phillips. Washington, D.C.: Brookings Institution Press.

—. 1999. "Changes in the black-white gap in achievement test scores." *Sociology of Education* 72(2):111-35.

Hemphill, F. Cadelle, Alan Vanneman, and Taslima Rahman. 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, Andrew D., and Sean F. Reardon. 2012. "Estimating Achievement Gaps From Test Scores Reported in Ordinal 'Proficiency' Categories." *Journal of Educational and Behavioral Statistics* 37(4):489-517.

Ho, Andrew Dean. 2008. "The problem with "proficiency": Limitations of statistics and policy under No Child Left Behind." *Educational Researcher* 37(6):351-60.

—. 2009. "A nonparametric framework for comparing trends and gaps across tests." *Journal of Educational and Behavioral Statistics* 34:201-28.

Ho, Andrew Dean, and Edward H. Haertel. 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.

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

Lankford, Hamilton, Susanna Loeb, and James Wyckoff. 2002. "Teacher sorting and the plight of urban schools: A descriptive analysis." *Educational Evaluation and Policy Analysis* 24(1):37-62.

Lauen, Douglas Lee, and S. Michael Gaddis. 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, Jaekyung. 2006. "Tracking achievement gaps and assessing the impact of NCLB on the gaps: An in-depth look into national and state reading and math outcome trends." Cambridge, MA: The Civil Rights Project at Harvard University.

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

Murnane, Richard J., John B. Willett, and Frank Levy. 1995. "The Growing Importance of Cognitive Skills in Wage Determination." *Review of Economics and Statistics* 78(2):251-66.

Neal, Derek A. 2005. "Why has Black-White skill convergence stopped?": University of Chicago.

—. 2006. "Why has Black-White skill convergence stopped?" Pp. 511-76 in *Handbook of the Economics of Education*, edited by Eric A. Hanushek and Finis Welch. New York: Elsevier.

Neal, Derek A., and William R. Johnson. 1996. "The role of premarket factors in black-white wage differences." *The Journal of Political Economy* 104(5):869-95.

Posselt, Julie, Ozan Jaquette, Michael Bastedo, and Rob Bielby. 2010. "Access without equity: Longitudinal analyses of institutional stratification by race and ethnicity, 1972-2004." in *Annual meeting of the Association for the Study of Higher Education*. Indianapolis, IN.

Reardon, S.F., and C. Galindo. 2009. "The Hispanic-White achievement gap in math and reading in the elementary grades." *American Educational Research Journal* 46(3):853.

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

Reardon, Sean F., and Joseph Robinson. 2007. "Patterns and Trends in Racial/Ethnic and Socioeconomic Academic Achievement Gaps." in *Handbook of Research in Education Finance and Policy*, edited by Helen Ladd and Edward Fiske.

Rothstein, Richard. 2004. "A wider lens on the black-white achievement gap." *Phi Delta Kappan* October:104-10.

Vanneman, A., L. Hamilton, J. Baldwin Anderson, and T. Rahman. 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.

Wong, Manyee, Thomas D. Cook, and Peter M. Steiner. forthcoming. "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." in *Institute for Policy Research Working Paper Series*. Evanston, IL: Northwestern University.

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

	Cohort (Year of Kindergarten Entry)															
Grade	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006
State Data																
2										4	8	8	12	12	12	12
3					3	8	25	50	60	83	119	174	190	201	197	178
4		2	2	28	39	49	58	66	104	123	175	190	198	196	179	6
5			3	10	29	53	62	83	121	178	193	201	198	177	7	
6		8	12	32	40	46	69	86	176	192	202	201	178	5		
7	6	6	18	30	32	64	102	180	193	202	195	174	7			
8	37	46	72	85	116	143	176	192	202	199	175	6				
NAEP Data																
4	88		82		84		95	204		204		204		204		
8	82		95	204		204		204		204						

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 (black-white and Hispanic-white gaps) for each state.

Table 2: Estimated Achievement Gap Trends

		Black-White Gaps			Hispanic-White Gaps		
		Pooled Subjects	Math Only	Reading Only	Pooled Subjects	Math Only	Reading Only
All data (V)							
Base Model	Intercept	0.788 ***	0.853 ***	0.723 ***	0.648 ***	0.660 ***	0.639 ***
		(0.028)	(0.027)	(0.028)	(0.029)	(0.028)	(0.030)
	Cohort	-0.008 ***	-0.010 ***	-0.007 **	-0.009 ***	-0.010 ***	-0.007 ***
		(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
	Grade	-0.003	-0.003	-0.003	-0.007 ***	-0.004 *	-0.009 ***
		(0.002)	(0.002)	(0.003)	(0.002)	(0.002)	(0.003)
	Residual SD	0.065	0.059	0.053	0.069	0.055	0.064
	SD(Intercept)	0.179	0.175	0.187	0.196	0.189	0.207
	SD(cohort)	0.017	0.018	0.017	0.014	0.015	0.015
	SD(grade)	0.017	0.019	0.019	0.014	0.015	0.017
	N	5074	2533	2541	5075	2541	2534
State data (V)							
Base Model	Intercept	0.707 ***	0.742 ***	0.673 ***	0.594 ***	0.568 ***	0.620 ***
		(0.025)	(0.024)	(0.027)	(0.029)	(0.027)	(0.030)
	Cohort	-0.009 ***	-0.012 ***	-0.006 *	-0.010 ***	-0.012 ***	-0.008 *
		(0.003)	(0.003)	(0.003)	(0.002)	(0.002)	(0.003)
	Grade	0.002	0.002	0.002	-0.003	-0.001	-0.006
		(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.004)
	Residual SD	0.060	0.053	0.048	0.066	0.050	0.059
	SD(Intercept)	0.173	0.169	0.181	0.198	0.192	0.209
	SD(cohort)	0.018	0.017	0.018	0.017	0.015	0.022
	SD(grade)	0.019	0.021	0.021	0.018	0.016	0.025
	N	3995	1998	1997	3996	2006	1990
NAEP data (V)							
Base Model	Intercept	0.841 ***	0.936 ***	0.745 ***	0.684 ***	0.733 ***	0.640 ***
		(0.037)	(0.038)	(0.037)	(0.034)	(0.034)	(0.035)
	Cohort	-0.009 ***	-0.010 ***	-0.009 ***	-0.007 ***	-0.008 ***	-0.005 **
		(0.001)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
	Grade	-0.009 ***	-0.011 ***	-0.007 *	-0.007 **	-0.004	-0.010 **
		(0.002)	(0.002)	(0.003)	(0.002)	(0.003)	(0.003)
	Residual SD	0.074	0.070	0.069	0.065	0.061	0.067
	SD(Intercept)	0.259	0.263	0.259	0.240	0.235	0.241
	SD(cohort)	0.008	0.007	0.009	0.008	0.008	0.007
	SD(grade)	0.008	0.004	0.015	0.009	0.009	0.009
	N	1079	535	544	1079	535	544
NAEP data (d)							
Base Model	Intercept	0.828 ***	0.928 ***	0.728 ***	0.673 ***	0.724 ***	0.626 ***
		(0.036)	(0.038)	(0.036)	(0.033)	(0.033)	(0.034)
	Cohort	-0.009 ***	-0.010 ***	-0.008 ***	-0.007 ***	-0.007 ***	-0.005 **
		(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
	Grade	-0.010 ***	-0.011 ***	-0.007 *	-0.007 ***	-0.004	-0.010 ***
		(0.002)	(0.002)	(0.003)	(0.002)	(0.003)	(0.003)
	Residual SD	0.074	0.069	0.069	0.063	0.059	0.064
	SD(Intercept)	0.256	0.264	0.254	0.236	0.232	0.235
	SD(cohort)	0.008	0.008	0.010	0.008	0.008	0.007
	SD(grade)	0.009	0.004	0.015	0.009	0.009	0.010
	N	1079	535	544	1079	535	544
State data (proficiency gap)							
Base Model	Intercept	23.156 ***	24.649 ***	21.718 ***	19.320 ***	18.512 ***	20.114 ***
		(1.082)	(1.074)	(1.145)	(1.107)	(1.090)	(1.180)
	Cohort	-0.337 *	-0.444 **	-0.315 *	-0.599 ***	-0.646 ***	-0.563 ***
		(0.139)	(0.150)	(0.137)	(0.123)	(0.135)	(0.138)
	Grade	0.240 +	0.317 *	0.105	-0.120	0.053	-0.337 *
		(0.145)	(0.154)	(0.183)	(0.122)	(0.123)	(0.161)
	Residual SD	0.128	0.125	0.121	0.116	0.112	0.116
	SD(Intercept)	7.613	7.450	7.964	7.795	7.579	8.229
	SD(cohort)	0.884	0.890	0.759	0.752	0.774	0.762
	SD(grade)	0.776	0.576	0.879	0.545	0.319	0.655
	N	3990	1997	1993			

