Under Pressure:
Job Security, Resource Allocation, and Productivity in Schools Under *NCLB*

Randall Reback
Barnard College

Jonah Rockoff
Columbia Business School and NBER

Heather L. Schwartz
RAND Corporation

September 12, 2012*

Abstract

The most sweeping federal education law in decades, the *No Child Left Behind* (NCLB) *Act*, requires states to administer standardized exams and to punish schools that do not make Adequate Yearly Progress (AYP) for the fraction of students passing these exams. While the literature on school accountability is well-established, there exists no nationwide study of the strong short-term incentives NCLB imposes for schools on the margin of failing AYP. We assemble the first comprehensive, national school-level dataset containing detailed performance measures used to calculate AYP, and demonstrate that idiosyncrasies in state policies create numerous cases where schools near the margin for satisfying their *own* state's AYP requirements would have almost certainly failed or almost certainly made AYP if they were located in *other* states. Using this variation as a means of identification, we examine the impact of NCLB on the behavior of school personnel and students' academic achievement in nationally representative samples. We find that accountability pressure from NCLB reduces teachers' perceptions of job security, especially among relatively inexperienced teachers. We also find evidence of reading and math specialist teachers working longer hours, generalist teachers (who teach multiple subjects) working fewer hours, a shift away from whole-class instruction, and reduced instruction in science and social studies. However, we also find that NCLB pressure has either neutral or positive effects on students' enjoyment of learning and their achievement gains on low-stakes exams in reading, math, and science.

---

* Email: (Reback) rr2165@columbia.edu, (Rockoff) jr2331@columbia.edu, (Schwartz) hschwart@rand.org.

We thank Abigail Payne, Steve Rivkin and David Figlio for their detailed comments, as well as participants at the American Economics Association meetings, the CALDER/Urban Institute NCLB Research Conference, the Association for Public Policy and Management conference, the Chinese University of Hong Kong, Teachers College, the International Workshop on Applied Economics of Education, and the American Education Finance Association conference for many thoughtful suggestions. This research project was made possible by funding from the Institute for Education Sciences and the Spencer Foundation, as well as seed grants from the Columbia University Institute for Social and Economic Research and Policy and Barnard College, and support from the Paul Milstein Center for Real Estate at Columbia Business School. The authors are solely responsible for any opinions or errors in the paper. Molly Alter, Daisy Chu, Ben Lockwood, Julia Zhou, Sean Tom, and especially Elizabeth Davidson provided outstanding research assistance. The authors thank the U.S. Department of Education for providing access to the restricted-use versions of the Early Childhood Longitudinal Survey and Schools and Staffing Survey. To comply with restricted-use data reporting requirements, all sample sizes in this paper have been rounded to the nearest ten.

On January 8, 2002, President George W. Bush signed into law the *No Child Left Behind* (NCLB) *Act*, which many consider the most significant federal intervention into education in the United States since the authorization of the Elementary and Secondary Education Act in 1965. Under NCLB, states were required to adopt school accountability systems based on student proficiency on statewide math and reading exams, and to measure proficiency within student subgroups (e.g., students from low income families, students with limited English proficiency). States must impose escalating sanctions on schools that fail to satisfy Adequate Yearly Progress (AYP) requirements for exam proficiency, including allowing students to transfer to other public schools, forcing schools to pay for students from low-income families to enroll in after-school tutoring programs, and, ultimately, closing or restructuring persistently failing schools.[1]

Most prior empirical research on school accountability focuses on the impacts of state and local systems, many of which preceded No Child Left Behind (e.g., Ladd & Zelli, 2002; Hanushek & Raymond, 2005; Chakrabarti, 2007; Rouse et al., 2007; Chiang, 2009; Rockoff & Turner, 2010). Several studies find evidence that accountability pressure causes schools to reallocate resources in ways that raise average student achievement. However, schools have also been found to shift resources towards students and subjects that are most critical to the accountability rating (e.g., Booher-Jennings, 2005; Reback, 2008; Neal & Whitmore Schanzenbach, 2010), teach to the test (Jacob, 2005; Figlio & Rouse, 2006), remove low performing students from the testing pool (Figlio & Getzler, 2006; Figlio, 2006, Cullen & Reback, 2006), or cheat (Jacob & Levitt, 2003).

Knowledge about the impacts of NCLB is still nascent. Among the few studies that apply rigorous methods, most examine only student performance on high stakes tests in one state or one city (Springer, 2008; Krieg, 2008; Ladd & Lauen, 2010; and Neal & Whitmore Schanzenbach, 2010). These studies have found that students enrolled in schools failing AYP tend to make greater than expected gains on high-stakes tests, though there is conflicting evidence concerning heterogeneous effects on students at different parts of the performance spectrum. Only two prior studies examine the impact of NCLB incentives in multiple states. Ballou and Springer (2008) examine variation in the grade levels tested for NCLB across seven states and find that students generally perform better on low-stakes exams during

[1] States must also publish school report cards, and schools' AYP status may affect school prestige and local property values (see Figlio and Lucas, 2004).

years when they took high-stakes tests, particularly for students near the margin of passing their high-stakes exams. Dee and Jacob (2011) find that students in states with no prior accountability policies experienced greater increases on the National Assessment of Educational Progress in some grades and subjects after NCLB was introduced.

This paper expands on this literature by providing the first nationwide study of the impact of NCLB pressure on teachers and students. We investigate the links between the accountability pressure under NCLB and a wide array of outcomes measured for nationally representative samples. To this end, we assemble a new dataset on the determination of AYP status for schools nationwide during the introduction of NCLB, and use these data to measure the degree to which schools faced moderate or severe risks of failing. We exploit the fact that each state selects its own standardized tests and rules for satisfying AYP, generating numerous cases where a school near the margin for satisfying its *own* state's AYP requirements would have almost certainly failed or almost certainly passed AYP if it were located in *another* state.[2]

This variation in state policy allows us to implement a cross-sectional identification strategy similar to a difference-in-differences approach. Specifically, we compare differences in outcomes between schools on and away from the AYP margin *within* a state to differences between similar schools that, because they are in other states, are not on the AYP margin. Our strategy is similar to Tyler et al. (2000), who estimate the labor market value of a General Educational Development (GED) certificate using cross-sectional variation across states in the score needed to pass the GED exam. But our strategy differs in two ways from Tyler et al. First, we are interested in the pressure faced by schools that were at risk of failure, rather than the impact of passing or failing *per se*.[3] Second, Tyler et al. make use of a single nationwide GED test, but, because each state uses a different set of NCLB exams and AYP rules, we must evaluate schools' probabilities of failing AYP on a state-by-state basis.

We illustrate this type of variation in Table 1, which presents summary statistics for two similar pairs of schools, one pair in New Jersey and the other in Pennsylvania. While the pairs appear very

[2] As we demonstrate below, states vary widely in the percent of schools that fail to make AYP, and much of this variation is due to policy parameters (e.g., rules regarding the minimum enrollment for subgroups to count towards AYP, the grade levels that count towards AYP, and confidence or "safe harbor" adjustments to proficiency rates for schools that would otherwise have not made AYP) rather than academic achievement.
[3] In the context of school accountability literature, our work is akin to Reback's (2008) study of accountability pressure, rather than studies of schools receiving failing designations, such as Rouse et al. (2007), Chiang (2009), Rockoff and Turner (2010), or Chakrabarti (2011).

similar on observable dimensions, only in New Jersey do the two schools differ in our measure of NCLB pressure. Thus, our identification is akin to measuring the difference in student outcomes between the New Jersey pair and subtracting away the difference in outcomes between the analogous Pennsylvania pair. In our actual specification, we use continuous sets of controls, rather than a simple differences-in-differences framework, but the idea behind the identification is the same.

Our analysis takes advantage of nationally representative data from the Schools and Staffing Survey (SASS) and the Early Childhood Longitudinal Survey (ECLS), which, serendipitously, were collected on teachers and students exposed to the initial years of NCLB. We find that accountability pressure from NCLB reduces teachers' perceptions of job security, especially among relatively inexperienced teachers. We also find evidence of reading and math specialist teachers working longer hours, generalist teachers (who teach multiple subjects) working fewer hours, and all types of teachers shifting time away from whole-class instruction. The topics of instruction sacrificed include science and social studies lessons.

We find short-term NCLB pressure has either positive or neutral effects on student achievement in math, reading, and science on *low-stakes* examinations. Students enrolled in schools with a moderate risk of failing to make AYP score 0.08 standard deviations higher in reading than comparable students in similar schools that were well above the margin for making AYP. Estimated effects for math and science are statistically insignificant. We also find no evidence of differential achievement effects of NCLB pressure on low-stakes exams for students in particularly crucial subgroups or students with scores close to the passing threshold on their states' high-stakes examinations. In addition, achievement gains from short-term NCLB pressure do not come at the expense of students' reported enjoyment of learning or their anxiety over testing—if anything, enjoyment increases and anxiety decreases.

The paper proceeds as follows. Section 2 describes the NCLB data we have collected as well as the SASS and ECLS survey data. We present our methodology and results for predictions of AYP failure probabilities in Section 3, and our estimated effects of NCLB on teachers and students in Section 4. Section 5 concludes by discussing how these results may inform current policy decisions.

2. Data and Descriptive Analysis

2.1 Data Description

To measure NCLB pressure faced by schools (and which subgroups and test subjects caused that pressure), our analysis requires a comprehensive, national database of schools' NCLB-related outcomes. Because NCLB did not require states to report these data to the federal government, we painstakingly collected them from individual school report cards or state-level data files wherever available, and supplemented remaining states' data with two existing but incomplete public datasets.[4] These variables include whether each subgroup met testing participation requirements. We present the categories of data collected and their sources in Appendix 1.

We examine teacher-level outcomes from the 2003-2004 wave of the SASS and student-level outcomes from the spring 2004 wave of the ECLS, when most students in the ECLS were in the fifth grade. Both of these surveys are sponsored and distributed by the National Center for Education Statistics. We use the non-public-use versions of these data in order to link teachers and students to school-level measures of short-term pressure to make AYP.

The SASS surveyed teachers in all 50 states and allows researchers to construct nationally-representative samples with the use of sampling weights.[5] For consistency with our examination of student outcomes in the ECLS, we limit the sample of teachers to those working in regular public schools that served at least five fifth graders as of 2001-2002. The first panel of Table 2 provides summary statistics on the outcome variables we create from SASS survey questions.[6]

The ECLS followed students for nine years, collecting data in both the fall and the spring of the school years 1998-1999 and 1999-2000 (kindergarten and first grade), and in the spring of the school years 2001-2002, 2003-2004, and 2006-2007 (third grade, fifth grade, and eighth grade). The ECLS has the widest coverage and array of student-level outcomes of any nationally representative longitudinal dataset covering years before and after the passage of NCLB. Indeed, the timing of the ECLS survey is

[4] These two sources of NCLB-related data are the Council of Chief State of School Officers' School Data Direct (http://www.schooldatadirect.org/) and the American Institutes for Research National AYP and Identification Database (http://www.air.org/publications/naypi.data.download.aspx). Whereas the first source includes AYP data in most states for the years 2002-2003 through the current year, the latter dataset includes states' yes/no determinations regarding 2003-2004 and 2004-2005 subgroups and schools' passage of AYP participation and proficiency targets. In addition to missing data for some states, these sources also contain discrepancies with states' school report cards. We prioritized school report card data where available since they are the final interface between schools and the public and should reflect final adjustments such as school appeals to states' determinations of AYP.

[5] The SASS also surveyed administrators but these questions were not relevant to NCLB pressure. Although the ECLS surveyed teachers, the SASS offers a much larger sample size, surveys teachers across all grades levels, and asks them pertinent survey questions about their time use, attitudes toward their job, and future career plans.

[6] We recoded teachers' reported work-related hours and instructional hours as missing if their reported 60 or more instructional hours, a suspiciously high level of reported instructional time given the typical five day school week.

serendipitous, as this cohort was tested just prior to the first year of NCLB and again two years later. The ECLS sample was designed to be representative of kindergartners, their classrooms, and their schools in the school year 1998-1999, (and representative of first grade students in 1999-2000). It includes students from 40 relatively populous states.[7]

The student-level data collection procedures in the ECLS result in samples of students that are not necessarily representative of the student populations at their schools, particularly due to tracking procedures for students making non-structural school transfers.[8] In our analysis of ECLS data, we use sampling weights to make our estimates nationally representative. However, we test the robustness of our results to removing child-level sampling weights and to removing students who made non-structural enrollment changes. This alternative specification does not change our main conclusions and has small effects on the precision of our estimates, increasing precision in some cases and reducing it in others.

