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

Randall Reback
Barnard College
rr2165@columbia.edu

Jonah Rockoff
Columbia Business School
jr2331@columbia.edu

Heather L. Schwartz
RAND Corporation
hschwart@rand.org

PRELIMINARY DRAFT: December, 2010*
PLEASE DO NOT CITE WITHOUT AUTHORS' PERMISSION

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 national study of the strong short-term incentives created by NCLB for schools on the margin of failing AYP. To examine the impact of NCLB on the behavior of school personnel and the academic achievement of students, we create the first comprehensive, national, school-level data set concerning detailed performance measures used to calculate AYP and merge these data with nationally representative samples of teachers and students. Our identification is based on idiosyncrasies in state policies, which 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. We find that accountability pressure due to NCLB lowered teachers' perceptions of job security and increased their work hours, particularly for untenured teachers. We also find that NCLB pressure has either neutral or positive effects on students' enjoyment of learning and their achievement gains in reading, math, and science.

* We would like to thank participants at CALDER/Urban Institute NCLB Research Conference, the Association for Public Policy and Management conference, the International Workshop on Applied Economics of Education, and the American Education Finance Association conference for their comments and 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 from Barnard College. 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 accountability systems that use student proficiency on statewide exams in math and reading to determine whether public schools satisfy Adequate Yearly Progress (AYP). Moreover, AYP status must be based on proficiency within student subgroups, (e.g., students from low income families, students with limited English proficiency), when a sufficient number of students in the subgroup enroll in the school. Failure to satisfy AYP triggers sanctions that escalate over time, including allowing students to transfer to other public schools and allowing students from low-income families to enroll in after-school tutoring programs funded by federal revenue that would have otherwise gone to the school.[1]

While school accountability has received much attention from economists, there is no national study of the impact of the incentives generated by NCLB on school personnel and students. We aim to fill this gap by investigating the links between the accountability incentives under NCLB and a wide array of outcomes for nationally representative samples of teachers and students. To this end, we assemble a new data set 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 schools near the margin for satisfying their *own* state's AYP requirements would have almost certainly failed or almost certainly passed AYP if they were located in *other* states.[2] This allows us to implement a difference-in-differences style approach, comparing differences in outcomes for schools on and away from the AYP margin within the same state to the difference in outcomes between similar schools in other states, neither of which is on the AYP margin.

We measure outcomes using non-public versions of two nationally representative datasets—the Schools and Staffing Survey (SASS) and the Early Childhood Longitudinal Survey (ECLS). We find clear evidence that accountability pressure from NCLB reduces perceptions of job security and increases

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

work hours, particularly among untenured teachers. Untenured teachers in schools likely to fail AYP report concerns about job security, regardless of subject area, while untenured teachers in schools with moderate chances of failure express concerns only if they teach in a high stakes grade and subject area. Reported hours worked per week also increase for untenured teachers teaching high stakes subjects in schools with a high or moderate risk of failing to make AYP.

Our analysis also suggests that short-term NCLB pressure has either positive or neutral effects on student achievement in math, reading, and science. Students enrolled in schools with a moderate risk of failing to make AYP score 0.07 standard deviations higher on low-stakes readings exams compared with students in comparable schools that were well above the margin for making AYP. The estimated effect for math performance is also positive (0.04 standard deviations) but not statistically significant at the .10 level. Achievement gains from short-term NCLB pressure do not come at the expense of performance in a low-stakes subject (science), reported enjoyment of learning, or reported anxiety over testing. In fact, students in schools facing stronger short-term incentives to raise school-wide math proficiency rates report significantly higher levels of enjoyment of math.

The paper proceeds as follows. In Section 2, we present a framework for how schools might be expected to respond to incentives under an accountability system like NCLB and discuss prior related empirical work. Section 3 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 4, and our estimated effects of NCLB on teachers and students in Section 5. Section 6 concludes.

2. Conceptual Framework and Related Literature

We first present a framework for how schools respond to a system of accountability such as No Child Left Behind. Schools have various resources they can use to improve student skills, and all of these resources have associated costs. Subject to a budget constraint, schools choose an allocation of resources (e.g., school staff, curriculum, facilities, parental involvement, etc.), based on preferences about the relative importance of helping students improve different types of skills and the relative importance of helping different types of students make improvements. There are also competing demands that constrain the amount and allocation of school resources; school staff members care about their own leisure time and local taxpayers care about their consumption of other goods and services.

More formally, we classify schools' resources into four types: the first (denoted u) helps to improve all skills for all students (e.g., the overall effort level of teachers), the second (denoted a_s) is skill-specific and serves all students (e.g., math lessons that equally help all students learn math), the third type (denoted b_i) is student-specific and serves all skills (e.g., providing individual students with lessons to improve study-skills), and the fourth (denoted c_{is}) is skill-specific and student-specific (e.g., individual math tutoring). Suppose there are two categories of student skills that schools aim to improve, one which is measured on standardized tests ($s=m$), and another which is not ($s=z$). Schools place weights (denoted γ_{is}) on skill acquisition for each student, depending on the preferences of school staff and the community. Finally, let l denote the type of resources that improve consumption of goods and services that are valued by the community and school staff but are unrelated to skill acquisition (e.g., teacher leisure). Schools with N students and total resources equal to K will choose an allocation of resources to maximize:

(1)
$$U(l)+\sum_{s=m,z}\sum_{i=1}^{N}\gamma_{is}\, f_{is}(u,a_s,b_i,c_{is})$$

$$\text{subject to } \sum_{s=m,z}\gamma_{is}=1 \text{ and } u+\sum_{s=m,z}a_s+\sum_i\left(b_i+\sum_{s=m,z}c_{is}\right)=K-l$$

In the equation above, the function U determines the value received by community members and school staff for non-skill resources (l) , and the function f_{is} maps other resources into the performance of student i in skill s. Schools choose an optimal allocation of resources given this objective function and budget constraint, which we will call "business as usual."

A system of accountability and ratings such as NCLB introduces benefits or costs that depend on the fraction of students who pass standardized tests. Suppose that an additional resource (denoted d_i) is available which increases the probability that student i passes the standardized tests but does not improve skill acquisition. The school now chooses an allocation of resources to maximize:

(2)
$$U(l)+\sum_{s=m,z}\sum_{i=1}^{N}\gamma_{is}\, f_{is}(u,a_s,b_i,c_{is})+V\left(\frac{1}{N}\sum_{i=1}^{N}g_i(u,a_m,b_i,c_{im},d_i,\varepsilon_i)\right)$$

$$\text{subject to } \sum_{s=m,z}\gamma_{is}=1 \text{ and } u+\sum_{s=m,z}a_s+\sum_i\left(b_i+\sum_{s=m,z}c_{is}+d_i\right)=K-l$$

In this equation, ε_i is idiosyncratic noise due to imperfect test measurement (with mean zero and known variance), the function g_i maps resources and test measurement error into whether student i passes the standardized tests, and the function V maps the school-wide pass rate into benefits or costs.[3] Note that resources that do not improve measured skills (a_z and c_{iz}) do not enter in the function g_i.