Robust standard errors are in parentheses. + p<.10; * p<.05; ** p<.01; *** p<.001.

Table 3: Estimated Assocation of White-Black Achievement Gaps With Years of Exposure to NCLB

	Pooled Subjects			Math Only			Reading Only		
	Pre-2003 Cohorts	Post-2002 Data	All Observations	Pre-2003 Cohorts	Post-2002 Data	All Observations	Pre-2003 Cohorts	Post-2002 Data	All Observations
All data (V)									
Base Model	0.007	-0.005	-0.006 *	0.005	-0.002	-0.005	0.011 *	-0.008 *	-0.008 **
	(0.004)	(0.004)	(0.003)	(0.005)	(0.005)	(0.004)	(0.006)	(0.003)	(0.003)
With Covariates	0.008	-0.004	-0.005	0.008	-0.002	-0.004	0.010 +	-0.007 *	-0.006 *
	(0.006)	(0.004)	(0.005)	(0.007)	(0.005)	(0.006)	(0.006)	(0.003)	(0.003)
NAEP data (V)									
Base Model	0.008 +		0.007	-0.003		-0.003	0.020 *		0.018 *
	(0.005)		(0.005)	(0.006)		(0.006)	(0.008)		(0.008)
With Covariates	0.008 +		0.008 +	-0.003		-0.003	0.016 *		0.018 *
	(0.004)		(0.004)	(0.006)		(0.006)	(0.008)		(0.008)
State data (V)									
Base Model	0.008	-0.002	-0.007 *	0.018 *	-0.001	-0.006	0.011	-0.006 +	-0.009 **
	(0.006)	(0.004)	(0.003)	(0.008)	(0.005)	(0.004)	(0.010)	(0.003)	(0.003)
With Covariates	0.011	-0.002	-0.006 *	0.020 +	-0.001	-0.006	0.014	-0.005	-0.007 *
	(0.007)	(0.004)	(0.003)	(0.012)	(0.005)	(0.004)	(0.010)	(0.003)	(0.003)
State data (proficiency gap)									
Base Model	0.037	-0.692 **	-0.978 ***	-0.572	-0.606 *	-0.987 ***	0.149	-0.841 **	-0.982 ***
	(0.417)	(0.265)	(0.225)	(0.537)	(0.273)	(0.275)	(0.435)	(0.308)	(0.254)
With Covariates	-0.002	-0.678 *	-0.949 ***	-0.674	-0.610 *	-0.989 ***	-0.004	-0.840 *	-0.943 ***
	(0.454)	(0.282)	(0.234)	(0.539)	(0.293)	(0.287)	(0.459)	(0.330)	(0.270)

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 4: Estimated Assocation of White-Hispanic Achievement Gaps With Years of Exposure to NCLB

	Pooled Subjects			Math Only			Reading Only		
	Pre-2003 Cohorts	Post-2002 Data	All Observations	Pre-2003 Cohorts	Post-2002 Data	All Observations	Pre-2003 Cohorts	Post-2002 Data	All Observations
All data (V)									
Base Model	-0.006	0.010 *	0.002	-0.002	0.012 **	0.005	-0.012	0.009 +	0.001
	(0.006)	(0.005)	(0.004)	(0.006)	(0.005)	(0.004)	(0.009)	(0.005)	(0.005)
With Covariates	-0.005	0.010 +	0.003	-0.001	0.012 *	0.005	-0.010	0.009	0.002
	(0.007)	(0.005)	(0.004)	(0.011)	(0.005)	(0.005)	(0.009)	(0.006)	(0.007)
NAEP data (V)									
Base Model	-0.001		0.000	-0.009		-0.008	0.008		0.009
	(0.006)		(0.005)	(0.008)		(0.008)	(0.009)		(0.007)
With Covariates	-0.002		0.001	-0.011		-0.008	0.009		0.011
	(0.006)		(0.005)	(0.008)		(0.008)	(0.009)		(0.007)
State data (V)									
Base Model	-0.012	0.014 **	0.006	0.004	0.015 ***	0.009 *	-0.016	0.013 *	0.003
	(0.012)	(0.005)	(0.005)	(0.011)	(0.004)	(0.004)	(0.015)	(0.005)	(0.005)
With Covariates	-0.006	0.014 **	0.006	0.009	0.015 **	0.009 +	-0.011	0.013 *	0.003
	(0.016)	(0.005)	(0.005)	(0.014)	(0.005)	(0.005)	(0.015)	(0.005)	(0.005)
State data (proficiency gap)									
Base Model	-0.889 *	-0.008	-0.359	-0.795	0.119	-0.199	-1.201 **	-0.087	-0.456
	(0.417)	(0.309)	(0.263)	(0.485)	(0.273)	(0.248)	(0.448)	(0.413)	(0.348)
With Covariates	-0.750 +	-0.011	-0.375	-0.759	0.083	-0.246	-1.090 *	-0.107	-0.505
	(0.415)	(0.310)	(0.268)	(0.512)	(0.275)	(0.252)	(0.456)	(0.414)	(0.353)

