

Shining a Light or Fumbling in the Dark?
The Effects of NCLB's Subgroup-Specific Accountability on Student Achievement

Douglas Lee Lauen
Assistant Professor
Carolina Institute of Public Policy
Department of Public Policy
University of North Carolina at Chapel Hill
dlauen@unc.edu

S. Michael Gaddis
Department of Sociology
University of North Carolina at Chapel Hill
mgaddis@email.unc.edu

Forthcoming in *Educational Evaluation and Policy Analysis*
October 20, 2011
Pre-Publication Draft

The theory of action behind NCLB is that “shining a light” on subgroup performance will increase reading and math test scores for minority and disadvantaged students. Using a panel of all students in grades 3-8 in North Carolina from 2000-2008 (N=1.7 students in 1800 schools), we estimate double- and triple-differenced models with school fixed effects to examine whether subgroup-specific accountability threats increase high-stakes test scores. We find that these sanctions have positive effects for minority and disadvantaged students. Larger positive effects emerge for the lowest achieving schools rather than schools near the margin of passing. We also find some evidence of adverse effects for low and high achievers in math, but not in reading, a finding we attribute to the combination of increases in the rigor of state standards in math and an responses to an accountability metric based on test score status rather than growth. We discuss the implications of our findings for the design of educational accountability systems.

Acknowledgements: We thank Cynthia Coburn, David Guilkey, Jennifer Jennings, Randall Reback, Steven Rivkin and three anonymous referees for their insightful feedback. We also thank seminar participants at the UC Berkeley Institute of Human Development, the 2010 University of Wisconsin's Institute of Research on Poverty Summer Research Workshop and the 2010 American Sociological Association meeting in Atlanta, GA. for their comments. We are grateful for the North Carolina Educational Research Data Center at Duke University for providing the data for this project.

INTRODUCTION

Educational accountability policy seeks to increase pressure on teachers, administrators and students to increase efficiency, effectiveness and equity in a system assumed to be in dire need of reform. These policies emerged during the late-1980s and 1990s as part of the standards based reform movement (Linn, 2000; O'Day & Smith, 1993). During this time, most states in the U.S. set academic standards, established curriculum frameworks outlining what students should know and be able to do at each grade level, and many implemented high-stakes testing to determine whether schools were meeting performance targets. Most early accountability systems required schools to meet overall proficiency goals and permitted exemptions for special education and limited English proficient students, which created an incentive for schools to ignore low performing subgroups in order to meet average performance targets.

The No Child Left Behind Act of 2001 (NCLB) was designed to avoid some of the unintended effects of prior accountability systems. NCLB intends to shine a light on the traditionally under-served populations that were in some cases shunted to the side under state accountability systems. The key feature of NCLB is subgroup accountability, which requires racial and ethnic subgroups, students with disabilities, limited English proficiency students, and economically disadvantaged students to pass proficiency standards for a school to avoid sanctions. Under NCLB rules, schools fail a metric called Adequate Yearly Progress (AYP) if any one of its subgroups fails performance targets. Schools failing AYP for two consecutive years are deemed “in need of improvement” (INI). High poverty INI schools receiving federal funds (Title I schools) face an escalating set of sanctions, including offering students public

school choice, tutoring, and the remote possibility of takeover or reconstitution as a charter school. Non-Title I schools face only public reporting of AYP status.

With NCLB long overdue for reauthorization, the Obama administration has begun granting waivers from some of the act's requirements. In a letter to chief state school officers, Secretary of Education Arne Duncan wrote that the department would take unilateral action to grant "flexibility regarding specific requirements of NCLB in exchange for rigorous and comprehensive State-developed plans designed to improve educational outcomes for all students, close achievement gaps, increase equity, and improve the quality of instruction."¹ This study aims to inform policymakers about the effectiveness of a key feature of NCLB, subgroup-specific accountability pressure, as the Obama administration moves forward with granting waivers to states, Congress debates the reauthorization of NCLB, and as state and local officials work to design the next generation of educational accountability systems.

This study examines whether subgroup-specific accountability pressure from NCLB has increased the math and reading achievement of students in focal subgroups and whether treatment effects are moderated by school type and student position in the prior test score distribution. Prior research suggests that the incentive to improve may be a function of distance to the cutoff (Brown & Clift, 2010; Reback, 2008; Reback, Rockoff, & Schwartz, 2011). In other words, schools at the margin of passing AYP may have greater incentives to improve compared with schools well above or well below passing thresholds. Moreover, schools under pressure from NCLB face pressure to raise test score levels, which may provide incentives to triage students – that is, to focus attention on students near test score cutoffs at the expense of students

¹ Letter from Education Secretary Arne Duncan to Chief State School Officers, September 23, 2011, <http://www2.ed.gov/policy/gen/guid/secletter/110923.html>, last accessed October 20, 2011.

well below and well above such cutoffs (Booher-Jennings, 2005; Ladd & Lauen, 2010; Neal & Schanzenbach, 2010). Whether NCLB represents an improvement over previous state-level accountability approaches such as North Carolina's is an open question and depends on the existence, strength, and effectiveness of existing systems (Dee & Jacob, 2011; Wong, Cook, & Steiner, 2009).

Specifically, we ask whether schools in North Carolina respond to subgroup-specific accountability pressure from NCLB by raising the scores of minority and academically and economically disadvantaged students above and beyond what was occurring prior to NCLB. Is the subgroup-specific effect distinguishable from school failure due to other subgroups? Do larger accountability pressure effects emerge for schools at the margin of passing AYP? Finally, does subgroup-specific pressure promote triage at the expense of low or high achieving students?

We review the empirical literature on the effects of accountability pressure on student achievement. Then we present the findings of our study, which uses population-level administrative data from multiple cohorts of all elementary and middle school aged children in North Carolina between 2000 and 2008. Our study represents a test of the effects of NCLB in a state with a strong tradition of educational accountability, but with no subgroup-based accountability prior to NCLB. Our dataset is unique in that it includes not only consistent test score data over time, but also yearly subgroup- and test-subject-specific variables on school performance under NCLB, and analogous variables constructed for the three years prior to the implementation of NCLB, which permits us to compare the effects of subgroup-specific accountability pressure both before and after NCLB. We estimate the effects of subgroup-specific accountability pressure on student achievement with school fixed effect difference-in-

difference and triple-difference models. Our approach eliminates between-school confounding by estimating effects on successive cohorts of students from the same school.

THEORETICAL AND EMPIRICAL BACKGROUND

School accountability evaluates the performance of schools on student performance metrics. This sometimes involve rewards and sanctions, which can be explicit or implicit. Explicit consequences imposed on schools by central decision makers include awarding bonuses, mandated assistance teams, and school restructuring. Implicit consequences flow from the provision of information from school ratings which can build community pressure on schools to improve (Figlio & Loeb, 2011).

We define accountability pressure as the pressure educators may feel due to the explicit consequences embedded in state and federal accountability policy. Qualitative research on the effects of accountability mandates on schools documents a variety of responses by teachers and principals, ranging from open rebellion, conflict, and compliance (Booher-Jennings, 2005; Coburn, 2004; Diamond, 2007; Hallett, 2010). It is clear from the recent and growing literature on school responses to institutional pressure that there is sufficient evidence to suggest that school-based responses to accountability pressure are likely and that quantitative investigation of such responses could be fruitful.

Scholars have hypothesized that accountability pressure affects student test scores through increased effort and productivity by educators (Reback, Rockoff, and Schwartz 2011), greater alignment between curriculum frameworks and the content of what is taught in classrooms (Brown and Clift 2010), teaching to the test and even the test format (McNeil, 2000; Pedulla et al., 2003), educational triage (Booher-Jennings, 2005), more intensive use of data to

identify struggling students, differentiated instruction, after-school tutoring, and double dose instruction. The heavy reliance on high-stakes testing has prompted scholars to caution against using test scores as performance measures (Amrien & Berliner, 2002; Darling-Hammond, 2004; Linn, 2000). These critiques often invoke Donald T. Campbell who wrote, “The more any quantitative indicator is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor” (1979, p. 85). This has become known as Campbell's law, which predicts that school accountability systems will induce unintended responses by educators in an effort to meet the specified level of the indicator. Scholarly evidence finds that accountability pressure has a number of negative consequences such as a narrowing of the curriculum (Au, 2007; Jacob, 2005), gaming the system by reclassifying students to remove them from the testing pool (Cullen & Reback, 2006; Figlio & Getzer, 2006; Jacob, 2005), and outright cheating (Jacob & Levitt, 2003).

This study examines whether subgroup-specific accountability pressure increases the math and reading achievement of students in focal subgroups and whether treatment effects are moderated by school position relative to targets and student position in the prior test score distribution. This study is not designed to specifically test mechanisms, but rather to estimate effects using quantitative data. Therefore, this review will focus on the effects of educational accountability policy on student test scores. Accountability pressure on schools tends to increase student test scores (Figlio & Loeb, 2011). Cross-state studies conducted on the systems that preceded NCLB report positive associations between state-level accountability pressure and gains on state NAEP scores, a so-called “low-stakes” test with no accountability sanctions linked

to it (Carnoy & Loeb, 2002; Hanushek & Raymond, 2005). Within-state and within-district studies on accountability systems that preceded NCLB also find positive effects of high-stakes accountability on student test scores (Chiang, 2009; Figlio & Rouse, 2006; Jacob, 2005; Jacob & Lefgren, 2004; Reback, 2008).

Both cross-state and within-state studies of NCLB accountability also show positive effects on student test scores. An interrupted time series (ITS) study that compares states with high-stakes accountability systems prior to NCLB to those with no high-stakes accountability reports sizeable gains (.26SD) in 4th grade math NAEP scores and smaller, but significant, gains in 8th grade math NAEP scores (.14SD) (Dee & Jacob, 2011). Another cross-state ITS study uses three strategies to compare pre to post NCLB trends² and finds significant gains in 4th and 8th grade math NAEP scores (Wong, et al., 2009). A third study uses plausibly exogenous variation across states to identify the effect of short term accountability pressure using school-specific AYP data linked to the nationally representative Early Childhood Longitudinal Survey sample. This study shows that students in schools near the margin of meeting AYP targets had significantly higher achievement gains in reading (.07SD) and higher and positive, but insignificant, gains in math (Reback, et al., 2011).

The research summarized thus far suggests that schools respond to the incentives embedded in accountability systems to raise scores on both high-stakes and low-stakes tests. However, the evidence base on whether accountability pressure from NCLB improves poor and minority student achievement is thin and contradictory. To our knowledge only two studies address the effects of NCLB accountability pressure on minority or disadvantaged students. Dee

² Wong, Cook, and Steiner (2009) use national NAEP data to compare trends in Catholic public schools, and state NAEP data to compare states with lower proficiency standards to those with higher proficiency standards, and states with explicit sanctions to those without such sanctions.

and Jacob (2011), discussed above, find that the impact of NCLB on 4th grade math NAEP gain scores is generally larger for black, Hispanic, and poor students than for white and non-poor students. Research from Florida using a school fixed effects difference-in-difference model finds no effect of NCLB pressure on test scores for black, Hispanic, and poor students (Figlio, Rouse, & Schlosser, 2009). Neither Dee and Jacob (2011) nor Figlio, Rouse & Schlosser (2009) estimate the effects of own-group failure on black, Hispanic, or poor student performance. Therefore, it is difficult to pinpoint whether accountability pressure triggered by subgroup performance is having the intended effect.