The provisions of NCLB essentially impose costs on schools with pass rates below a certain threshold (AYP). This structure tends to make the value of the school-wide pass rate have an "all or nothing" quality. Formally, let V take the following form:

$$V = \begin{cases} \overline{V} \; if \left(\dfrac{1}{N} \displaystyle\sum_{i=1}^{N} g_i\left(u, a_m, b_i, c_{im}, d_i, \varepsilon_i\right) \right) \geq P^* \\ \underline{V} \; if \left(\dfrac{1}{N} \displaystyle\sum_{i=1}^{N} g_i\left(u, a_m, b_i, c_{im}, d_i, \varepsilon_i\right) \right) < P^* \end{cases} \quad where \; \overline{V} > \underline{V} \tag{3}$$

In other words, the school is worse off if the pass rate falls below some threshold P^*, but all other variation in the pass rate above or below that threshold does not have immediate consequences related to NCLB. The "all or nothing" structure has important implications for the accountability pressure faced by different schools. Because input allocations and the variance of test measurement error are known, schools will form expectations about their probabilities of making AYP. If a school has a very high probability of making AYP under its optimal pre-NCLB resource allocation**,** it will face very little pressure to improve and, consequently, resource allocation under NCLB should resemble "business as usual." In contrast, if a school expects to be close to the margin of making AYP under its optimal pre-NCLB resource allocation, the school and its community will face considerable pressure to improve student pass rates. This is a key identifying assumption in our methodology.[4]

There are several ways in which an accountability system such as NCLB may change resource allocation decisions. It may induce schools and communities to reduce resources devoted to consumption of goods and services (l) and direct them towards improving student skills. Accountability pressure may also induce schools to spend fewer resources promoting non-tested skills (a_z and c_{iz}), more resources promoting tested skills (a_m and c_{im}), and more resources targeted to particular students (b_i, c_{im}) for whom

[3] Alternatively, the function V could enter into the school's budget constraint, rather than the utility function, but the qualitative results from this alternative framework would be the same. Note that, for simplicity, V is based on the overall student proficiency rate on a single test, whereas NCLB holds schools accountable for proficiency rates for the overall student population and additional subgroups of students on both math and reading tests.

[4] Schools with a very low probability of making AYP in the current year will also face pressure to improve over a longer period of time. Our empirical work focuses on comparisons of schools near the margin with schools with high probabilities of making AYP, but in some cases we also test for effects of schools having low probabilities of making AYP.

extra resources will most improve their probability of passing the standardized exams. Finally, schools may allocate resources to activities that improve pass rates (d_i) but not skill acquisition.

The extent to which these incentives change the allocation of resources in a school will depend on schools' preferences and the functions that determine how resources affect skill acquisition and exam pass rates. Schools may respond by exhorting their staff to perform more effectively and/or to exert greater effort, attempting to raise student achievement for all students without negatively affecting any student. In less ideal circumstances, schools may shift resources in ways that improve their chances of making AYP but at the expense of the acquisition of skills by some students or the acquisition of non-tested skills.

Most empirical research on school accountability incentives focuses on 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). These studies find evidence that accountability pressure causes schools to reallocate resources in ways that raise average student achievement. However, this research has also found that schools shift resources towards students and subjects that are most critical to the schools' 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). Feng et al. (2010) also find that schools with poor accountability ratings subsequently experience higher rates of teacher turnover.

Knowledge about the impacts of NCLB is still nascent. Among the few studies that apply rigorous methods, most examine only one state or one city (Springer, 2008; Krieg, 2008; Ladd & Lauen, 2010; and Neal & Whitmore Schanzenbach, 2010). These studies have found that low scoring students enrolled in schools failing AYP tend to make greater than expected test score gains, but there is conflicting evidence concerning heterogeneous effects on students at different parts of the performance spectrum. Only two 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 years they took high-stakes tests, particularly for students near the margin of passing their high-stakes exam. Dee and Jacob (2009) 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.

3. Data and Descriptive Analysis

3.1 Data Description

Our analysis focuses on the initial years of NCLB implementation following its passage in January 2002. To measure NCLB pressure faced by schools during these years (and which subgroups and 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 NCLB datasets.[5] We present the categories of data collected and their sources in Appendix 1.

We examine the effects of NCLB on teacher-level outcomes measured in the 2003-2004 wave of the SASS and student-level outcomes measured in the spring 2004 wave of the ECLS. Each of these surveys is sponsored and distributed by the National Center for Educational Statistics. We have gained access to the non-public-use versions of these surveys' data sets, so we can link individual schools to our constructed measures of NCLB pressure based on data we collected from states. The SASS surveyed teachers in all 50 states and, with the use of sampling weights, allows researchers to construct nationally-representative samples.[6] The first panel of Table 1 provides summary statistics on the outcome variables we create from SASS survey questions.[7]

The ECLS followed students for nine years, from kindergarten until most had completed eighth grade. Data collection took place in both the fall and the spring of the school years 1998-1999 and 1999-

[5] 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.

[6] The SASS surveyed administrators but we did not feel these questions were 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.

[7] For consistency with our examination of student outcomes, we limit the sample of teachers to those working in regular public schools that served at least five fifth graders as of 2001-2002. We replaced teachers' reported 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.

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 longitudinal dataset covering years before and after the passage of NCLB. Indeed, the timing of the ECLS survey is 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 nationally representative of kindergartners, their classrooms, and their schools in the school year 1998-1999, (and representative of first grade students in 1999-2000), and it includes students from 40 relatively populous states.[8]

Of particular interest to us in the ECLS is student performance on a series of standardized tests in reading, math, and science. Unlike the tests that states administer under NCLB, the ECLS tests were low stakes, and they were given un-timed in an adaptive manner (i.e., subsequent questions are selected based on a student's performance on preceding questions) to prevent floor or ceiling effects and increase 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 it had no consequences for NCLB. This reduces concerns about teaching to the test or strategic responses to survey questions. Also, the ECLS tests are not directly related to NCLB, so measurement error or other shocks to high-stakes test scores that do not reflect real achievement should not induce mean reversion in our dependent variables.

The second panel of Table 1 provides descriptive statistics for our ECLS outcome measures. We limit the sample of students to those attending regular public schools in the spring of the school year 2003-2004 that also served at least five fifth grade students as of 2001-2002. We standardize students'

[8] It used a multistage probability sample design, first selecting broad geographic areas (e.g., a county), then selecting schools within that area, and finally selecting students within those schools. On average, 23 kindergarten students were sampled from each school. The ECLS includes students who were retained within the same grade or skipped a grade level, but has some attrition so that the ECLS may not be perfectly representative of the national student population in the later years of data collection. In the school year 1999-2000, the sample remained representative by surveying a randomly-selected 50 percent sub-sample of students who transferred from their original school and adding another random sample of first graders in the same schools where transfer students were followed. However, this "freshening" of the sample was not repeated in the third, fifth, and eighth grades. Approximately, one-quarter of children changed schools between kindergarten and first grade, and half of the children in the ECLS had changed schools at least once between kindergarten and third grade. While accountability pressure from NCLB might have affected student mobility between 3rd and 5th grade, this does not appear to substantially influence our estimates below. We observe qualitatively similar estimates from similar models that, instead of using child-level sampling weights, simply remove students who made non-structural changes in their schools (i.e., 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).

scores within subject and year to have a mean of zero and a standard deviation of one.[9] In additional to standardized exams, we examine students' reported enjoyment of math and reading, as well as reported anxiety over standardized tests.[10]

Table 2 provides descriptive statistics on control variables using in our regression analyses. We show them for our two samples: public school teachers from the SASS and public school students from the ECLS. Along with control variables from the surveys themselves, we use as control variables school characteristics from the Common Core of Data (CCD) compiled by the National Center for Educational Statistics (NCES), and aggregated student test performance variables from the National Longitudinal School-Level State Assessment Score Database (compiled by American Institutes for Research).[11] We standardize test performance variables within states to have a mean of zero and standard deviation 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.