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 5. Estimated Association of White-Black Achievement Gaps With Years of NCLB Exposure and Its Interaction with Proportion of Black Students in Schools Subject to Accountability

	Pooled Subjects			Math Only			Reading Only		
	Pre-2003 Cohorts	Post-2002 Data	All Observations	Pre-2003 Cohorts	Post-2002 Data	All Observations	Pre-2003 Cohorts	Post-2002 Data	All Observations
All data (V)									
Exposure	0.035 ***	0.009	0.004	0.039 **	0.020 +	0.014	0.038 ***	0.002	-0.003
	(0.008)	(0.007)	(0.006)	(0.012)	(0.012)	(0.009)	(0.011)	(0.007)	(0.006)
Exposure*Proportion Accountable	-0.048 ***	-0.023 **	-0.016 *	-0.056 ***	-0.039 **	-0.033 **	-0.047 ***	-0.015	-0.005
	(0.012)	(0.009)	(0.008)	(0.017)	(0.014)	(0.012)	(0.013)	(0.011)	(0.009)
NAEP data (V)									
Exposure	0.031 **		0.028 *	0.026 +		0.028 *	0.037 *		0.026
	(0.012)		(0.012)	(0.014)		(0.014)	(0.019)		(0.019)
Exposure*Proportion Accountable	-0.037 *		-0.032 *	-0.045 *		-0.048 *	-0.031		-0.014
	(0.017)		(0.016)	(0.021)		(0.020)	(0.026)		(0.025)
State data (V)									
Exposure	0.030 **	0.010	0.000	0.052 **	0.016	0.006	0.027 +	0.006	-0.004
	(0.011)	(0.007)	(0.006)	(0.018)	(0.010)	(0.009)	(0.016)	(0.007)	(0.006)
Exposure*Proportion Accountable	-0.038 *	-0.021 *	-0.011	-0.056 *	-0.029 *	-0.021 +	-0.034 *	-0.017 +	-0.006
	(0.017)	(0.009)	(0.008)	(0.025)	(0.012)	(0.013)	(0.016)	(0.011)	(0.010)
State data (proficiency gap)									
Exposure	0.468	-0.266	-0.550 +	-0.609	0.217	-0.259	0.852	-0.612	-0.822 *
	(0.538)	(0.335)	(0.319)	(0.833)	(0.384)	(0.374)	(0.577)	(0.409)	(0.364)
Exposure*Proportion Accountable	-1.115 *	-0.728 +	-0.737 +	-0.608	-1.329 *	-1.220 *	-1.519 **	-0.427	-0.295
	(0.521)	(0.439)	(0.427)	(0.877)	(0.529)	(0.539)	(0.546)	(0.526)	(0.482)

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

	Pooled Subjects			Math Only			Reading Only		
	Pre-2003 Cohorts	Post-2002 Data	All Observations	Pre-2003 Cohorts	Post-2002 Data	All Observations	Pre-2003 Cohorts	Post-2002 Data	All Observations
All data (V)									
Exposure	0.000	0.017 **	0.011 +	-0.002	0.018 **	0.012 *	-0.003	0.016 *	0.009
	(0.010)	(0.006)	(0.006)	(0.011)	(0.006)	(0.006)	(0.012)	(0.007)	(0.007)
Exposure*Proportion Accountable	-0.014	-0.015	-0.016	-0.003	-0.014	-0.016	-0.020	-0.013	-0.014
	(0.017)	(0.012)	(0.013)	(0.019)	(0.012)	(0.012)	(0.018)	(0.011)	(0.011)
NAEP data (V)									
Exposure	-0.012		-0.005	-0.023		-0.013	0.006		0.005
	(0.010)		(0.010)	(0.016)		(0.016)	(0.017)		(0.017)
Exposure*Proportion Accountable	0.016		0.009	0.021		0.008	0.006		0.010
	(0.013)		(0.014)	(0.023)		(0.023)	(0.022)		(0.022)
State data (V)									
Exposure	-0.003	0.020 **	0.014	0.010	0.022 ***	0.019 ***	-0.011	0.014 *	0.008
	(0.016)	(0.006)	(0.009)	(0.016)	(0.006)	(0.005)	(0.017)	(0.007)	(0.006)
Exposure*Proportion Accountable	-0.011	-0.012	-0.017	-0.009	-0.015	-0.020 +	-0.012	-0.002	-0.010
	(0.025)	(0.012)	(0.019)	(0.019)	(0.011)	(0.011)	(0.019)	(0.011)	(0.010)
State data (proficiency gap)									
Exposure	-0.245	0.342	0.072	-0.508	0.205	0.001	-0.355	0.498	0.155
	(0.519)	(0.378)	(0.342)	(0.623)	(0.432)	(0.379)	(0.603)	(0.404)	(0.351)
Exposure*Proportion Accountable	-1.418	-0.796	-0.980	-0.770	-0.199	-0.473	-1.978 +	-1.383 *	-1.452 *
	(0.892)	(0.637)	(0.628)	(0.942)	(0.763)	(0.711)	(1.137)	(0.654)	(0.628)

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.