In the ECLS data, we are particularly interested in measures of student performance on a series of standardized tests in math, reading, and science. Unlike the tests that states administer under NCLB, the ECLS tests were low-stakes, un-timed, and adaptive (i.e., subsequent questions are selected based on a student's performance on preceding questions), thus preventing floor or ceiling effects and increasing test reliability. Students and schools became involved in the ECLS survey well before NCLB, and likely were familiar with the ECLS surveyors and understood that these tests were not high-stakes. This reduces concerns about teachers teaching to the ECLS test or strategic responses to ECLS survey questions. Also, by examining tests unrelated to NCLB, we avoid problems of mean reversion due to measurement error or other shocks to high-stakes test scores that do not reflect real achievement but would nevertheless affect the accountability pressure faced by the school.

The second panel of Table 2 provides descriptive statistics for our ECLS outcome measures. Since most surveyed students in the ECLS spring 2004 wave were fifth graders, we limit the sample of students to those attending regular public schools in the spring of the school year 2003-2004 that also

[7] It used a multistage probability sample design, first selecting broad geographic areas (e.g., a county), then schools within each area, and finally students within schools. On average, 23 kindergarteners were sampled in each school.
[8] The ECLS includes students who were retained within the same grade or skipped a grade level, but has some attrition. In the school year 1999-2000, a randomly-selected 50 percent sub-sample of students who transferred from their original school was surveyed, and another random sample of first graders in the same schools where transfer students were followed was added. However, this "freshening" of the sample was not repeated in the third, fifth, and eighth grades, and the ECLS simply sampled 50 percent of students who transferred schools for non-structural reasons (e.g., students who switched schools for reasons other than moving from a K-4th grade school to a 5th-8th grade school in the same district).

served at least five fifth grade students as of 2001-2002. We standardize students' scores within subject and year so that the national mean score equals zero and the national standard deviation equals one.[9] In addition to standardized exams, we examine students' reported enjoyment of math and reading, as well as reported anxiety over standardized tests.[10]

Table 3 provides descriptive statistics on control variables used in our regression analyses. We show statistics separately for our samples of public school teachers from the SASS and for public school students from the ECLS. Along with variables from these two surveys, we also use school characteristics from the Common Core of Data (CCD), compiled by the National Center for Education Statistics (NCES), and pre-NCLB aggregated student test performance variables from the National Longitudinal School-Level State Assessment Score Database (compiled by American Institutes for Research).[11] We standardize pre-NCLB test performance variables within states to have a mean of zero and standard deviation of one.

In addition to our analysis of the SASS and ECLS data, we examine a set of survey responses from the Implementing Standards-Based Accountability (ISBA) study, conducted by the RAND Corporation. As part of ISBA, principals and math teachers in three states (Pennsylvania, Georgia, and California) were surveyed regarding their views on NCLB-related policies and the implementation of these policies in their schools. While these data are not public, researchers at RAND generously provided us with cross-tabulations of survey responses on a number of items, broken down by our measure of NCLB pressure. We discuss our measure of pressure and present the ISBA results in Section 4.

2.2 Descriptive Analysis of AYP Outcomes under NCLB

[9] The ECLS data report t-scores of students' IRT-based "theta scores," which are estimates of students' skill levels. These t-scores are already constructed so that the national (cross-sectional) mean equals 50 and the national standard deviation equals 10, so we simply subtract 50 from these scores and then divide by 10 to convert them to Z-scores. Our sample means and standard deviations for these variables are not exactly equal to 0 and 1, respectively, because we must exclude a small fraction of states and schools with missing data, and because we use longitudinal student sample weights rather than cross-sectional sample weights.

[10] Answers to these specific questions, rather than an index based on a larger set of items, were obtained via special application to the National Center for Education Statistics. Due to copyright restrictions we cannot report the exact wording of these questions. For interest in and enjoyment of math and reading, we create dependent variables by summing the subject-specific numeric values for four relevant questions. We use only one question regarding feelings of test anxiety and create an indicator for reporting that such feelings were "mostly" or "very" true.

[11] Tennessee did not report school level demographic information to the federal government after 1998-1999. Rather than drop Tennessee from our analysis, we use data from 1998-1999 in lieu of data from 2001-2002.

For a school to make AYP, each of its numerically significant student subgroups must meet a test *proficiency rate* threshold in both math and reading in addition to a test *participation* cutoff of 95 percent. Secondary schools must also meet thresholds for graduation rates, and primary schools must also perform sufficiently well on a state-selected "additional indicator," which is typically the attendance rate. Beyond this set of parameters, states have a great deal of flexibility in setting a number of other rules and regulations. Specifically, states must:

- select *standardized tests* in math, reading, and (starting in 2007-2008) science;
- select which *grade levels* to test (until 2005-2006) [12];
- establish *proficiency rate thresholds*, i.e., the percent of students that must score proficient or higher. Thresholds apply to the whole school as well as individual subgroups;
- determine whether to calculate proficiency rates using all students *across* tested grade levels within each school or *within* tested grade levels;[13]
- determine whether to calculate subgroup proficiency rates using *multiple years of testing*;
- define *continuous enrollment*, where only continuously enrolled students count towards calculation of subgroup size as well as test participation and proficiency rates;
- select the minimum number of students that must be enrolled in tested grade levels for a student subgroup to be *numerically significant* and thus count towards a school's AYP determination;
- determine the size of *confidence intervals* applied to student subgroups' raw proficiency rates, where larger intervals effectively lower proficiency thresholds needed to make AYP;
- determine the nature of *safe harbor* provisions that allow schools to make AYP in spite of a subgroup not meeting the required proficiency rate that year; and,
- decide upon the *appeals process* for schools to appeal their AYP status from the state.

Even this long list does not fully capture all the minutiae of NCLB rulemaking. For example, while most states consider the performance of five ethnic subgroups (Asian/Pacific Islander, black,

[12] From 2003 to 2005, states were allowed to choose which tested grade levels counted towards AYP determination, so long as at least one level in each of three grade spans (3-5, 6-9, and 10-12) were included. Only beginning in 2005-2006 did states have to assess the math and reading proficiency of all third through eighth graders and at least one level for grades 10 to 12.

[13] While most states determine subgroup size using students across all tested grades within a school, eight states (Arizona, Colorado, Maine, New York, New Jersey, Rhode Island, Tennessee, and Washington) further disaggregate subgroup size and subgroup results to the grade or grade span level.

Hispanic, Native American, and white) in their AYP determinations, California and Alaska added additional subgroups (Filipino and Alaskan Native, respectively) while Asian/Pacific Islander is not an AYP subgroup in Texas.

These seemingly esoteric decisions have real implications for whether schools fail to meet the targets set for them under NCLB, as can be seen in the remarkable amount of variation in the fraction of schools in each state that made AYP. In 2003, most states' failure rates fell between 20 and 40 percent, but the range extended from roughly 1 percent in Iowa to 82 percent in Florida (see Figure 1).

Importantly for our study, variation in the fraction of schools making AYP was mostly a function of states' rulemaking choices and bears little relation to measures of statewide academic achievement. For example, the fraction of schools failing to make AYP by state is not significantly correlated with the fraction of students in the state deemed proficient on the state's own exams, because required proficiency rates were often set at the 20th percentile of baseline (spring 2002) school performance.[14] More importantly, as shown in Figure 2, there is little relationship between the fraction of schools failing to make AYP in a state and the state's average student achievement as measured on the National Assessment of Educational Progress (NAEP), a federal exam that has been administered to nationally representative samples of students in grades 4 and 8 for several decades.[15] States with the highest NAEP proficiency rates have slightly lower AYP failure rates than other states, but this relationship is not statistically significant and NAEP proficiency rates explain very little of the cross-state variation in AYP failure rates.

We have been unable to find any single aspect of NCLB design that can explain the wide variation in failure rates. However, by testing a number of factors we have come to the conclusion that the interaction of four features significantly influences the fraction of schools failing AYP: (1) state rules for the numerical significance of student subgroups (i.e., how large subgroups must be to count towards AYP); (2) within-school heterogeneity, which influences how many student subgroups are numerically significant; (3) the generosity of the state's confidence intervals; and (4) the generosity of the state's safe harbor provisions. The complex manner in which these policy details interact increases our confidence

[14] However, there was even wide variation in how states calculated the 20th percentile. For example, some states based the 20th percentile measure on baseline school-wide pass rates and some used grade-specific and/or subject-specific baseline pass rates.
[15] Note that we plot AYP failure rates for schools serving fifth grade students, which is the type of schools we analyze in SASS and ECLS. In Figure 1, AYP failure rates are shown for all schools that receive AYP designations.

that the wide differences in the apparent leniency of NCLB requirements across states can help identify its impact on schools and students.

3. Predicting the Probability of Failing AYP

In the first stage of our analysis, we use our assembled data set to determine which student subgroups and, by extension, which schools were on the margin of failing to make AYP in the first two years that NCLB was in effect. We begin by estimating state- and subject-specific probit regressions to generate predictions of the likelihood that each numerically-significant student subgroup would pass AYP proficiency targets in the spring of both 2003 and 2004. To do so, we use school demographic characteristics (listed in Table 3) and subgroup-level/school-level test performance variables from the school year 2001-2002, which is the year after the passage of NCLB but prior to the first AYP determinations.[16]

We conduct regressions separately by state, so that coefficients capture the nuances of how states' NCLB rules affect schools' chances of making AYP. Regressions are run at the student subgroup level and are restricted to those that were numerically significant in either 2003 or 2004. This means a single school will have as many AYP predictions per subject (math or reading) as it has numerically significant student subgroups. For states that further disaggregate subgroup results to the grade or grade span level, we also define subgroups at this disaggregated level—with separate observations for each subgroup-by-grade-level combination. Our variables differ somewhat across states due to variation in NCLB regulations. To be as consistent as possible, we applied a set of rules (described in Appendix 2) for how to specify our regressions conditional on the available data for that state.

For each subject *s*, we estimate state-specific regressions of the following form:

$$(1)\qquad AYP_{jks03\text{-}04} = \begin{cases} 1 & \text{if } \alpha_q + X_{jks02}\beta_1 + N_{jks04}\beta_2 + XN_{jks}\beta_3 + W_{j02}\beta_4 + M_{jks03-04}\beta_5 + \zeta_{jks} > 0 \\ 0 & otherwise \end{cases}$$

[16] In the vast majority of states, student test performance during the 2001-2002 school year did not directly affect the proficiency rates used to formulate schools' AYP determinations during 2002-2003 or 2003-2004. A few states incorporated 2001-2002 proficiency rates into 2002-2003 AYP determinations by generating two-year or three-year average proficiency rates for student subgroups; the remaining states used contemporaneous proficiency rates. Most states calculated a "safe harbor" provision whereby a school could make AYP if the only subgroup not meeting its target proficiency rate demonstrated sufficient improvement from the prior year. In 2002-2003, this would be based on performance relative to 2001-2002.

where $AYP_{jks03\text{-}04}$ denotes whether subgroup k at school j met its AYP proficiency rate targets in 2003 and 2004 in subject s. X_{jks02} is a vector of test score variables for subgroup k based on performance on statewide exams in subject s during the school year 2001-2002, N_{jks04} is a vector of student subgroup size variables in subject s for subgroup k in 2004, XN_{jks} represents interactions of test score and subgroup size variables, W_{j02} is a vector of control variables for school-level demographics from the school year 2001-2002 (listed in Table 3), and $M_{jks03\text{–}04}$ is a vector of two dichotomous indicators for whether student subgroup j was numerically significant in subject s in only 2002-2003 or only 2003-2004, and ζ_{jks} is a normally distributed disturbance term. The X_{jks02} vector includes cubic terms for the test performance in subject s among students in subgroup k at school j.[17] The subgroup size variables (N_{jks04}) and interactions with test score measures (XN_{jks}) are included to account for states' confidence interval adjustments and the mechanical decrease in the error variance of student pass rates as the number of tested students within subgroup k increases. In particular, the N_{jks04} vector contains cubic terms for the inverse of the square root of the number of accountable test-taking students in subject s in subgroup k in school j during the school year. We exclude subgroups from our sample if they were too small to be accountable under AYP in *both* 2003 and 2004. Appendix 2 provides detailed descriptions of each predictor and its data source.