Moreover, many, but not all, existing studies assume homogeneity in the responses of schools to accountability pressure. Due to accountability rules, however, not all schools face the same degree of pressure. Therefore, it is reasonable to assume differential effects of accountability by school type. First, schools at the margin of passing an AYP target may have greater incentive to increase student achievement than schools well below or well above passing (Brown & Clift, 2010; Reback, 2008; Reback, et al., 2011). Schools well below thresholds may not have the resources to attain targets and may perceive the costs of attaining targets as too great (Brown & Clift, 2010). Schools well above targets have weaker incentives to improve under a status-based system such as NCLB. In addition to assuming homogeneity in the incentives faced by school types, many studies assume all students will get equal benefits from accountability pressure, when in fact some accountability systems may create incentives to “triage” students (Booher-Jennings, 2005; Ladd & Lauen, 2010; Neal & Schanzenbach, 2010). There is a growing, but contradictory, literature on the differential effects of accountability pressure based on student position in the prior test score distribution. In accountability systems based on holding schools

accountable for test score levels (rather than growth or gain), schools may face incentives to adopt triage practices, which involve assessing students' likelihood to pass or fail a test and then diverting resources to those within reach of a passing test score. In a qualitative study from Texas, Booher-Jennings (2005) found that the ideology of data-driven decision making led teachers to focus instruction on the bubble kids, that is, those students just below grade level proficiency. Larger accountability-induced test score gains for students in the middle of a test score distribution than for low and high achieving students is consistent with the triage hypothesis. Quantitative studies of NCLB accountability from Washington State (Krieg, 2008) and of district-level and NCLB accountability from Chicago (Neal & Schanzenbach, 2010) report larger increases for students in the middle of the test score distribution than for low or high achieving students. Other quantitative studies find no evidence of educational triage favoring students in the middle of the test score distribution (Dee & Jacob, 2011; Ladd & Lauen, 2010; Reback, 2008; Reback, et al., 2011; Springer, 2008). The present study will contribute to the research on the distributional effects of accountability policy by estimating the distributional effects of NCLB accountability for black, Hispanic, and poor students.

In summary, the accumulated evidence suggests that high-stakes accountability has positive effects on tests scores. There is relatively little quantitative research on the effects of accountability pressure on poor and minority students, and on the differential effects on different types of schools. There is no research yet on subgroup-specific accountability pressure from NCLB. This study seeks to first establish whether subgroup-specific accountability pressure from NCLB has had positive effects on poor and minority students, and second examine the differential effects by school and student type. We compare the effects of subgroup-specific

accountability pressure under NCLB with the effects of low subgroup performance prior to the implementation of NCLB in a state which had no subgroup-level accountability in place prior to NCLB. We do this by deriving a measure of subgroup status during the pre-NCLB period based on the 2003 NCLB target, subgroup size, and performance. Schools in the pre-NCLB period had subgroups that would have failed these targets if NCLB had been in place. Thus, this comparison highlights whether subgroup-specific accountability test score increases, if any, can be attributed to the implementation of NCLB or to a previously existing desire on the part of educators to improve the test scores of low performing students. The answers to these questions will address not only important policy questions relevant to the reauthorization of the Elementary and Secondary Education Act, but also the growing literature on how schools respond to accountability pressure.

RESEARCH QUESTIONS AND HYPOTHESES

This study aims to inform scientific and policy debate about the effects of educational accountability policy, specifically the subgroup-specific accountability pressure embedded in NCLB. We first ask, *does subgroup-specific accountability pressure from NCLB raise the test scores of black, Hispanic, and economically disadvantaged students (ED) above and beyond what occurred prior to NCLB (RQ1)?* Based on prior research on the positive effects of accountability pressure on high and low stakes tests, we hypothesize that subgroup-specific accountability pressure will have a larger positive effect on reading and math test scores post-NCLB than pre-NCLB for blacks, Hispanics and ED students (H1). We distinguish between subgroup-specific accountability pressure from pressure due to other subgroups failing to meet targets by disaggregating the treatment effect into components based on focal and non-focal

subgroup failure. We then ask, *is the subgroup-specific effect distinguishable from school failure due to other subgroups* (RQ2)? We hypothesize that subgroup-specific accountability pressure will have a larger positive effect when the focal subgroup fails an accountability target than when at least one non-focal subgroup fails a target and the focal subgroup meets a target (H2). Testing this hypothesis will help us better understand whether schools respond to subgroup-specific accountability threats by increasing achievement of students in focal subgroups or whether achievement increases by students in focal subgroups can be attributed to more general threats about the low performance of students in many different subgroups.

The first two research questions of the study aim to establish whether schools respond to NCLB's subgroup-specific accountability pressure. The third and fourth examine whether there are differential effects by school position relative to test score targets and by student position in the test score distribution. We examine *whether larger accountability pressure effects emerge for schools at the margin of passing AYP* (RQ3)? A rational model of effort investment predicts that school personnel exert more effort when targets appear to be in reach, the benefits are attainable, and the costs are not too high. Therefore, we posit that subgroup-specific accountability effects will be larger in schools at the margin of passing AYP for a focal subgroup than for schools well above and well below AYP targets (H3a). The social and administrative pressure, on the other hand, may be greater on schools well below test score targets. In addition, district and state resources may help defray the costs low performing schools face in attempting to raise student achievement. As a result we pose an additional hypothesis for this research question: subgroup-specific accountability effects will be larger in schools well below passing AYP for a focal subgroup than for schools at the margin and well above AYP targets (H3b). Given prior theory

and research on status-based accountability systems (those that hold schools accountable for metrics like percent of students at grade level), we ask *does subgroup-specific pressure promote triage at the expense of low or high achieving students* (RQ4)? Due to the mixed results of prior research, with some studies finding larger gains for students in the middle of the distribution and some finding larger gains for low achieving students, we pose two hypotheses: subgroup-specific accountability effects will be larger for students at the margin of grade level than for students well above or well below grade level (H4a) and subgroup-specific accountability effects will be larger for students well below grade level than for students at the margin or above grade level (H4b).

DATA

This project uses test score and related data from multiple cohorts of students in grades 3-8 in North Carolina between 2000 and 2008.³ The full dataset contains about 5.7 million student-year observations (1.7 million students in more than 1800 schools). The dependent variables are math and reading end of grade test scores from North Carolina's state assessment system. North Carolina state test scores are produced from a three-parameter logistic IRT model and are scored on a developmental scale to allow computation of growth across grade levels and to measure growth throughout the achievement distribution. These scores are linked to the statewide curriculum and NCLB. As such, we are examining the effects of subgroup-specific accountability on so-called high-stakes test scores, rather than a low-stakes test such as NAEP. We standardize scores by grade level, subject, and year to facilitate interpretation of the magnitude of coefficients account and to account for the fact that the math test was rescaled in

³ We thank the North Carolina Education Research Data Center, housed at Duke University, for providing the data for this project.

2001 and 2006 and the reading test was rescaled in 2003 and 2008. Because the 2006 math rescaling raised standards and increased test difficulty, which could have had differential effects on low achieving students, we examine the impact of excluding 2007 and 2008 from our estimates of educational triage, discussed below.

The treatment effect is a difference-in-difference (DD) estimate, which is the coefficient on the interaction of a post-NCLB indicator and a one-year lag of a subgroup-specific accountability pressure variable (see equation 1, below). In our most basic specification we define subgroup-specific accountability pressure as failing Adequate Yearly Progress (AYP) under NCLB for a focal subgroup. In each post-NCLB year in each school, a subgroup meets the target, fails the target, or is not accountable for the target (due to too few students in the school's subgroup). Since accountability ratings are reported over the summer on the basis of spring testing, we examine the effects of the prior year's rating on next year's score. Specifying the variable in this manner sets up a dichotomy of accountability pressure: under pressure from the previous year's performance of a focal subgroup (failed) or not under pressure from the previous year's performance of a focal subgroup (met).

Under NCLB, schools are accountable for the performance of all students and for nine subgroups: white, black, Hispanic, Asian, Native American, multi-racial, special education, limited English proficiency, and economically disadvantaged (ED). The subgroups for which schools are in practice held accountable, however, depend on the minimum subgroup sizes set by each state and the distribution of various subgroups across a state's school population. North Carolina's minimum subgroup size is 40 students. In addition, for a student to count as a member of a subgroup, she must have attended the school for at least 140 days at the time of spring

testing. For this study, we focus on three of the most numerically significant subgroups present in the North Carolina student population: black, Hispanic, and ED. The percentages of schools accountable for black, Hispanic, and ED subgroups are 63, 21, and 92 respectively (table A1). About 45 percent and 10 percent of schools are accountable for special education and LEP students, respectively; however, inconsistencies in how special education and limited English proficient students are coded by administrative entities across time prevent us from including these subgroups in this paper. Less than three percent of schools are accountable for Asian, Native American, or multi-racial subgroups.

To compare the effects of subgroup-specific accountability pressure pre- to post-NCLB, we derive subgroup-specific accountability variables for the pre-NCLB period by determining for each school whether a subgroup had 40 or more members and whether the subgroup's average passing percentage in math exceeded 2003 NCLB targets (74.6% in math and 68.9% in reading). Any group with fewer than 40 members was coded as "not accountable." Any group with more than 40 members was coded as "met" if its members exceeded the 2003 target and "failed" if its members fell below this target.⁴

Figure 1 displays yearly percentages of schools failing the math AYP target (panel A) failing the reading AYP target (panel B) and not being accountable (panel C) for the three subgroups we examine. Panels A and B use as the denominator the number of schools accountable for a subgroup. This is necessary to place each subgroup on a common scale due to the relatively small number of schools accountable for Hispanics. Recall that the years from

⁴ Our approach does not take into account safe harbor (which gives schools credit for a failing subgroup if the students in the subgroup have particularly strong test score growth) or the confidence interval (which gives schools credit for being within the 95% confidence interval of a target), however, our proxies are accurate. For example, our approach correctly classifies 95% of schools' reported 2003 black subgroup accountability ratings.

2003-2008 use the *actual* subgroup variables used to determine AYP status, while the years 2000-2002 use *derived* variables to provide a pre-NCLB comparison (the solid line separates the pre- and post-NCLB periods). The yearly AYP target – percent of students at grade level – is shown in parentheses next to each year along the x-axis. As in all states, the targets increased periodically to eventually attain the NCLB-mandated target of 100% proficiency by 2014. Between 2003 and 2008 the math target changes three times and the reading target changes twice.

Panels A and B show variation over time in treatment status, which is critical for the DD model used in this study. Trends in accountability pressure are not flat in either the pre-NCLB period or the post-NCLB period, which makes identification of yearly school level responses from this variation feasible. For example, among those schools accountable for the black subgroup, the percentage of schools failing the math AYP target due to black subgroup failure varies from a low of 16% in 2004 to a high of 92% in 2006. The percentage of schools failing a subgroup target generally increases when the target increases. The exception to this general pattern is 2006 in math and 2008 in reading when the target fell and the percentage of schools failing subgroup targets increased rather than decreased. In these years the state implemented more difficult standards and assessments and received permission from the federal Department of Education to lower the targets for determining AYP. Evidently these lower targets were not low enough to ensure comparability with prior years. To account for the varying difficulty in tests over time, we standardize test scores by grade level and year.

Even though NCLB had not yet been implemented, the percentage of schools failing to meet the initial 2003 target for each subgroup was on a downward trend prior to 2003. Therefore,

it is likely that prior to NCLB, educators in North Carolina had been implementing interventions to address the low performance of students in these subgroups. Therefore, an identification strategy that relies on a difference-in-difference is preferable to a simple pre-post comparison. The percentage of schools not accountable for Hispanics is much higher than for blacks or ED students. This percentage also declined a great deal between 2000 and 2008. It is clear from numerous sources that the Hispanic population has been growing in North Carolina. Our dataset on students enrolled in grades three through eight contains about 23,000 Hispanic students in 2000 and 67,000 students in 2008, nearly a three-fold increase. During the same period, the number of whites declined from approximately 375,000 to 364,000 (a decline of 3%), and the number of blacks remained constant at about 183,000. Due to the rapid growth in the Hispanic population in North Carolina, many schools were newly accountable for Hispanics during the post-NCLB period, which is a threat to the validity of a standard DD estimation strategy. For this reason, we prefer a school fixed effects DD model, which we discuss below.