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

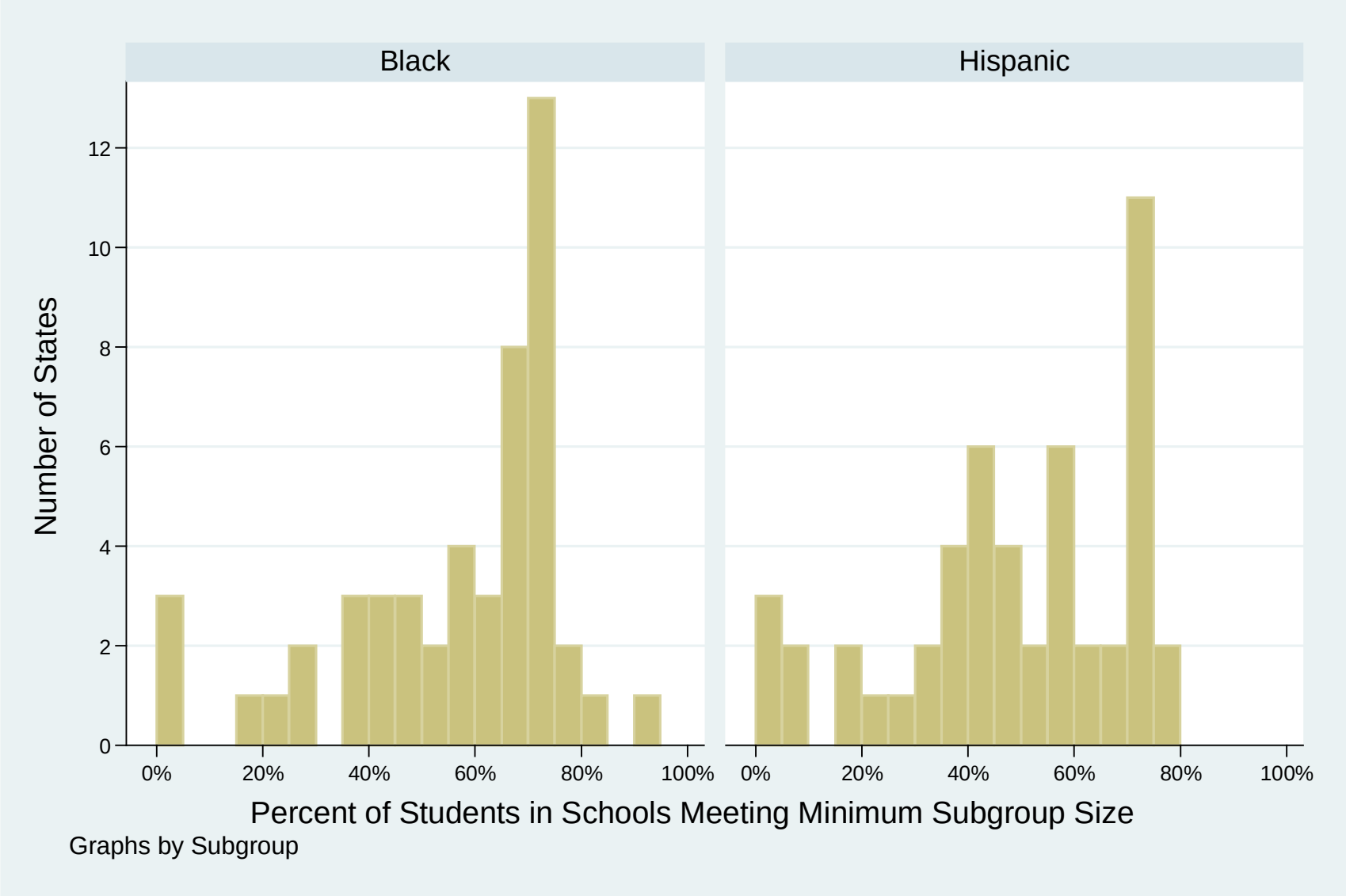


Figure 2: Exposure to NCLB, by cohort and grade

	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
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

T=0, N=0: Pre-2003 cohort; not subject to NCLB in current year

T=1, N=0: Pre-2003 cohort; subject to NCLB in current year

T=1, N=1: Post-2002 cohort; subject to NCLB in current year

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

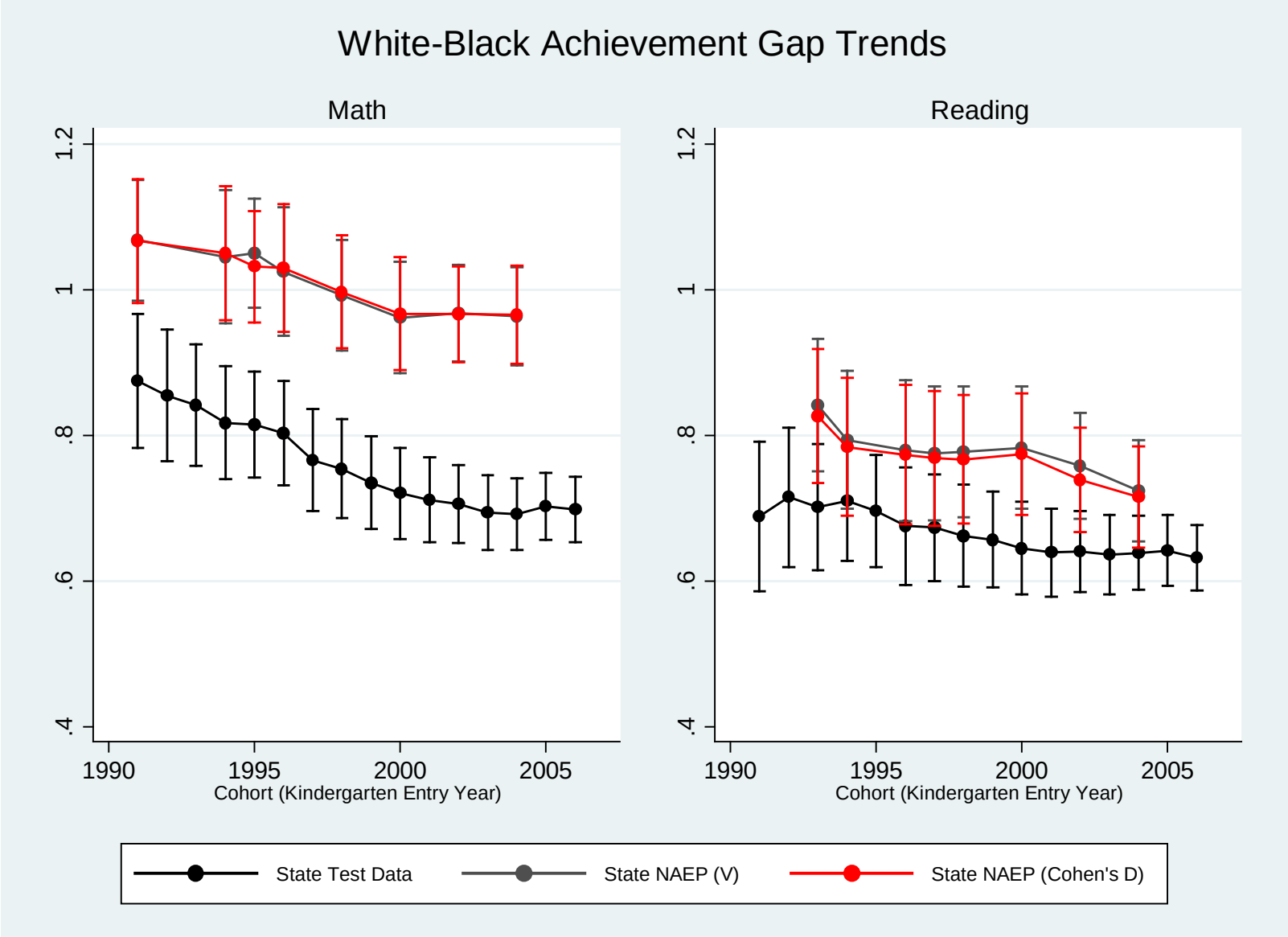


Figure 4: White-Hispanic Achievement Gap Trends, Math and Reading, 1991-2006 Cohorts