3.2 Descriptive Analysis of AYP Outcomes under NCLB

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," typically the attendance rate. Beyond these

---

[9] The ECLS data report t-scores of students' IRT-based "theta scores," which are estimates of students' skill levels. We calculate Z-scores of these t-scores without adjusting for sample weights, so that the mean score of our measure equals zero for each subject in each period.

[10] Answers to these specific questions, rather than an index based on a larger set of items, are only available in the non-public-use version of the ECLS. 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.

general 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. These thresholds apply to proficiency rates for 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 to *within* each tested grade level;[13]
- determine whether to calculate proficiency rates using *multiple years of testing*;
- define *continuous enrollment*, where only continuously enrolled students count towards calculation of subgroup size 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 generosity of *confidence intervals* applied to student subgroups' raw proficiency rates, which 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 only five ethnic subgroups (Asian/Pacific Islander, black, Hispanic, Native American, and white), California and Alaska add additional subgroups (Filipino and Alaskan Native, respectively) while Asian/Pacific Islander is not an AYP subgroup in Texas.

All of these seemingly esoteric decisions have real implications for whether schools fail to meet the targets set for them under NCLB, and there was a remarkable amount of variation in the fraction of

[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.

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 and 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 a single aspect of NCLB design that determines school failure rates. However, by testing a number of factors we have come to the conclusion that interaction of four features significantly influences the likelihood that a school fails AYP: (1) state rules for the numerical significance of student subgroups; (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 manner in which these policy details interact increases our confidence that the wide differences in the leniency of AYP requirements across states can help identify the causal impact of NCLB incentives on schools and students.

4. Predicting the Probability of Failing AYP

In the first stage of our analysis, we use our data to determine which student subgroups and, by extension, which schools were on the margin of failing to make AYP in the first two years during which NCLB was in effect. To do so, we use explanatory variables from the school year 2001-2002—after the

[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.

passage of NCLB but prior to the first AYP determinations—to predict whether student subgroups met AYP targets in math and reading for the school years 2002-2003 and 2003-2004. We use these predictions to assess schools' overall probabilities of making AYP.

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. The independent variables in these models include the 2001-2002 school demographic characteristics (listed in Table 2) and 2001-2002 subgroup-level/school-level test performance variables.[16] We conduct regressions separately by state so that coefficients capture the nuances of how states' NCLB rules affect their 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.[17] Because of the variation in NCLB rules across states, our variables differ somewhat across some states. To be as consistent as possible, we applied a set of rules for how to specify our regressions conditional on the available data, and Appendix 2 describes these rules.

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

$$(4)\ AYP_{jks03\text{-}04} = \begin{cases} 1 \quad if \quad \alpha_q + X_{jks02}\beta_{1q} + N_{jks04}\beta_{2q} + X_{jks02}N_{jks04}\beta_{3q} + W_{j02}\beta_{4q} + M_{jks03-04}\beta_{5q} + \zeta_{jks} > 0 \\ \qquad\qquad 0 \ otherwise \end{cases},\ j \in q$$

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, W_{j02} is a vector of control variables for school-level demographics from the school year 2001-2002 (listed in Table 2), and $M_{jks03-04}$ 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

[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.

[17] 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.

includes cubic terms for the test performance in subject *s* among students in subgroup *k* at school *j*.[18] The N_{jks04} vector and interactions between N_{jks04} and X_{jks02} 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 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. 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 source.

We focus our sample on 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 classifications in the Common Core of Data and (c) served at least five students in grade 5.[19] 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 tests or AYP determinations for subgroups). 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 unreliable. Fortunately, these states are relatively small, and more than 92 percent of the U.S. population resides in one of the 41 states with sufficient data for our analyses.

4.1 Defining the AYP Margin

We use predicted subgroup-level AYP pass probabilities from the state- and subject-specific regressions in Equation 4 to construct measures of accountability pressure under NCLB. Our measures are based on the following logic. Schools where all subgroups have high chances of passing state

[18] 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.

[19] 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.

proficiency targets in both math and reading likely faced little NCLB pressure. In contrast, schools where *any* 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 run.

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 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 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 subgroup in the school has a moderate chance of making AYP in that subject, and (b) no 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. While these cutoffs are admittedly ad hoc, our results are not very sensitive to using other cutoffs ranging from 35-65 percent to 15-85 percent.

Table 3 summarizes our measures of NCLB pressure over the years 2003 and 2004 for schools in 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 years were: 87 percent for schools above the margin, 38 percent for schools on the margin, and 7 percent for schools below the margin. These results provide evidence that our first stage specification has sufficient power to identify substantial variation in which subgroups and 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 test whether our estimates are related to teachers' and administrators' reported sense of accountability pressure, which we do below.

The results reported in Table 3 also reveal that, with the exception of white and economically disadvantaged students, most student subgroups were typically not numerically significant and 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. Moreover, this rate varied across states depending on their minimum subgroup size requirements, again underscoring the importance of these detailed regulations. 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 and 82 percent in Florida.

Among subgroups that were numerically significant and thus accountable under NCLB, the fraction we predict to have a moderate or low chance of making AYP varies considerably. The fraction of significant subgroups we predict to have a moderate chance of passing proficiency targets is highest for disabled and limited English proficient students in reading (30 and 37 percent, respectively) and highest for disabled and Black students in math (26 and 27 percent, respectively). For disabled student subgroups, relatively high fractions (about 15 percent) are also predicted to have low chances of passing proficiency targets in each subject. This is also true of Native American student subgroups in both subjects (17 percent in math, 22 percent in reading) and Asian student subgroups in reading (25 percent). In contrast, extremely low fractions of white student subgroups are predicted to have a moderate or low chance of passing proficiency targets in either math or reading.

4.2 Variation in Predicted NCLB Pressure across States

Our identification strategy is predicated on the idea that similar schools faced different levels of pressure to improve under NCLB based only on 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. To illustrate both of these ideas, we take our primary measure of NCLB pressure—whether a school was on the AYP margin—and plot cumulative distributions of the percent of schools on the margin across 41 states, separating schools by quartile of within-state school-wide test score performance in the school year 2001-2002. These results (Figure 3, top panel) show that we place more schools on the AYP margin within the lowest performance quartile, but that being on the margin is not the exclusive territory of low scoring schools. The median state has 60 percent of its lowest quartile schools on the AYP margin, but also has 25 percent, 10 percent, and 5 percent of schools on the AYP margin for schools in the second, third, and top performance quartiles, respectively.

These cumulative distributions also illustrate that in some states with "tough" NCLB rules we put many relatively high performing schools on the AYP margin. In 20 percent of states, the percent of schools on the AYP margin in the lowest through highest performing quartiles, respectively, were at least 80 percent, 50 percent, 25 percent, and 10 percent. In contrast, for the 20 percent of states that appear to have the *lowest* amount of NCLB pressure, the percent of schools on the AYP margin in the lowest through highest performing quartiles were, respectively, no greater than 40, 12, 5, and 3 percent.