METHODS

To answer our first research question, *does subgroup-specific accountability pressure from NCLB raise the test scores of black, Hispanic, and economically disadvantaged students (ED) above and beyond what was occurring prior to NCLB*, we use a difference-in-difference (DD) model. This model tests whether subgroup-specific accountability pressure had a larger effect on student test scores in the post-NCLB period than in the pre-NCLB period:

$$A_{ist} = \beta_0 + \beta_1 T_{s,t-1} + \beta_2 Post_t + \beta_3 T_{s,t-1} X Post_t + e_{ist} \quad (1)$$

In this equation the standardized achievement score⁵ for student i in school s at time t is regressed on T , a school-level, time-varying, subgroup-specific accountability threat variable, coded 1 if student i 's school failed the AYP subgroup target in the prior year, and 0 otherwise; $Post$, an indicator coded 1 if the year is 2003-2008, and 0 if the year is 2000 to 2002; and the interaction of T and $Post$. The treatment effect in equation 1 is β_3 , the additional effect of NCLB sanctions in the post period. If NCLB sanctions are effective, we would expect a larger failed AYP effect in the post period than in the pre-period. Recall that the accountability pressure variables in the pre-period were derived for the purposes of our study. This specification tests whether the gap between schools that met and failed AYP targets closed or widened between 2000 and 2008:

$$\hat{\beta}_3 = (\bar{y}_{T=1,post} - \bar{y}_{T=0,post}) - (\bar{y}_{T=1,pre} - \bar{y}_{T=0,pre}) \quad (2)$$

If β_3 is positive, then the subgroup-specific accountability pressure is stronger in the post-period than in the pre-period. In other words, the gap between schools that failed and made AYP closed between the pre and post period.

Estimating equation 1 with OLS will produce consistent parameter estimates under two conditions: 1) no time-varying confounders of the outcome, and 2) no time-invariant between-school confounding. The difference-in-difference estimation approach reduces, but does not eliminate, between-school confounding. A standard OLS DD estimate is based on a change in the gap from a pooled estimate of a heterogeneous group of schools: schools that never failed, always failed, and sometimes failed AYP targets. This is a reasonable approach assuming that all these types of schools were affected equally by the imposition of NCLB and there were no major

⁵ We examine test score levels rather than gains because NCLB holds schools accountable for test score levels and not gains. We standardize scores to account for differences in the difficulty of tests over time and to report effect size estimates consistent with other studies.

compositional changes in the student population over time. Because NCLB was based on a status measure of accountability (test score level), however, it is unlikely that NCLB would have similar effects on schools with high achievement levels (the never-failers) and schools with lower test scores (the sometimes and always failers). Moreover, due to the large increase in the Hispanic population, it is not reasonable to assume that school composition is constant over time. More than 300 schools were newly accountable for Hispanics in the post period (42.8% of all schools accountable for Hispanics), which is likely to be a more challenging educational problem than cycling in and out of accountability status with a more stable proportion of Hispanics. For these reasons, we include a vector of school fixed effects, ϕ_s , and a vector of student characteristics, X^6 , to equation 1:

$$A_{ist} = \beta_0 + \beta_1 T_{s,t-1} + \beta_2 Post_t + \beta_3 T_{s,t-1} X Post_t + \lambda \phi_s + \beta X_{ist} + e_{ist} \quad (3)$$

We separately estimate equation 3 for three numerically significant subgroups (black, Hispanic, and economically disadvantaged) and specify T to refer to focal subgroup of interest. In other words, in the black equation, we define T as failing AYP for the black subgroup. A positive and statistically significant coefficient on β_3 would be evidence in support of H1, that subgroup-specific accountability pressure will have a larger positive effect on reading and math test scores post-NCLB than pre-NCLB for blacks, Hispanics and ED students. Our specification tests whether, for example, black students benefit more in the post-period from subgroup-specific pressure than did black students in the pre-period. Due to between-school segregation and NCLB minimum subgroup size rules, it is much more difficult to address the question of whether

⁶ Student-level variables included in the X vector include race, gender, gifted, special education, and LEP designations, and structural (i.e., switching to a school due to a change in grade configuration) and non-structural (i.e., switching to a school due to a residential move) school moves. Descriptive statistics for accountability pressure variables and controls are shown in table 1.

historically disadvantaged students benefit *more* from subgroup-specific accountability than do non-poor whites. This is an important question, and one that should be addressed in future research, but is secondary to the more basic question we pose.⁷

We include school fixed effects to remove between-school heterogeneity that would otherwise bias the accountability effect. This approach estimates the effect of subgroup-specific accountability pressure on successive cohorts of students in the same school. Omitting the school fixed effect and estimating equation 1 would involve using both within- and between-school variation to determine whether students in schools facing subgroup-specific pressure from NCLB have higher test scores than students in schools not facing such pressure. This approach could produce biased estimates because it would fail to account for exogenous unobserved differences between schools. The specification shown in equation 3 uses only within-school variation across cohorts to produce estimates of subgroup-specific accountability pressure. Therefore, those schools consistently failing or consistently attaining subgroup performance targets contribute nothing to the subgroup-specific accountability pressure estimates.⁸ While this approach ignores

⁷ Our research question asks whether subgroup-specific accountability pressure increases achievement for students in a focal subgroup. A regression with students from multiple subgroups included would answer a different question: do students from subgroup x gain more when their subgroup fails than students in subgroup y ? A model of subgroup accountability for whites, blacks, Hispanics, and ED students would require a sample of schools accountable for all four subgroups. Due to between-school patterns of segregation, only 21% of schools in North Carolina are simultaneously accountable for whites, blacks, Hispanics, and ED students. Moreover, our models require variation on the subgroup failure variables across the two periods. Thus, the initially small subset of schools is reduced even further: with math as the outcome, 4% of all schools for the white subgroup, 13% for the black subgroup, 13% for the Hispanic subgroup, and 12% for the ED subgroup. (Percentages are similar for reading.) Both the resulting comparison group of schools and the group of schools which contribute to the estimated coefficients raise serious doubts about generalizability to the population of students in North Carolina. Our current black subgroup models include 74% of all schools in the state. The analogous figures for the Hispanic and economically disadvantaged student models are 27 and 96 percent, respectively. Represented in our current models are all schools in the state that are accountable for these subgroups, an important aspect of aligning this research with policy evaluation.

⁸ During the NCLB period (2003-2008), a school could conceivably fail from 0 to 6 times for each target subgroup. The percentages of schools never failing due to the low performance (including not accountable years) of blacks, Hispanics, and ED students are 39, 76, and 19, respectively. The percentages of schools failing all six years for the low performance of these same subgroups are 6, 2, and 7 percent, respectively. Summing by subgroup produces the following percentages of schools with no within-school variation in accountability status for each of the subgroups:

the performance of schools with static treatment statuses, we view comparing the performance of cohorts exposed to subgroup-specific accountability pressure to cohorts not exposed to such pressure who attended the same school at different points in time as more valid than comparing the performance of students in different schools.⁹ Time-varying school-level confounders that differ in the pre and post periods remain threats to the validity of our findings. Including school fixed effects buys us some, but not complete, protection from all possible sources confounding.

Defining accountability pressure as a simple dichotomy could produce misleading estimates if there are spillover effects across subgroups. The dichotomous treatment specification makes no distinction between a subgroup failing alone and a subgroup failing along with other subgroups. So we also estimate models with more nuanced treatment categories to test RQ2, *is the subgroup-specific effect distinguishable from school failure due to other subgroups?* One might expect the strongest accountability effects on a focal subgroup when a subgroup fails alone and weaker effects when multiple subgroups fail. This turns out to be a complex problem due to the high degree of overlap between race/ethnicity, poverty, and other subgroup designations. To address the issue of multiple subgroup failure and spillover effects across subgroups, we test for differences across the following groups (with failed no targets as baseline): only the focal subgroup fails (defined as T1, in the equation below); school fails AYP because of the focal subgroup along with at least one other subgroup (T2); and the school met AYP for the focal subgroup, but failed for at least one other subgroup (T3):

45, 78, and 26 (table A2).

⁹ An alternative to including school fixed effects is to include school-specific trends. Including school-specific trends, however, greatly increases the risk of controlling for endogenous factors. Moreover, with our data the most flexible specification of school-specific trends would involve including dummies for nine time points, as many as 1700 schools, and $9 \times 1700 = 15,300$ time \times school interactions, which is a matrix that is too large for most statistical software and computers to invert with the required precision.

$$A_{ist} = \beta_0 + \beta_1 T1_{s,t-1} + \beta_2 T2_{s,t-1} + \beta_3 T3_{s,t-1} + \beta_4 Post_t + \beta_5 T1_{s,t-1} XPost_t + \beta_6 T2_{s,t-1} XPost_t + \beta_7 T3_{s,t-1} XPost_t + \lambda \phi_s + \beta X_{ist} + e_{ist} \quad (4)$$

Above we hypothesized that subgroup-specific accountability pressure will have a larger positive effect when the focal subgroup fails an accountability target than when at least one non-focal subgroup fails a target and the focal subgroup meets a target (H2). We test this hypothesis by testing whether $\beta_5 > \beta_7$ and whether $\beta_6 > \beta_7$. It may also be the case that when a focal subgroup fails and no other subgroups fail we might find a larger positive effect than when a focal subgroup fails along with at least one other subgroup (i.e., $\beta_5 > \beta_6$), but because of the overlapping nature of subgroups (especially black and ED and Hispanic and ED), it may difficult to test this hypothesis precisely with these data.¹⁰

The basic DD specification shown in (3) might understate the effect of NCLB because it ignores school proximity to passing thresholds. Schools well below the standard could face reduced incentives due to the high cost-benefit ratio they may face in attempting to meet standards and schools just above the standard also respond to accountability pressure for fear of falling below it in the future. Therefore, we might expect the strongest effect among schools “at

¹⁰ Ideally, we would like to estimate the following DDD model: $A = \beta_0 + \beta_1 sgf + \beta_2 allf + \beta_3 post + \beta_4 sgf Xallf + \beta_5 sgf Xpost + \beta_6 allf Xpost + \beta_7 sgf Xallf Xpost + e$, where *sgf* is school failed for the a focal subgroup, and *allf* is school failed for the overall student body. This model would indicate whether failing for both the focal subgroup and the overall student body has an incremental effect over and above the focal subgroup failing alone. Unfortunately, there are no schools in the pre-NCLB period that met the math target for the black subgroup and failed overall, making estimation of β_7 impossible. Estimating this equation for reading would rely on 229 black students in only two schools in the pre-period that met for the black subgroup and failed overall. For economically disadvantaged subgroup equations in both reading and math, there are fewer than 400 poor students in the pre-period in five schools that met for the ED target and failed for all students. The situation for Hispanics is slightly better (1,383 students in 19 schools in the pre-period for math and 579 students in 7 schools in the pre-period in reading). Even if these models resulted in believable significant estimates, the limited and unusual comparison group would result in problems with generalizability. If instead we omit the three-way interaction and only estimate models that focus on the focal subgroup failure and student body failure separately, we find mostly significant effects for both coefficients. However, these models are misspecified, as the combined effect of failure is forced into the estimation of each category separately.

the margin” (Brown & Clift, 2010; Reback, 2008). To test this hypothesis, we specify subgroup-specific accountability as follows: at least 10% below the math target (T1, below), between 10% below and 10% above the math target (i.e., at the margin, T2, below), and at least 10% above the math target (omitted baseline category in equation below):

$$A_{ist} = \beta_0 + \beta_1 T1_{s,t-1} + \beta_2 T2_{s,t-1} + \beta_3 Post_t + \beta_4 T1_{s,t-1} XPost_t + \beta_5 T2_{s,t-1} XPost_t + \lambda \phi_s + \beta X_{ist} + e_{ist} \quad (5)$$

This specification is designed to test RQ3, *do larger accountability pressure effects emerge for schools at the margin of passing AYP?* If the effects of subgroup-specific threats are larger in the post-period for schools at the margin than for schools well below the margin (H3a), then $\beta_4 < \beta_5$. If the effects of these threats are larger in the post period for schools well below the target than for schools at the margin (H3b), then $\beta_4 > \beta_5$.¹¹

Finally, accountability threats could have differential effects based on a student’s position in the prior year’s test score distribution. Because NCLB is a status-based system based on percent of students at grade level, schools may have an incentive to triage students for interventions. Literature on the “bubble kid” phenomenon suggests that schools may focus interventions on students near test score cutoffs for grade level proficiency (e.g. Booher-Jennings, 2005; Brown & Clift, 2010). This would produce larger treatment effects for students near grade level than students well below or well above grade level. To answer our final research

¹¹ Specifying accountability pressure as distance to a cutoff raises the important question of distance to which cutoff, last year’s or this year’s? Assuming foreknowledge of next year’s target, it is possible that schools above the current year’s target, but below next year’s target would also be under accountability pressure. Complicating matters, however, is that AYP targets in North Carolina sometimes changed even after they were announced. For instance, after math standards were raised in the 2005-2006 school year and the percentage of students at or above grade level fell a great deal, targets for the upcoming two school years were changed and not announced until a few months into the 2006-2007 school year (see <http://www.ncpublicschools.org/docs/accountability/reporting/abc/2006-07/aypstatus.pdf>).