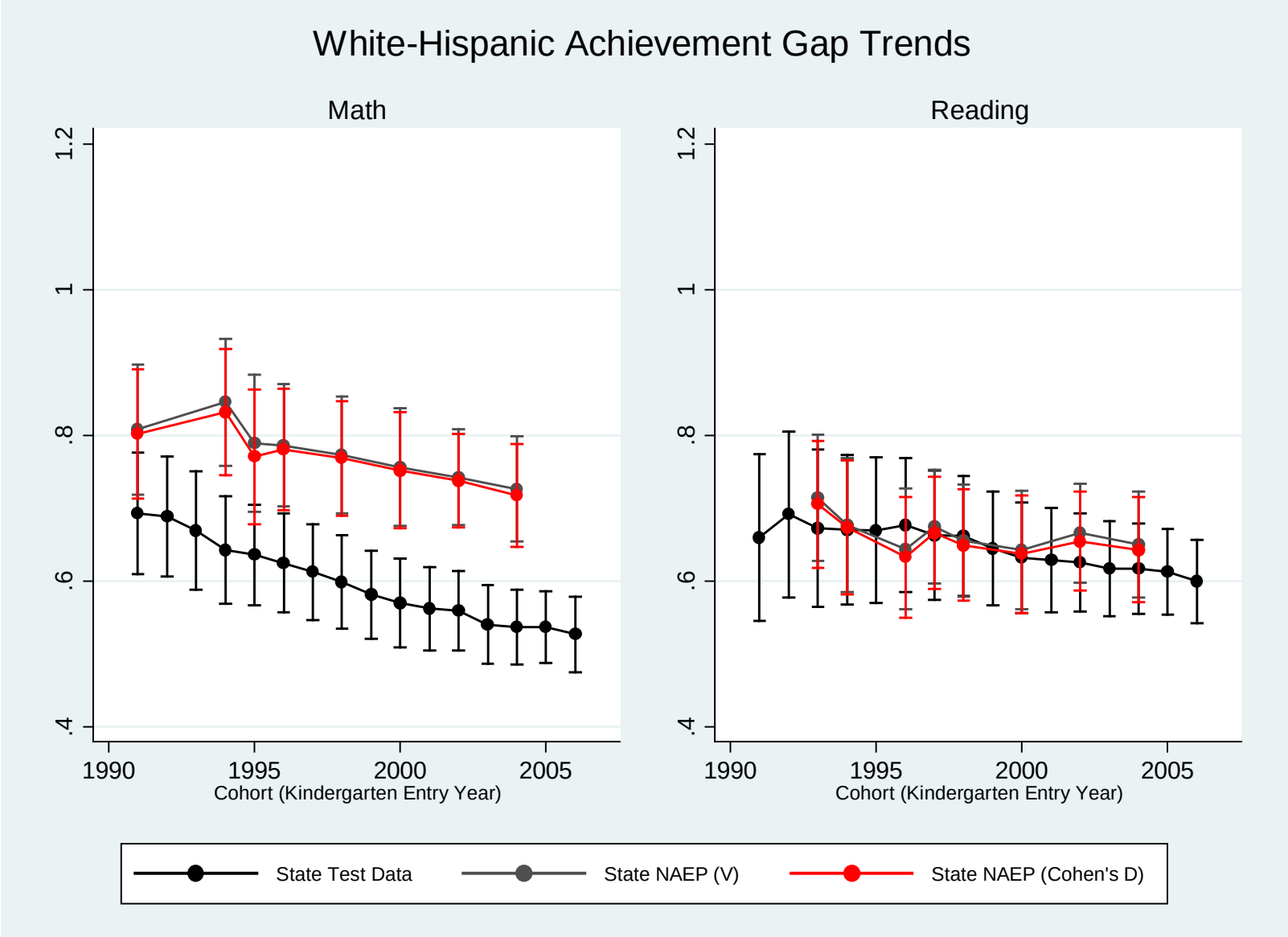


Figure 5: Estimated Annual Effect of NCLB on Achievement Gaps, by State (Empirical Bayes Estimates)

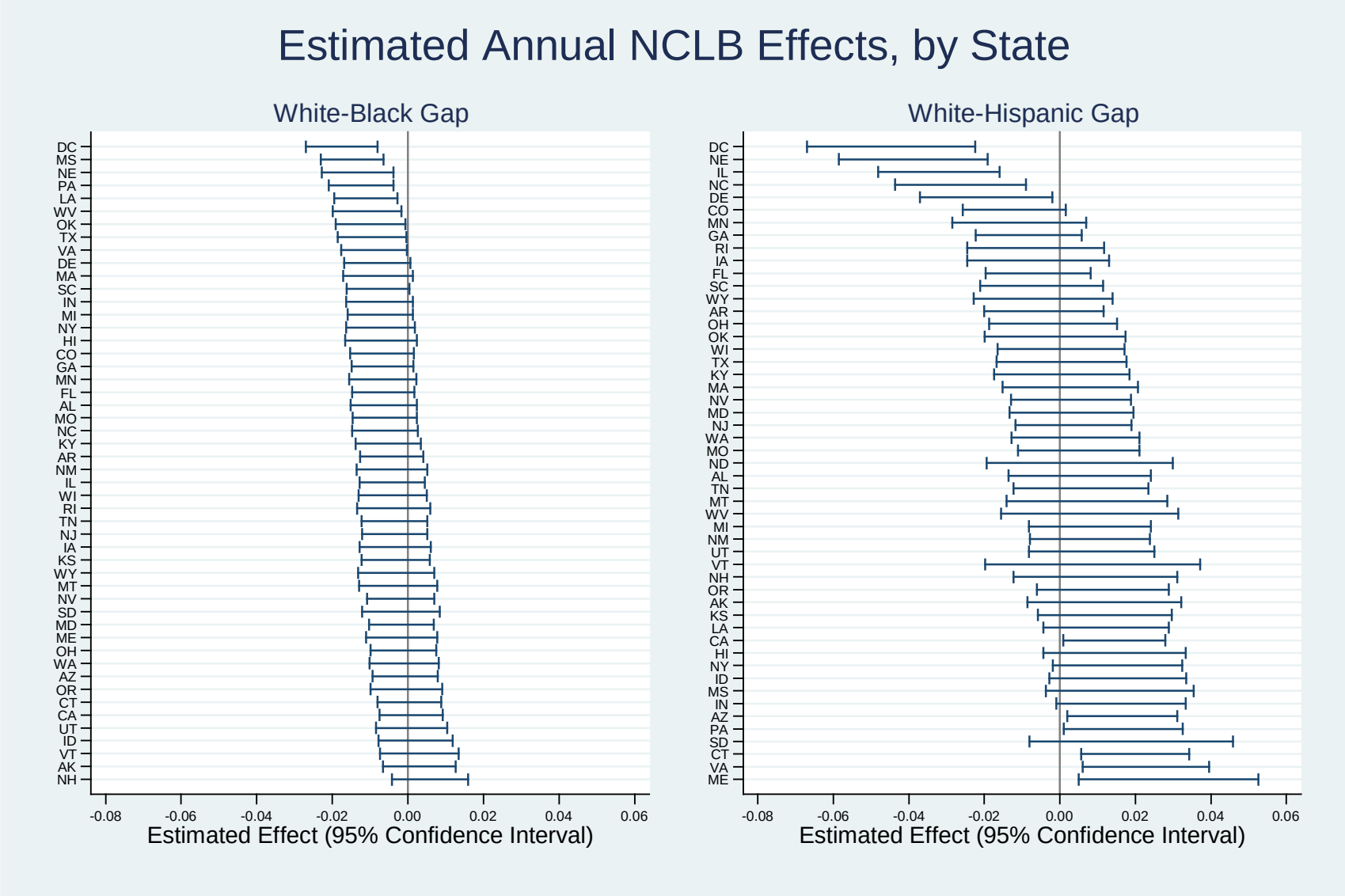


Figure 6: Estimated State-Specific NCLB Annual Effect on White-black Achievement Gap, by Proportion of Black Students in Schools Meeting State Minimum Subgroup Size Threshold
Estimates from data pooled across test subjects, data sources, and all cohorts/years