When we plot similar cumulative distributions for the percentage of schools we place *below* the AYP margin, we see less variation but similar qualitative results (see Figure 3, bottom panel). Far more schools are below the margin in the bottom quartile of school test performance. Nevertheless, we place hardly any schools below the margin in a few states, while in some states we place a substantial fraction of schools below the margin in the second or third quartiles of within-state performance, i.e., despite not scoring very poorly overall, they have little chance of making AYP in the short run.

4.3 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 three states during the school year 2003-2004 that focused on various aspects of NCLB.[20] We lack micro-data from this survey, which are not publicly available, and present these results as suggestive. We pursue a more rigorous methodology in Section 5.

For principals, we are only able to examine schools on the margin (21 schools) or above the margin (104 schools) of AYP, because no principals were surveyed at any school that we predict had a low probability of making AYP. Among principals working in schools we classified as being above the AYP margin, 96 percent felt they would make AYP in the school year 2003-2004, relative to only 71 percent in the marginal group. 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 4, Panel A). Principals in schools on the AYP margin were between 9 and 14 percentage points more likely to say that they had: 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

[20] As mentioned in Section 3, the RAND Corporation collected these data as part of their Implementing Standards-Based Accountability (ISBA) study and provided us with these cross-tabulations.

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 teacher surveys, we can examine teachers working in schools below the margin (19 teachers), on the margin (224 teachers), and above the margin (1,074 teachers) of AYP. Teachers were asked about various actions, such as teaching test-taking strategies, focusing on students who are close to proficient, 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 the marginal schools were between 11 and 19 percentage points more likely than teachers working in schools above the AYP margin to have taken these actions, while teachers below the margin were between 3 and 20 percentage points more likely to have taken these actions than teachers in the marginal group. All of the differences between the schools above the margin and either of the other two groups are statistically significant at the one percent level, and help confirm that our constructed measures of NCLB pressure align with principals' and teachers' reported perceptions.

5. 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 5:

$$(5) \qquad Y_{ij} = \delta_q + Q_{ij}\rho_1 + W_{j02}\rho_2 + X_{j02}\rho_3 + \lambda(Margin\_AYP_j) + \gamma(BelowMargin\_AYP_j) + \zeta_{ij}\,.$$

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 the W_{j02} vector of school-level control variables is the same as in Equation 4. The X_{j02} vector is similar to X_{jks02} in Equation 4; it contains school-wide student proficiency in reading and math during the school year 2001-02, normalized within the state to have a mean of zero and standard deviation of one, and the square and cube of this performance measure. The coefficient of interest is λ, which represents the average impact of being in a school that is on the AYP margin. This specification also includes an indicator for being in schools below the AYP margin, so λ should be interpreted as the change in the outcome variable for a school on the AYP margin compared with a school with a high probability of making AYP.[21]

Because Equation 5 uses covariates estimated from our first stage, we measure standard errors based on a two-sample bootstrap adjusted for school-level clustering. We use 1,000 Monte Carlo simulations of both the first-stage 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.

5.1 Impacts on Teachers

We examine the effect of accountability pressure on teachers' attitudes and work hours using the SASS data. To promote comparability with the ECLS analysis of fifth grade students, we restrict the sample to teachers working in public schools serving at least five fifth graders in the school year 2001-2002. Many of these teachers do not teach students or subjects tested under NCLB, so we augment Equation 5 with an indicator for whether the teacher taught math or reading in a tested grade level, and interact this with the indicators for whether the school was on or below the AYP margin.[22] The Q_{ij} vector includes the teacher-level control variables listed in Table 2, with both linear and squared terms for teachers' years of experience.

The first column of Table 5 (Panel A) displays the estimated effects of NCLB pressure on whether teachers are concerned that student test performance at their school will affect their job security. Teachers of high-stakes grades/subjects are more likely to be concerned about their job security if they are teaching at a school that is on the AYP margin. Compared to teachers of high-stakes grades/subjects at schools that are above the AYP margin, these teachers are 4.9 percentage points more likely to report concern over their job security related to student test performance—a large increase considering that only 7.5 percent of teachers reported this concern overall. Like the ISBA results above, this finding supports the notion that our measure of NCLB pressure is valid and captures significant variation in school staff members' perceptions of pressure.

---

[21] Because few schools are predicted as having a very low probability of satisfying AYP, our estimates of λ remain qualitatively similar if we drop the "below margin" indicator from the regressions.

[22] Some teachers cover multiple grade levels, so we set the "high stakes grade/subject" indicator variable equal to one if the teacher covers either math or reading and more than half of the teacher's covered grade levels were tested for NCLB in that teacher's state during the spring of 2004.

Given that untenured teachers may react differently to NCLB pressure than tenured teachers, Panel B of Table 5 displays estimates from similar models that further restrict the sample to untenured teachers.[23] Untenured teachers working in schools below the AYP margin, even those not teaching high-stakes grades/subjects, tend to be very concerned about how student test performance will affect their job security (Column 1). This makes sense, given that the consequences of failing to make AYP (decreases in enrollment or school closure) would make untenured teachers most vulnerable to losing their jobs

We next examine how NCLB pressure affects teachers' long term career plans. We construct an indicator variable from teachers' survey responses concerning whether they plan to teach until retirement. Teachers working in high-stakes grades/subjects are 12 percentage points less likely to plan to teach until retirement if they are working in a school below the AYP margin rather than a school above the AYP margin (Table 5, Panel A, Column 2). This result is especially strong for untenured teachers, suggesting that these teachers may be discouraged by the challenge of raising student proficiency rates at schools that are unlikely to make AYP (Table 5, Panel B, Column 2). This discouragement appears limited to schools with low chances of making AYP in the short run. If a school is on the AYP margin, untenured teachers are at least as likely to have long-term teaching career plans as their counterparts teaching in schools with higher probabilities of making AYP.

The third column of Table 5 presents results concerning how NCLB pressure affects teachers' total weekly work hours, measuring several months ahead of NCLB testing. The largest and most significant differences in work hours occur for untenured teachers who work in high-stakes areas at schools on the AYP margin. These teachers report working about four hours more per week than their untenured co-workers teaching low-stakes grades/subjects at their schools, and over two hours more per week than untenured teachers of high stakes grades/subjects working in schools with high chances of making AYP. Given that the standard deviation of work hours is under 10 hours (see Table 1), these increases in hours worked by untenured teachers are substantial.

We also estimate the impact of NCLB pressure on teachers' self-reported number of *instructional* hours per week, a subset of their total work hours. Unlike total hours, instructional hours do not increase for any type of teacher in response to NCLB pressure (Table 5, Column 4). Moreover, instructional hours

[23] The SASS does not measure tenure, so we created an indicator for whether a teacher's total years of experience (measured in the SASS) exceeds the state's required number of years for tenure. Data on state requirements come from 2002-2003, see Brunner and Imazeki (forthcoming), and we thank Eric Brunner for providing them.

actually *decline* for teachers working in schools below the AYP margin and for untenured teachers working in schools that are on the margin, though not differentially for those teaching high-stakes grades/subjects. Untenured teachers at schools on the AYP margin spend almost two fewer hours per week on instruction than untenured teachers at schools above the AYP margin. The fact that untenured teachers in schools facing short-term NCLB pressure report working more hours but spending less time on instruction means that part of their work day must be shifting towards other activities such as student assessment, grading, lesson planning, or other non-instructional activities. This heightens the concern that NCLB pressure may have negative effects on student achievement, particularly for material not covered on high stakes exams, and provides further motivation for our analysis of the ECLS data.

