American Economic Journal: Economic Policy 2014, 6(3): 207–241
http://dx.doi.org/10.1257/pol.6.3.207

Under Pressure: Job Security, Resource Allocation, and Productivity in Schools under No Child Left Behind†

By Randall Reback, Jonah Rockoff, and Heather L. Schwartz*

We conduct the first nationwide study of incentives under the No Child Left Behind (NCLB) Act, which requires states to punish schools failing to meet target passing rates on students' standardized exams. States' idiosyncratic policies created variation in the risk of failure among very similar schools in different states, which we use to identify effects of accountability pressure. We find NCLB lowers teachers' perceptions of job security, shifts time towards specialist teachers in high-stakes subjects and away from whole-class instruction, and has positive or neutral effects on students' enjoyment of learning and achievement in reading, math, and science. (*JEL* H52, H75, I21, I28, J45)

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 are required to adopt school accountability systems based on student proficiency on statewide math and reading exams, and to measure proficiency within student subgroups (e.g., students from low-income families, students with limited English proficiency). States must impose escalating sanctions on schools that fail to satisfy Adequate Yearly Progress (AYP) requirements for exam proficiency, including allowing students to transfer to other public schools, forcing schools to pay for students from low-income families to enroll in after-school tutoring programs, and, ultimately, closing or restructuring persistently failing schools.[1]

*Reback: Department of Economics, Barnard College, Columbia University, 3009 Broadway, New York, NY 10027 (e-mail: rr2165@columbia.edu); Rockoff: Columbia Business School, 3022 Broadway, Uris Hall 603, New York, NY 10027 (e-mail: jr2331@columbia.edu); Schwartz: Rand Corporation, Poydras Center, 650 Poydras Street, Suite 1400, New Orleans, LA 70130 (e-mail: hschwart@rand.org). We thank Abigail Payne, Steve Rivkin, David Figlio, and three anonymous referees for their detailed comments, as well as participants at the American Economics Association meetings, the CALDER/Urban Institute NCLB Research Conference, the Association for Public Policy and Management conference, the Chinese University of Hong Kong, Teachers College, the International Workshop on Applied Economics of Education, and the American Education Finance Association conference for many thoughtful suggestions. This research project was made possible by funding from the Institute for Education Sciences and the Spencer Foundation, as well as seed grants from the Columbia University Institute for Social and Economic Research and Policy and Barnard College, and support from the Paul Milstein Center for Real Estate at Columbia Business School. The authors are solely responsible for any opinions or errors in the paper. Molly Alter, Daisy Chu, Ben Lockwood, Julia Zhou, Sean Tom, and especially Elizabeth Davidson provided outstanding research assistance. The authors thank the US 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 10.

†Go to http://dx.doi.org/10.1257/pol.6.3.207 to visit the article page for additional materials and author disclosure statement(s) or to comment in the online discussion forum.

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

Most prior empirical research on school accountability focuses on the impacts of state and local systems, many of which preceded No Child Left Behind (e.g., Ladd and Zelli 2002; Hanushek and Raymond 2005; Rouse et al. 2007; Chiang 2009; Rockoff and Turner 2010; Chakrabarti 2013). Several studies find evidence that accountability pressure causes schools to reallocate resources in ways that raise average student achievement. However, schools have also been found to shift resources towards students and subjects that are most critical to the accountability rating (e.g., Booher-Jennings 2005; Reback 2008; Neal and Schanzenbach 2010), teach to the test (Jacob 2005; Figlio and Rouse 2006), remove low-performing students from the testing pool (Figlio and Getzler 2006; Figlio 2006; Cullen and Reback 2006), or cheat (Jacob and Levitt 2003).