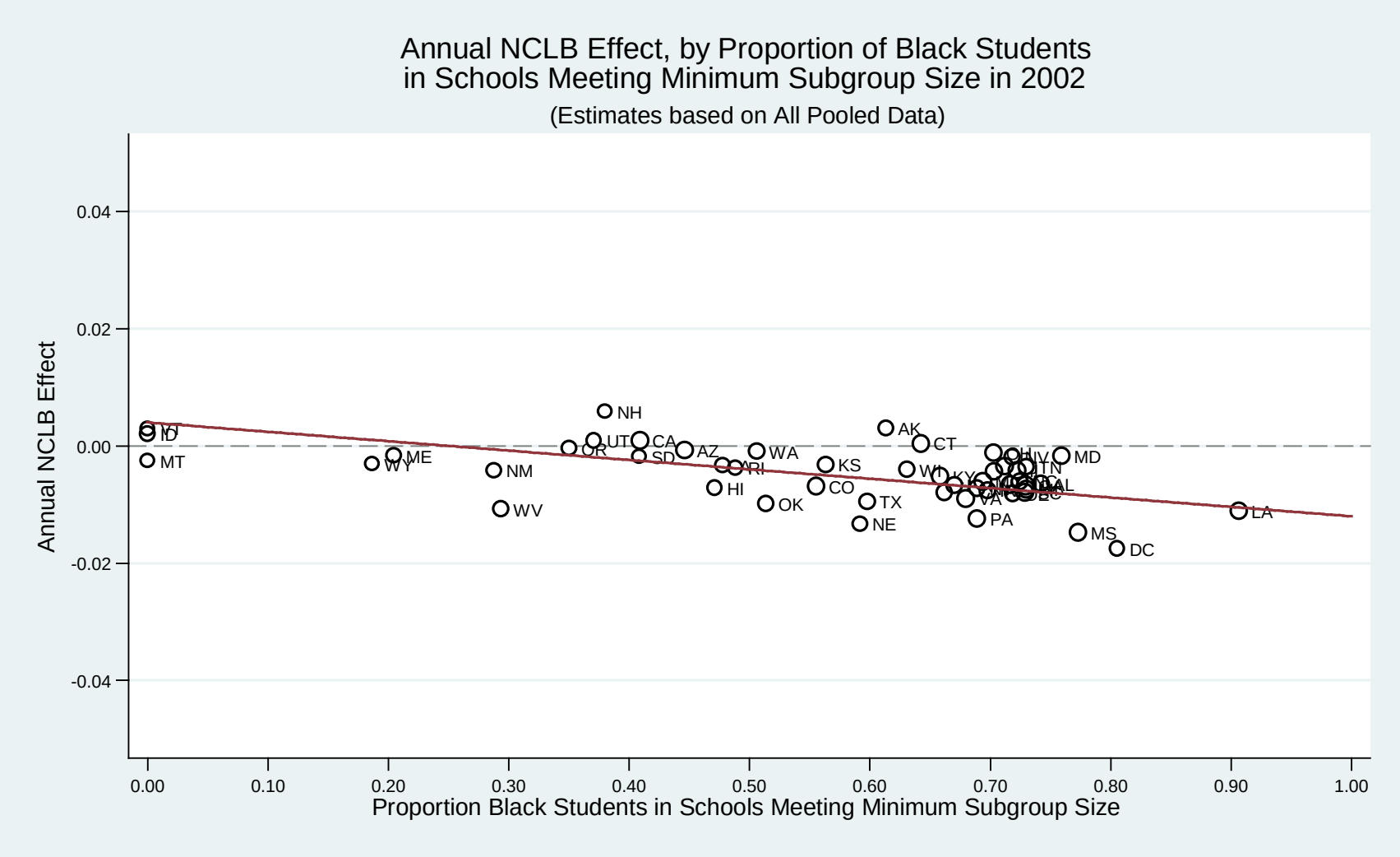


Figure 7: Estimated State-Specific NCLB Annual Effect on White-black Achievement Gap, by Proportion of Black Students in Schools Meeting State Minimum Subgroup Size Threshold
Estimates from pre-2003 cohorts, data pooled across test subjects and data sources

