Left (Less Far) Behind?
Academic Achievement Gaps in the Era of No Child Left Behind

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

Stanford University

Draft: September 11, 2012

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.

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 "dosage" of NCLB that a cohort has experienced 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 dosage-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 lead 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 black-white test score gap now explains virtually all of the black-white difference in college enrollment (including enrollment at the most selective colleges and universities) and most or all of the black-white 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.

Evidence on the Effect of No Child Left Behind on 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. However, NCLB provides states with some latitude in determining the minimum subgroup size in a school required for reporting of disaggregated results; as a result, the proportion of minority students in schools whose test results are publicly disaggregated varies among states, which may lead to differential pressure among states to focus on narrowing achievement gaps. Indeed, 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.

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.

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)[1] 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

[1] 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.

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.

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.

Hypothesis 2: States in which more minority students are in schools that meet the state's minimum subgroup size reporting threshold will have smaller gaps.

NCLB allows states to set a minimum number of students in each demographic subgroup above which schools must report the test scores of that subgroup. In schools with populations of black or Hispanic students numbering less than the minimum subgroup size, the test scores of those groups are not reported and are not used (in their disaggregated form) to determine sanctions. As a result, such schools are not held accountable for the performance of students in these subgroups, except insofar as they contribute to overall school test performance. Therefore, 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.

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.[2]

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

[2] We do not estimate the effects of NCLB on proficiency gaps in this draft of the paper, but plan to do so in a future draft.

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 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. For the purposes of this paper, we look at only data reported in 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 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, but 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

[3] See Appendix for details on the estimation of V. [TO BE ADDED IN A FUTURE DRAFT]
[4] "State NAEP," the form used in this study, is taken by fourth- and eighth-graders every two years and includes a 50-state representative sample of around 2,500 students from approximately 100 schools in each state-grade-subject cohort.

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 focused attention on black and Hispanic students. 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—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 2003 (during the first year of NCLB implementation) where their group met the minimum subgroup size threshold. This proportion varies among states not only because of differences between states in the minimum subgroup size, but also because of between-state variation in the number of black or Hispanic students and in the levels of racial segregation among schools. We expect that, if the subgroup-specific reporting and sanctions of NCLB lead to narrower achievement gaps, we will see more rapid narrowing of gaps in states where large proportions of black or Hispanic students are in schools meeting the subgroup size thresholds than in states where fewer minority students are such schools. 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 black-white and Hispanic-white 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 black-white and Hispanic-white 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

The data described above report achievement in math and reading for blacks, Hispanics, and whites. Each data set has a distinct set of advantages, however. NAEP assessments provide a useful source of data because they have remained relatively unchanged over the last two decades, are low-stakes, and allow for cross-state comparison. NAEP data are limited because they are only available for a few grades and not every year; moreover, because they draw from a relatively small sample, it can be hard to discern significant differences across states and over time. State assessments are useful because they are administered to virtually all students in grades 2-8 each year and better reflect state curriculum and standards. They are limited because they are high stakes, are less likely to be commensurable across states, and yield outcomes in proficiency categories that are difficult to interpret. In addition, there are two additional potential objections to the use of state assessment data that relate to concerns about their internal validity.

In last decade, education researchers, policymakers, and practitioners have raised concerns regarding the "high stakes" nature of state assessments. Koretz (2001) has argued that there are two kinds of test preparation that might cause grade inflation, meaning "increases in scores that do not warrant the inference that students' mastery of the target of inference has improved by a commensurate amount" (p. 17). The least ambiguous form of grade inflation is cheating. This is a concern for high stakes tests because the incentives for teachers and administrators to do so are high when tests are used to enforce accountability standards that either reward or punish schools. In addition to the recent cheating scandal that took place in Atlanta, GA (Education Week, 2011), Jacob & Levitt (forthcoming) found that cheating occurred in at least 4-5 percent of elementary school classrooms each year in the Chicago Public Schools. While this might be a concern, particularly in some districts or states, the evidence remains inconclusive regarding the national prevalence of cheating.

The latter form of grade inflation is usually categorized as "teaching to the test." This is more ambiguous (Jacob 2007; Koretz 2005; Koretz 2002; Popham 2001). Teaching to the test could include organizing instruction around specific types of test-items, or reallocating instruction represented by a particular test type. There seems to be little empirical research on this topic (see Jacob, 2007 for a notable exception), and in fact it may be difficult to test for such a phenomenon. Nonetheless, if teachers are targeting instruction to the types of questions being tested, then the tests are not a true representation of the latent construct. Still, gains in tests that result from teaching to the test might reflect meaningful improvement. For example, perhaps lower-order, procedural skills are easier to target with instruction, and are thus more susceptible to such techniques. We still have reason to care about improvement on such skills, as they are important prerequisites to the development of other cognitive skills.