One concern with these results is that principals at schools facing NCLB pressure might strategically place teachers into high-stakes grades and subjects. However, we believe such behavior would likely bias our estimates toward zero. For example, if principals wishing to boost high-stakes test performance assigned their most talented teachers to the high-stakes areas, this would likely work against our finding that these teachers are relatively more concerned with their job security, unlikely to plan to teach until retirement, and work longer hours *only* if they do not have tenure.

While we are unable to identify the specific activities for which teachers devote more time, we can explore whether NCLB pressure caused schools to shift instructional time across subject areas using teachers' reports concerning their teaching content during the previous week. The SASS randomly selected teachers to survey, and their activities should therefore be an unbiased (albeit noisy) measure for those of all teachers in the school. To examine whether NCLB pressure shifted resources away from low-stakes subjects, we focus on whether the teacher taught at least one science lesson or at least one social studies lesson during the prior week. The SASS surveyed teachers in the fall, well ahead of NCLB testing, and survey responses should reflect general shifts in instruction rather than last-minute preparation for high-stakes tests.

The estimates displayed in Table 6 suggest that schools on the AYP margin slightly change the proportion of teachers offering science lessons. Compared to teachers at schools above the margin, these teachers are 4.0 percentage points less likely to have offered a science lesson. They are also 1.3 percentage points less likely to offer a social science lesson, but this estimate is not statistically significant. The effects on science and social studies offerings in schools below the AYP margin are even larger and more statistically significant. Compared to teachers at schools above the margin, they are 11 percentage points less likely to offer a science lesson and 6 percentage points less likely to offer a social studies lesson. These are large differences considering that 63 percent and 65 percent of teachers in this sample taught science and social studies lessons, respectively (see Table 1). Schools with little chance of making AYP in the short term may still try to shift instruction toward the high-stakes subjects in order increase their chances of eventually making AYP. [24]

5.2 Impacts on Students

Our student-level analysis of the ELCS data is also based on specifications similar to Equation 5. These regressions control for the variables listed in Table 2, state fixed effects, an indicator for whether the school is predicted to be below the AYP margin, and a third degree polynomial of the student's standardized math and reading performance in both the first and third grade waves of the ECLS.

Panel I of Table 7 displays estimates of the coefficient on whether the school was on 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.[25] Our estimates suggest that NCLB pressure has either neutral or positive effects on student achievement in both low- and high-stakes subjects. Students' reading scores are .073 of a standard deviation greater on average when schools are on the AYP margin for reading (Table 7, Panel I, Column 1). This estimate is statistically significant at the .05 level and it is a considerably large effect; 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). Students' math scores are 0.043 of a standard deviation greater on average when the school is on the AYP margin for math performance, though this estimate is not statistically significant at the 10 percent level. Although we are

[24] Teachers may be "generalists" (i.e., teach all subjects in a self-contained classroom) or specialists, and our estimates could be driven by schools below the margin using specialists who teach intensive amounts of science and social studies to compensate for decreased lessons from other instructors. However, if we examine hours taught in science or social studies conditional on teaching at least one lesson in the subject we find no evidence for this explanation.

[25] We focus on the most relevant subject(s) here due to power limitations for separating relevant-subject and cross-subject effects using the ECLS, which is smaller than the SASS. For brevity, we do not report estimated coefficients for "below the AYP margin" in Table 7; these estimates are never statistically significant at the .10 level and only one is even statistically significant at the .20 level (a negative estimate for the test anxiety model).

examining results for multiple dependent variables, a power test suggests that these three estimates are far too large to simply be due to chance.[26]

Importantly, our results also suggest that when schools face NCLB pressure, gains in achievement do not come at the expense of the average student's enjoyment of reading or math, anxiety over testing, or performance in science. Science scores are actually greater (an increase of 0.049 of a standard deviation) in schools on the AYP margin, though this estimate is not statistically significant. In schools on the AYP margin for reading, students' enjoyment of reading decreases by a statistically insignificant 0.031 standard deviations, while enjoyment of math increases by a statistically significant 0.148 standard deviations in schools on the AYP margin for math. In schools on the AYP margin for either subject, students report a small and statistically insignificant decrease in their anxiety over testing.

The framework presented in Section 2 motivates the idea that the impacts of NCLB may differ across students. 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 student's own subgroup. Panel II's models replace the single "on the AYP margin" variable with three mutually exclusive indicators for whether the school was on the AYP margin in the relevant subject for: (1) the overall student group, (2) the student's own subgroup (and *not* the overall student group as well), and (3) other subgroups (and *not* the student's own subgroup or the overall student group). Interestingly, the point estimates for all three subjects in Panel II are positive, regardless of whether students are members of subgroups whose performance is most critical to the schools' AYP ratings, and our results do not appear to be driven by targeting the most critical students. For example, the largest improvements in reading scores when schools are on the AYP margin for reading occur among students whose own subgroup(s) are *not* on the margin, while the largest increases in enjoyment of math occur when schools are on the AYP margin for the math performance of the entire student population. It is possible that schools do target resources toward students in critical subgroups in ways that help these students' high-stakes test performance more than their low-stakes test performance, but NCLB pressure does not appear to have negative effects on student achievement, regardless of whether students contribute to the most critical proficiency rates.

[26] To test the joint significance of these test score estimates, we simulated estimation of these three models after randomly reassigning schools to different AYP status. Out of 1,000 simulations, *none* produced three estimates that were, respectively, at least as large in absolute value of as the actual highest, second highest and third highest estimate reported in the first three columns of Panel I of Table 7.

To produce the largest increase in student proficiency rates, schools might also target resources to students who are likely to score close to the threshold of passing the exam (see Reback (2008) and Neal and Whitmore Schanzenbach (2010)). To investigate this issue, we classify a student as “on the bubble” for passing their state exam if their third grade ECLS test score was within 15 percentiles below or 5 percentiles above the national percentile equivalent of their states' NCLB exam passing threshold, as estimated by the National Center for Education Statistics (2007).[27] Panel III of Table 7 displays estimates from specifications that add an indicator for whether a student is "on the bubble" for passing the state's NCLB exam and an interaction of this indicator with whether the school was on the AYP margin. The interaction term coefficients are generally positive but statistically insignificant. As mentioned above, the evidence on heterogeneity in the impacts of accountability across students is mixed. Our results do not suggest that students on the bubble of passing the high-stakes exam perform very differently when their schools face strong NCLB pressure, although our estimates are too imprecise to rule out small effects.