question, *does subgroup-specific pressure promote triage at the expense of low or high achieving students* (RQ4), we define students as low achieving if they are at least 0.5 SD below grade level in the prior year, at the margin if they are between 0.5 SD below and 0.5 SD above, and high achieving if they are at least 0.5 SD above grade level. In this case, we have one treatment, T, and two student-level moderators, M1, low achieving, and M2, high achieving, with at the margin as the omitted baseline category:

$$\begin{aligned}
A_{ist} = & \beta_0 + \beta_1 T_{s,t-1} + \beta_2 Post_t + \beta_3 T_{s,t-1} X Post_t + \beta_4 M1_{i,t-1} + \beta_5 M2_{i,t-1} + \\
& \beta_6 M1_{i,t-1} X T_{s,t-1} + \beta_7 M2_{i,t-1} X T_{s,t-1} + \beta_8 M1_{i,t-1} X Post_t + \\
& \beta_9 M2_{i,t-1} X Post_t + \beta_{10} M1_{i,t-1} X T_{s,t-1} X Post_t + \beta_{11} M2_{i,t-1} X T_{s,t-1} X Post_t + \\
& \lambda \phi_s + \beta X_{ist} + e_{ist}
\end{aligned} \tag{6}$$

Triage theory predicts that the treatment effect of subgroup-specific accountability from NCLB would be stronger for students at the margin. Therefore, both β_{10} and β_{11} should be negative (H4a). If, on the other hand, educators do not triage and instead focus attention on the lowest achievers, then β_{10} should be positive and larger than β_3 and β_{11} (H4b). To test the robustness of our results on triage, we also estimate alternative specifications with more refined categorizations. In addition, because state standards increased during the post-NCLB period in math (in 2006, as mentioned above), which could adversely affect low achieving students, we test our triage models on the period 2000-2006, a period of generally low academic standards, and the period 2000-2008, a period that included both lower and higher academic standards. We hypothesize that adverse effects on low achieving students will be larger in the models covering 2000-2008 than in the models covering 2000-2006 due to the fact that incentives to triage will be

greater under a status-based accountability system with high standards than a status-based accountability system with low standards.

RESULTS

Main Treatment Effects

Table 2 combines the coefficients of interest and cluster-corrected standard errors¹² from separate subgroup regressions into one summary table. Models in columns 2 and 4 include school fixed effects and correspond to equation 3. We include pooled OLS results in columns 1 and 3 to show that in some cases coefficients differ between the two specifications. All regressions control for, but do not display the coefficients on student designations as academically gifted, special education, limited English proficient, male, whether the student had ever been retained, parental education, economic disadvantage (in race subgroup models), race (in economically disadvantaged subgroup models), and structural and non-structural school moves. A full set of tables is available from the authors upon request.

The *FailedSGXPostNCLB* interaction coefficient is the treatment effect. Recall that this DD estimate compares standardized test scores before and after the introduction of NCLB between schools that met the subgroup target and schools that failed the subgroup target. The first figure in column 1 shows a treatment effect on standardized math test scores of failing the black subgroup for black students of .07 standard deviations (SD). Including school fixed effects reduces this treatment effect to .04 SD. For reading, the effects of black subgroup failure for blacks are about .05 SD, with virtually no difference between models with school fixed effects

¹² We report cluster corrected standard errors, with school as the clustering unit. This choice produces larger standard errors than using student as the clustering unit. To reduce the risk of type I error, we report larger standard errors, although with the sample sizes available for analysis in this study, statistical significance is rarely at issue.

and without them. Effects of Hispanic subgroup failure on Hispanic achievement are larger than for black subgroup failure on blacks. Effects range from .09 to .13 SD. Treatment effects for ED students are similar in size to those in the black equation. These results, with accountability pressure defined as failed a subgroup target, indicate support for hypothesis H1, that schools responded to subgroup-specific threats by increasing test scores more in the post-NCLB period than in the pre-NCLB period. Because the school fixed effects specification provides protection from between-school confounding, and in the interests of brevity, we present only school fixed effects estimates in the tables to follow.

Figure 2 provides a visual depiction of the calculation of the DD estimate in math using the school fixed effects models shown in table 2. The first set of bars shows that on average black students in schools meeting subgroup targets outperformed black students in schools failing subgroup targets in both the pre and post periods, but that this gap shrunk in the post period. The pre-post difference in the met-failed gap for blacks is .043 SD. For Hispanics the pre gap is relatively large and negative and the post gap is weakly positive, so the difference between the two is larger at .134 SD.

Because the vast majority of blacks and Hispanics are also poor, if a school fails to meet AYP due to the black or Hispanic subgroup, they almost certainly also fail to meet the AYP for the ED subgroup. Therefore, almost no schools fail AYP because of the focal subgroup alone. For the same reason the treatment effects shown in table 3 based on equation (4) for failed focal subgroup alone are imprecisely estimated. While we cannot reliably distinguish the effect of failing a focal subgroup alone from the effect of meeting AYP for all subgroups and therefore cannot test the hypothesis that sole focal subgroup failure increases test scores more than

multiple subgroup failure, we can distinguish the effect of failing AYP for a focal subgroup along with other subgroups from the effect of meeting AYP for a focal subgroup while failing for other subgroups. As shown in table 3, the effect of failing AYP for the focal subgroup and other subgroups is positive and larger than the effect of meeting AYP for the focal subgroup and failing AYP due to other subgroups, which is evidence in support of hypothesis H2. The differences in the black and ED equations are small, but statistically distinguishable. The differences in the Hispanic equations are larger. For example, in math, the effect of failing the Hispanic subgroup and other subgroups on Hispanics is .23 SD; the effect of failing for other subgroups and not the Hispanic subgroup on Hispanics is .12 SD. This is strong evidence of a subgroup-specific accountability effect on poor and minority students as opposed to a more generalized failed AYP effect on these subgroups.

Treatment Effects by School Distance to Target

Defining accountability pressure as a trichotomy (more than 10% above the target, within 10% of the target, and more than 10% below the target) produces positive treatment effects for schools at the margin of the target and generally larger treatment effects for schools well below the target, especially in math (table 4 based on equation (5)). The effects for Hispanics in math are especially large. The DD estimate for schools at the margin for Hispanics is about one-tenth of a standard deviation; this estimate for schools well below the margin is more than one-quarter of a standard deviation. Larger effects for schools well below the target runs counter to the hypothesis that NCLB would have larger effects on schools at the margin of passing. We therefore reject hypothesis H3a and find support for hypothesis H3b. An alternative specification with distance to the target defined as 0-10% above target, 0-10% below target, and >10% below

the target reveals that differences from baseline (>10% above the target) increase monotonically with each step in the categorization. In other words accountability induced effects take the following general pattern: well above < just above < just below < well below (results available from authors upon request).

Treatment Effects and Educational Triage

Table 5 shows school fixed effects regression based on equation 6, with coefficients on the main treatment effects, which now represents the effect of subgroup-specific accountability for students near grade level, and the DDD coefficients for students well below and well above grade level. We divide this analysis into two time periods, 2000-2008 and 2000-2006 because of the large effect the 2006 math standards had on the estimation of the treatment effects on students in various points in the achievement distribution.¹³ Recall that the 2006 math standards and tests were so much more rigorous than the prior year standards and tests that the state received permission from the federal Department of Education to lower the 2006 AYP target. The percent of students at grade level in math fell from 88% in 2005 to 64% in 2006. Despite the lower AYP target, the percentage of schools failing AYP due to ED students, for example, increased from 33% in 2005 to 76% in 2006. The rigor of standards, the difficulty of tests, and the cut points used to define grade level proficiency have implications for the theory of educational triage because when few students are below grade level, the effort and cost to bring students up to grade level is lower than when many more students are below grade level.

If triage theory is correct, we expect to find negative DDD coefficients for low and high

¹³ The period 2000-2006 leaves out the new standards year in math because our study design uses lags of prior achievement as moderators of the lagged accountability effect. The new standards in math in 2006, therefore, would affect achievement from 2007 on. Similarly, the new standards in reading in 2008 would affect achievement from 2009 on. This study, however, only covers the period from 2000 to 2008.

achieving students, which would possibly indicate greater attention to students near grade level. In column 1, estimated on the years 2000-2008, the treatment effects for students near grade level in math are positive, statistically significant, and relatively large at over one-quarter SD for blacks, Hispanics and ED students. The DDD effects are consistently significant and negative for all three subgroups, indicating that the accountability treatment effect in math is about one-tenth SD lower for low and high achieving students than for students near grade level. These results provide strong support for hypothesis H4a. The estimates in column 2, however, estimated on the years 2000-2006, provide evidence that increasing the standards for grade level proficiency may have contributed to accountability-induced triage. The treatment effects for students near grade level in column 2 are positive and significant, but much smaller than in column 1, the effects for students below grade level are weakly positive rather than negative, and the effects for high achievers are small and negative among blacks and ED students and zero among Hispanics. These estimates provide much weaker support for hypothesis H4a. Among blacks and ED students, the difference between the treatment effects for high and low achievers is significant, but only about one-twentieth of a SD. The absence of negative effects for low achieving students during the period in which the rigor of math standards was relatively stable suggests that educational triage at the expense of low achieving students is not necessarily a function of NCLB induced pressure, but rather the unfortunate combination of an accountability system based on test score levels and a large increase in the portion of the student population below grade level due to an increase in standards.

Further evidence on this point comes from the reading results in columns 3 and 4. During both periods, treatment effects for students near grade level are quite close to zero, while effects

for students below grade level are positive for blacks and ED students, and effects for students above grade level are negative and small for all subgroups. These results provide evidence to reject hypothesis H4a and in support of H4b. Differences between the treatment effects for black and ED low and high achievers are about one-tenth SD. The signs of the coefficients in the Hispanic equations are consistent with the triage hypothesis, but are quite small and imprecisely estimated. As a robustness check we respecified the positional categories and reestimated equation 6 (see tables A4 and A5). These robustness checks show that our conclusions about triage in math and reading in both periods are not a result of the simple three-category specification shown in table 5.

In summary, this evidence suggests that accountability-induced triage could negatively affect low and high achieving students, but is not automatic consequence of systems like NCLB. Triage can emerge, but the effects likely depend on where grade level standards are set and the stability of these standards over time.

CONCLUSION

Schools in today's accountability era are subject to strong pressure to increase student achievement. Since *A Nation at Risk* (Gardner, 1983), it is clear that American schools have become more tightly coupled to state and federal policymaker's mandates of educational excellence and equity (Coburn, 2004; Diamond, 2007; Mehta, 2006). NCLB represents the strongest federal incursion in educational accountability policy to date. It set lofty goals, forced schools to track achievement gaps, mandated public reporting of performance measures for all schools, and put schools receiving federal funding on a sequence of escalating sanctions for failing AYP in consecutive years. It is not yet clear what shape the next generation of federal

educational accountability policy will take. It may allow more state flexibility, measure growth rather than status, hold teachers rather than school accountable, and it may or may not track test score gaps. But it is clear that state and local officials are forging ahead in designing accountability systems with the Obama administration encouraging these efforts by granting waivers from NCLB, which is now about five years overdue for reauthorization. As this policy activity moves forward, it is important for policy makers to understand the effects of subgroup-specific accountability pressure based on percent of students at or above test score proficiency, i.e., a “status” (rather than a growth) measure of proficiency.

Using data from third through eighth graders from the entire state of North Carolina between 2000 and 2008, we estimate a school fixed effects difference-in-difference (DD) model with derived variables in the pre-NCLB period for subgroup failure. Consistent with prior research, we report positive effects of NCLB accountability pressure on student test scores (Dee & Jacob, 2011; Reback, et al., 2011; Wong, et al., 2009). Unlike previous research, we examine the effects of school-specific, subgroup-specific threats. So ours is not a study of the “failed AYP effect,” but of the subgroup-level accountability embedded within it. We ask whether poor and minority student test scores increase if their subgroup failed AYP the year before. We report a positive subgroup-specific accountability effect that is distinguishable from a more generalized failed NCLB target effect. Using this model we report small accountability induced increases for blacks and poor students, larger effects for Hispanics, and effects in reading that are roughly comparable in size to effects in math.