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 (gr_g), cohort (coh_c), and the number of years the cohort has been exposed to NCLB by grade g (exp_{cg}):

$$G_{csg} = \lambda_s + \alpha_s(gr_g) + \eta(gr_g^2) + \beta(gr_g \cdot coh_c) + \gamma_s(coh_c) + \delta_s(exp_{cg}) + e_{csg}. \tag{1}$$

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:

$$\begin{gathered} \hat{G}_{csgt} = \lambda_s + \alpha_s(gr_g) + \eta(gr_g^2) + \beta(gr_g \cdot coh_c) + \gamma_s(coh_c) + \zeta(sub_t) + \delta_s(exp_{cg}) + e_{csgt} + \epsilon_{csgt}, \\ e_{csgt} \sim N[0, \sigma^2] \\ \epsilon_{vsgt} \sim N[0, \omega_{csgt}^2] = N[0, var(\hat{G}_{csgt})] \\ \begin{bmatrix} \lambda_s \\ \gamma_s \\ \alpha_s \\ \delta_s \end{bmatrix} \sim N\left[\begin{pmatrix} \lambda \\ \gamma \\ \alpha \\ \delta \end{pmatrix}, \begin{pmatrix} \tau_\lambda & \tau_{\lambda\gamma} & \tau_{\lambda\alpha} & \tau_{\lambda\delta} \\ \tau_{\gamma\lambda} & \tau_\gamma & \tau_{\gamma\alpha} & \tau_{\gamma\delta} \\ \tau_{\alpha\lambda} & \tau_{\alpha\gamma} & \tau_\alpha & \tau_{\alpha\delta} \\ \tau_{\delta\lambda} & \tau_{\delta\gamma} & \tau_{\delta\alpha} & \tau_\delta \end{pmatrix}\right] \end{gathered} \tag{2}$$

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; and exp_{cg} is the number of years that cohort c has been exposed to NCLB by the spring of grade g. The key parameter of interest is δ, the average annual effect of NCLB on the achievement gap within a cohort. The error term ϵ_{csgt} is the sampling error of $\hat{G}_{csgt}$; we set its variance ω^2_{csgt} to be equal to the square of the standard error of $\hat{G}_{csgt}$. We estimate the parameters of this model, as well as σ^2 and the $\boldsymbol{\tau}$ matrix, using 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 2003, and then increases linearly across grades (within a cohort) or across cohorts (within a grade) after 2002. 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 2002; equivalently, δ is the average within-state difference in the trend in the achievement gap across cohorts within a grade before and after 2002. Second, for years after 2002, $exp_{cg} = coh^*_c + gr_g$ for cohorts entering kindergarten prior to 2003, but $exp_{cg} = gr_g$ for later cohorts. Thus, after 2002, δ is the average within-state difference in the trend in the achievement gap across cohorts within a grade between pre-2003 cohorts and later cohorts.

Figure A1 in the Appendix helps to clarify these different sources of variation in exp_{cg}. 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. Because almost all of the available NAEP data fall in the yellow and green regions of Figure 1 (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.

We fit several versions of model (2), each using different subsets of our data. First, we fit the model using data from NAEP, limiting the data to pre-2003 cohorts. Second, we fit the model using state test data, limiting the data to pre-2003 cohorts, post-2002 data, and then using all state data. When using the state data, we estimate models using V as a measure of the gap and using the difference in proficiency rates, to assess whether NCLB affected distributional or proficiency gaps. We also fit the models for math and reading separately. In a final set of models, we pool all available NAEP and state test data, including both math and reading tests. In an additional set of models, we include a set of 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.

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. Our second set of models does not rely on this assumption. In these models, we adapt the approach used by Dee and Jacob (2011) and Wong, Cook, and Steiner (forthcoming), and test whether the coefficient δ_s differs between states where we expect NCLB would have had a larger effect and those where it would have had a smaller effect. Specifically, we interact exp_{cg} with a variable indicating the proportion of black (or Hispanic) students in Spring

2003 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, which would be 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. Third, both white-black and Hispanic-white achievement gaps have been narrowing, albeit slowly, over the last two decades; this pattern is consistent across each of the three different gap measures.