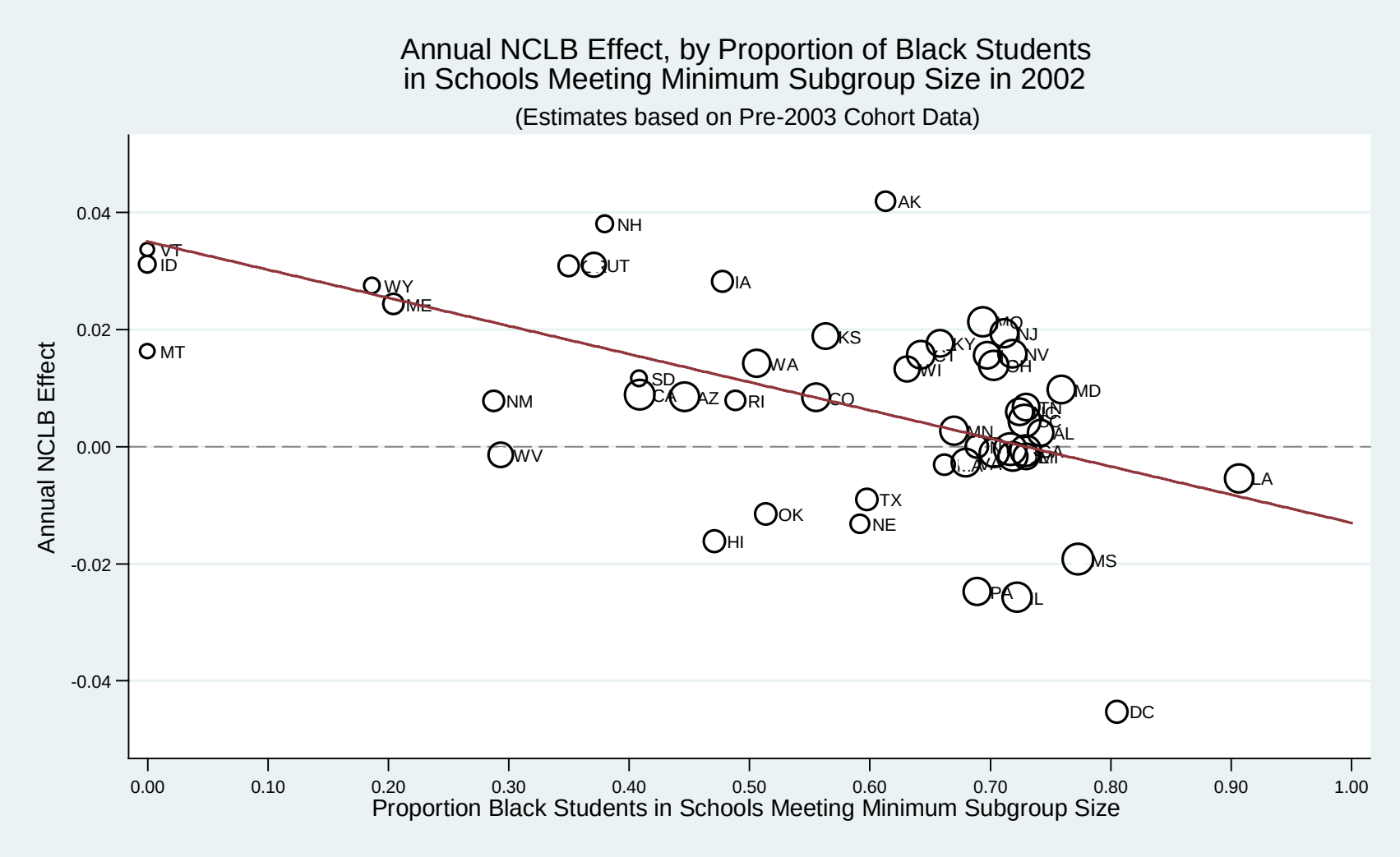
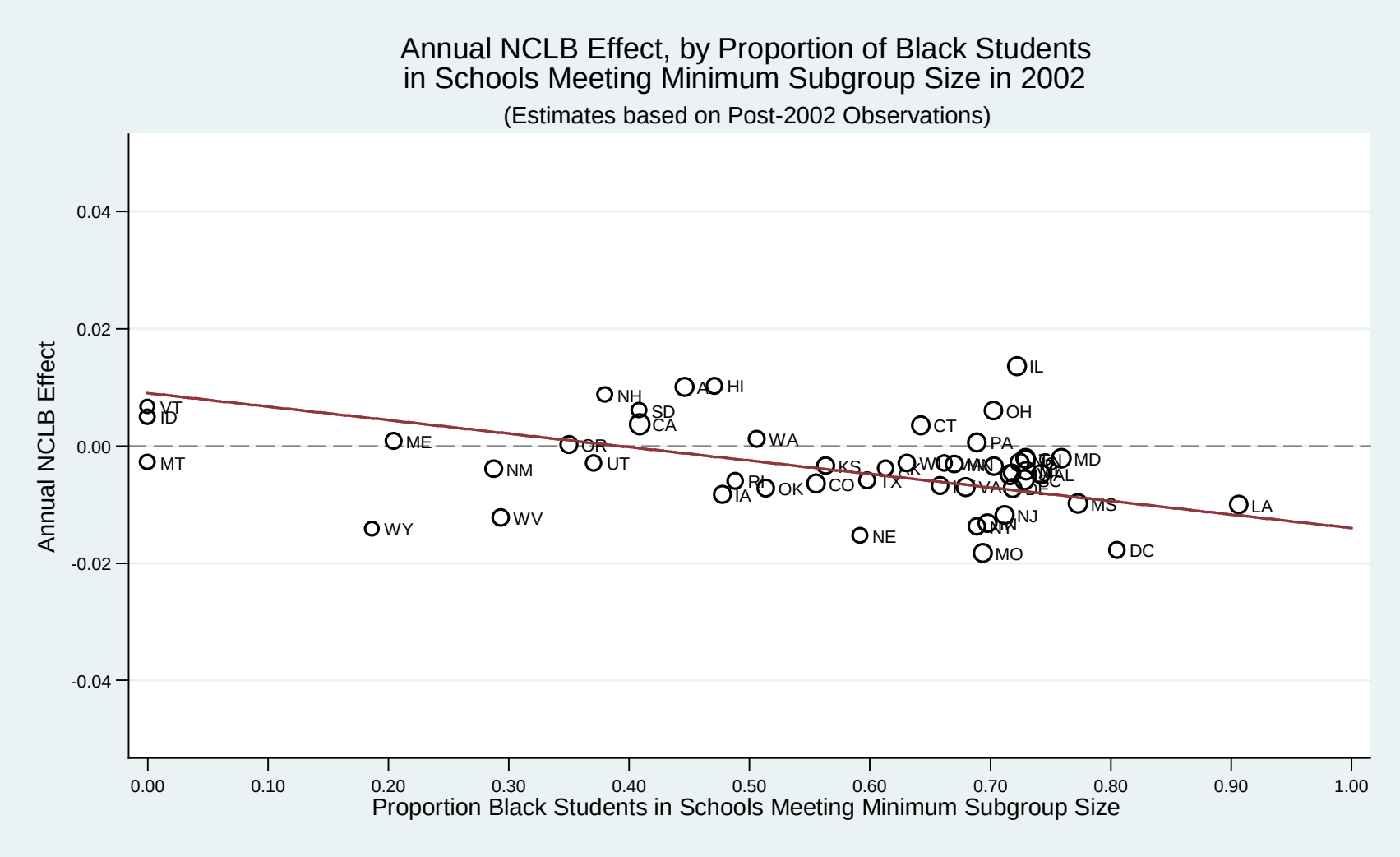


Figure 8: Estimated State-Specific NCLB Annual Effect on White-black Achievement Gap, by Proportion of Black Students in Schools Meeting State Minimum Subgroup Size Threshold
Estimates from post-2002 years, data pooled across test subjects and data sources



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 linear function 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}:

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

[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 γ_s is the linear 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 cohorts and grades (but not necessarily across states). A model that lets the effect of NCLB vary across grades would be

$$\delta_{csg} = \alpha + v_s + \beta(coh_c^*) + \eta(g) + \delta_{0s} T_{cg} + \delta_1\left(T_{cg} \cdot g\right) + \mathbf{w}_{csg}\mathbf{B} + e_{csg}.$$

[A5]

Here δ_{0s} is the NCLB effect on the gap during kindergarten in state s, and δ_1 is the average linear

change in the effect of NCLB across grades.