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 measured targets for student test performance. This represents a sweeping change for most areas of the nation, but our understanding of its impact has been hindered by a lack of national data on NCLB implementation and nationally comparable data on outcomes. Assembling an extensive national dataset of school and student subgroup performance on the examinations required under NCLB, we exploit extensive cross-state variation in NCLB rules and standards NCLB to examine how NCLB pressure affects school personnel and students. We find that teachers in schools with strong incentives to improve student test performance are more concerned about how student test performance will affect their job security, and untenured teachers in high-stakes grades/subjects at these schools work longer hours. We also find evidence that schools that face longer-term incentives to improve student

[27] 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.

proficiency (i.e., they will almost certainly fail state requirements in the short-run) allocate less time to science and social studies instruction. Relative to students in schools facing little NCLB pressure, students in schools facing strong short-term incentives to improve student proficiency raise achievement by 0.07 standard deviations on low-stakes reading exams, do at least as well on low-stakes math and science tests, do not report less enjoyment of reading or math, and do not report more test anxiety.

One of our most important and robust findings is that short-term NCLB pressure does not negatively affect student learning or their enjoyment of learning. On the other hand, our results also raise questions concerning whether NCLB pressure motivates both tenured and untenured teachers alike, whether talented teachers become discouraged after working in schools with little chance of making AYP, and whether schools neglect low-stakes subjects if their proficiency rates lag far below the standards. These questions loom larger every year as NCLB requires higher proficiency targets and the share of schools that fail to meet those standards rises. As Congress is likely to debate revisions to No Child Left Behind in the near future (see Dillon, *New York Times*, 2010), policymakers may wish to ensure that schools along the entire performance spectrum continually face incentives to improve along a wide array of outcomes. For example, Barlevy and Neal (2010) propose rewards for teacher performance designed so that teachers' incentives are independent of their students' prior achievement levels and the scaling of the students' test scores.

Policymakers may also want to consider the very premise upon which we identify the effects of NCLB pressure on schools, i.e., large differences in rules and regulations across states. The difficulty of meeting AYP is driven in great part by the minutiae of state rules, such as minimum significant subgroup size, the number of grade levels tested, and adjustments to raw proficiency rates (e.g., one-tailed vs. two-tailed confidence intervals, degree of confidence required, number of years allowed for test score averaging, and method for safe harbor adjustments). Although there is much support for increasing the consistency of standardized achievement tests across states, much of the 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' standards, then NCLB reforms must address the other sources of variation within state formulae. Ideally, accountability pressure should stem from student performance levels along the entire distribution of performance, rather than the idiosyncrasies of state rules.

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.

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. (2009). The impact of No Child Left Behind on student achievement. NBER Working Paper No. 15531. Retrieved November 20, 2009, from NBER's web site: http://www.nber.org/papers/w15531

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.

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.

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

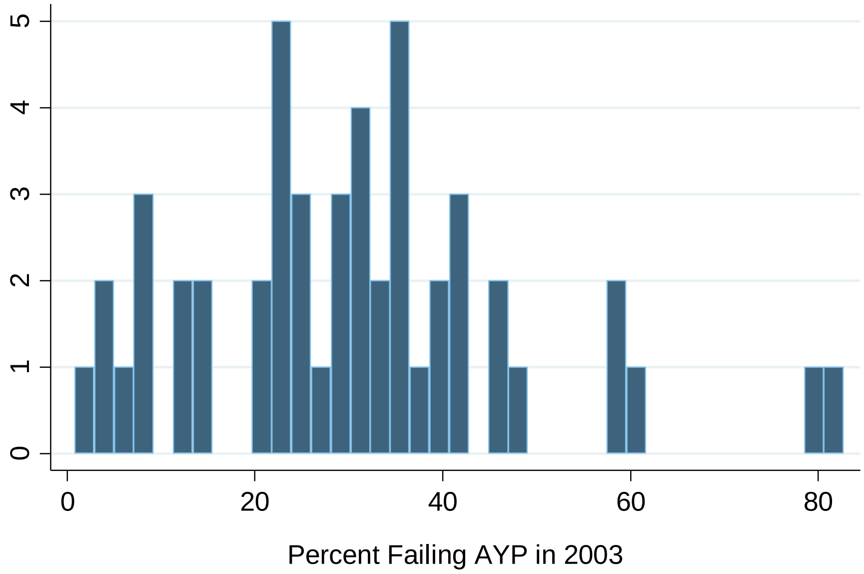
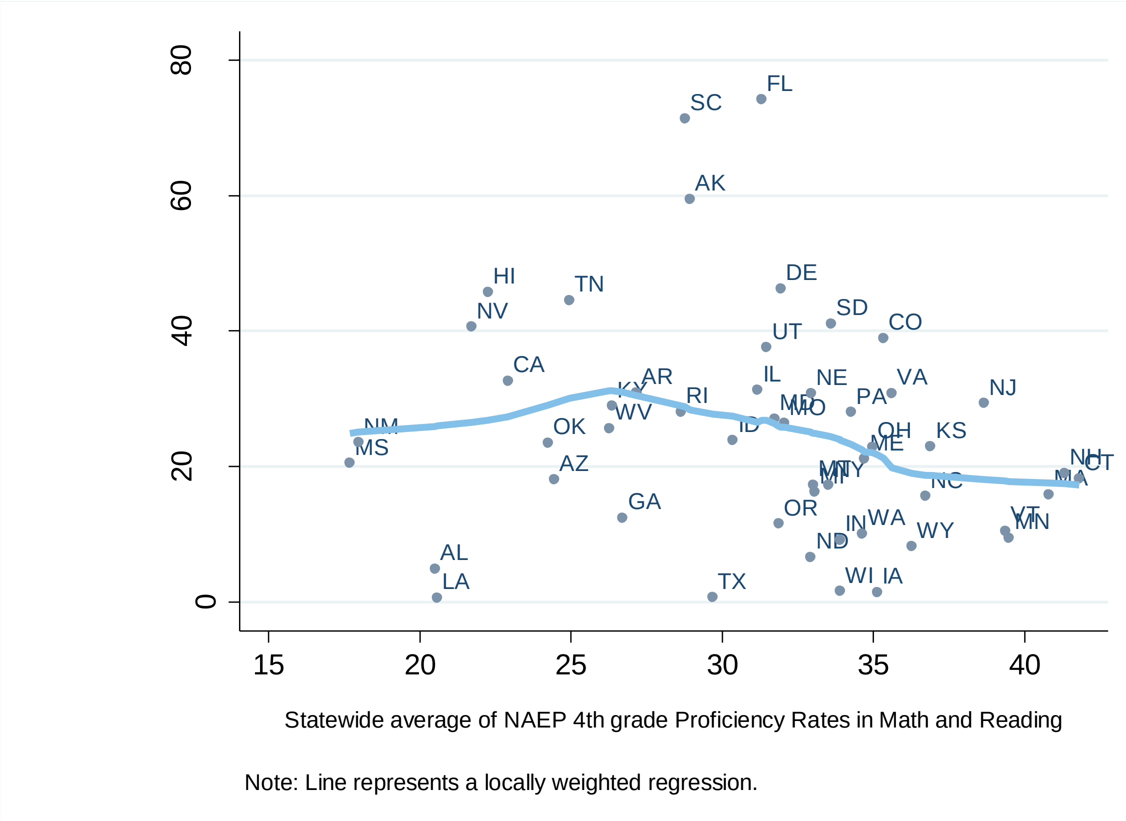