[5] The trends displayed in Figures 1-2 indicate the trend in the estimated cohort fixed effects (the $\hat{\Gamma}_s$'s) from the model $G_{csg} = \lambda_s + \alpha_s(grade_g - 4) + u_{\gamma s}(cohort_c - 2002) + \Gamma_c + e_{csg}$, where $\lambda_s \sim N[\lambda, \tau_\lambda]$; $\alpha_s \sim N[\alpha, \tau_\alpha]$; and $u_{\gamma s} \sim N[0, \tau_\gamma]$.

Figures 1-2 here

Figures 1-2 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 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 1 reports the results of these models. The table shows estimates from 24 models, each using different combinations of data sources, measures, and test subjects. Across all 24 models, however, the estimated cohort slope is always negative and statistically significant, ranging from estimates of -0.005 standard deviations per year to -0.011 standard deviations per year. 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 -0.007 (se = 0.002, $p < .001$) and -0.008 (se = 0.001, $p < .001$) standard deviations per year, respectively. At this rate, the white-black and white-Hispanic gaps will be eliminated in roughly 80-100 years.

Table 1 here

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.

Tables 2-3 here

The bottom left panels of Tables 2 and 3 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 ($\hat{\delta} = -0.004, se = 0.003$ for white-black gap; and $\hat{\delta} = 0.004, se = 0.005$ 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 2, there is some evidence that NCLB *widened* white-black achievement gaps (based on the generally positive, and often 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 3, 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 2 nor 3, 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 2 and 3 to test whether the effect of NCLB changes across grade levels (see Appendix 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 2 and 3 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 0.006 and 0.015, 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.

In Tables 4 and 5 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 4-5 here

Several patterns are evident in the coefficients reported in Table 4. 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, ND, 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. In some ways, this is consistent with the incentives inherent in NCLB: if a state has many white students, but few black students, whose test scores matter for accountability purposes, there may be a greater incentive to focus on improving test scores of white students, thereby widening achievement gaps. Conversely, if a substantial majority of black students are in schools where their scores are consequential, there may be a greater incentive to improve black students' performance, leading to narrower gaps. The results in Table 4 are broadly consistent with this pattern: the main effect of exposure is almost always in the direction of increasing achievement gaps, while the interaction term is always negative, and generally of equal or larger magnitude than the main effect, implying the effect of NCLB is to reduce gaps when most black students are in schools where there scores are consequential.

Interestingly, the story is quite different in Table 5, which shows the coefficients from the same models fit to the white-Hispanic gaps. In Table 5, the interaction term is never significant. Moreover, the main effect of exposure to NCLB is generally not associated with the size of the achievement gaps, except in the post-2002 data models, where the association is positive, similar to the results in Table 3. 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.

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.

Jacob, B.A. 2007. "Test-based accountability and student achievement: an investigation of differential performance on NAEP and state assessments." National Bureau of Economic Research.

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.

Koretz, D. 2005. "Alignment, high stakes, and the inflation of test scores." *Yearbook of the National Society for the Study of Education* 104(2):99-118.

Koretz, D.M. 2002. "Limitations in the use of achievement tests as measures of educators' productivity." *Journal of human resources*:752-77.

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.

Popham, W.J. 2001. "Teaching to the Test?" *Educational Leadership* 58(6):16-21.

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. HLM Models of Achievement Gap Trends