We restrict our sample to schools that were (a) operational from at least 2001-2002 through 2003-2004, (b) neither technical/vocational nor only for special education students according to the school classifications in the Common Core of Data, and (c) served at least five students in grade 5 as of the school year 2001-2002.[18] We are forced to omit nine states from the SASS sample and five states from the ECLS sample due to missing data (e.g., 2002 test scores or a state's AYP determinations for

[17] Because we focus on schools serving fifth grade, we prioritize using fifth grade students' 2001-2002 proficiency rates for these control variables. Because some states either did not test fifth graders in 2001-2002 or disaggregated 2002-2003/2003-2004 subgroup AYP status by grade level, the 2001-2002 test performance variables are in some cases based either in part or wholly on tests from other grades, typically grade 4 or grade 6; full details are provided in Appendix 2. In addition, subgroup-specific performance for 2001-2002 is unavailable for some states, in which case we use overall student test performance in subject s, and include interaction terms between test performance and the fraction of the overall student population at each school comprised of students in group k. In practice, we find that subgroup-specific and overall measures of pre-NCLB test score performance work equally well in predicting the likelihood that the schools' pass rates will be near the NCLB required cutoff in 2003-2004.

[18] We use the restriction of having five fifth graders because some schools that should serve grade 5 according to grade level ranges indicated in the CCD actually enrolled no fifth graders. In cases where we use test performance from a grade other than grade 5 in the X_{jks02} vector, the regressions also include subgroups from schools serving the tested grade even if the school does not serve grade 5. For example, if a state tested fourth graders but not fifth graders in 2001-2002, we use grade 4 test performance in X_{jks02} and include K-4 schools in our first stage. Full details are provided in Appendix 2.

subgroups were missing). Our numerous attempts at gathering these data from state departments of education have either been unsuccessful or, in most cases, states claim that the data simply do not exist or are too unreliable to release. Fortunately, these states have relatively small populations; more than 92 percent of the U.S. population resides in one of the 41 states with sufficient data for our analyses.

3.1 Defining the AYP Margin

We use the predicted probabilities at the subgroup level from the Equation 1 regressions to construct school-level measures of accountability pressure under NCLB. Our measures are based on the following logic. Schools where all numerically significant subgroups have high chances of passing state proficiency targets in both math and reading likely faced little NCLB pressure. In contrast, schools where *any* numerically significant subgroup was close to the margin of passing are likely to have faced accountability pressure. However, schools where *any* subgroup has a very low probability of passing are unlikely to be able to do anything to change their AYP outcome in the short term.

Following this logic, we construct the following school level measures of NCLB pressure:

(i) A school is classified as *above the AYP margin* if all numerically significant subgroups have a high chance of making AYP in both math and reading;

(ii) A school is classified as *below the AYP margin* if it has at least one numerically significant subgroup with a low chance of making AYP in either math or reading;

(iii) A school is classified as *on the AYP margin for a particular subject* if (a) at least one numerically significant subgroup in the school has a moderate chance of making AYP in that subject, and (b) no numerically significant subgroup in the school has a low chance of making AYP in either subject;

(iv) A school is classified as *on the AYP margin* if it is on the AYP margin for math or reading.

For all of our analyses below, we define a "moderate chance" of a subgroup making AYP as between 25 and 75 percent, a "high chance" as above 75 percent, and a "low chance" as less than 25 percent. Our results are not very sensitive to using other cutoffs ranging from between 35 and 65 percent to between 15 and 85 percent.

Table 4 provides descriptive statistics for our measures of NCLB pressure. In these 41 states, we classify 69.1 percent of schools above the AYP margin, 21.4 percent on the AYP margin, and 9.5 percent below the AYP margin. The actual rates with which schools made AYP in both 2003 and 2004 were 87 percent for schools above the margin, 38 percent for schools on the margin, and 7 percent for schools below the margin, demonstrating that our specification has sufficient power to identify substantial variation in which schools were at risk of failing to make AYP. However, our analyses below are predicated on the idea that the risks of AYP failure were foreseeable to school administrators and teachers. To the extent that measurement error causes us to misclassify which schools *believed* they were on the AYP margin, our estimated effects of NCLB pressure may be biased towards zero. This possibility motivates the need to examine whether our estimates are related to teachers' and administrators' reported sense of accountability pressure, which we do below.

The results reported in Table 4 also reveal that, with the exception of white and economically disadvantaged students, most student subgroups were typically not numerically significant and thus did not count towards AYP. For example, 70 percent of schools did not have a sufficient number of disabled (special education) students in either 2003 or 2004 to be held accountable for that group's performance. This rate varied across states depending on minimum subgroup size requirements, again underscoring the importance of AYP formulae. For example, disabled subgroups were accountable under NCLB in either 2003 or 2004 in just 7 percent of Arizona schools, compared with 61 percent in Massachusetts.

Among subgroups that were numerically significant and thus accountable, the fraction we predict to have a moderate or low chance of making AYP varies considerably. The subgroups most frequently predicted to have a *moderate* chance of passing in reading were disabled and limited English proficient (30 and 37 percent, respectively) and, in math, disabled and Black (26 and 27 percent, respectively). Disabled student subgroups also have relatively high fractions (about 15 percent) predicted to have *low* chances of passing proficiency targets in both subjects, as do Native American subgroups (17 percent in math, 22 percent in reading) and Asian subgroups in reading (25 percent). In contrast, White subgroups are nearly always predicted to have a high chance of passing proficiency targets.

3.2 Variation in Predicted NCLB Pressure across States

Our identification strategy is based on the idea that similar schools faced different levels of NCLB pressure because of the state in which they were located. However, it is still broadly true that schools with high average achievement had greater chances of making AYP than schools with low average achievement. Figures 3a, 3b, and 3c illustrate these points. To construct the y-axis variables in these figures, we pair each school with similar out-of-state schools to simulate the counterfactual NCLB pressure in math and reading that schools would face had they been located in other states. We find the most similar school in each other state based on an index of school characteristics, which includes student math and reading test scores from the spring of 2002, the number of total enrolled students in that year, the percent of students who are from low-income households, and the racial composition of students.[19] We then calculate the fraction of paired schools that are on the AYP margin and the fraction below the AYP margin, and we use these calculations as our estimates of counterfactual AYP pressure in other states. As shown in the figures, these rates are related to schools' prior academic achievement levels and the level of AYP pressure they face in their own state. Yet the figures also illustrate that schools frequently would face *different* pressure in other states than they do in their own state. Substantial cross-state variation in NCLB rules creates substantial cross-state variation in AYP pressure.

Along the lines of a "simulated instruments" approach, we include these out-of-state counterfactual accountability pressure variables as controls in our analysis below—the estimated fraction of states in which the school would be on the AYP margin and the fraction of states in which the school would be below the AYP margin. The motivation for their inclusion is to isolate variation in schools' treatment status that depends on how their own states' rules differ from those in other states. Their inclusion ultimately has very little impact on our findings, though, because their correlations with outcomes of interest are mostly captured by other control variables. In all of our analyses below, we include flexible controls for schools' relative performance on statewide examinations in the school year 2001-2002. Thus, our identification is not based on the general tendency of low-scoring schools to face more NCLB pressure.

3.3 Assessing our Measure of NCLB Pressure in the ISBA Surveys

To get an initial sense of the validity of our measures of NCLB pressure, we examine aggregate statistics from surveys of principals and math teachers in California, Pennsylvania, and Georgia by the

[19] We construct dissimilarity indexes based on 2002 reading and math test performance within-state Z-scores, as well as six demographic variables standardized at the national level: number of enrolled students in the school as of 2001-2002, percent of students who are from low income households, percent of students who are white, percent who are black, percent who are Hispanic, and percent who are Asian. The indexes use a weighted average of the differences in schools' values, with 20 percent weighting given to each of the test score variables, 20 percent weighting given to the number of enrolled students, 20 percent weighting given to the percent of students from low income households, and 20 percent total given to the four racial composition variables (5 percent each).

RAND Corporation in the school-year 2003-2004. While these cross-tabulations are only suggestive—the micro data from these surveys are not publicly available—these data are unique in that principals and teachers were asked specifically about NCLB pressure.

We examine principals' survey responses in 21 schools that we classified as on the AYP margin and 104 schools above the AYP margin; no principals were surveyed at any school that predicted to have a low chance of making AYP. Among principals working in schools above the AYP margin, 96 percent felt they would make AYP in the school year 2003-2004, relative to only 71 percent who worked in schools on the AYP margin. Indeed, among principals in schools above the AYP margin, 72 percent felt they would make AYP for *the next five years*, relative to only 48 percent in the marginal group (Table 5, Panel A). Principals in schools on the AYP margin were also between 9 and 14 percentage points more likely to say that they had taken the following actions: encouraged teachers to focus more time on tested subjects; distributed commercial test preparation materials; or distributed copies of previous state tests or test items. All of these differences in responses across principals in the two groups are statistically significant at approximately the one percent level.

Because of the larger number of teachers than principals surveyed, we can examine teachers working in schools we classify as below the margin (19 teachers), on the margin (224 teachers), and above the margin (1,074 teachers) of AYP. Relevant survey questions included probes about teaching test-taking strategies, focusing on students who are close to proficient on the high stakes test, emphasizing the topics and types of problems given on the state test, spending more time teaching content, and searching for more effective teaching methods. Teachers working in schools on the AYP margin were between 11 and 19 percentage points more likely than teachers working in schools above the AYP margin to report having taken these actions, while teachers in schools below the AYP margin were between 3 and 20 percentage points more likely to report having taken these actions than teachers in schools on the margin. All of the differences between responses from teachers in the schools above the margin and either of the other two teacher groups are statistically significant at the one percent level, and support the idea that our constructed measures of NCLB pressure align with principals' and teachers' reported perceptions.

4. Estimates of the Impact of Accountability Pressure Under NCLB

We use our measures of whether a school is below, on, or above the AYP margin to predict various outcomes for an individual *i* (i.e., a student or teacher) in school *j* and state *q*. Our basic regression specification is shown by Equation 2:

$$(2) \qquad Y_{ij} = \delta_q + Q_{ij}\rho_1 + W_{j02}\rho_2 + X_{j02}\rho_3 + \lambda M_j^{AYP} + \gamma B_j^{AYP} + \rho_4 \% M_j^{AYP} + \rho_5 \% B_j^{AYP} + \zeta_{ij} \;.$$

Y_{ij} is an outcome of interest, δ_q represents state fixed effects, Q_{ij} is a vector of (student- or teacher-level) control variables, and W_{j02} is a vector of school-level control variables (as in Equation 1). The X_{j02} vector is analogous to the X_{jks02} vector in Equation 1, with school-wide student achievement measures replacing subgroup-specific achievement measures. M_j^{AYP} and B_j^{AYP} are indicators for schools on or below the AYP margin, respectively, and $\% M_j^{AYP}$ and $\% B_j^{AYP}$ are controls for the simulated fraction of states where the school would be on or below the AYP margin.

The coefficients of interest are λ and γ, which represent the average impact of the NCLB pressure associated with being in a school on or below the AYP margin. In particular, we are most interested in the estimate of λ, which measures the causal effect of short-term accountability pressure under the assumption that, conditional on a host of observable school characteristics, the variation across states in whether a school falls on the AYP margin is exogenous. Because our measures of NCLB pressure are derived from first-stage probit regressions, we estimate standard errors using a two-sample bootstrap adjusted for school-level clustering. We use 1,000 Monte Carlo simulations of both the first- and second-stage models, randomly sampling coefficients from the first-stage model using the implied distribution from the variance-covariance matrix which allows for school clustering, and randomly sampling schools (with replacement) in the second-stage models.

4.1 Impacts on Teachers

We examine the effect of NCLB pressure on teachers using the SASS data. The Q_{ij} vector includes the teacher-level control variables listed in Table 3, with both linear and squared terms for teachers' age, years of teaching experience, and years of experience teaching at the same school. We present results for our main sample from the 2003-2004 SASS in the top panels of Tables 5 and 6, and then compare our estimates with a falsification sample based on the (pre-NCLB) 1999-2000 SASS.