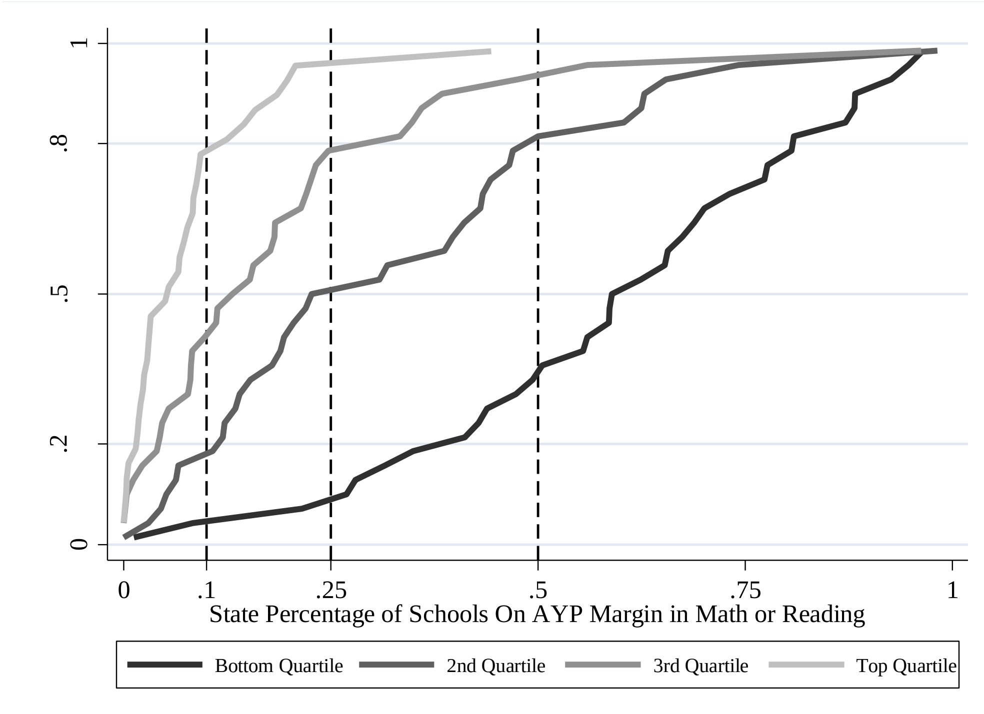
Figure 2: AYP Failure Rates vs. NAEP Proficiency Rates by State, 2003



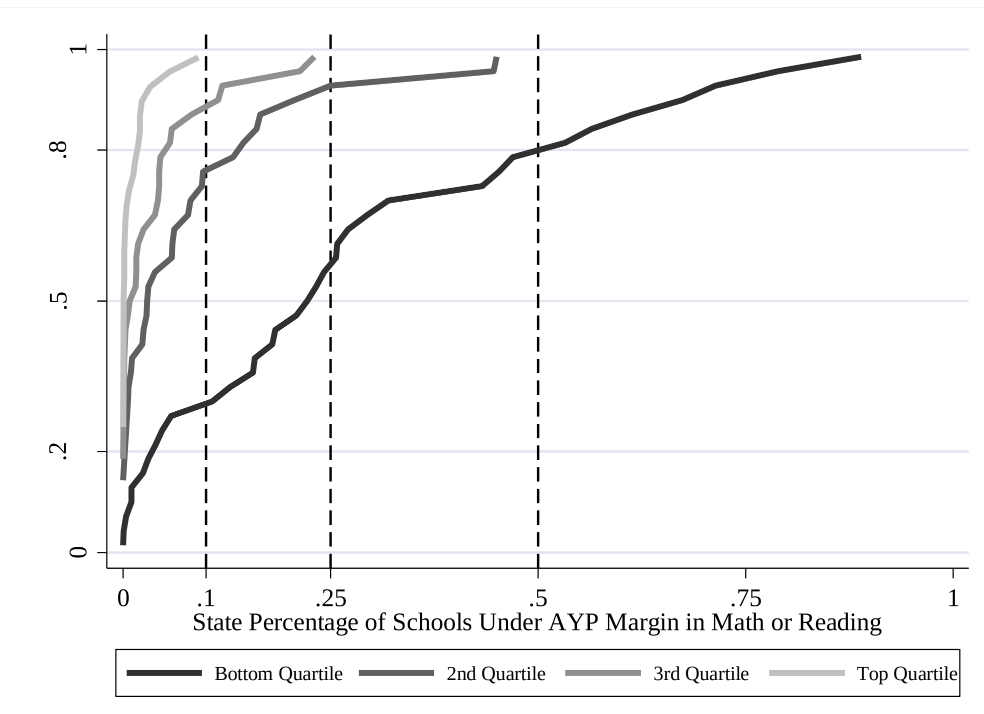
Failure rates are based on schools serving at least five fifth grade students.

Figure 3: State Variation in the Percentage of Schools Facing NCLB Pressure

Schools on the AYP Margin



Schools Below the AYP Margin



Note: These figures show cumulative distributions of the percentage of schools we consider on the margin of making Adequate Yearly Progress (top panel) and below the margin of making Adequate Yearly Progress (bottom panel) for 2003 and 2004 for the 41 states in our Schools and Staffing Survey analysis. Quartiles reflect schools' positions in their own state's distribution of student test performance during the school year 2001-2002.

Table 1: Summary Statistics for Dependent Variables

	Mean	SD
Teacher-level Dependent Variables from the SASS		
Concerned about Job Security due to Student Test Performance	7.5%	
Plan to Teach Until Retirement	78%	
Work Hours per weekeek†	52.4	8.91
Instructional Hours per weekeek†	29.1	5.17
Gave at Least One Science Lesson Last Week	63%	
Gave at Least One Social Studies Lesson Last Week	65%	
Untenured Teachers Only:		
Concerned about Job Security due to Student Test Performance	11%	
Plan to Teach Until Retirement	73%	
Work Hours per week†	53.8	9.46
Instructional Hours per week†	29.5	5.40
Student-level Dependent Variables from the ECLS		
5th Grade Reading Score (Standardized)	.009	.967
5th Grade Math Score (Standardized)	.028	.982
5th Grade Science Score (Standardized)	.081	.950
Enjoyment of Reading (Standardized)	-.002	1.01
Enjoyment of Math (Standardized)	.037	1.01
Anxiety about standardized tests	42%	

Notes to Table 1: 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 7,870 teachers for the SASS sample and approximately 6,860 students for the ECLS sample, (rounded to the nearest 10 due to restricted-use data reporting requirements). Standardized variables are Z-scores that were standardized prior to weighting and prior to limiting the sample to states with adequate data, so that the standardized variables' means and standard deviations above differ from zero and one respectively.