			Black-White Gaps			Hispanic-White Gaps		
			Pooled Subjects	Math Only	Reading Only	Pooled Subjects	Math Only	Reading Only
NAEP data (V)								
	Base Model	Intercept	0.820 ***	0.916 ***	0.726 ***	0.665 ***	0.714 ***	0.623 ***
			(0.029)	(0.030)	(0.029)	(0.028)	(0.028)	(0.029)
		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	-0.009 ***	-0.011 ***	-0.006 +	-0.006 **	-0.003	-0.010 ***
			(0.002)	(0.002)	(0.003)	(0.002)	(0.003)	(0.003)
		Residual SD	0.073	0.069	0.068	0.064	0.059	0.067
		SD(Intercept)	0.198	0.203	0.199	0.196	0.189	0.199
		SD(cohort)	0.008	0.008	0.009	0.007	0.008	0.007
		SD(grade)	0.008	0.004	0.014	0.008	0.010	0.008
		N	1057	524	533	1057	524	533
NAEP data (d)								
	Base Model	Intercept	0.808 ***	0.907 ***	0.710 ***	0.654 ***	0.704 ***	0.610 ***
			(0.028)	(0.029)	(0.029)	(0.028)	(0.027)	(0.029)
		Cohort	-0.009 ***	-0.009 ***	-0.008 ***	-0.006 ***	-0.007 ***	-0.005 **
			(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
		Grade	-0.010 ***	-0.012 ***	-0.007 *	-0.007 **	-0.003	-0.010 ***
			(0.002)	(0.002)	(0.003)	(0.002)	(0.003)	(0.003)
		Residual SD	0.073	0.068	0.069	0.062	0.057	0.064
		SD(Intercept)	0.196	0.201	0.197	0.193	0.185	0.197
		SD(cohort)	0.009	0.008	0.010	0.008	0.008	0.007
		SD(grade)	0.008	0.004	0.015	0.008	0.010	0.009
		N	1057	524	533	1057	524	533
State data (V)								
	Base Model	Intercept	0.693 ***	0.729 ***	0.659 ***	0.581 ***	0.555 ***	0.606 ***
			(0.021)	(0.021)	(0.022)	(0.025)	(0.024)	(0.028)
		Cohort	-0.008 ***	-0.011 ***	-0.005 *	-0.009 ***	-0.011 ***	-0.007 *
			(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.003)
		Grade	0.003	0.002	0.002	-0.003	-0.001	-0.004
			(0.003)	(0.003)	(0.003)	(0.002)	(0.002)	(0.003)
		Residual SD	0.060	0.053	0.048	0.065	0.050	0.059
		SD(Intercept)	0.144	0.142	0.149	0.175	0.171	0.183
		SD(cohort)	0.015	0.015	0.015	0.012	0.012	0.017
		SD(grade)	0.018	0.021	0.020	0.016	0.016	0.021
		N	3889	1944	1945	3890	1952	1938
All data (V)								
	Base Model	Intercept	0.771 ***	0.837 ***	0.706 ***	0.633 ***	0.646 ***	0.623 ***
			(0.022)	(0.022)	(0.023)	(0.025)	(0.025)	(0.027)
		Cohort	-0.007 ***	-0.009 ***	-0.006 **	-0.008 ***	-0.009 ***	-0.006 ***
			(0.002)	(0.002)	(0.002)	(0.001)	(0.002)	(0.002)
		Grade	-0.003	-0.003	-0.003	-0.006 ***	-0.004	-0.008 ***
			(0.002)	(0.002)	(0.003)	(0.002)	(0.002)	(0.002)
		Residual SD	0.064	0.059	0.053	0.069	0.055	0.064
		SD(Intercept)	0.147	0.145	0.152	0.171	0.167	0.179
		SD(cohort)	0.016	0.016	0.015	0.011	0.013	0.011
		SD(grade)	0.016	0.019	0.018	0.013	0.015	0.015
		N	4946	2468	2478	4947	2476	2471

Table 2. HLM Models of Black-White Achievement Gaps by Years of NCLB Exposure

	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
NAEP data									
Base Model	0.009 *		0.008 +	0.000		-0.001	0.021 **		0.018 *
	(0.005)		(0.004)	(0.006)		(0.006)	(0.008)		(0.008)
With Covariates	0.008 +		0.008 +	-0.002		-0.002	0.017 *		0.018 *
	(0.004)		(0.004)	(0.006)		(0.006)	(0.008)		(0.008)
State data									
Base Model	0.008	-0.002	-0.007 *	0.020 **	-0.001	-0.007	0.013	-0.005	-0.008 **
	(0.005)	(0.004)	(0.003)	(0.008)	(0.005)	(0.004)	(0.009)	(0.003)	(0.003)
With Covariates	0.010 +	-0.001	-0.006 +	0.021 *	-0.001	-0.006	0.018 +	-0.004	-0.006 *
	(0.005)	(0.004)	(0.003)	(0.009)	(0.005)	(0.004)	(0.010)	(0.003)	(0.003)
All data									
Base Model	0.007 +	-0.004	-0.005 *	0.007	-0.002	-0.005	0.012 *	-0.008 *	-0.007 *
	(0.004)	(0.004)	(0.003)	(0.005)	(0.005)	(0.004)	(0.005)	(0.003)	(0.003)
With Covariates	0.009 *	-0.003	-0.004	0.009	-0.001	-0.004	0.013 *	-0.006 +	-0.005 +
	(0.004)	(0.004)	(0.003)	(0.005)	(0.005)	(0.004)	(0.005)	(0.003)	(0.003)

Each cell indicates the coefficient on the variable indicating the number of years of exposure to NCLB. Each coefficient is from a separate model.