We first examine teachers' views on job security and satisfaction, limiting our sample to those teaching in high-stakes grades and subjects tested under NCLB.[20] The first column of Table 6 displays estimated effects of NCLB pressure on whether teachers agreed with the statement: "I worry about the security of my job because of the performance of my students on state and/or local tests." Compared to teachers of high-stakes grades/subjects at schools above the AYP margin, those in schools on the AYP margin or under the AYP margin are, respectively, 4.6 percentage points and 9.7 percentage points more likely to report concern over their job security related to student test performance. The latter difference is statistically significant and fairly large considering that 38 percent of teachers reported this concern overall. Point estimates for the pre-NCLB SASS sample are of the opposite sign, supporting the notion that our measure of NCLB pressure is valid and captures significant variation in school staff members' perceptions of pressure.

Less experienced teachers should be most sensitive to issues of job security, given the prevalence of seniority-based layoff and transfer rules (see Boyd et al., 2011). We therefore estimate regressions separately based on whether a teacher had more or less than 10 years of experience—roughly the sample median. The effects of accountability on perceived job security are much stronger for teachers with fewer than ten years of experience; relatively inexperienced teachers in schools on the AYP margin or under the AYP margin are, respectively, 10.1 percentage points and 17.0 percentage points more likely to report concern over their job security related to student test performance (Table 6, column 2). These effects are statistically significant at the .10 and .05 levels respectively; they are also different from the falsification estimates based on the pre-NCLB SASS sample, with p-values of approximately .10 for these differences.

Roughly three quarters of teachers in our sample state that they plan to teach "until retirement," which we view as a measure of job satisfaction. We find that teachers in schools on or below the AYP margin are less likely to plan to teach until retirement than their counterparts teaching in schools with high probabilities of making AYP (Table 6, Column 4), and these effects are especially strong for veteran

[20] In cases of teachers covering multiple grades, we include the teacher if more than half of the teacher's covered grade levels were tested for NCLB in that teacher's state during the spring of 2004. One concern in this analysis is that we limit the sample to high-stakes teachers, and principals at schools facing NCLB pressure might strategically place teachers into high-stakes grades and subjects. While we cannot test this hypothesis, we believe such behavior would most likely create a bias against our findings. If principals facing strong pressure wish to boost high-stakes test performance, then they should assign their most talented teachers to high-stakes areas. Yet we find these teachers are more concerned about their job security, are less likely to plan to teach until retirement, and work fewer hours.

teachers. Again, estimates from the falsification estimates based on the pre-NCLB sample wave are of the opposite sign and significantly different from the main estimates, supporting the idea that NCLB pressure had an impact on teachers' career plans. These results comport with findings by Feng et al. (2010) that schools in Florida that received low accountability ratings subsequently experienced higher rates of teacher turnover.

While the SASS does not ask about plans to teach next year, we are able to examine short-run turnover for approximately 540 teachers from the NCES Teacher Follow-up Survey, which tracks a subsample of teachers from the prior year's SASS wave. Using the same specification as above, we estimated a linear probability model predicting whether teachers applied to non-teaching jobs or left the profession the next year. We estimate coefficients of 0.16 and 0.09 for being on and below the AYP margin, respectively, with the former coefficient statistically significant at the five percent level. This provides additional support to the notion that NCLB pressure affected teachers' job security and job satisfaction.

Table 7 present results concerning how NCLB pressure affects teachers' self-reported total weekly work hours, whole-class instructional hours, and coverage of science and social studies.[21] The SASS surveyed teachers in the fall, well ahead of NCLB spring testing, and survey responses should reflect general shifts in instruction rather than last-minute preparation for high-stakes tests. Weekly reported whole-class instructional hours, which average 29 hours, are a subset of total reported weekly work hours, which average 53. Most teachers in these elementary and middle schools are generalists, meaning they cover multiple subjects and teach in self-contained classrooms where the same students remain for most of the day. Other teachers are specialists, such as math instructors who see several different groups of students during the same day. Given the different roles of these teachers, we divide our sample between generalist or specialist teachers when examining work time.

We find that greater accountability pressure is associated with a decrease in work hours among generalists but an increase among specialists, while whole-class instructional hours fall for both types of teachers. About half of the generalists' decrease in work hours is accounted for by fewer hours per week devoted to whole-class instruction. We cannot estimate impacts on instructional hours for the pre-NCLB

[21] Specifically, the SASS asks teachers about hours spent "on all teaching and school-related activities" and hours spent "delivering instruction to a class of students."

wave of the SASS because this question was not asked, but point estimates for changes in total work hours in the pre-NCLB sample are insignificant, small, and significantly different than the main estimates for generalist teachers. Thus, in schools facing NCLB pressure, the overall trend is to have generalist teachers working less while their specialist colleagues work longer hours on activities other than whole-class instruction. We can only speculate on the mechanisms for these time reallocations; one possibility is that specialists spend more time working with small groups of students and teachers on test preparation, student assessments, and tutoring. Regular teachers may tend to work slightly fewer hours due to the increased role of specialists and other factors, such as decreased teacher autonomy.

Finally, we use the SASS data to explore shifts in instructional time across subject areas, using self-reports of whether the teacher taught a science lesson or a social studies lesson during the prior week. Unlike the estimates discussed above, we now include teachers in the sample regardless of whether they taught a high-stakes subject, which, due to the random sampling of teachers in the SASS, allows us to capture both shifts in the subject composition taught by generalists and shifts in the employment of specialists teaching low-stakes subjects.[22] The estimates suggest that schools facing accountability pressure reduce the frequency of science and social studies lessons (Table 7, columns 5 and 6). Compared to teachers at schools above the margin, teachers are 8.5 percentage points less likely to have taught a science lesson in the last week and 4.9 percentage points less likely to have taught a social studies lesson. The former estimate is statistically significant at the .05 level but the latter estimate is not highly significant. The effects on science and social studies offerings in schools below the AYP margin are even larger: a greater than 14 percentage point reduction in the likelihood of either type of lesson. These effects are substantial, considering that 59 percent and 62 percent of teachers in this sample taught science and social studies lessons, respectively, in the previous week. Our falsification results based on the pre-NCLB produce very small and statistically insignificant estimates for these models, suggesting that our main estimates are capturing responses to NCLB pressure. Schools may shift instructional time from science and social studies to reading and math in order to increase their chances of making AYP.

[22] The results are similar if we instead limit the sample to generalist teachers: an estimated coefficient of being on the AYP margin of -0.09 (.04 bootstrapped standard error) for science lessons and -0.05 (.04 bootstrapped standard error) for social studies. This suggests that schools do not ramp up their use of social studies and science specialist teachers to compensate for the lower frequency of social studies and science lessons taught by generalists.

5.2 Impacts on Students

Using the ECLS data, we now turn to whether NCLB pressure affected student achievement, enjoyment of material, and test anxiety. Our specification is again shown by Equation 2, and student-level controls are listed in Table 3. Importantly, we include third degree polynomials of the student's standardized math and reading performance in both the first and third grade waves of the ECLS. Thus, in addition to controls at the school level, our identification comes only from comparisons of students with very similar prior learning trajectories. In these regressions, we focus on the AYP margin for the most relevant subject(s): math for math test performance or enjoyment, reading for reading test performance or enjoyment, and “either math or reading” for science test performance and for anxiety about standardized tests. We lack power to separate relevant-subject and cross-subject effects using the ECLS.

Our estimates reveal that NCLB pressure has either neutral or positive effects on student achievement growth in both low- and high-stakes subjects (Panel I of Table 8). When schools are on the AYP margin and thus have strong short-term incentives to raise high-stakes test performance, their students perform better on low-stakes reading tests and perform just as well on low-stakes math and science tests. Students' reading scores are .08 of a standard deviation greater on average when schools are on the AYP margin. This estimate is statistically significant at the .05 level and is meaningfully large since previous estimates of the impact of accountability pressure on *high-stakes* tests are typically between 0.1 and 0.2 standard deviations (e.g., Rouse et al., 2007; Rockoff and Turner, 2010). Although we are examining results for multiple dependent variables, a power test suggests that these three estimated effects for schools on the AYP margin are far too large to occur by chance.[23] Students also perform no worse on low-stakes exams when their schools are below the AYP margin rather than above this margin, suggesting the instructional shifts observed in Table 7 may not be harmful for general learning, at least not in the short term.

Importantly, our results also suggest that when schools face NCLB pressure, gains in achievement do not decrease students’ enjoyment of reading or math and are likely to decrease anxiety

[23] To test the joint significance of these test score estimates, we simulated estimations of these three models after randomly reassigning schools to different AYP status. Only 0.3 percent (3 out of 1,000) of these simulations produced a set of counterfactual estimates that had absolute values, from greatest to least, greater than the values of the actual highest, second highest and third highest estimate reported in the first three columns of the first row of Panel I of Table 8. Only 1.2 percent of these simulations produced *any* estimates that were greater than the actual estimated effect for reading test scores in the first row of Panel I of Table 8.

over testing. Respective point estimates for the impact of NCLB pressure on students' enjoyment of reading and math are 0.03 standard deviations (statistically insignificant) and 0.13 standard deviations (significant at the 10 percent level), respectively. We also find statistically significant *decreases* of 9 and 14 percentage points in rates of students' reported anxiety over testing for schools on and below the AYP margin, respectively. While schools under pressure may focus more on testing—using practice exams, motivational techniques, etc.—these actions appear to alleviate student anxiety rather than exacerbate it.

The previous school accountability literature motivates the idea that the impacts of NCLB may differ across students within a school. We first examine whether our estimates depend on whether schools faced strong pressure to raise proficiency rates for the overall student population or for the focal student's own subgroup(s). Estimates presented in Panel II of Table 7 come from specifications where we replace the single "on the AYP margin" variable with three mutually exclusive indicators for whether the school was on the AYP margin due to: (1) the overall student group, (2) any one of the student's own subgroups (and *not* the overall student group as well), and (3) other subgroups (*not* the student's own subgroup or the overall student group). The point estimates in Panel II for reading performance are never negative, regardless of whether the students are members of subgroups whose performance is most critical to the schools' AYP ratings. The estimates for math and science performance are all statistically insignificant. The largest gains on low-stakes reading exams occur for students whose own performance will likely not affect their schools' chances of making AYP, but these differences in performance by group are not statistically significant. The large estimates for this group suggest that all students may improve their reading skills when their schools spend more time on reading instruction due to accountability pressure. In summary, our analysis using the highly reliable examinations in the ECLS provides assurance that NCLB pressure does not systematically lead to adverse achievement outcomes.

5.2 Heterogeneous Effects

While NCLB pressure does not appear to lead to adverse average effects on students, the effects could be negative for particular types of students or particular state policy settings. We next examine heterogeneity across students based on their proximity to the passing threshold, their families' socio-economic status, and whether their state implemented NCLB on top of an existing test-based school accountability system.

AYP is based on proficiency rates, and schools might therefore direct resources to students who are likely to score close to the threshold of passing the exam (Reback, 2008; Neal and Whitmore Schanzenbach, 2010). We classify students as "on the bubble" for passing their state exam if their third grade test scores were estimated to be within 15 percentiles below or 5 percentiles above their states' NCLB exam passing threshold.[24] We then re-estimate the specification above adding an indicator for "on the bubble" and an interaction term between this indicator and whether the school is on the AYP margin. In each case, the estimated coefficients for this interaction term are statistically insignificant. For reading, math, and science performance, these estimates are -.02 (.06 standard error), -.002 (.10 standard error) and .07 (.06 standard error) respectively. Students on the bubble of passing high-stakes exams thus do not perform very differently on *low-stakes* exams when their schools face strong NCLB pressure, although our estimates are too imprecise to rule out moderate differences.

One concern with school responses to accountability pressure is that schools might reduce their focus on real learning for traditionally underperforming students, such as students from relatively poor households. NCLB's use of subgroup-specific pass rates is intended to mitigate these effects, but they might still occur—especially for low-stakes rather than high-stakes test performance. We examine this notion by adding to our main specification an indicator for having family income of $35,000 or less and an interaction of this indicator with whether the school is on the AYP margin. The estimated coefficients for this interaction terms are all statistically insignificant, and they are very small for the reading and math performance models.[25] This provides reassuring evidence that greater short-term accountability pressure did not widen achievement gaps based on students' families' socio-economic status.