Knowledge about the impacts of NCLB is still nascent. Among the studies that apply rigorous methods, most are limited to examining student performance on high-stakes tests in one state or one city (Springer 2008; Krieg 2008; Ladd and Lauen 2010; Neal and Schanzenbach 2010; Hemelt 2011; Cooley Fruehwirth and Traczynski 2012). These studies have found that students enrolled in schools failing AYP tend to make greater than expected gains on high-stakes tests, though there is conflicting evidence concerning heterogeneous effects on students at different parts of the performance spectrum. Only two prior studies examine the impact of NCLB incentives in multiple states. Ballou and Springer (2008) examine variation in the grade levels tested for NCLB across seven states and find that students generally perform better on low-stakes exams during years when they took high-stakes tests, particularly for students near the margin of passing their high-stakes exams. Dee and Jacob (2011) find that students in states with no prior accountability policies experienced greater increases on the National Assessment of Educational Progress in some grades and subjects after NCLB was introduced.

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

This variation in state policy allows us to implement a cross-sectional identification strategy similar to a difference-in-differences approach. Specifically, we compare within-state *differences* in outcomes between schools on and away from the AYP failure margin to *differences* between similar schools within *other* states that are both far from the AYP failure margin. Our strategy bears resemblance to that used by Tyler, Murnane, and Willett (2000) to estimate the labor market

[2] These data are available for download at www.gsb.columbia.edu/nclb. Schools in these data are identified using the standard National Center for Education Statistics (NCES) ID number, which is easily linked to other datasets such as the SASS and ECLS data used in our analysis.

value of a General Educational Development (GED) certificate. Their work uses cross-sectional variation across states in the score needed to pass the GED exam, comparing differences between students within a state, one of which failed the GED exam, to differences between students with the same test scores in other states that both passed the exam. Our identification is also cross sectional, though for some outcomes we can control for prior levels and trends.

While Tyler, Murnane, and Willett (2000) take advantage of a single nationwide GED exam, NCLB exams are different in each state. We therefore evaluate schools' probabilities of failing AYP on a state-by-state basis, which we explain in detail below. Another important difference between our focus and that of Tyler, Murnane, and Willett (2000), as well as recent school accountability studies that use regression discontinuity methods, is that we are interested in the pressure faced by schools that were at risk of failure, rather than the impact of actually failing to make AYP.[3] While the impact of a failing designation is of interest, we regard the increase in accountability pressure more generally to be the more significant change induced by NCLB.

Our analysis takes advantage of nationally representative data from the Schools and Staffing Survey (SASS) and the Early Childhood Longitudinal Survey (ECLS), which, serendipitously, were collected on teachers and students exposed to the initial years of NCLB.[4] We find that accountability pressure from NCLB reduces teachers' perceptions of job security, especially among relatively inexperienced teachers. We also find evidence that reading and math specialist teachers work longer hours, generalist teachers (i.e., those teaching multiple subjects) work fewer hours, and all types of teachers shift time away from whole-class instruction. The topics of instruction sacrificed include science and social studies lessons.

We find short-term NCLB pressure has either positive or neutral effects on student achievement in math, reading, and science on low-stakes examinations. Students enrolled in schools near the AYP failure margin score more than 0.06 standard deviations higher in reading than comparable students in similar schools that were well above the margin for making AYP. Estimated effects for math and science achievement are also positive, though we cannot confidently reject the hypothesis of zero effects. We also find no evidence of differential achievement effects of NCLB pressure on low-stakes exams for students in particularly crucial subgroups or students with scores close to the passing threshold on their states' high-stakes examinations. In addition, achievement gains from short-term NCLB pressure do not come at the expense of students' reported enjoyment of learning or their anxiety over testing.

The paper proceeds as follows. Section I 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 II, and our estimated effects of NCLB on teachers and students in Section III. Section IV concludes by discussing how these results may inform current policy debates.

[3] Studies that focus on the effect of actually receiving a failing designation that use regression discontinuity methods include Rouse et al. (2007); Chiang (2009); Rockoff and Turner (2010); Hemelt (2011); Cooley Fruehwirth and Traczynski (2012); and Chakrabarti (2014).

[4] In addition, the SASS wave just prior to NCLB allows us to conduct placebo tests, while the ECLS is a panel dataset that allows us to control for students' levels and trends in achievement prior to NCLB.