We extend this analysis to examine differential effects based on school distance to a subgroup target and student distance to grade level. Contrary to predictions that schools at the

margin of passing would have larger effects, we report that schools well below cutoffs have larger accountability-induced effects than schools at the margin of passing, a finding that is robust to alternative specifications of distance to the margin. The effects for Hispanic subgroup failure in schools well below the margin are particularly large at .27 SD in math and .19 SD in reading. We also report strong evidence of triage for all subgroups based on student's distance to grade level, but only in math and only with data from 2000-2008, a period that included a large increase in the rigor of math standards in 2006. These tougher math standards increased the percent of students below grade level, which may have increased the risk of triage. During the period 2000-2006, we find much weaker evidence favoring students near grade level in math. In reading, a subject that had relatively consistent standards between 2000 and 2007, we find no consistent evidence of triage favoring students near grade level in either the 2000 to 2008 period or the 2000 to 2006 period. This evidence suggests that accountability-induced triage from NCLB is not an automatic consequence of a status-based approach to accountability, but rather a risk factor along with the rigor of proficiency levels and the stability of these levels across time.

NCLB first rated schools in 2003. In some states, these school ratings were the first visible signs of educational accountability; in many, however, school accountability had been in place for years. For example, North Carolina, where this study is conducted, has been recognized as a strong accountability state.¹⁴ Between 1997 and 2007, North Carolina teachers received bonuses of \$750 or \$1500 for meeting or exceeding school-wide student test score growth targets.¹⁵ Schools also receive public designations based on test score levels. Schools with low

¹⁴ Carnoy & Loeb (2002) rated the strength of state accountability systems on a 0 to 5 scale (with 5 being high) based on a 2000 survey. North Carolina's accountability system a five out of five, one of only eight states to receive such a rating. Twenty three states received a rating of no or weak accountability (0 or 1).

¹⁵ The recent budget crisis has adversely affected North Carolina's bonus program. In 2008, teacher bonuses were

average test scores and less than expected growth (one year of growth for all students, on average) face the possibility of mandated state assistance. Due to the different definitions of progress under the North Carolina ABC's accountability program and the federal NCLB accountability program, schools face differences in the type and amount of pressure faced when failing to meet AYP. Some schools are under the simultaneous "double-threat" of state and federal accountability pressure, while other schools receive mixed signals about students' progress.

In supplementary analysis not presented in this paper, but available from the authors upon request, we address the potential confounding of state and federal accountability due to North Carolina's long-standing accountability system based on year-to-year student growth. We find little evidence that additional pressure from North Carolina's growth model of accountability either confounds or moderates NCLB treatment effects. With one exception, the effects are statistically insignificant and have inconsistent signs. (A small positive effect emerges for blacks in math, which indicates that on average blacks in schools that face the double threat of state and federal accountability outperform blacks in schools that face only the threat of federal accountability.) The absence of negative effects is worth noting. The imposition of an accountability system based on test score levels and subgroups was quite different than North Carolina's pre-existing state accountability system based on growth. This evidence suggests that while accountability systems based on growth and status can no doubt send conflicting signals to parents and educators, they do not necessarily lead to worse outcomes for the minority and disadvantaged students in the lowest achieving schools.

shaved by 30% and cut altogether in 2009 and 2010.

This study has important limitations. An ideal test of subgroup-specific accountability pressure would compare results from two scenarios: NCLB with only overall proficiency pressure and NCLB with only subgroup-specific pressure. Of course, NCLB sanctions schools for failing to meet overall proficiency targets and for failing to meet subgroup targets, so such a comparison is impossible to test. Moreover, individual students can be members of multiple subgroups, which complicates estimating discrete subgroup effects. Our results, however, indicate a more targeted effect of subgroup-specific accountability threats. We show that subgroup-specific accountability pressure on a focal subgroup has a larger effect when a school fails AYP for the focal subgroup along with at least one other subgroup than when a school met AYP for the focal subgroup and at least one other subgroup failed to meet AYP. Another limitation of quantitative studies of accountability effects based on administrative data, such as the present one, is the inability to test mechanisms. Qualitative research on the effects of high stakes testing on schools and teachers reports a variety of effects on curriculum and instruction. A meta analysis of qualitative studies, for example, found that accountability pressure generally narrows the curriculum to tested subjects, fragments subject area knowledge into test-related pieces, and increases teacher-centered pedagogy, but sometimes can expand curriculum content, promote the integration of knowledge and student-centered pedagogy (Au, 2007). Au (ibid) argues that the effects of accountability on curriculum may depend on the types of assessments used: if tests reward drill and kill, teachers will prepare students for drill and kill; if tests reward synthesis, solving multistep problems, and higher order thinking skills, on the other hand, teachers tend to teach these skills in the classroom. In other words, what gets tested gets taught. Other qualitative studies have found that accountability pressure leads to greater alignment

between curriculum frameworks and the content of what is taught in classrooms (Brown & Clift, 2010) and educational triage either based on student position in the test score distribution (Booher-Jennings, 2005) or based on school accountability status or distance to meeting a yearly target (Brown & Clift, 2010; Diamond & Spillane, 2004). Our study, however, raises doubts about whether accountability pressure is a sufficient cause of educational triage. It may be a necessary cause, but the rigor of and changes in state standards probably plays an important role as well. In addition, our findings run counter to claims in both qualitative and quantitative research that schools at the margin of passing would have the largest accountability induced effects. We find that the lowest performing schools have larger accountability-induced effects than schools at the margin or above accountability targets.

A within-state design limits the generalizability of our study. Our estimates are not nationally representative and results from other states may vary. Ours is the only study to our knowledge that is specifically designed to test the short run effects of yearly, school-specific, subgroup-specific accountability threats on the achievement of students in poor and minority subgroups, however, there are no national data that can examine this question. We have chosen a state with a strong accountability system that pre-dated NCLB. Given the findings of Dee and Jacob (2011), which show larger effects in weak accountability states than in strong accountability states, one could view our results as lower bound estimates given North Carolina's strong emphasis on consequential accountability (e.g., teacher bonuses for school-wide performance) prior to NCLB (Carnoy & Loeb, 2002; Dee & Jacob, 2011; Hanushek & Raymond, 2005; Lee & Wong, 2004).

By separately analyzing the effect of own-subgroup accountability pressure on subgroups,

we leave many important complexities to be explored in future research. For example, do blacks gain more than whites when the black subgroup fails? Do Hispanics gain less in schools where both blacks and Hispanics fail than when only Hispanics fail? Do schools take a sequential approach to improvement? Perhaps those in the smallest subgroups get the most attention initially and those in larger subgroups get attention after progress has been made with the first group. Perhaps the most difficult challenge is to meet AYP targets with special needs students. Studies of reclassification and achievement gains will be particularly important for this vulnerable subpopulation. Finally, we report larger effects for Hispanics than for blacks and economically disadvantaged students. Uncovering the reasons for this difference may provide important insights into successful interventions with low performing subgroups.

This study shows some positive effects of subgroup-specific accountability on student achievement on a high stakes test. This study is not designed to examine the validity of these achievement benefits or whether these benefits came at the expense of other, non-tested, subjects such as science or social studies. Ideally, an “audit test” would be available to determine whether these high stakes test score results generalized to another test with the same content. Unfortunately no such test is available in North Carolina. Future research using both high and low stakes tests should explore the validity and reliability of accountability-induced achievement effects.

A widespread concern is that NCLB has created an incentive for states to lower proficiency standards as a way to game the system – a so-called “race to the bottom.” North Carolina’s case is an interesting counterexample. North Carolina has traditionally been a low standards state, but rather than lowering standards during the post-NCLB period, it instead raised

them. Between 2000 and 2005 the percent of third through eighth grade students at grade level (i.e., “proficient”) math increased from 80% to 88%. In response to national reports showing the wide disparity between North Carolina’s percent proficient on the state tests and percent proficient on NAEP,¹⁶ state policymakers implemented more difficult math standards in 2006 and received permission from the federal DOE to lower the AYP target. In 2006, percent at grade level in math fell from 88 to 64. In 2008, the state implemented tougher reading standards, again received permission to lower the AYP target and again percent at grade level fell. More recently, NC was chosen as a Round II Race to the Top state. In 2012-2013, the state will be among the first to implement common core standards. So the trend in NC has been to increase rigor, not water it down. Unfortunately the increase in rigor may have imposed short term costs on low achieving and high achieving students in math as schools struggled to get many more students up to grade level proficiency. This is not to say that states should not raise standards. These short run costs must be weighed against the potential of raising achievement for all students in the long run and should be taken into account as policymakers consider large and sudden changes in the rigor of state standards.

NCLB’s focus on racial minorities, and economically and educationally disadvantaged subgroups is in keeping with the traditional federal role of compensatory education and enhancing equality of educational opportunity. Our results show positive effects of accountability based on a “status” measure of achievement, AYP, and that subgroup-specific accountability pressure has had positive effects for poor and minority students in the lowest achieving schools, although an increase in rigor in state standards appeared to contribute to test score declines for

¹⁶ Lou Fabrizio, Federal Liaison, North Carolina Department of Public Instruction, personal communication, May 20, 2011.

the lowest achieving and highest achieving students in schools facing accountability pressure. On balance, therefore, our findings provide some empirical justification for a continued federal role in tracking test-score gaps and subgroup performance as a part of educational accountability. While these findings suggest benefits of tracking subgroup performance, it is also clear that the method used by NCLB has serious flaws, chief among them making it much more likely for diverse schools with many subgroups to be sanctioned. The method for tracking subgroup-level performance could no doubt be improved. For example, schools could track and be held accountable for the performance growth of students at various points in the achievement distribution, rather than for test score levels of nine separate subgroups. This is but one possibility scholars and policymakers may want to consider in the design of the next generation of accountability systems.