In many cases, NCLB was layered on top of states' pre-existing school accountability systems, and the impact of NCLB pressure might be dampened in states that already put schools under pressure based on student test performance according to the state's own rating system. On the other hand, schools

[24] The National Center for Education Statistics (2007) estimates NAEP score equivalents associated with the passing threshold for most states' NCLB exams, and we obtained national percentile equivalents for these NAEP scores. We are unable to do this for eight ECLS states that were not included in the National Center for Education Statistics (2007) publication. Using ranges smaller than 20 percentiles would lead to highly imprecise estimates, and we use a wider range below the cutoffs than above the cutoffs because schools may have anticipated their capacity to improve student performance over time—i.e., most states experienced upward trends in proficiency rates over the first few years of NCLB. For reading and math outcomes our indicator is subject specific; for science tests and test anxiety we use an indicator for being on the bubble in either math or reading.

[25] For models with reading, math, or science test performance as the dependent variable, the estimated coefficients of the interaction terms are -0.02 (.06 standard error), 0.004 (.07), and -0.05 (.06) respectively. These estimates continue to be statistically insignificant if one instead uses a smaller income cutoff of $20,000 instead of $35,000.

in states with pre-NCLB accountability systems might have more experience quickly mobilizing their resources to meet performance targets. Dee and Jacob's research (2011) supports the former hypothesis, since they find that states lacking strong accountability systems prior to NCLB had stronger upward trends in 4th grade students' math performance comparing pre- and post-NCLB cohorts. They do not find a similar trend for reading performance.

Using Dee and Jacob's classifications, we add to our main specification an interaction term between the state having a strong pre-NCLB accountability system and the school being on the margin for making AYP. Our estimates are consistent with Dee and Jacob's findings—math AYP pressure had a more positive effect on students in states lacking strong accountability systems prior to NCLB, and this difference is significant (p=.07). We do not find statistically significant differences in effects of pressure on reading or science performance across states with or without strong pre-NCLB accountability. While our estimates are too imprecise to determine the extent to which these schools contributed to their states' upward math performance trajectories, these results suggest that the upward trends in math performance observed by Dee and Jacob (2011) may have been driven by schools facing the greatest short-term pressure to increase math pass rates.

6. Conclusion

As a result of the No Child Left Behind act, virtually every public school in the U.S. is now accountable for meeting targets for student test performance. To further our understanding of the impact of NCLB on such a wide set of schools, we assembled an extensive national data set of school and student subgroup performance on the examinations required under NCLB, and we exploit the extensive cross-state variation in states' rules and standards to examine how the threat of failing under NCLB affects school personnel and students. In schools facing NCLB pressure, we find that teachers report greater concern over how student test performance will affect their job security and they expect to leave the profession sooner. We find changes in work hours suggesting a shift towards teachers who specialize in single subjects and away from instruction of low-stakes subjects like science and social studies.

Nevertheless, we find that students perform at least as well academically in schools facing strong short-term pressure from NCLB as those in comparable schools that do not face such pressure. In schools facing stronger short-term incentives to improve student proficiency rates on high-stakes exams, students raise their achievement by 0.08 standard deviations on low-stakes reading exams, do similarly on low-stakes math and science tests, report more enjoyment of math, do not report less enjoyment of reading, and report less test anxiety. We can also rule out substantial negative effects for various types of students—students in/out of subgroups that have greater influence on whether their schools meet NCLB requirements, students who will likely score close/far from the passing score on their states' high-stakes exam, and students who are from poorer/wealthier families.

Our finding that short-term NCLB pressure does not negatively affect student learning or enjoyment of learning is quite important, given widely held concerns about the use of test-based accountability systems. On the other hand, our results also raise questions concerning whether NCLB pressure discourages the work effort and career length of teachers working in schools with little chance of meeting student performance standards. In addition, our finding that these schools reduce instruction in low-stakes subjects may have negative consequences in the longer term.

These issues loom larger every year as NCLB standards become more stringent and more schools fail to meet those standards. NCLB requires states to set targeted student proficiency rates at 100% by 2014. In view of ballooning AYP school failure rates, the U.S. Department of Education has been granting waivers to states so that schools can avoid AYP failure designations in spite of less than perfect proficiency rates (U.S. Dept. of Education, 2012). These waivers are conditional on broad education policy reforms that some states view as too costly to implement (L.A. Times, 11/12/11). Even across the states that have received waivers, there is variation in the types of performance targets used to determine AYP and whether AYP designations are used at all.

Policymakers may also want to consider the large differences in rules and regulations across states, which we as researchers used to identify the effects of NCLB pressure on schools. Thus far, the minutiae of state rules rather than student proficiency (as measured by national exams) have largely determined the difficulty of meeting AYP. Although a majority of states have adopted "Common Core State Standards" that may increase the consistency of student achievement tests across states, most of the current variation in AYP failure rates across states is not driven by the difficulty of state exams. If policymakers would like to establish more uniformity across states' school accountability standards, then federal policy reforms must address the often overlooked sources of variation within state formulae.

Ideally, accountability pressure should stem from the performance of students along the entire distribution of achievement, rather than the idiosyncrasies of opaque state AYP rules. Congress has an opportunity to revisit the design of school accountability systems when they enact revisions to NCLB, which will likely occur after the 2012 Presidential election. When so doing, policymakers may wish to ensure that schools along the entire performance spectrum face more continuous incentives to improve along a wide array of outcomes. Revised accountability systems could also provide schools with incentives to improve student performance regardless of students' prior achievement levels and the scaling of their test scores.[26]

[26] See Barlevy and Neal (2010) for a detailed discussion and analysis of one example of such a system. They propose a system for rating teacher performance, but a similar system could instead be used to assess school performance.

References

Ballou, D. & Springer, M.G. (2008). Achievement Trade-Offs and No Child Left Behind. Working Paper. Urban Institute.

Barlevy, G. & Neal, D. (2010). Pay for Percentile. mimeo, University of Chicago, October 2010.

Booher-Jennings, J. (2005). Below the Bubble: "Educational Triage" and the Texas Accountability System. *American Educational Research Journal 42:* 231-268.

Brunner, E. & Imazki, J. (forthcoming). Probation Length and Teacher Salaries: Does Waiting Pay Off? forthcoming in *Industrial Relations and Labor Review*.

Chakrabarti, Rajashri. (2007). Vouchers, Public School Response, and the Role of Incentives: Evidence from Florida. *Federal Reserve Bank of New York Staff Reports*, no. 306.

Chakrabarti, Rajashri. (2011). Incentives and Responses under No Child Left Behind: Credible Threats and the Role of Competition. *Federal Reserve Bank of New York Staff Reports*, no. 525.

Chiang, H. (2009). How accountability pressure on failing schools affects student achievement. *Journal of Public Economics 93*, 1045-1057.

Cullen, J.B. & Reback, R. (2006). Tinkering toward accolades: School gaming under a performance accountability system. In T. Gronberg & D. Jansen (Eds), *Advances in Applied Microeconomics, 14*.

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

Dillon, S. (2010). Obama to Seek Sweeping Change in 'No Child' Law. *New York Times*, February 1, 2010.

Feng, L., Figlio, D., and Sass T. (2010) "School Accountability and Teacher Mobility," NBER Working Paper 16070.

Figlio, D. (2006). Testing, crime, and punishment. *Journal of Public Economics 90*, 837-851.

Figlio, D. & Getzler, L. (2006). Accountability, ability, and disability: Gaming the system? In T. Gronberg & D. Jansen (Eds), *Advances in Applied Microeconomics*, *14*.

Figlio, D. & Lucas, M. (2004). What's in a grade? School report cards and the housing market. *American Economic Review* 94(3), 591-604.

Figlio, D. & Rouse, C. (2006). Do accountability and voucher threats improve low-performing schools? *Journal of Public Economics 90*, 239-255.

Figlio, D. & Winicki, J. (2005). Food for thought? The effects of school accountability plans on school nutrition. *Journal of Public Economics 89*, 381-394.

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

Jacob, B. (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.

Jacob, B. & Levitt, S. (2003). Rotten apples: An investigation of the prevalence and predictors of teacher cheating. *Quarterly Journal of Economics 118*(3), 843-877.

Krieg, J. 2008. Are students left behind? The distributional effects of No Child Left Behind. *Education, Finance and Policy 3*(2), 250-281.

Ladd, H.F. & Lauen, D.L. (2009). Status versus growth: The distributional effects of school accountability policies. Working paper. Urban Institute.

Ladd, H.F. & Zelli, A. (2002). School-based accountability in North Carolina: The responses of school principals. *Educational Administration Quarterly 38*(4), 494-529.

Los Angeles Times, (2011). No Child Left Behind waiver could cost $2 billion, report says. by Howard Blume, 11/12/2011.

National Center for Education Statistics. (2007). *Mapping 2005 State Proficiency Standards Onto the NAEP Scales* (NCES 2007-482). U.S. Department of Education. Washington, DC: Author.

Neal, D. & Whitmore Schanzenbach, D. (2010). Left behind by design: Proficiency counts and test-based accountability. *Review of Economics and Statistics.* 92(2): 263-283.

Reback, R. (2008). Teaching to the rating: School accountability and the distribution of student achievement. *Journal of Public Economics* 92, 1394-1415.

Rockoff, J., and Turner, L. (2010). Short-run Impacts of Accountability on School Quality, *American Economic Journal, Economic Policy* 2(4): 119-147.

Rouse, C., Hannaway, J., Goldhaber, D., & Figlio, D. (2007). Feeling the Florida heat? How low-performing schools respond to voucher and accountability pressure. *National Bureau of Economic Research*, working paper 13681.

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.

Tyler, J.H., Murnane, R.J., and Willett, J.B. (2000). Estimating the Labor Market Signaling Value of the GED. *Quarterly Journal of Economics* 115(2), 431-468.

U.S. Department of Education, (2012). *ESEA Flexibility*. report downloaded from http://www.ed.gov/esea/flexibility on July 24, 2012.