I. Data and Descriptive Analysis

A. *Data Description*

To measure NCLB pressure nationwide, our analysis requires a comprehensive database of schools' NCLB-related performance metrics. 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 publicly available datasets.[5] We present the categories of data collected and their sources in the Appendix.

We also use school characteristics from the 2001–2002 Common Core of Data (CCD) compiled by the National Center for Education Statistics (NCES) and pre-NCLB aggregated student test performance variables from the National Longitudinal School-Level State Assessment Score Database (compiled by American Institutes for Research).[6] States' standardized tests are not measured on the same scale, and we standardize pre-NCLB school average test performance to have a mean of zero and standard deviation of one within each state. These test performance variables and school characteristics variables are used in our predictions of school-level NCLB pressure (Section II) and as control variables in our analysis of teacher- and student-level outcomes (Section III).

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

The fortunate timing of the SASS and ECLS data allow us to study NCLB on a national level, but it also limits our focus to the first two years of the implementation of NCLB. While the longer-run effects of NCLB are certainly of great interest, the impacts of the initial roll-out of NCLB allow for better identification because of the availability of test scores both before and after NCLB implementation. The consequences associated with continued failure to make AYP were escalating, but this escalation was explicit from the beginning, so that early failures should have been regarded as substantially increasing the risk of future consequences such as school closure.

[5] These two sources of NCLB-related data are the NLSLSAS Database and NAYPI Database. 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 since they are the final interface between schools and the public and should reflect final adjustments such as schools' appeals to AYP determinations.

[6] 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 the 1998–1999 CCD in lieu of data from the 2001–2002 CCD.

The SASS surveyed teachers in all 50 states and provides nationally representative samples with the use of sampling weights.[7] For consistency with our examination of student outcomes in the ECLS, we focus on regular, full-time teachers (i.e., omitting substitute teachers, teacher's aides, etc.) working in high-stakes grade levels in traditional public schools that served (at least five) fifth graders in the school year 2001–2002. This leaves roughly 3,000 teachers in our sample.[8] Panel A of Table 1 provides summary statistics for the outcome variables we create from SASS survey questions.

The ECLS followed students for nine years, collecting data in both the fall and the spring of the school years 1998–1999 and 1999–2000 (kindergarten and first grade), and in the spring of the school years 2001–2002, 2003–2004, and 2006–2007 (third grade, fifth grade, and eighth grade). The ECLS has the widest coverage and array of student-level outcomes of any nationally representative longitudinal dataset covering years before and after the passage of NCLB. The timing of the ECLS survey is fortuitous because this cohort was tested just prior to the first year of NCLB and again two years later.

The ECLS includes students from 40 relatively populous states and 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.[9] However, data collection procedures in later waves result in samples that are not necessarily representative of the student populations at each school, particularly due to procedures for tracking students making nonstructural school transfers.[10] In our analysis of ECLS data, we use data on roughly 6,870 students.[11] While we take advantage of sampling weights to make our estimates nationally representative, our main conclusions do not change if we remove child-level sampling weights or remove students who made nonstructural enrollment changes. Attrition can influence the interpretation of our results if students experience heterogeneous effects from

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

[8] Of the more than 40,000 public school teachers surveyed in the 2003–2004 wave of the SASS, roughly 39,000 of these were "regular, full-time" teachers, roughly 30,000 of these taught in traditional public schools that were open in 2001–2002 and have available NCLB outcomes data for 2003 and 2004, roughly 9,000 of these worked in traditional public schools serving at least five fifth grade students, and roughly 3,000 of these served high-stakes grades (with test results used for spring 2004 AYP determinations). In cases of teachers covering multiple grades, we include the teacher if more than half of the teacher's covered grade levels were tested for NCLB in that teacher's state during the spring of 2004. As a falsification test, we also examine outcomes from the prior wave of the SASS (1999–2000). Our sample sizes for this prior wave are slightly larger, partly because schools must have been in operation during the school year 2001–2002 to be included in our analysis while some of the schools in the 2003–2004 wave were new.