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 $expgr_{cg} = \sum_{k=0}^{g}(T_{ck} \cdot k)$. And fourth, 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

$$
\begin{aligned}
G_{csg} &= G_{cs(-1)} + \sum_{k=0}^{g} \delta_{csk} \\
&= [\lambda_s + \gamma_s(coh_c^*) + \mathbf{X}_{cs}\mathbf{A} + \nu_{cs}] \\
&\qquad + \sum_{k=0}^{g} [\alpha + v_s + \beta(coh_c^*) + \eta(k) + \delta_{0s} T_{ck} + \delta_1 (T_{ck} \cdot k) + \mathbf{w}_{csk}\mathbf{B} + e_{csk}] \\
&= [\lambda_s + \gamma_s(coh_c^*) + \mathbf{X}_{cs}\mathbf{A} + \nu_{cs}] + (g+1)\left(\alpha + v_s + \beta(coh_c^*)\right) + \eta\left(E_g\right) + \delta_{0s}\left(exp_{cg}\right) \\
&\qquad + \delta_1\left(expgr_{cg}\right) + \mathbf{W}_{csg}\mathbf{B} + \sum_{k=0}^{g} e_{csk} \\
&= \lambda_s + \gamma_s(coh_c^*) + \alpha_s\left(gr_g\right) + \beta\left(gr_g \cdot coh_c^*\right) + \eta\left(E_g\right) + \delta_{0s}\left(exp_{cg}\right) + \delta_1\left(expgr_{cg}\right) + \mathbf{X}_{cs}\mathbf{A} \\
&\qquad + \mathbf{W}_{csg}\mathbf{B} + e'_{csg}
\end{aligned}
\qquad \text{[A6]}
$$

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} and δ_1 by using a random coefficients model to regress G_{csg} on coh^*, gr, $gr \cdot coh^*$, E, $\mathbf{X}$, $\mathbf{W}$, and exp:[8]

$$
\begin{aligned}
\hat{G}_{csg} &= (\lambda + u_{\lambda s}) + \left(\gamma + u_{\gamma s}\right)(coh_c^*) + (\alpha + u_{\alpha s})\left(gr_g\right) + \beta\left(gr_g \cdot coh_c^*\right) + \eta\left(E_g\right) + \mathbf{X}_{cs}\mathbf{A} + \mathbf{W}_{csg}\mathbf{B} \\
&\qquad + (\delta + u_{\delta s})\left(exp_{cg}\right) + e'_{csg} + \epsilon_{csg} \\
&\qquad\qquad e'_{csg} \sim N[0, \sigma^2]
\end{aligned}
$$

[8] Note that we do not include the variable $expgr_{cg}$ in our model here for parsimony. We fit models including this term, but the coefficient on $expgr_{cg}$ was never significant in any model, so we have dropped it.

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

$$\begin{bmatrix} u_{\lambda s} \\ u_{\gamma s} \\ u_{\alpha s} \\ u_{\delta s} \end{bmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_\lambda & \tau_{\lambda\gamma} & \tau_{\lambda\alpha} & \tau_{\lambda\delta} \\ \tau_{\gamma\lambda} & \tau_\gamma & \tau_{\gamma\alpha} & \tau_{\gamma\delta} \\ \tau_{\alpha\lambda} & \tau_{\alpha\gamma} & \tau_\alpha & \tau_{\alpha\delta} \\ \tau_{\delta\lambda} & \tau_{\delta\gamma} & \tau_{\delta\alpha} & \tau_\delta \end{pmatrix}\right].$$

[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 is the key parameter of interest—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_{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 $\boldsymbol{\tau}$ matrix, using the HLM v7 software.

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:

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

[A8]

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

Figure 1: 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
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

T=0, N=0: Pre-2003 cohort; not subject to NCLB in current year
T=1, N=0: Pre-2003 cohort; subject to NCLB in current year
T=1, N=1: Post-2002 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):

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

[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.

Difference-in-Differences Models

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. The identification of δ in Model [A7] depends on the assumption that there is no other factor that affected all states' achievement gap trends in a similar way following 2002. As a check on this assumption, we adapt the approach used by Dee and Jacob (2011) and Wong, Cook, and Steiner (forthcoming), and compare the coefficient δ in states where we expect NCLB would have had a larger effect to those where it would have had a smaller effect. Specifically, we define P_s as the proportion of students of a subgroup in state s who were in schools meeting the minimum subgroup size reporting threshold, and fit the model

$$\hat{G}_{csg} = (\lambda_0 + \lambda_1 P_s + u_{\lambda s}) + (\gamma_0 + \gamma_1 P_s + u_{\gamma s})(coh_c^*) + (\alpha + u_{\alpha s})(gr_g) + \beta(gr_g \cdot coh_c^*) + \eta(E_g)$$
$$+ \mathbf{X}_{cs}\mathbf{A} + \mathbf{W}_{csg}\mathbf{B} + (\delta_0 + \delta_1 P_s + u_{\delta s})(exp_{cg}) + e'_{csg} + \epsilon_{csg}$$
$$e'_{csg} \sim N[0, \sigma^2]$$
$$\epsilon_{csg} \sim N[0, \omega_{csg}^2] = N[0, var(\hat{G}_{csg})]$$

$$\begin{bmatrix} u_{\lambda s} \\ u_{\gamma s} \\ u_{\alpha s} \\ u_{\delta s} \end{bmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_{\lambda} & \tau_{\lambda\gamma} & \tau_{\lambda\alpha} & \tau_{\lambda\delta} \\ \tau_{\gamma\lambda} & \tau_{\gamma} & \tau_{\gamma\alpha} & \tau_{\gamma\delta} \\ \tau_{\alpha\lambda} & \tau_{\alpha\gamma} & \tau_{\alpha} & \tau_{\alpha\delta} \\ \tau_{\delta\lambda} & \tau_{\delta\gamma} & \tau_{\delta\alpha} & \tau_{\delta} \end{pmatrix}\right]. \quad \text{[A11]}$$

Of interest in this model are the parameters δ_0 and δ_1.

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 51 states (including Washington, D.C.) 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 51 states (including Washington, D.C.) from the Center on Education Policy (CEP) website. 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.

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 X 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

To be added.

Appendix Tables and Figures

Table A1: Results Using Overlapping for NAEP and State Data (Pooled Subject, All Observations)

		Black-White Gaps NAEP Data (V)		Black-White Gaps State Data (V)		Hispanic-White Gaps NAEP Data (V)		Hispanic-White Gaps State Data (V)	
Estimated Achievement Gap Trends									
Base Model	Intercept	0.843	***	0.703	***	0.687	***	0.597	***
		(0.035)		(0.025)		(0.033)		(0.028)	
	Cohort	-0.007	**	-0.010	***	-0.006	*	-0.010	**
		(0.003)		(0.003)		(0.003)		(0.003)	
	Grade	-0.008	*	0.002		-0.008	*	-0.003	
		(0.003)		(0.003)		(0.004)		(0.004)	
	Residual SD	0.065		0.063		0.060		0.070	
	SD(Intercept)	0.246		0.173		0.235		0.197	
	SD(cohort)	0.013		0.017		0.011		0.019	
	SD(grade)	0.013		0.018		0.013		0.019	
	N	639		639		644		644	
Exposure Models									
With Covariates									
	Exposure	0.013		0.003		-0.001		0.011	
		(0.008)		(0.005)		(0.008)		(0.008)	
*Exposure*Proportion Accountable*									
With Covariates									
	Exposure	0.033	*	0.003		-0.002		0.027	*
		(0.016)		(0.012)		(0.015)		(0.011)	
	Exposure*	-0.031		0.000		0.005		-0.033	+
	Proportion Accountable	(0.022)		(0.016)		(0.023)		(0.018)	

Robust standard errors are in parentheses. + p<.10; * p<.05; ** p<.01; *** p<.001.