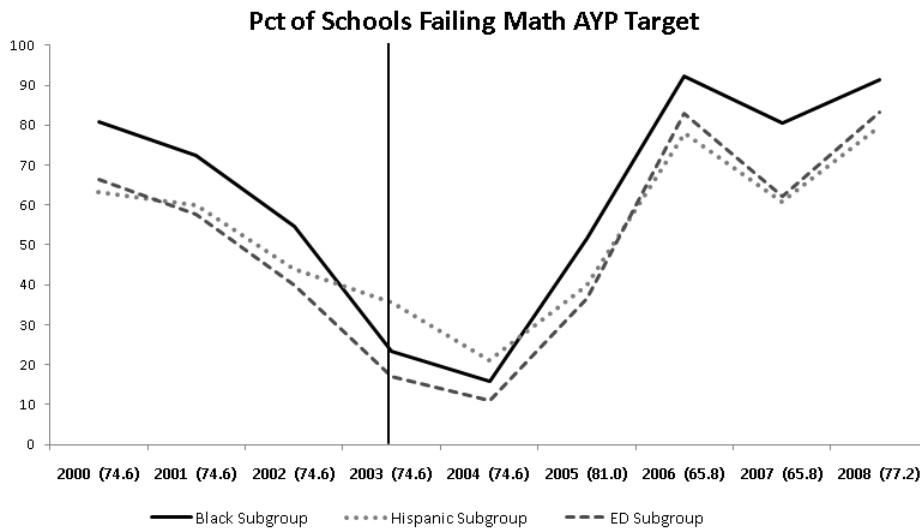
REFERENCES

- Amrien, A. L., & Berliner, D. C. (2002). High-Stakes Testing, Uncertainty, and Student Learning. *Education Policy Analysis Archives*, 10(18).
- Au, W. (2007). High-stakes testing and curricular control: A qualitative metasynthesis. *Educational Researcher*, 36(5), 258.
- Booher-Jennings, J. (2005). Below the bubble: "Educational triage" and the Texas Accountability System. *American Educational Research Journal*, 42(2), 231-268.
- Brown, A. B., & Clift, J. W. (2010). The Unequal Effect of Adequate Yearly Progress: Evidence From School Visits. *American Educational Research Journal*, 47(4), 774-798.
- Campbell, D. T. (1979). Assessing the Impact of Planned Social Change. *Evaluation and Program Planning*, 2, 67-90.
- Carnoy, M., & Loeb, S. (2002). Does External Accountability Affect Student Outcomes? A Cross-State Analysis. *Educational Evaluation and Policy Analysis*, 24(4), 305-331.
- Chiang, H. (2009). How accountability pressure on failing schools affects student achievement. *Journal of Public Economics*, 93(9-10), 1045-1057. doi: 10.1016/j.jpubeco.2009.06.002
- Coburn, C. E. (2004). Beyond decoupling: Rethinking the relationship between the institutional environment and the classroom. *Sociology of Education*, 77(3), 211-244.
- Cullen, J. B., & Reback, R. (2006). Tinkering Towards Accolades: School Gaming under a Performance Accountability System. In T. J. Gronberg & D. W. Jansen (Eds.), *Advances in Applied Microeconomics* (Vol. 14, pp. 1-34). Oxford, UK: Elsevier.
- Darling-Hammond, L. (2004). Standards, Accountability, and School Reform. *Teachers College Record*, 106(6), 1047-1085.
- Dee, T. S., & Jacob, B. (2011). The Impact of No Child Left Behind on Student Achievement. *Journal of Policy Analysis and Management, Forthcoming*.
- Diamond, J. B. (2007). Where the rubber meets the road: Rethinking the connection between high-stakes testing policy and classroom instruction. *Sociology of Education*, 80(4), 285-313.
- Diamond, J. B., & Spillane, J. (2004). High-Stakes Accountability in Urban Elementary Schools: Challenging or Reproducing Inequality? *Teachers College Record*, 106(6), 1145-1176.
- Figlio, D. N., & Getzer, L. (2006). Accountability, Ability, and Disability: Gaming the System? In T. J. Gronberg & D. W. Jansen (Eds.), *Advances in Applied Microeconomics* (Vol. 14, pp. 35-50). Oxford, UK: Elsevier.
- Figlio, D. N., & Loeb, S. (2011). School Accountability. In E. A. Hanushek, S. Machin & L. Woessmann (Eds.), *Handbook of the Economics of Education* (Vol. 3, pp. 383-421). North-Holland, The Netherlands: Elsevier.
- Figlio, D. N., & Rouse, C. E. (2006). Do Accountability and Voucher Threats Improve Low-Performing Schools? *Journal of Public Economics*, 90((1-2)), 239-255.
- Figlio, D. N., Rouse, C. E., & Schlosser, A. (2009). *Leaving No Child Behind: Two Paths to School Accountability*. Working Paper. Working Paper.
- Gardner, D. P. (1983). *A Nation at Risk: The Imperative for Educational Reform*. Washington, D.C.: U.S. Government Printing Office.
- Hallett, T. (2010). The Myth Incarnate: Recoupling Processes, Turmoil, and Inhabited Institutions in an Urban Elementary School. *American Sociological Review*, 75(1), 52-74.

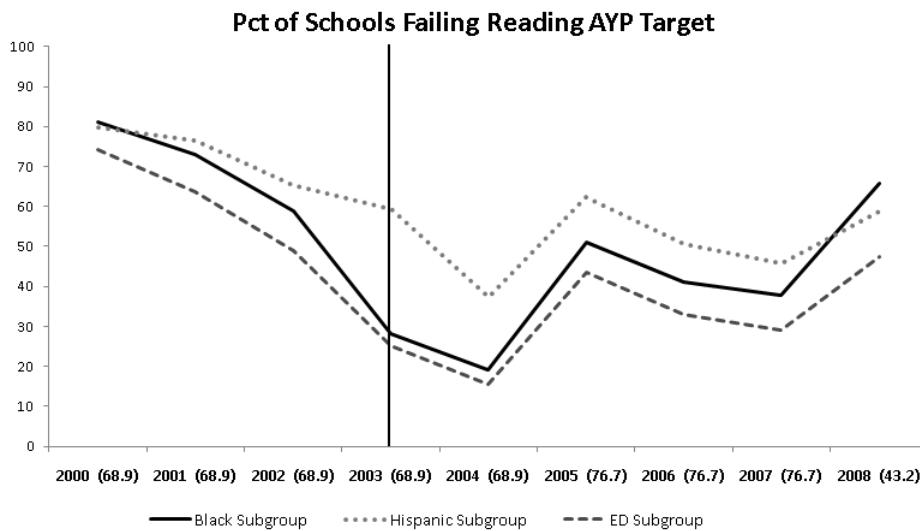
- Hanushek, E. A., & Raymond, M. E. (2005). Does School Accountability Lead to Improved Student Performance? *Journal of Policy Analysis and Management*, 24(2), 297-327.
- Jacob, B. A. (2005). Accountability, incentives and behavior: the impact of high-stakes testing in the Chicago Public Schools. *Journal of Public Economics*, 89(5-6), 761-796. doi: 10.1016/j.jpubeco.2004.08.004
- Jacob, B. A., & Lefgren, L. (2004). The impact of teacher training on student achievement - Quasi-experimental evidence from school reform efforts in Chicago. *Journal of Human Resources*, 39(1), 50-79.
- Jacob, B. A., & Levitt, S. D. (2003). Rotten apples: An investigation of the prevalence and predictors of teacher cheating. *Quarterly Journal of Economics*, 118(3), 843-877.
- Krieg, J. M. (2008). Are Students Left Behind? The Distributional Effects of the No Child Left Behind Act. *Education Finance and Policy*, 3(2), 250-281.
- Ladd, H. F., & Lauen, D. L. (2010). Status versus growth: The distributional effects of school accountability policies. *Journal of Policy Analysis and Management*, 29(3), 426-450.
- Lee, J., & Wong, K. K. (2004). The Impact of Accountability on Racial and Socioeconomic Equity: Considering Both School Resources and Achievement Outcomes. *American Educational Research Journal*, 41(4), 797-832.
- Linn, R. L. (2000). Assessments and Accountability. *Educational Researcher*, 29(2), 4-16.
- McNeil, L. M. (2000). *Contradictions of School Reform: The Educational Costs of Standardized Testing*. London: Routledge.
- Mehta, J. (2006). *The Transformation of American Educational Policy, 1980-2001*. Unpublished Dissertation.
- Neal, D., & Schanzenbach, D. W. (2010). Left Behind by Design: Proficiency Counts and Test-Based Accountability. *Review of Economics and Statistics*, 92(2), 263-283.
- O'Day, J. A., & Smith, M. S. (1993). Systemic Reform and Educational Opportunity. In S. Fuhrman (Ed.), *Designing Coherent Education Policy: Improving the System* (pp. 250-312). San Francisco: Jossey Bass.
- Pedulla, J. L., Abrams, L., Madaus, G., Russell, M., Ramos, M., & Miao, J. (2003). Perceived Effects of State-Mandated Testing Programs on Teaching and Learning: Findings from a National Survey of Teachers. Boston, MA: National Board on Education Testing and Public Policy.
- Reback, R. (2008). Teaching to the rating: School accountability and the distribution of student achievement. *Journal of Public Economics*, 92(5-6), 1394-1415. doi: 10.1016/j.jpubeco.2007.05.003
- Reback, R., Rockoff, J., & Schwartz, H. L. (2011). *Under Pressure: Job Security, Resource Allocation, and Productivity in Schools Under NCLB*. Working Paper.
- Springer, M. G. (2008). The influence of an NCLB accountability plan on the distribution of student test score gains. *Economics of Education Review*, 27(5), 556-563. doi: 10.1016/j.econedurev.2007.06.004
- Wong, M., Cook, T. D., & Steiner, P. M. (2009). *No Child Left Behind: An Interim Evaluation of its Effects on Learning using two Interrupted Time Series each with its own Non-Equivalent Comparison Series*. Working paper.

Figure 1. Percent of Schools Failing and Not Accountable for AYP Targets. Note: percent of schools failing is out of total number of schools accountable for the focal subgroup. Yearly AYP targets are noted in parentheses on the x-axis. The vertical line denotes the pre- and post-NCLB periods.

A.



B.



C.

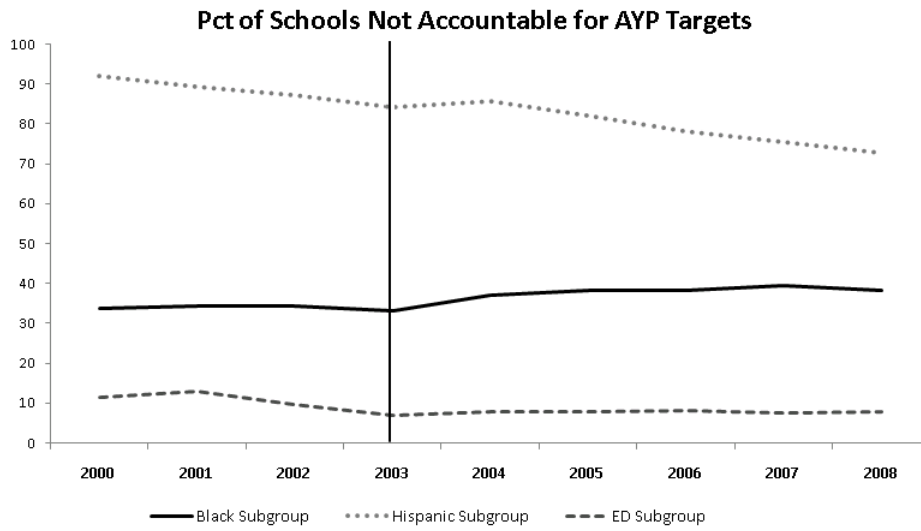
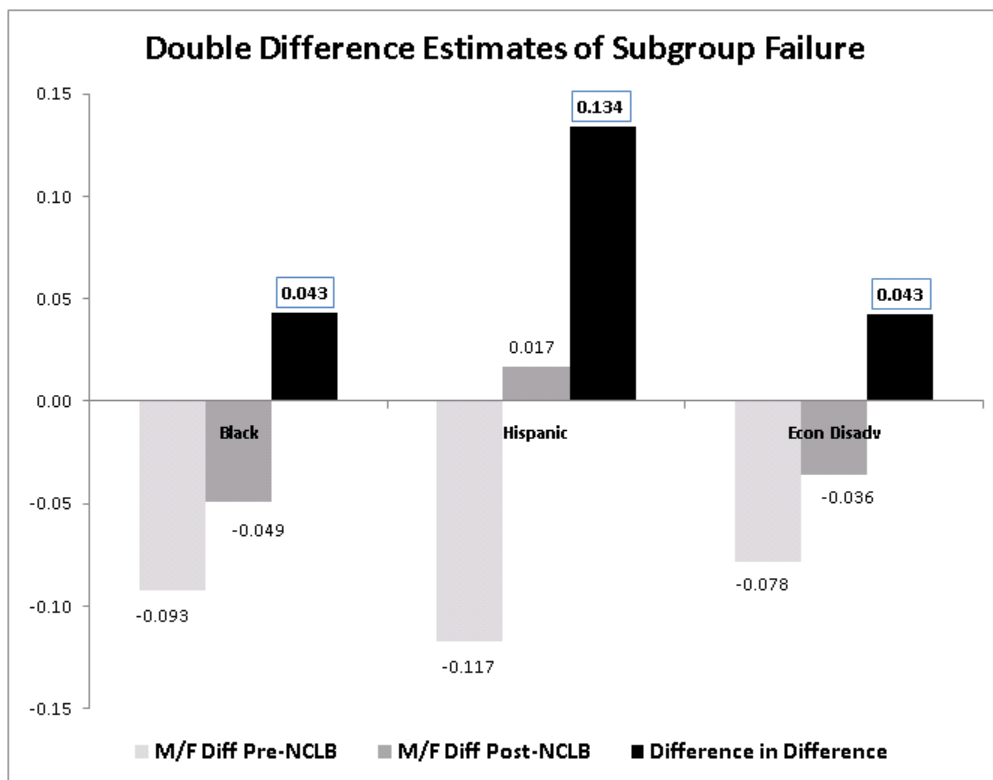


Figure 2.



Note: Coefficients are from school fixed effects models for math (see Table 2, Model 2).