Table 3. HLM Models of Hispanic-White Achievement Gaps by Years of NCLB Exposure

	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
NAEP data									
Base Model	-0.002		-0.001	-0.010		-0.009	0.005		0.007
	(0.006)		(0.005)	(0.008)		(0.007)	(0.009)		(0.007)
With Covariates	-0.001		0.002	-0.008		-0.006	0.010		0.011
	(0.006)		(0.005)	(0.008)		(0.007)	(0.009)		(0.007)
State data									
Base Model	-0.010	0.015 ***	0.008 +	0.006	0.016 ***	0.010 *	-0.012	0.014 **	0.004
	(0.012)	(0.004)	(0.004)	(0.011)	(0.004)	(0.004)	(0.015)	(0.005)	(0.005)
With Covariates	-0.006	0.014 *	0.007	0.011	0.014 **	0.009 +	-0.010	0.013 *	0.003
	(0.013)	(0.007)	(0.005)	(0.010)	(0.005)	(0.005)	(0.015)	(0.005)	(0.005)
All data									
Base Model	-0.004	0.012 **	0.004	0.000	0.013 **	0.006	-0.011	0.010 *	0.003
	(0.005)	(0.004)	(0.004)	(0.006)	(0.004)	(0.004)	(0.008)	(0.005)	(0.005)
With Covariates	-0.003	0.011 *	0.004	0.001	0.013 **	0.006	-0.008	0.010 +	0.003
	(0.005)	(0.005)	(0.005)	(0.006)	(0.004)	(0.004)	(0.009)	(0.005)	(0.005)

Each cell indicates the coefficient on the variable indicating the number of years of exposure to NCLB. Each coefficient is from a separate model.

Table 4. Estimated Association of White-Black Achievement Gaps With Years of NCLB Exposure and Its Interaction with Proportion of Black Student 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
NAEP data									
Exposure	0.033 ***		0.029 **	0.010		0.011	0.055 ***		0.045 **
	(0.010)		(0.009)	(0.011)		(0.011)	(0.016)		(0.016)
Exposure*Proportion Accountable	-0.042 **		-0.036 **	-0.024		-0.025	-0.059 **		-0.044 *
	(0.013)		(0.013)	(0.016)		(0.015)	(0.019)		(0.019)
State data									
Exposure	0.020 *	0.006	-0.001	0.027 *	0.006	-0.001	0.023 *	0.003	-0.002
	(0.008)	(0.006)	(0.005)	(0.011)	(0.007)	(0.006)	(0.011)	(0.006)	(0.006)
Exposure*Proportion Accountable	-0.022 *	-0.013 +	-0.009	-0.022	-0.013 +	-0.009	-0.025 *	-0.013	-0.009
	(0.011)	(0.007)	(0.006)	(0.013)	(0.007)	(0.007)	(0.011)	(0.009)	(0.009)
All data									
Exposure	0.026 ***	0.006	0.005	0.021 **	0.010	0.006	0.032 ***	0.002	0.002
	(0.006)	(0.005)	(0.005)	(0.008)	(0.007)	(0.006)	(0.009)	(0.006)	(0.006)
Exposure*Proportion Accountable	-0.035 ***	-0.017 **	-0.016 **	-0.032 **	-0.021 **	-0.021 **	-0.036 ***	-0.014	-0.013
	(0.008)	(0.006)	(0.006)	(0.011)	(0.007)	(0.007)	(0.010)	(0.009)	(0.009)

Note: All models include controls for grade, cohort, and time-varying economic and school composition and segregation covariates.

Table 5. Estimated Association of White-Hispanic Achievement Gaps With Years of NCLB Exposure and Its Interaction with Proportion of Hispanic Student 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
NAEP data									
Exposure	-0.001		0.001	-0.008		-0.003	0.008		0.005
	(0.010)		(0.010)	(0.015)		(0.015)	(0.015)		(0.014)
Exposure*Proportion	0.002		0.000	-0.001		-0.008	0.005		0.010
Accountable	(0.013)		(0.012)	(0.020)		(0.019)	(0.020)		(0.019)
State data									
Exposure	-0.013	0.015 **	0.009 +	0.000	0.014 **	0.011 *	-0.017	0.012 +	0.005
	(0.013)	(0.005)	(0.005)	(0.013)	(0.005)	(0.004)	(0.016)	(0.007)	(0.006)
Exposure*Proportion	0.009	0.001	-0.002	0.015	0.002	-0.001	0.004	0.003	-0.003
Accountable	(0.009)	(0.007)	(0.006)	(0.011)	(0.007)	(0.007)	(0.012)	(0.009)	(0.008)
All data									
Exposure	-0.006	0.012 *	0.005	-0.004	0.013 **	0.006	-0.011	0.010	0.004
	(0.007)	(0.005)	(0.004)	(0.009)	(0.005)	(0.004)	(0.010)	(0.006)	(0.006)
Exposure*Proportion	0.003	0.000	-0.001	0.007	0.001	-0.001	0.001	0.000	-0.001
Accountable	(0.008)	(0.007)	(0.006)	(0.011)	(0.007)	(0.007)	(0.010)	(0.008)	(0.008)

Note: All models include controls for grade, cohort, and time-varying economic and school composition and segregation covariates.