Figure 1: Distribution of AYP Failure Rates across States, 2003

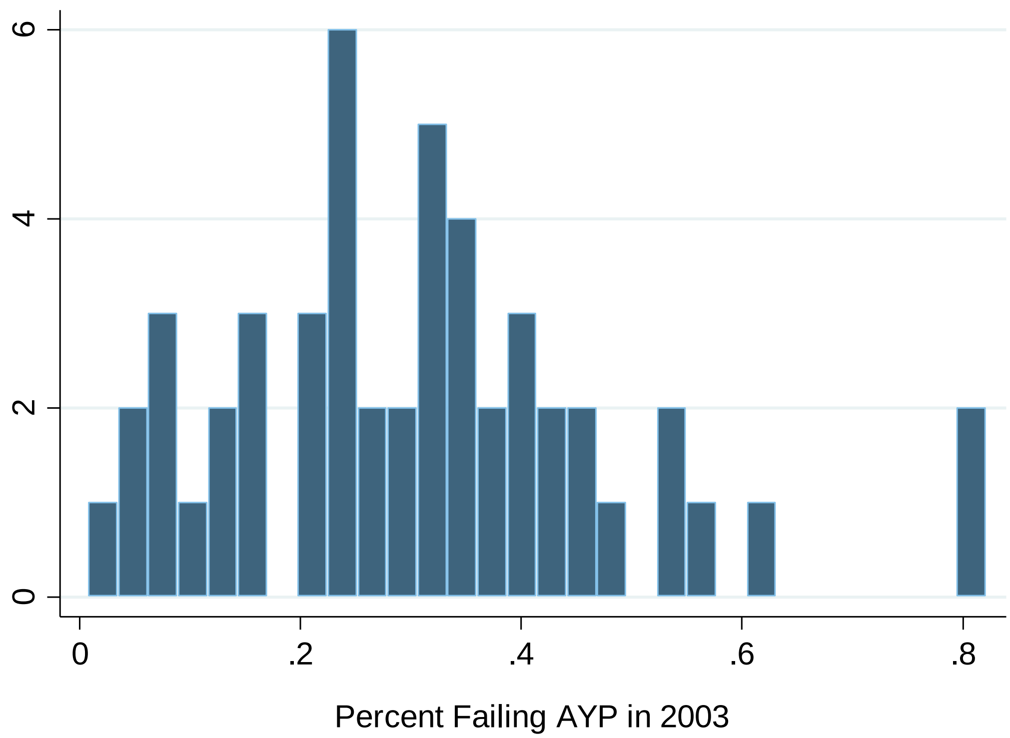
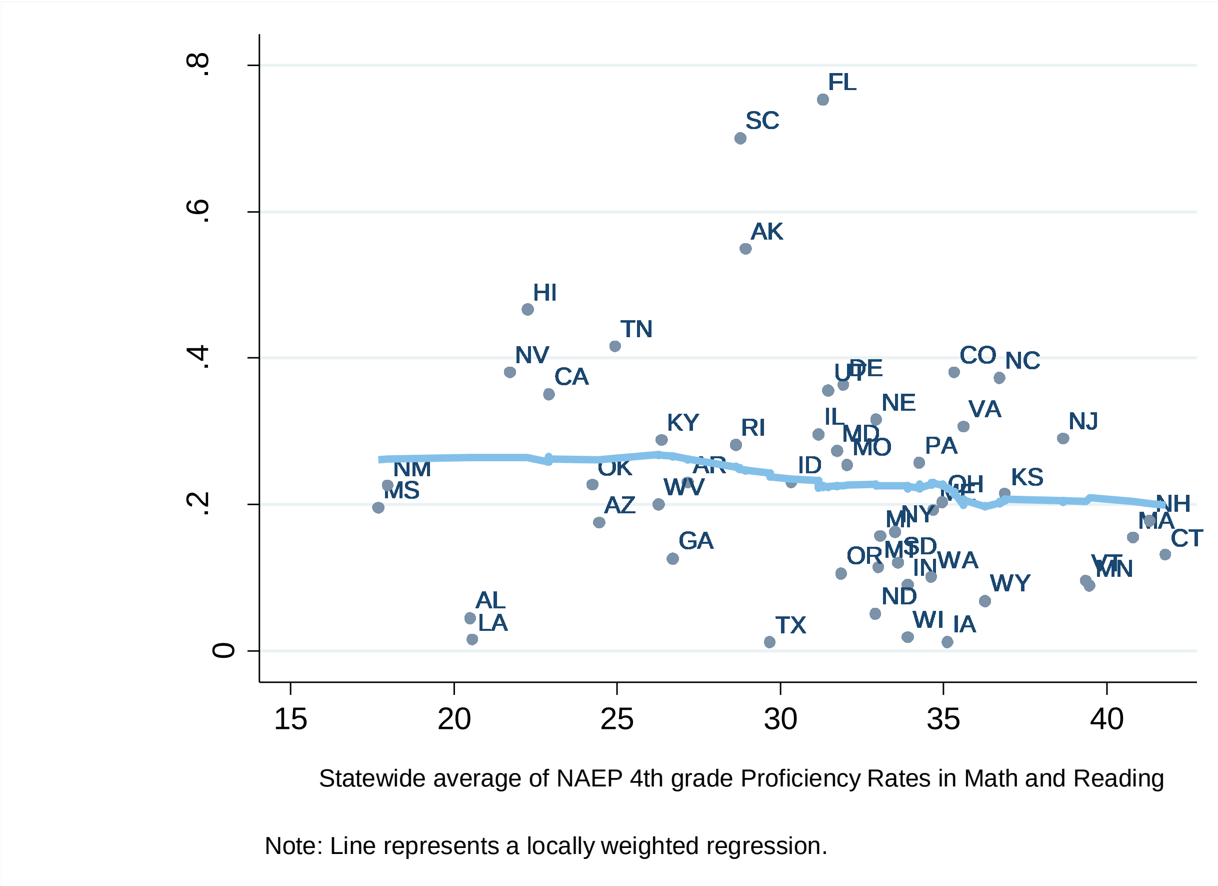


Figure 2: AYP Failure Rates vs. NAEP Proficiency Rates by State, 2003



Failure rates are based on schools serving at least five fifth grade students.

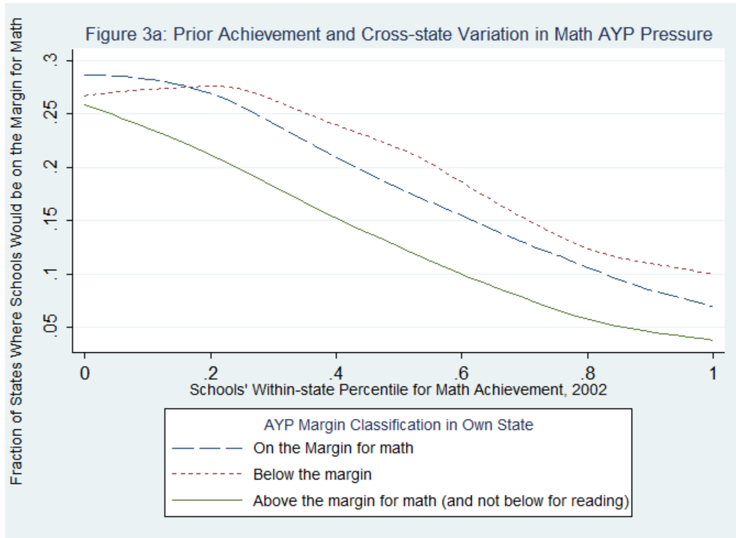


Figure 3a: Prior Achievement and Cross-state Variation in Math AYP Pressure
Fraction of States Where Schools Would be on the Margin for Math
.05
.1
.15
.2
.25
.3
0
.2
.4
.6
.8
1
Schools' Within-state Percentile for Math Achievement, 2002
AYP Margin Classification in Own State
On the Margin for math
Below the margin
Above the margin for math (and not below for reading)

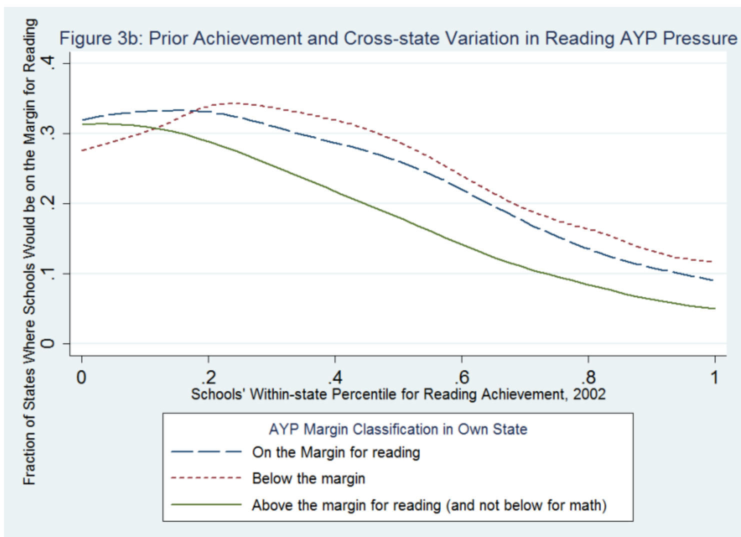


Figure 3b: Prior Achievement and Cross-state Variation in Reading AYP Pressure
Fraction of States Where Schools Would be on the Margin for Reading
0
.1
.2
.3
.4
0
.2
.4
.6
.8
1
Schools' Within-state Percentile for Reading Achievement, 2002
AYP Margin Classification in Own State
On the Margin for reading
Below the margin
Above the margin for reading (and not below for math)

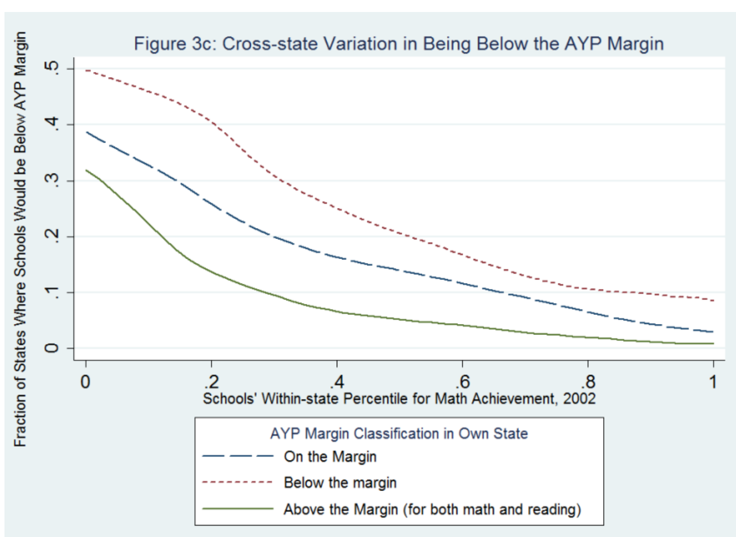


Figure 3c: Cross-state Variation in Being Below the AYP Margin
Fraction of States Where Schools Would be Below AYP Margin
0
.1
.2
.3
.4
.5
0
.2
.4
.6
.8
1
Schools' Within-state Percentile for Math Achievement, 2002
AYP Margin Classification in Own State
On the Margin
Below the margin
Above the Margin (for both math and reading)

Table 1: An Example of Difference-in-Differences in NCLB Pressure

School name	State	Title I eligible	Percent poor	Percent white	Percent black	Percent Hispanic	Student Enrollment	**Our Classification: On Margin for Making AYP?**
Hamilton School	PA	yes	89%	5%	83%	10%	454	No
Richland Elementary	PA	no	18%	94%	1%	3%	472	No
Bradley Elementary	NJ	yes	87%	4%	81%	15%	418	YES
Upper Pittsgrove	NJ	no	16%	93%	4%	3%	419	No

Table 2: Summary Statistics for Dependent Variables

	Mean	SD
<u>Teacher-level Dependent Variables from the SASS</u>		
Concerned about Job Security due to Student Test Performance	38.0%	48.5%
Plan to Teach Until Retirement	76.0%	42.7%
Gave at Least One <u>Science</u> Lesson Last Week	58.9%	49.2%
Gave at Least One <u>Social Studies</u> Lesson Last Week	62.0%	48.6%
Less Experienced Teachers (<10 Years):		
Concerned about Job Security due to Student Test Performance	42.5%	49.4%
Plan to Teach Until Retirement	68.4%	46.5%
More Experienced Teachers (at least 10 Years):		
Concerned about Job Security due to Student Test Performance	34.6%	47.6%
Plan to Teach Until Retirement	82.3%	38.2%
Generalists		
Work Hours per typical week†	52.8	9.3
Whole-class Instructional Hours per typical week†	29.3	5.1
Specialists		
Work Hours per typical week†	53.2	7.6
Whole-class Instructional Hours per typical week†	29.5	5.3
<u>Student-level Dependent Variables from the ECLS</u>		
5th Grade Reading Score (Standardized)	.008	.97
5th Grade Math Score (Standardized)	.027	.98
5th Grade Science Score (Standardized)	.065	.96
Enjoyment of Reading (Standardized)	-.002	1.01
Enjoyment of Math (Standardized)	.037	1.01
Anxiety about standardized tests	42%	49%

Notes to Table 2: Means and standard deviations using relevant sample weights provided by the SASS and ECLS to produce nationally representative estimates. The sample is restricted to observations used in the main analyses: teachers in 41 states for the SASS sample and students in 35 states in the ELCS sample. The sample sizes are approximately 2,800 teachers for the SASS sample (high-stakes teachers only) and approximately 6,870 students for the ECLS sample, each rounded to the nearest 10 due to restricted-use data reporting requirements. Standardized variables are Z-scores that were standardized based on the national, cross-sectional student distribution; their means and standard deviations above differ from zero and one respectively because some states/students are omitted due to missing data and because we use longitudinal sampling weights rather than cross-sectional sampling weights.

†We set teachers' work-related hours and instructional hours to missing if their reported instructional hours were 60 hours or greater, a suspiciously high level of reported instructional time given the typical five day school week. The work hours per week variable is based on teachers' self-reported hours spent on "all teaching and other school-related activities during a typical full week."