Table 1. Descriptive Statistics

	Obs	Mean	SD	Min	Max
Dependent Variables					
Standardized math score	5,531,106	0	1	-4.66	3.65
Standardized reading score	5,509,798	0	1	-4.11	3.13
Student Background					
Black student	5,716,284	0.294	0.456	0	1
Hispanic student	5,716,284	0.069	0.254	0	1
Male student	5,716,284	0.511	0.500	0	1
Student was designated gifted	5,716,284	0.137	0.342	0	1
Student received special education services	5,694,479	0.135	0.342	0	1
Student showed Limited English Proficiency	5,704,907	0.041	0.181	0	1
Student was ever retained	5,331,476	0.073	0.260	0	1
Student made a structural school move	5,716,223	0.134	0.341	0	1
Student made a non-structural school move	5,716,223	0.081	0.272	0	1
Student was economically disadvantaged	5,116,304	0.454	0.498	0	1
Parent has some postsecondary education	5,227,372	0.212	0.408	0	1
Parent has bachelor's degree or higher	5,227,372	0.275	0.447	0	1
Student was well below grade level in test score distribution	5,531,106	0.094	0.291	0	1
Student was marginal in test score distribution	5,531,106	0.266	0.442	0	1
Student was well above grade level in test score distribution	5,531,106	0.640	0.480	0	1
Accountability Pressure Main Effects (lagged one year)					
Black subgroup failed math target	3,675,191	0.602	0.489	0	1
Hispanic subgroup failed math target	1,291,258	0.543	0.498	0	1
Econ disadv subgroup failed math target	4,771,252	0.512	0.500	0	1
Black subgroup failed reading target	3,675,754	0.468	0.499	0	1
Hispanic subgroup failed reading target	1,290,099	0.548	0.498	0	1
Econ disadv subgroup failed reading target	4,771,291	0.424	0.494	0	1
Time					
Post-NCLB	5,716,284	0.564	0.496	0	1
Year	5,716,284	2004	2.580	2000	2008

Table 2. Estimated DD Effects of Subgroup-Specific NCLB Accountability Pressure on Standardized Test Scores

	Math		Reading	
	(1) Pooled OLS	(2) School FE	(3) Pooled OLS	(4) School FE
Blacks				
Failed Black SG*PostNCLB	0.0723 ^{***}	0.0432 ^{***}	0.0545 ^{***}	0.0542 ^{***}
	(0.00973)	(0.00824)	(0.0100)	(0.00788)
N	1,193,725	1,193,725	1,189,405	1,189,405
Hispanics				
Failed Hispanic SG*PostNCLB	0.128 ^{***}	0.134 ^{***}	0.0903 ^{**}	0.108 ^{***}
	(0.0269)	(0.0233)	(0.0307)	(0.0252)
N	143,722	143,722	141,687	141,687
Econ Disadv				
Failed ED*PostNCLB	0.0691 ^{***}	0.0425 ^{***}	0.0524 ^{***}	0.0485 ^{***}
	(0.00760)	(0.00653)	(0.00805)	(0.00647)
N	1,866,282	1,866,282	1,854,949	1,854,949

Note: Each coefficient is from a separate model using only students from the focal subgroup. All models control for student characteristics: gifted, special education, limited English proficiency, sex, if the student was ever retained, parental education, student poverty status, and whether the student made a school move (both district mandated and otherwise). Economically disadvantaged subgroup models do not include the student poverty covariate but include race instead. Cluster-corrected standard errors in parentheses.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 3. Estimated DD Effects of Subgroup-Specific NCLB Accountability Pressure on Standardized Test Scores – Alternative Specification Taking into Account Other Subgroup Status

	(1)	(2)
	Math	Reading
	School FE	School FE
Blacks		
Failed Black SG only*PostNCLB	0.0387 ⁺ (0.0204)	-0.00766 (0.0227)
Failed Black SG & at least 1 other SG Failed*PostNCLB	0.0645 ^{****a} (0.0107)	0.0732 ^{***a} (0.00990)
Met Black SG & at least 1 other SG Failed*PostNCLB	0.0346 ^{**a} (0.0112)	0.0270 ^{**a} (0.0101)
N	1,193,725	1,189,405
Hispanics		
Failed Hisp SG only*PostNCLB	0.0412 (0.0958)	-0.0116 (0.0838)
Failed Hisp SG & at least 1 other SG Failed*PostNCLB	0.232 ^{***a} (0.0402)	0.200 ^{***a} (0.0431)
Met Hisp SG & at least 1 other SG Failed*PostNCLB	0.118 ^{**a} (0.0414)	0.109 ^{*a} (0.0430)
N	143,722	141,687
Econ Disadv		
Failed ED SG only*PostNCLB	0.0242 (0.0168)	0.0287 [*] (0.0143)
Failed ED SG & at least 1 other SG Failed*PostNCLB	0.0653 ^{***a} (0.00795)	0.0696 ^{***a} (0.00774)
Met ED SG & at least 1 other SG Failed*PostNCLB	0.0378 ^{***a} (0.00770)	0.0320 ^{***a} (0.00726)
N	1,866,282	1,854,949

Note: Each coefficient is from a separate model using only students from the focal subgroup. All models control for student characteristics: gifted, special education, limited English proficiency, sex, if the student was ever retained, parental education, student poverty status, and whether the student made a school move (both district mandated and otherwise). Economically disadvantaged subgroup models do not include the student poverty covariate but include race instead. Cluster-corrected standard errors in parentheses.

⁺ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$; a indicates two category effects are significantly different from each other ($p < 0.05$)

Table 4. Estimated DD Effects of Subgroup-Specific NCLB Accountability Pressure on Standardized Test Scores – Alternative Specification Using School’s Distance between Subgroup Performance and AYP Target

	(1) Math School FE	(2) Reading School FE
Blacks		
Black SG at Margin of Meeting AYP +/- 10% around cutoff*PostNCLB	0.0275 ^{* a} (0.0127)	0.0361 ^{***} (0.0108)
Black SG Below Margin of Meeting AYP > 10% below cutoff*PostNCLB	0.0716 ^{*** a} (0.0148)	0.0511 ^{**} (0.0162)
N	1,198,598	1,194,222
Hispanics		
Hispanic SG at Margin of Meeting AYP +/- 10% around cutoff*PostNCLB	0.0813 ^{** a} (0.0279)	0.107 ^{** a} (0.0370)
Hispanic SG Below Margin of Meeting AYP > 10% below cutoff*PostNCLB	0.272 ^{*** a} (0.0306)	0.189 ^{*** a} (0.0368)
N	153,508	151,419
Econ Disadv		
ED SG at Margin of Meeting AYP +/- 10% around cutoff*PostNCLB	0.0348 ^{*** a} (0.00785)	0.0391 ^{*** a} (0.00749)
ED SG Below Margin of Meeting AYP > 10% below cutoff*PostNCLB	0.0844 ^{*** a} (0.0104)	0.0688 ^{*** a} (0.0140)
N	1,866,566	1,855,211

Note: Each coefficient is from a separate model using only students from the focal subgroup. All models control for student characteristics: gifted, special education, limited English proficiency, sex, if the student was ever retained, parental education, student poverty status, and whether the student made a school move (both district mandated and otherwise). Economically disadvantaged subgroup models do not include the student poverty covariate but include race instead. Cluster-corrected standard errors in parentheses. In each trichotomy model, the accountability pressure variables are specified with reference to the distance between the subgroup performance and the AYP target, where the reference category is > 10% above cutoff*PostNCLB.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$; a indicates middle and low subgroup cutoff effects are significantly different from each other ($p < 0.05$)

Table 5. Estimated DDD Effects of Subgroup-Specific NCLB Accountability Pressure on Standardized Test Scores – by Student Position in the Prior Test Score Distribution

	Math School FE		Reading School FE	
	(1) 2000-2008	(2) 2000-2006	(3) 2000-2008	(4) 2000-2006
Black SG				
Failed Black SG*PostNCLB (at margin) ($\leq +0.5$ SD & ≥ -0.5 SD)	0.305 ^{***} (0.0102)	0.0745 ^{***} (0.0101)	-0.0016 (0.0081)	-0.0028 (0.0081)
Failed Black SG*PostNCLB*LowScore (< -0.5 SD)	-0.0953 ^{*** a} (0.0126)	0.0030 ^a (0.0133)	0.0768 ^{***a} (0.0122)	0.0792 ^{*** a} (0.0130)
Failed Black SG*PostNCLB*HighScore (> +0.5 SD)	-0.143 ^{*** a} (0.00915)	-0.0436 ^{*** a} (0.0097)	-0.0322 ^{***a} (0.0091)	-0.0310 ^{** a} (0.0096)
N	1,003,075	780,377	997,921	775,760
Hispanic SG				
Failed Hispanic SG*PostNCLB (at margin) ($\leq +0.5$ SD & ≥ -0.5 SD)	0.260 ^{***} (0.0280)	0.0670 [*] (0.0270)	0.0260 (0.0246)	0.0339 (0.0248)
Failed Hispanic SG*PostNCLB*LowScore (< -0.5 SD)	-0.0900 [*] (0.0368)	0.0209 (0.0412)	-0.0177 (0.0379)	-0.0263 (0.0392)
Failed Hispanic SG*PostNCLB*HighScore (> +0.5 SD)	-0.0623 [*] (0.0254)	-0.0026 (0.0265)	-0.0298 (0.0292)	-0.0352 (0.0304)
N	117,796	69,996	115,862	68,733
Economically disadvantaged SG				
Failed ED SG*PostNCLB (at margin) ($\leq +0.5$ SD & ≥ -0.5 SD)	0.270 ^{***} (0.00867)	0.0670 ^{***} (0.0086)	-0.0071 (0.0064)	-0.0116 ⁺ (0.0065)
Failed ED SG*PostNCLB*LowScore (< -0.5 SD)	-0.0889 ^{*** a} (0.00992)	0.0246 ^{* a} (0.0109)	0.0661 ^{***a} (0.0100)	0.0749 ^{*** a} (0.0110)
Failed ED SG*PostNCLB*HighScore (> +0.5 SD)	-0.108 ^{*** a} (0.00751)	-0.0331 ^{*** a} (0.0084)	-0.0268 ^{***a} (0.0068)	-0.0237 ^{** a} (0.00734)
N	1,534,749	1,166,207	1,522,845	1,156,343

Note: Each coefficient is from a separate model using only students from the focal subgroup. All models control for student characteristics: gifted, special education, limited English proficiency, sex, if the student was ever retained, parental education, student poverty status, and whether the student made a school move (both district mandated and otherwise). Economically disadvantaged subgroup models do not include the student poverty covariate but include race instead. Cluster-corrected standard errors in parentheses.

⁺ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$; a indicates low and high distribution effects are significantly different from each other ($p < 0.05$)

Table A1. School-Years That Schools are Accountable or Not Accountable for Subgroup in Math for All Subgroups

	Accountable		Not Accountable	
Economically Disadvantaged	9,952	92.38%	821	7.62%
White	9,114	84.60%	1,659	15.40%
Black	6,770	62.84%	4,003	37.16%
Special Education	4,860	45.11%	5,913	54.89%
Hispanic	2,214	20.55%	8,559	79.45%
LEP	1,156	10.73%	9,617	89.27%
Asian	363	3.37%	10,410	96.63%
Native American	268	2.49%	10,505	97.51%
Multi-Racial	131	1.22%	10,642	98.78%

Table A2. Number of Schools Failing a Subgroup Target by Number of Years Failed

	0	1	2	3	4	5	6
Black	763	162	175	333	297	117	119
	38.81%	8.24%	8.90%	16.94%	15.11%	5.95%	6.05%
Hispanic	1,484	149	103	92	74	34	30
	75.48%	7.58%	5.24%	4.68%	3.76%	1.73%	1.53%
Special Ed	857	255	167	135	132	99	321
	43.59%	12.97%	8.49%	6.87%	6.71%	5.04%	16.33%
Economically Disadvantaged	368	275	282	488	306	118	129
	18.72%	13.99%	14.34%	24.82%	15.56%	6.00%	6.56%

Note: Data from NCLB period (2003-2008) only. Years during which a subgroup is not accountable are treated as non-failing years.