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

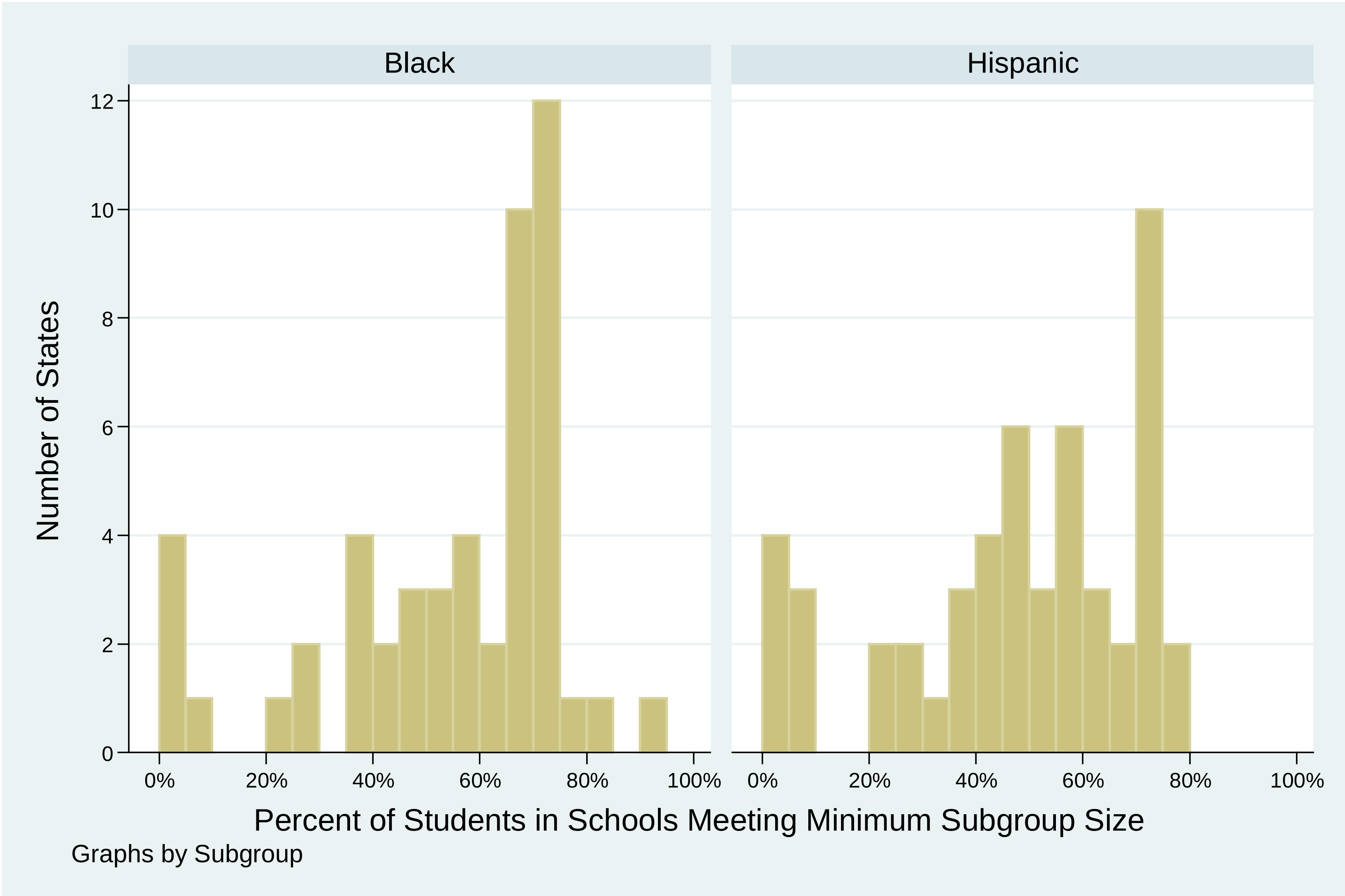


Figure 2: White-Black Achievement Gap Trends, Math and Reading, 1991-2006 Cohorts