Table 3: Descriptive Statistics for Control Variables

Variable	SASS Sample Mean	SASS Sample SD	ECLS Sample Mean	ECLS Sample SD
School characteristics				
% of states where schools would be on margin, reading			20%	13%
% of states where schools would be on margin, math			16%	11%
% of states where schools would be on margin, either	23%	14%		
% of states where schools would be below the margin	17%	20%	14%	18%
Within-state Z-score for 2001-2002 reading	0.007	0.949	0.125	0.957
Within-state Z-score for 2001-2002 math	0.043	0.925	0.100	0.960
Eligible for Title I	69%		60%	49%
Number of enrolled students	587	258	586	252
Percent Asian students	4%	9%	5%	10%
Percent Hispanic students	19%	28%	16%	24%
Percent African American students	18%	26%	19%	26%
Percent economically disadvantaged students	47%	30%	44%	30%
Number LEP students in the grade			5	13
Missing Number of LEP students in the grade			14%	35%
Teacher characteristics (from the SASS)				
Age	41.4	10.9		
Total years of experience	13.1	9.6		
Total years of experience at same school	6.9	7.5		
Female	87%	34%		
White, non-Hispanic	78%	41%		
Black	10%	29%		
Hispanic	9%	28%		
Has full certification	88%	33%		
Has Master’s in education	41%	49%		
Has Master’s in other field	3%	17%		
Completed an undergraduate certification program	46%	50%		
Teaches grades 4 or 5	61%	49%		
Teaches grades 6 or higher	8%	27%		
Family characteristics (from the ECLS)				
Two parent household			67%	47%
Mother's education level unknown			9%	28%
Mother has at least a high school diploma			89%	31%
Mother possesses a B.A.			31%	46%
Family income missing			16%	36%
Family income under $20,000			15%	36%
Family income $20,000 -$35,000			18%	38%
Family income $35,000 - $50,000			14%	35%
Family income $50,000 - $75,000			14%	35%
Family income $75,000 - $100,000			11%	32%
Student characteristics (from the ECLS)				
Reading Z-score in spring 2000			0.017	0.950
Math Z-score in spring 2000			0.027	0.920
Reading Z-score score in spring 2002			-0.002	0.981
Math Z-score in spring 2002			0.028	0.971
Reading Enjoyment Z-score in spring 2002			-0.011	1.031
Math Enjoyment Z-score in spring 2002			0.030	1.018
African American			18%	39%

Hispanic	20%	40%
Asian	3%	17%
Other	5%	21%
Female	48%	50%
Date of birth (measured in days)	3/18/93	140

N = approximately 2,800 teachers for the SASS sample and approximately 6,870 students for the ECLS sample.

Table 4: Predictions of AYP Outcomes

Panel A: School-wide Outcomes

	On the AYP Margin	Below the AYP Margin	Above the AYP Margin
Percent of Schools	21.4%	9.5%	69.1%
Percent Actually Made AYP 2003 and 2004	37.9%	7.4%	86.5%

Panel B: Subgroup Outcomes

		Conditional on Numerical Significance			
	Numerically Significant Subgroup	Predicted Moderate Chance		Predicted Low Chance	
		Math	Reading	Math	Reading
Overall School Population	92.8%	7.2%	9.0%	2.1%	2.5%
Actually made AYP in subject in '03 and '04		51.9%	52.4%	10.7%	8.9%
Economically Disadvantaged	60.5%	14.2%	17.4%	3.7%	4.6%
Actually made AYP in subject in '03 and '04		54.0%	53.0%	12.8%	12.7%
Limited English Proficient	20.0%	18.6%	36.7%	4.8%	10.6%
Actually made AYP in subject in '03 and '04		58.3%	49.9%	13.5%	19.5%
Disabled	30.0%	26.1%	30.0%	13.9%	15.8%
Actually made AYP in subject in '03 and '04		52.0%	53.1%	14.1%	12.3%
White	69.5%	1.2%	0.9%	0.1%	0.0%
Actually made AYP in subject in '03 and '04		55.2%	61.5%	15.8%	25.0%
Black	29.7%	26.9%	23.2%	9.5%	7.8%
Actually made AYP in subject in '03 and '04		51.5%	52.5%	16.7%	15.3%
Hispanic	28.7%	10.9%	18.6%	1.2%	2.7%
Actually made AYP in subject in '03 and '04		56.8%	54.6%	13.8%	15.0%
Asian/Pacific Islander/Filipino	12.3%	0.7%	3.5%	0.0%	25.2%
Actually made AYP in subject in '03 and '04		54.6%	53.6%	33.3%	8.8%
Native American / Alaskan Native	5.9%	14.7%	14.5%	17.1%	22.1%
Actually made AYP in subject in '03 and '04		56.7%	52.4%	5.7%	7.3%

Notes to Table 4: This sample includes all public schools used to estimate Equation 1. These schools provide 2001-2002 student test performance data for the relevant grade level, typically fifth grade. For more details on chosen grade levels, please consult the "Student test performance in focal subject in 2001-2002" row in Appendix 2.

Table 5: Evidence on NCLB Pressure from the ISBA Survey in California, Georgia, and Pennsylvania

Panel A: Principals	Above AYP Margin (N=104)	On AYP Margin (N=21)	
Do you agree with the following statement:			
My school can attain the AYP targets for 2003-2004	96.1%	71.4%	
My school can attain the AYP targets for the next five years	71.6%	47.6%	
Has your school and/or district done any of the following:			
Encouraged or required teachers to spend more time on tested subjects and less time on other subjects	49.0%	61.9%	
Distributed commercial test preparation materials	67.0%	81.0%	
Distributed released copies of the state test or test items	76.9%	85.7%	
Panel B: Math Teachers	Above AYP Margin (N=1074)	On AYP Margin (N=224)	Below AYP Margin (N=19)
As a result of the state mathematics test:			
I focus more effort on students who are close to proficient	25.9%	41.3%	52.6%
I spend more time teaching general test-taking strategies	52.6%	66.7%	73.7%
I look for particular styles and formats of problems in the state test and emphasize those in my instruction	66.5%	79.9%	100.0%
I focus more on topics emphasized in the state test	69.4%	81.3%	84.2%
I spend more time teaching content	54.1%	73.4%	79.0%
I search for more effective teaching methods	72.7%	83.9%	94.4%

Notes to Table 5: Percentages shown in this table refer to the percentage of respondents who agreed with the corresponding statement. Above, on, and below the AYP margin correspond to our classifications of how likely the school was to make AYP in 2003 and 2004. See Section 4 of the paper for details. No principal surveyed was in a school classified by our analysis as below the AYP margin. All of the differences in rates between the groups above the AYP margin and either of the other two groups are statistically significant at approximately the .01 level or better. Differences in rates between teachers in schools above the AYP margin and those in schools on the AYP margin are statistically significant at the .05 level for "I focus more effort on students who are close to proficient," and at the .01 level for "I look for particular styles..." and "I search for more effective teaching methods."

Table 6: Effects of NCLB Pressure on Teacher Attitudes

Dependent Variable:	Concerned about Job Security due to Student Test Score Performance			Plan to Teach Until Retirement		
	(1)	(2)	(3)	(4)	(5)	(6)
Teachers:	All	Less Experienced (<10 Years)	More Experienced (>= 10 years)	All	Less Experienced (<10 Years)	More Experienced (>= 10 years)
Main Sample: NCLB Sample Wave						
On AYP Margin	.046	.101*	.001	-.057*	-.0002	-.082**
	(.038)	(.057)	(.051)	(.033)	(.057)	(.035)
Below AYP Margin	.097*	.170**	-.002	-.154***	-.108	-.161***
	(.051)	(.076)	(.070)	(.048)	(.083)	(.053)
Falsification Sample: Pre-NCLB Sample Wave						
On AYP Margin	-.021	-.071	.037	.039	.028	.043
	(.042)	(.068)	(.050)	(.035)	(.061)	(.038)
Below AYP Margin	-.030	-.025	-.016	.066	.083	.058
	(.057)	(.093)	(.068)	(.049)	(.086)	(.056)
Differences between Actual and Falsification Estimates						
On AYP Margin	.066	.172*	-.036	-.096**	-.028	-.125**
	(.057)	(.089)	(.071)	(.048)	(.084)	(.052)
Below AYP Margin	.127*	.194	.014	-.226***	-.191	-.219***
	(.076)	(.120)	(.098)	(.070)	(.119)	(.077)
N, Main Sample	2800	1270	1520	2800	1270	1520
N, Falsification Sample	3200	1310	1890	3200	1310	1890

Notes to Table 6: The main sample includes teachers sampled in the 2003-2004 wave of the Schools and Staffing Survey (SASS) who were working in high-stakes grades/subjects. The falsification samples includes teachers sampled in the 1999-2000 wave of the SASS who were working in grades/subjects that later became high stakes for NCLB. For the falsification regressions, schools are classified as “On AYP Margin” or “Below AYP Margin” if they later had that status during the main sample period. To facilitate comparisons with the ECLS analysis in the remainder of the paper, we limit both samples to teachers in public schools that serve 5th grade students. All models control for the independent variables listed in the "SASS sample" column of Table 3, and also control for state fixed effects, the fraction of states in which the school would likely be “On the AYP Margin,” the fraction of states in which the school would likely be “Below the AYP Margin,” a squared term for the number of Limited English proficient students in the grade, a squared term for the teacher’s years of experience, and squared and cubic terms for schools' within-state standardized 2001-2002 test score performance in both math and reading. All models use the SASS cross-sectional sample weights to make the estimates nationally representative. Bootstrapped standard errors, adjusted for school-level clustering using 1,000 Monte Carlo simulations of both the first-stage and second-stage models, are displayed in parentheses below each estimate. Sample sizes are rounded to the nearest 10 to comply with restricted-data reporting requirements.
*** significant at .01 level; ** significant at .05 level; * significant at .10 level.

Table 7: Effects of NCLB Pressure on Teachers' Work Hours and Instruction

Dependent Variable:	Work Hours in a Typical Week		Whole-class Instructional Hours in a Typical Week		Taught at Least One Lesson During the Last Week in… Science	Social Studies
	(1)	(2)	(3)	(4)	(5)	(6)
Teachers:	Generalists	Specialists	Generalists	Specialists	All	All
Main Sample: NCLB Sample Wave						
On AYP Margin	-1.84**	4.15**	-0.99**	-1.52	-.085**	-.049
	(0.84)	(1.81)	(0.44)	(1.18)	(.038)	(.036)
Below AYP Margin	-2.65**	4.89**	-1.76**	-2.20	-.157***	-.146***
	(1.27)	(2.31)	(0.72)	(1.50)	(.053)	(.049)
Falsification Sample: Pre-NCLB Sample Wave						
On AYP Margin	0.28	0.63	N/A	N/A	.016	.036
	(0.66)	(2.16)			(.036)	(.034)
Below AYP Margin	0.79	-0.40			-.022	.011
	(0.86)	(2.73)			(.048)	(.049)
Differences between Actual and Falsification Estimates						
On AYP Margin	-2.11**	3.52			-.101**	-.085*
	(1.07)	(2.82)			(.052)	(.050)
Below AYP Margin	-3.44**	5.29			-.135*	-.158**
	(1.53)	(3.57)			(.072)	(.069)
N, Main Sample	2300	490	2300	490	3370	3370
N, Falsification Sample	2710	480	N/A	N/A	3870	3870

Notes to Table 7: See Notes to Table 6. To capture the full effects of shifting of instruction across subjects, the models used for columns 5 and 6 above do not limit the sample to teachers of high-stakes subjects. The Pre-NCLB Sample Wave of the Schools and Staffing Survey did not include a question about instructional hours in a typical week. Specialist teachers are defined as those whose classroom organization is reported as departmentalized instruction or elementary enrichment (i.e., not a self-contained classroom). All other teachers are classified as generalists.
*** significant at .01 level; ** significant at .05 level; * significant at .10 level.