†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 2: Descriptive Statistics for Control Variables

	SASS Sample		ECLS Sample	
Variable	**Mean**	**SD**	**Mean**	**SD**
School characteristics				
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%	
Number of enrolled students	587	258	587	251
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%	
Teacher characteristics (from the SASS)				
Total years of experience	13.9	10.1		
Teaches Math	77%			
Teaches Reading (or English)	83%			
Teaches a high-stakes subject/grade	34%			
Teaches grades 2 or 3	41%			
Teaches grades 4 or 5	42%			
Teaches grades 6 or 7	14%			
Teaches grades 8 or 9	7%			
Teaches grade 10 or higher	3%			
Teaches grades 2 or 3 and grades 4 or 5	15%			
Teaches grades 4 or 5 and grades 6 or 7	7%			
Teaches grades 6 or 7 and grades 8 or 9	5%			
Teaches grades 8 or 9 and grade 10 or higher	2%			
Family characteristics (from the ECLS)				
Two parent household			67%	
Mother's education level unknown			9%	
Mother has at least a high school diploma			89%	
Mother possesses a B.A.			31%	
Family income missing			16%	
Family income under $20,000			15%	
Family income $20,000 -$35,000			18%	
Family income $35,000 - $50,000			14%	
Family income $50,000 - $75,000			14%	
Family income $75,000 - $100,000			11%	
Student characteristics (from the ECLS)				
Reading Z-score in spring 2000			0.017	0.950
Math Z-score in spring 2000			0.029	0.919
Reading Z-score score in spring 2002			-0.001	0.981
Math Z-score in spring 2002			0.029	0.970
African American			18%	
Hispanic			20%	
Asian			3%	
Other			5%	
Female			48%	
Date of birth (measured in days)			3/18/93	140

N = approximately 7,870 teachers for the SASS sample and approximately 6,860 students for the ECLS sample.

Table 3: 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 3: This sample includes all public schools used to estimate Equation 4. 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 4: 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 2003-04	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 4: 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 administrator 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 the 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 5: Effects of NCLB Pressure on Teacher Attitudes and Work Hours

	Concerned about Job Security due to Student Test Performance		Plan to Teach Until Retirement		Work Hours in a Typical Week		Instructional Hours in a Typical Week	
All Teachers								
On the AYP Margin	-0.015		0.004		0.03		-0.42	
	(.014)		(.025)		(.57)		(.32)	
Below the AYP Margin	0.015		0.002		-1.24		-1.20	**
	(.023)		(.038)		(.81)		(.41)	
Teach High-stakes	0.007		-0.005		-0.73		-0.33	
	(.014)		(.023)		(.45)		(.26)	
On the AYP Margin	0.054	**	-0.039		0.04		-0.32	
*Teach High-stakes	(.021)		(.034)		(.76)		(.43)	
Below the AYP Margin	0.046		-0.115	**	1.21		0.27	
*Teach High-stakes	(.031)		(.045)		(1.0)		(.52)	
Untenured Teachers Only								
On the AYP Margin	0.017		0.114	*	-0.74		-1.93	**
	(.042)		(.065)		(1.5)		(.73)	
Below the AYP Margin	0.124	**	0.116		0.04		-2.21	**
	(.062)		(.085)		(2.25)		(.94)	
Teach High-stakes	0.022		0.096		0.87		0.19	
	(.048)		(.059)		(1.24)		(.63)	
On the AYP Margin	0.020		-0.089		3.07	*	0.34	
*Teach High-stakes	(.075)		(.078)		(1.87)		(1.11)	
Below the AYP Margin	-0.050		-0.286	**	2.70		-1.21	
*Teach High-stakes	(.089)		(.128)		(2.65)		(1.16)	

Notes to Table 5: Each column displays estimates from two separate teacher-level regressions using data from the 2003-2004 wave of the Schools and Staffing Survey (SASS). The top panel uses a sample of both tenured and non-tenured teachers, and the bottom panel restricts the sample to tenured teachers. Teachers' tenure status is not reported directly in the SASS, so we impute it using teachers' reported years of experience and their states' tenure policies (see footnote 24). All models control for the independent variables with summary statistics listed in the "SASS sample" column of Table 2, and also control for state fixed effects, for a squared term for the number of Limited English proficient students in the grade, for a squared term for the teacher's years of experience, and for 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, 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.

Table 6: NCLB Pressure and Instruction in Low-stakes Subjects

	Teacher gave at least one science lesson last week		Teacher gave at least one social studies lesson last week	
On the AYP Margin	-0.040	*	-0.013	
	(.022)		(.021)	
Below the AYP Margin	-0.104	**	-0.062	**
	(.034)		(.030)	

Notes to Table 6: Each column displays estimates from a teacher-level regression using data from the 2003-2004 wave of the Schools and Staffing Survey (SASS). These models control for the independent variables with summary statistics listed in the "SASS sample" column of Table 2, except for the indicators for whether the teachers covered math or reading and the indicator for whether the teachers covered a high stakes grade/subject. Similar to Table 5, the models also control for state fixed effects, for a squared term for the number of Limited English proficient students in the grade, for a squared term for the teacher's years of experience, and for 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, 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.

Table 7: 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 (35 states)						
On the AYP Margin	0.073**	0.043	0.049	-0.031	0.148**	-0.051
	(.032)	(.036)	(.033)	(.064)	(.073)	(.035)
Panel II (35 states)						
On the AYP Margin based on the performance of…						
Overall student group	0.010	0.092	0.040	0.061	0.282**	0.005
	(.049)	(.068)	(.056)	(.113)	(.14)	(.059)
Student's subgroup (not overall)	0.053	0.014	0.074	-0.012	0.109	-0.038
	(.053)	(.067)	(.049)	(.096)	(.143)	(.05)
Other subgroup(s) (not overall or student's subgroup)	0.108**	0.030	0.034	-0.078	0.099	-0.086 **
	(.039)	(.046)	(.037)	(.081)	(.088)	(.042)
Panel III (27 states)						
On the AYP Margin * Student on the bubble for Passing	-0.035	0.080	0.075	0.113	0.081	0.018
	(.061)	(.095)	(.056)	(.148)	(.191)	(.064)
On the AYP Margin	0.056	-0.013	0.034	0.002	0.146	-0.059
	(.036)	(.043)	(.037)	(.074)	(.090)	(.042)

Notes to Table 7: 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 coefficient on whether the school was on 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. Panel III's models use the same independent variable as in Panel I, but add an interaction term with a dummy variable for whether the student is on the bubble for passing the state's NCLB exam in the relevant subject; this dummy variable also enters the model separately and its creation is described in the text of Section 5. All models control for the variables listed in the "ECLS sample" column of Table 2, plus state fixed effects, an indicator for whether the school is predicted to be below the margin for making AYP, 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 6,860 students for the first two panels and 5,630 students for Panel III, (rounded to the nearest 10 to comply with data reporting requirements). The smaller sample for Panel III is due to missing information concerning the difficulty of six states' NCLB exams. The estimates in Panels I and II remain fairly similar if we restrict the sample to the roughly 5,630 observations used to estimate the models of Panel III. All models weight observations using the student-level longitudinal sample weights provided in the ECLS data. Bootstrapped standard errors, 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

	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 Appendix 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 even 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).

Model Specification and Data Construction for State Probits Estimating Likelihood of 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 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 5^{th} 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.

Model Specification and Data Construction for State Probits Estimating Likelihood of 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.
Indicator for whether more than one grade level of 2001-2002 proficiency data used	Constructed	Equals 1 in states where more than one grade level of 2001-2002 proficiency rate data was used. Equals 0 in states where only one grade level of proficiency rate data is used to predict the dependent variable.
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.	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).

Model Specification and Data Construction for State Probits Estimating Likelihood of Making AYP in 2003 and 2004		
Variable description	**Data sources**	**Variable coding**
group interaction terms)		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.
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.)