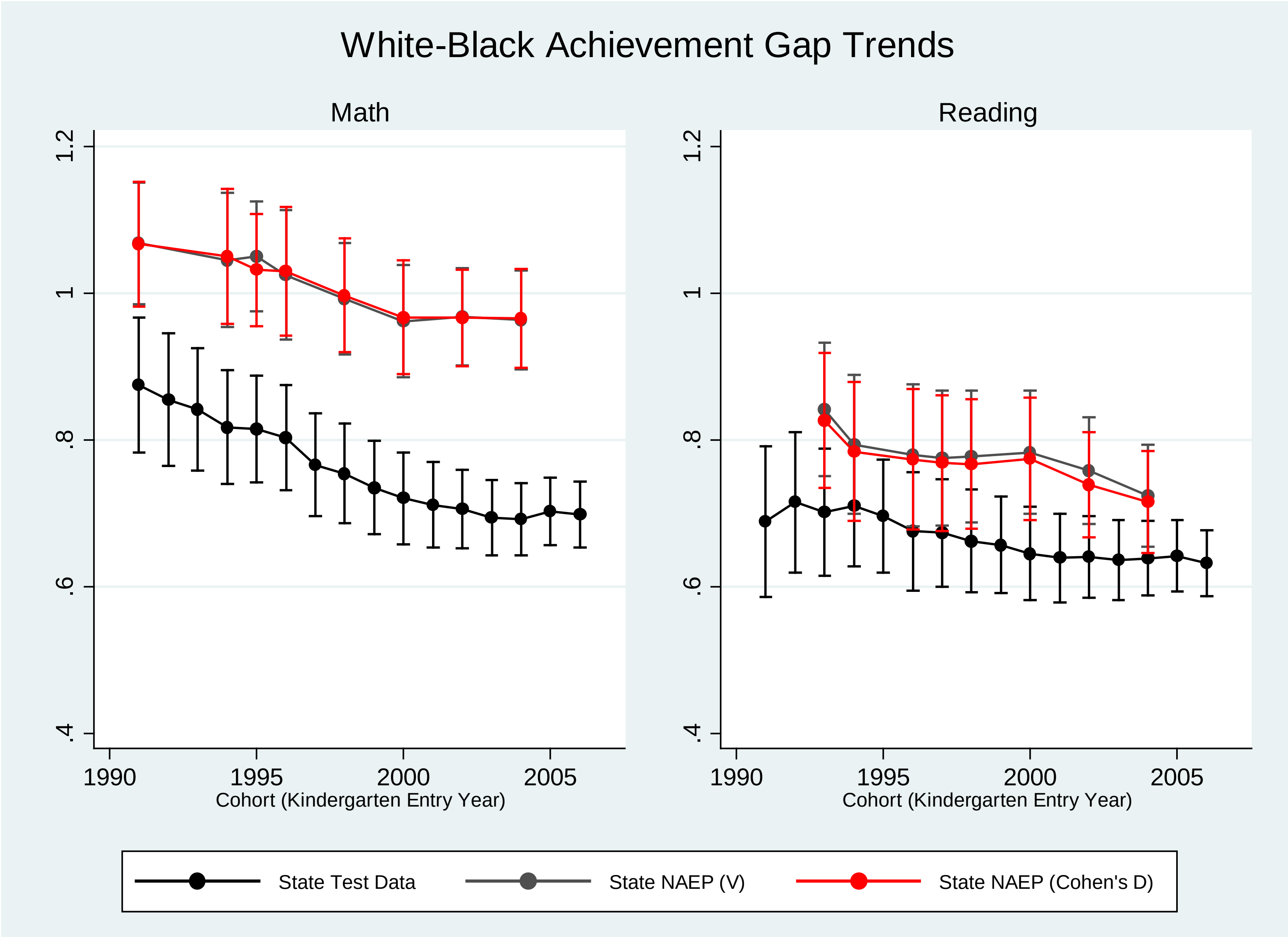
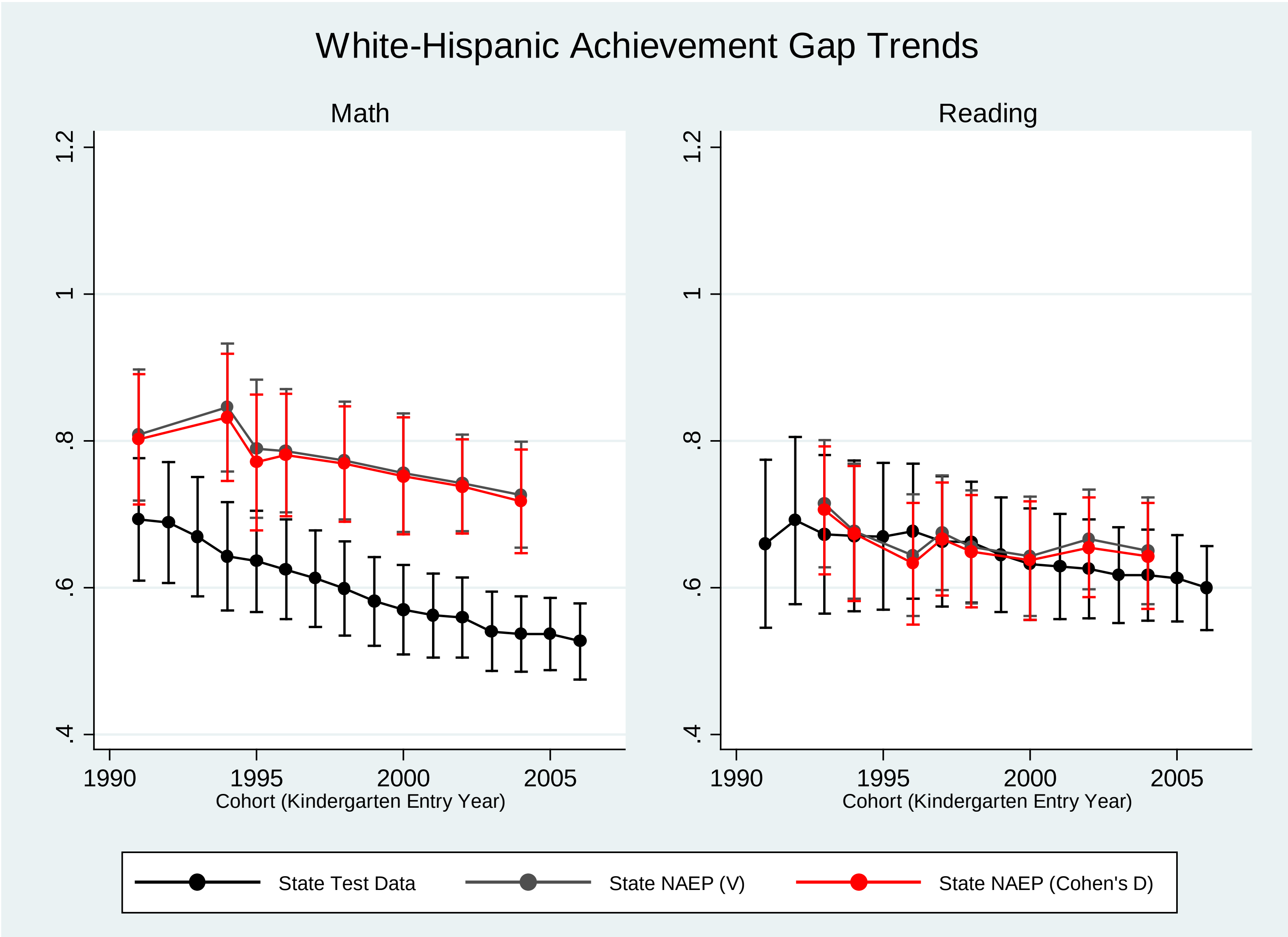