Table A3. Estimated Effects of Subgroup-Specific NCLB Accountability Pressure on Standardized Test Scores – Alternative Specification Using School’s Distance between Subgroup Performance and AYP Target

	(1) Math School FE	(2) Reading School FE
Blacks		
0 to 10% above cutoff*PostNCLB	0.0012 ^{abc} (0.0132)	0.0040 ^{ab} (0.0112)
0 to 10% below cutoff*PostNCLB	0.0355 ^{**abc} (0.0134)	0.0291 ^{*ab} (0.0117)
≥ 10% below cutoff*PostNCLB	0.0752 ^{***abc} (0.0148)	0.0499 ^{**ab} (0.0161)
N	1,198,598	1,194,222
Hispanics		
0 to 10% above cutoff*PostNCLB	0.0535 ^{+abc} (0.0298)	0.0701 ^{+bc} (0.0403)
0 to 10% below cutoff*PostNCLB	0.116 ^{***abc} (0.0317)	0.121 ^{**bc} (0.0393)
≥ 10% below cutoff*PostNCLB	0.273 ^{***abc} (0.0308)	0.188 ^{***bc} (0.0372)
N	153,508	151,419
Econ Disadv		
0 to 10% above cutoff*PostNCLB	0.0156 ^{+abc} (0.0081)	0.0221 ^{**bc} (0.0076)
0 to 10% below cutoff*PostNCLB	0.0492 ^{***abc} (0.0091)	0.0292 ^{***bc} (0.0088)
≥ 10% below cutoff*PostNCLB	0.0869 ^{***abc} (0.0104)	0.0663 ^{***bc} (0.0138)
N	1,866,566	1,855,211

Note: Each coefficient is from a separate model using only students from the focal subgroup. All models control for student characteristics: gifted, special education, limited English proficiency, sex, if the student was ever retained, parental education, student poverty status, and whether the student made a school move (both district mandated and otherwise). Economically disadvantaged subgroup models do not include the student poverty covariate but include race instead. Cluster-corrected standard errors in parentheses. In each model the reference category is > 10% above SG cutoff*PostNCLB.

+ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$; a indicates just above and just below subgroup cutoff effects are significantly different from each other, b indicates just above and well below subgroup cutoff effects are significantly different from each other, and c indicates just below and well below subgroup cutoff effects are significantly different from each other ($p < 0.05$)

Table A4. Estimated DDD Effects of Subgroup-Specific NCLB Accountability Pressure on Standardized Test Scores – by Student Position in the Prior Test Score Distribution

	Math School FE		Reading School FE	
	(1) 2000-2008	(2) 2000-2006	(1) 2000-2008	(2) 2000-2006
Black SG				
Failed Black SG*PostNCLB (At Margin) ($\leq +0.25$ SD & ≥ -0.25 SD)	0.304 ^{***} (0.0107)	0.0629 ^{***} (0.0105)	-0.0141 (0.0092)	-0.0174 ⁺ (0.00956)
Failed Black SG*PostNCLB*WellBelowGradeLvl (< -0.75 SD)	-0.0735 ^{***} (0.01618)	0.0322 ⁺ (0.0177)	0.0840 ^{***} (0.0151)	0.0885 ^{***} (0.0161)
Failed Black SG*PostNCLB*SomewhatBelowGradeLvl (≤ -0.25 SD & ≥ -0.75 SD)	0.0041 (0.0102)	0.000680 (0.0105)	0.0387 ^{***} (0.0111)	0.0413 ^{***} (0.0122)
Failed Black SG*PostNCLB*SomewhatAboveGradeLvl ($\leq +0.75$ SD & $\geq +0.25$ SD)	-0.0176 [*] (0.0077)	-0.0132 (0.00818)	-0.0269 ^{**} (0.0091)	-0.0183 ⁺ (0.0101)
Failed Black SG*PostNCLB*WellAboveGradeLvl ($> +0.75$ SD)	-0.1481 ^{***} (0.0109)	-0.0335 ^{**} (0.0115)	-0.0142 (0.0109)	-0.00515 (0.0116)
N	1,003,075	780,377	997,921	775,760
Hispanic SG				
Failed Hispanic SG*PostNCLB (At Margin) ($\leq +0.25$ SD & ≥ -0.25 SD)	0.261 ^{***} (0.0296)	0.0419 (0.0292)	0.0075 (0.0293)	0.0212 (0.0301)
Failed Hispanic SG*PostNCLB*WellBelowGradeLvl (< -0.75 SD)	-0.133 ^{***} (0.0474)	0.0232 (0.0542)	0.0013 (0.0466)	-0.00314 (0.0496)
Failed Hispanic SG*PostNCLB*SomewhatBelowGradeLvl (≤ -0.25 SD & ≥ -0.75 SD)	-0.0152 (0.0331)	0.0543 (0.0371)	0.0269 (0.0394)	0.00217 (0.0431)
Failed Hispanic SG*PostNCLB*SomewhatAboveGradeLvl ($\leq +0.75$ SD & $\geq +0.25$ SD)	0.0125 (0.0280)	-0.00418 (0.0305)	-0.0119 (0.0294)	-0.00905 (0.0315)
Failed Hispanic SG*PostNCLB*WellAboveGradeLvl ($> +0.75$ SD)	-0.0666 [*] (0.0291)	0.0248 (0.0307)	-0.0117 (0.0337)	-0.0236 (0.0356)
N	117,796	69,996	115,862	68,733
Economically disadvantaged SG				
Failed ED SG*PostNCLB (At Margin) ($\leq +0.25$ SD & ≥ -0.25 SD)	0.278 ^{***} (0.0093)	0.0506 ^{***} (0.00904)	-0.0248 ^{***} (0.0077)	-0.0311 ^{***} (0.00805)
Failed ED SG*PostNCLB*WellBelowGradeLvl (< -0.75 SD)	-0.0847 ^{***} (0.0130)	0.0638 ^{***} (0.0148)	0.0883 ^{***} (0.0121)	0.102 ^{***} (0.0133)
Failed ED SG*PostNCLB*SomewhatBelowGradeLvl (≤ -0.25 SD & ≥ -0.75 SD)	-0.0121 (0.0086)	0.0129 (0.00923)	0.0309 ^{***} (0.0095)	0.0333 ^{**} (0.0104)
Failed ED SG*PostNCLB*SomewhatAboveGradeLvl ($\leq +0.75$ SD & $\geq +0.25$ SD)	-0.0105 (0.0065)	-0.0157 [*] (0.00697)	-0.0046 (0.0075)	-0.000211 (0.00824)
Failed ED SG*PostNCLB*WellAboveGradeLvl ($> +0.75$ SD)	-0.124 ^{***} (0.0087)	-0.0259 ^{**} (0.00954)	-0.0064 (0.0083)	0.00310 (0.00901)
N	1,534,749	1,166,207	1,522,845	1,156,343

Note: Each coefficient is from a separate model using only students from the focal subgroup. All models control for student characteristics: gifted, special education, limited English proficiency, sex, if the student was ever retained, parental education, student poverty status, and whether the student made a school move (both district mandated and otherwise). Economically disadvantaged models do not include the student poverty covariate but include race instead. Cluster-corrected standard errors in parentheses

⁺ $p < 0.10$, ^{*} $p < 0.05$, ^{**} $p < 0.01$, ^{***} $p < 0.001$

Table A5. Estimated DDD Effects of Subgroup-Specific NCLB Accountability Pressure on Standardized Test Scores – by Student Position in the Prior Test Score Distribution

	Math School FE		Reading School FE	
	(1) 2000-2008	(2) 2000-2006	(1) 2000-2008	(2) 2000-2006
Blacks				
Failed Black SG*PostNCLB (Just Below Grade Lvl) (≤ 0.00 SD & > -0.25 SD)	0.326*** (0.0127)	0.0837*** (0.0131)	-0.0064 (0.0122)	-0.00552 (0.0127)
Failed Black SG*PostNCLB*WellBelowGradeLvl (≤ -0.75 SD)	-0.0947*** (0.0171)	0.0116 (0.0186)	0.0763*** (0.0166)	0.0767*** (0.0176)
Failed Black SG*PostNCLB*SomewhatBelowGradeLvl (≤ -0.25 SD & > -0.75 SD)	-0.0174 (0.0117)	-0.0200 (0.0123)	0.0310* (0.0130)	0.0295* (0.0141)
Failed Black SG*PostNCLB*JustAboveGradeLvl (≤ 0.25 SD & > 0.00 SD)	-0.0399*** (0.0113)	-0.0399*** (0.0119)	-0.0135 (0.0139)	-0.0191 (0.0150)
Failed Black SG*PostNCLB*SomewhatAboveGradeLvl (≤ 0.75 SD & > 0.25 SD)	-0.0396*** (0.0103)	-0.0339* (0.0110)	-0.0347** (0.0118)	-0.0302* (0.0129)
Failed Black SG*PostNCLB*WellAboveGradeLvl (> 0.75 SD)	-0.170*** (0.0109)	-0.0542*** (0.0142)	-0.0220 (0.0135)	-0.0170 (0.0143)
N	1,003,075	780,377	997,921	775,760
Hispanics				
Failed Hispanic SG*PostNCLB (Just Below Grade Lvl) (≤ 0.00 SD & > -0.25 SD)	0.212*** (0.0355)	-0.00420 (0.0377)	0.0233 (0.0405)	0.0417 (0.0425)
Failed Hispanic SG*PostNCLB*WellBelowGradeLvl (≤ -0.75 SD)	-0.0838+ (0.0501)	0.0691 (0.0578)	-0.0147 (0.0543)	-0.0239 (0.0569)
Failed Hispanic SG*PostNCLB*SomewhatBelowGradeLvl (≤ -0.25 SD & > -0.75 SD)	0.0340 (0.0378)	0.100* (0.0427)	0.0108 (0.0493)	-0.0187 (0.0529)
Failed Hispanic SG*PostNCLB*JustAboveGradeLvl (≤ 0.25 SD & > 0.00 SD)	0.0767+ (0.0403)	0.0711 (0.0443)	-0.0285 (0.0519)	-0.0344 (0.0568)
Failed Hispanic SG*PostNCLB*SomewhatAboveGradeLvl (≤ 0.75 SD & > 0.25 SD)	0.0617+ (0.0350)	0.0415 (0.0392)	-0.0279 (0.0401)	-0.0299 (0.0436)
Failed Hispanic SG*PostNCLB*WellAboveGradeLvl (> 0.75 SD)	-0.0173* (0.0349)	0.0707+ (0.0386)	-0.0277 (0.0429)	-0.0444 (0.0460)
N	117,796	69,996	115,862	68,733
Econ Disadv				
Failed ED SG*PostNCLB (Just Below Grade Lvl) (≤ 0.00 SD & > -0.25 SD)	0.288*** (0.0104)	0.0664*** (0.0105)	-0.0316** (0.0098)	-0.0355*** (0.0103)
Failed ED SG*PostNCLB*WellBelowGradeLvl (≤ -0.75 SD)	-0.0947*** (0.0135)	0.0482* (0.0154)	0.0951*** (0.0132)	0.106*** (0.0145)
Failed ED SG*PostNCLB*SomewhatBelowGradeLvl (≤ -0.25 SD & > -0.75 SD)	-0.0212* (0.0095)	-0.00280 (0.0104)	0.0377*** (0.0110)	0.0376** (0.0121)
Failed ED SG*PostNCLB*JustAboveGradeLvl (≤ 0.25 SD & > 0.00 SD)	-0.0202* (0.0090)	-0.0336*** (0.00988)	0.0137 (0.0117)	0.0113 (0.0126)
Failed ED SG*PostNCLB*SomewhatAboveGradeLvl (≤ 0.75 SD & > 0.25 SD)	-0.0198* (0.0083)	-0.0314*** (0.00895)	0.0023 (0.0099)	0.00415 (0.0108)
Failed ED SG*PostNCLB*WellAboveGradeLvl (> 0.75 SD)	-0.133*** (0.0103)	-0.0417*** (0.0112)	0.0005 (0.0103)	0.00748 (0.0111)
N	1,534,749	1,166,207	1,522,845	1,156,343

Note: Each coefficient is from a separate model using only students from the focal subgroup. All models control for student characteristics: gifted, special education, limited English proficiency, sex, if the student was ever retained, parental education, student poverty status, and whether the student made a school move (both district mandated and otherwise). Economically disadvantaged models do not include the student poverty covariate but include race instead. Cluster-corrected standard errors in parentheses

+ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$