Table 8: Effects of NCLB Pressure on Student Learning and Motivation

	Reading Score	Math Score	Science Score	Enjoyment of Reading	Enjoyment of Math	Anxious About Standardized Tests
Panel I						
On the AYP Margin	0.08**	0.00	0.04	0.03	0.13*	-0.09***
	(0.04)	(0.04)	(0.03)	(0.07)	(0.08)	(0.04)
Below the AYP Margin	0.03	0.01	0.00	0.02	-0.02	-0.14***
	(0.05)	(0.05)	(0.05)	(0.09)	(0.09)	(0.05)
Panel II						
On the AYP Margin based on the performance of…						
Overall student group	0.03	0.06	0.02	0.11	0.23	-0.03
	(0.05)	(0.07)	(0.06)	(0.12)	(0.14)	(0.06)
Student's subgroup (not overall)	0.06	-0.03	0.06	0.01	0.04	-0.05
	(0.05)	(0.07)	(0.05)	(0.11)	(0.15)	(0.05)
Other subgroup(s) (not overall or student's subgroup)	0.11***	-0.02	0.04	0.00	0.12	-0.15*
	(0.04)	(0.05)	(0.04)	(0.08)	(0.09)	(0.04)
Below the AYP Margin	0.03	0.01	0.00	0.03	-0.01	-0.13***
	(0.05)	(0.05)	(0.05)	(0.09)	(0.09)	(0.05)

Notes to Table 8: Each column displays estimates from three student-level models using data from the Early Childhood Longitudinal Survey-Kindergarten Cohort (ECLS). Panel I displays estimates of the coefficients for whether the school was on the AYP margin or below the AYP margin in the relevant subject: math for math test performance or enjoyment, reading for reading test performance or enjoyment, and *either* math or reading for science test performance or anxiety about standardized tests. To decompose the first panel results by the type of subgroup(s) that were on the AYP margin, Panel II's models use three mutually exclusive indicators that sum to "On the AYP Margin" variable. All models control for the variables listed in the "ECLS sample" column of Table 3, plus state fixed effects, an indicator for whether the school is predicted to be below the margin for making AYP, the fraction of states in which the school would likely be "On the AYP Margin," the fraction of states in which the school would likely be "Below the AYP Margin," and squared and cubic terms for the student's standardized math and reading performance in both the first and third grade waves of the ECLS. Dependent variables are from the fifth grade wave of the ECLS. Sample sizes are approximately 6,870 (rounded to the nearest 10 to comply with data reporting requirements). All models weight observations using the student-level longitudinal sample weights provided in the ECLS data. Bootstrapped standard errors, adjusted for school-level clustering using 1,000 Monte Carlo simulations of both the first-stage and second-stage models, are displayed in parentheses below each estimate.
** significant at .05 level; * significant at .10 level.

Appendix 1. Sources of Collected AYP Data

Table A.1

	Available in existing databases	We have collected	Not available	State Abbreviations Where Data are Not Available
States in 2002-2003				
School made AYP	24	44	0	—
Subgroup made AYP	5	38	9[i]	AL[ii], IA, ME, NE, NM, ND, OK, WI, WY
Percent proficient by subgroup	16	41	5	AL, ME, NE, NH, WV
Number of students in subgroup	2	34	15	AL, CO, DE, HI, ID, IA, ME, MS, NE, ND, OH, OK, SD, WV, WY
States in 2003-2004				
School made AYP	48	46	0	—
Subgroup made AYP	39	40	4	IA, NE, NM, ND
Percent proficient by subgroup	16	44	3	AL, NE, NH
Number of students in subgroup	1	37	10	CO, ID, IA, ME, MS, NE, ND, OH, SD, WY

Notes to Table A.1: Existing databases refer to School Data Direct and the National AYP and Identification Database
Number of states per row can exceed 50 because we collected data in states included in existing databases.
(i) For schools in Arizona, New Jersey, and Pennsylvania, due to otherwise missing data, we impute whether some subgroups made AYP in 2002-2003 using their 2002-2003 proficiency rates and their states' published standards.
(ii) Although Alabama did not publish whether student subgroups made AYP in 2002-2003, we can include Alabama schools in our analyses because Alabama (incorrectly) did not base schools' AYP status in 2002-2003 on student subgroup performance.

Appendix 2: Predicting the Probability of Making AYP

We run state-specific regressions using the data described below to generate predictions of the likelihood that each numerically-significant student subgroup and (by extension) their school would pass AYP in the spring of both 2003 and 2004 in the subjects of reading and math. To be as consistent as possible in our state-by-state predictions of which student subgroups were on the AYP margin, we applied a set of rules to the construction of data to generate subgroup-level AYP failure predictions. The table on the following page explains the data construction in detail.

We use a specific subgroup's 2001-2002 proficiency rate wherever available to predict that subgroup's likelihood of making AYP in 2003 and 2004 (note these are cross-sectional measures of a subgroup's performance). For privacy protection, the 2001-2002 test score data is typically missing for groups below a state-determined minimum size (e.g., fewer than 20 students). Thus, for schools where subgroup enrollment grew between 2001-2002 and 2004, there might be AYP determinations for a subgroup in 2004 but no 2001-2002 proficiency rate. (In the rare case, the 2001-2002 suppression rules redacted data for groups larger than minimum subgroup size requirements for AYP accountability.) To retain these cases in our sample, we specified an alternate version of the probit regression, where we assign the school-wide 2001-2002 proficiency rate to all student subgroups within the school regardless of whether we possessed subgroup-specific 2001-2002 proficiency rates. In this case, we add an interaction term with a variable measuring the fraction of the school-wide population composed of students in the relevant subgroup. We then use predictions from the alternate probit version in cases when predictions were missing from the main specification.

Sometimes entire subgroups were dropped from probit regressions when there was not any within-subgroup variation in the subject in the state (e.g., there were only 11 numerically-significant Asian subgroups in 2004 among Washington's elementary schools and all 11 passed AYP their math and reading proficiency targets). In cases where subgroups' success or failure was perfectly determined, we overwrote their missing probabilities of making AYP with predicted probabilities obtained from OLS regressions that used the same set of predictors. This practice was of little consequence, because subgroups in these cases were always classified as having either low or high likelihoods of making AYP (they never fall in the moderate category).

Table A.2: Model Specification and Data Construction for State Probits Estimating Likelihood of Subgroups Making AYP in 2003 and 2004		
Variable description	**Data sources**	**Variable coding**
Dependent variable		
Subject-specific subgroup AYP proficient indicator Subjects are math and reading. Student subgroups are: school-wide; African American; Asian/Pacific Islander; Hispanic; White; Native American; Limited English Proficient; Disabled; Economically Disadvantaged; Filipino (when used by state); Asian (when used by state); Pacific Islander (when used by state); and Alaskan Native (when used by state).	Wherever available, school report card data from states' departments of education listing state's own determinations of whether student subgroups passed their proficiency targets in the years 2002-2003 and 2003-2004. State's final yes/no determinations typically account for all forms of adjustment of subgroup raw proficiency rates (e.g., 2- or 3-year averaging; confidence intervals; safe harbor; and appeals). When not available from state DOE sources, data is from SchoolDataDirect.org or the National AYP and Identification Database (for 2003-2004 only). In two states which lacked 2002-2003 proficiency target data from all three sources of data, we constructed the variable using each state's published raw subgroup proficiency rates, which we adjusted using the state's documented confidence interval methods (if applicable) to determine whether each subgroup passed, failed, or was not applicable. This approximation method had greater than 90% accuracy when tested it in two populous states with complete data.	Equals 0 if the subgroup failed its AYP subject-specific proficiency target in either 2002-2003 or 2003-2004. Equals 1 if the subgroup (a) passed its AYP proficiency target in the given subject in 2002-2003 and 2003-2004, or (b) passed in one year and numerically insignificant in the other year. Equals missing if the subgroup was numerically insignificant in both years (according to the state's own definition of numerical significance). For states that further break out AYP proficiency targets by grade level or grade span, subgroup indicators are specific to each accountable grade level/span, using the same rules for creating values of missing, zero, or one. Two states did not use subgroup-level pass rates to determine schools' AYP status in 2002-2003. In each case, only 2004 subgroup-level AYP proficiency target data was used to construct the dependent variable. Two states only published whether the subgroup passed AYP in each subject overall (a measure that includes both the subgroup's proficiency rate and its participation rate for that subject). In these cases, we used this overall subject measure in lieu of proficiency-only indicators.
Independent variables		
Subgroup test performance in focal subject in 2001-2002 (entered into model as linear, squared, and cubed terms)	National Longitudinal School-Level State Assessment Score Database	When available, we use the subgroup's unadjusted 5th grade proficiency rate on the statewide test administered in 2001-2002 for the focal subject. (We selected grade 5 because our second stage of analysis examines ECLS student outcomes in 2003-2004, when the majority of ECLS students are fifth graders.) For states not reporting performance for particular subgroups, we use the overall student performance in the focal subject in the selected grade level in that school. As described in the text, we supplement those models with interaction terms between the test performance variable and the fraction of students who are members of that subgroup. For 6 states where proficiency rates are unavailable, we instead use the reported percentile rank scores or scale scores.

Table A.2: Model Specification and Data Construction for State Probits Estimating Likelihood of Subgroups Making AYP in 2003 and 2004		
Variable description	**Data sources**	**Variable coding**
		For states that did not test grade 5 in 2001-2002, we use the next closest lower tested grade level (i.e., grade 4, grade 3) or, if that is unavailable, the next closest higher tested grade (i.e., grade 6, grade 7). The models then include observations for all schools in that state with test performance variables in the relevant grade levels. When these models include test performance from two different grade levels (e.g., 4th and 6th), we also include a dichotomous dummy variable indicating whether the test variable values come from students in the higher grade. In states that further break out subgroups' AYP proficiency targets by grade levels or grade spans, we run separate models for each high-stakes grade for schools serving 5th graders. Depending on availability, we use 2001-2002 test performance variables from either the same grade, the next lowest grade, or the next highest grade.
Pct. that the student subgroup comprised of the denominator for its 2001-2002 proficiency rate value (entered as a main effect, and interacted with the three 2002 proficiency rate terms)	National Longitudinal School-Level State Assessment Score Database Where student subgroup size not present in State Assessment Score database, data is from the Common Core of Data.	Equals 1 when the subgroup's own proficiency rate available from 2001-2002. Otherwise, ranges from 0 to 1, and is equal to the ratio of enrolled students in the given subgroup in 2001-2002 within the school (from CCD) to the total number of enrolled students in the school. Since data about the number of LEP students and disabled students is not available at the school level in the CCD, we substituted in 2003-2004 AYP subgroup size ratios for the LEP and disabled subgroups. If this subgroup size data not available in a state for 2003-2004, then we use district-level LEP and disabled ratios (applicable to three states).
Size of the student subgroup in 2003-2004 (entered as 1/sqrt(size), and this term is also interacted with the three 2002 proficiency rate terms and the three 2002 proficiency rate x 2002 pct. group interaction terms)	Wherever available, school report card data from state departments of education that list student subgroup size (using AYP definitions). Where not available from state sources, then drawn from 2003-2004 data in the National Longitudinal School-Level State Assessment Score Database or the 2003-2004 Common Core of Data.	This variable is derived from the state's count of continuously enrolled students per student subgroup accountable under NCLB (note that states' definitions of "continuous enrollment" for the purposes of AYP accountability differ somewhat from state definitions for state accountability systems or just cross-sectional enrollment counts as of the fall in the school year). Where state sources are not available, size is estimated using 2004 State Assessment Score data about number of students tested per subgroup. If this source is not available for the state, we used grade-specific CCD enrollment data and district-level LEP and disabled ratios and applied them to school-by-grade-level membership.

Table A.2: Model Specification and Data Construction for State Probits Estimating Likelihood of Subgroups Making AYP in 2003 and 2004		
Variable description	**Data sources**	**Variable coding**
Indicators for years held accountable	The same data source used to obtain the dependent variable.	Two dichotomous variables indicating whether the subgroup was only numerically significant in 2003 (but not 2004) in the focal subject and, vice versa, numerically significant in 2004 (but not 2003) in the focal subject. The omitted category is the subgroup is numerically significant in both 2003 and 2004.
Subgroup indicators	Constructed	A series of dichotomous variables indicating the student subgroup to which the observation belongs. The omitted category is the campus-wide student group.
School-level characteristics in 2001-2002: (a) percent of students who are black (b) percent of students who are Hispanic (c) percent of students who are Asian (d) percent of students who qualify for a free- or reduced-price meal (e) whether the school is Title I eligible (f) total student membership	Common Core of Data 2001-2002 school-level data	We constructed the racial and economic demographic using total student membership as the denominator. In cases where categories of school-level data were missing from 2002 state files, the variables were constructed using the next closest year in which those variables were present in CCD files (2000-2001, then 2002-2003, then 1999-2000, etc.)