White-Black Achievement Gap Trends
Math
Reading
1.2
1
.8
.6
.4
1990
1995
2000
2005
Cohort (Kindergarten Entry Year)
State Test Data
State NAEP (V)
State NAEP (Cohen's D)

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



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) 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

$$yr_{cg}^* = coh_c^* + gr_g. \qquad \text{[A1]}$$

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 black-white [or Hispanic-white] income, poverty, and unemployment ratios in state s

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

$$G_{cs(-1)} = \lambda_s + \gamma_s(coh_c^*) + \mathbf{X}_{cs}\mathbf{A} + \nu_{cs}. \tag{A2}$$

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}:

$$G_{csg} = G_{cs(g-1)} + \delta_{csg}. \tag{A3}$$

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}):

$$\delta_{csg} = \alpha + v_s + \beta(coh_c^*) + \eta(g) + \delta_s T_{cg} + \mathbf{w}_{csg}\mathbf{B} + e_{csg}, \tag{A4}$$

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

$$\delta_{csg} = \alpha + v_s + \beta(coh_c^*) + \eta(g) + \delta_{0s}T_{cg} + \delta_1\left(T_{cg} \cdot g\right) + \mathbf{w}_{csg}\mathbf{B} + e_{csg}. \tag{A5}$$

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

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

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:[6]

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

[6] 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:

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

[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

$$\frac{\partial G}{\partial gr} = \begin{cases} \alpha_s' + \beta coh^* + 2\eta' gr & if \quad T = 0 \\ \alpha_s' + \beta coh^* + 2\eta' gr + \delta & if \quad T = 1 \end{cases}$$

[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

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

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

[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.