[9] It used a multistage probability sample design, first selecting broad geographic areas (e.g., a county), then schools within each area, and finally students within schools. On average, 23 kindergarteners were sampled in each school.

[10] The ECLS includes students who were retained within the same grade or skipped a grade level, but has some attrition. In the school year 1999–2000, a randomly selected 50 percent subsample of students who transferred from their original school was surveyed, and another random sample of first graders in the same schools where transfer students were followed was added. However, this "freshening" of the sample was not repeated in the third, fifth, and eighth grades, and the ECLS simply sampled 50 percent of students who transferred schools for nonstructural reasons (e.g., students who switched schools for reasons other than moving from a kindergarten–fourth grade school to a fifth–eighth grade school in the same district).

[11] All reported sample sizes are rounded to the nearest 10 due to restricted-use data reporting requirements. In a falsification test, we also examine test score growth from the fall to spring of kindergarten using a sample of 5,760 students.

Table 1—Summary Statistics for Dependent Variables

	Main sample			Falsification sample		
	Mean	SD	*N*	Mean	SD	*N*
Panel A. Teacher-level dependent variables from the SASS						
Concerned about job security due to student test performance[†]	38.0%		2,800	39.7%		3,200
Plan to teach until retirement[††]	76.0%		2,800	73.3%		3,200
Gave at least one science lesson last week	58.9%		3,370	57.0%		3,870
Gave at least one social studies lesson last week	62.0%		3,370	61.2%		3,870
Less experienced teachers (*<10 years*):						
Concerned about job security due to student test performance[†]	42.5%		1,270	45.7%		1,310
Plan to teach until retirement[††]	68.4%		1,270	64.3%		1,310
More experienced teachers (*at least 10 years*):						
Concerned about job security due to student test performance[†]	34.6%		1,520	35.5%		1,890
Plan to teach until retirement[††]	82.3%		1,520	79.6%		1,890
Generalists						
Work hours per typical week[†††]	52.8	9.3	2,300	49.0	7.8	2,710
Whole-class instructional hours per typical week[†††]	29.3	5.1	2,300		N/A	
Specialists						
Work hours per typical week[†††]	53.2	7.6	500	49.6	8.5	480
Whole-class instructional hours per typical week[†††]	29.5	5.3	500		N/A	
Panel B. Student-level dependent variables from the ECLS						
Fifth grade reading score (standardized)	0.008	0.97	6,870		N/A	
Fifth grade math score (standardized)	0.027	0.98	6,870			
Fifth grade science score (standardized)	0.065	0.96	6,870			
Enjoyment of reading (standardized)	−0.002	1.01	6,870			
Enjoyment of math (standardized)	0.037	1.01	6,870			
Has anxiety about standardized tests	42%		6,870			

Notes: Means and standard deviations using relevant sample weights provided by the SASS and ECLS to produce nationally representative estimates. Reported sample sizes are rounded to the nearest 10 to comply with nonpublic data use reporting requirements. The samples are restricted to observations used in the main analyses: teachers in 41 states for the SASS sample and students in 35 states in the ECLS sample. Sample sizes are larger for the two SASS dependent variables related to science or social studies instruction, because those models include teachers from high-stakes grades regardless of their subject areas. For reading, math, and science scores, the ECLS data report *t*-scores of students' IRT-based "theta scores," which are estimates of students' skill levels. These *t*-scores are already constructed so that the national (cross-sectional) mean equals 50 and the national standard deviation equals 10, so we simply subtract 50 from these scores and then divide by 10 to convert them to *Z*-scores. Standardized variables are *Z*-scores that were standardized based on the national, cross-sectional student distribution; their means and standard deviations above differ from zero and one respectively because some states/students are omitted due to missing data and because we use longitudinal sampling weights rather than cross-sectional sampling weights.

[†]This variable measures whether teachers responded that they "somewhat agree" or "strongly agree" with the statement: "I worry about the security of my job because of the performance of my students on state and/or local tests." The other two possible responses were "somewhat disagree" or "strongly disagree."

[††]This variable describes teachers' responses to the question "How long do you plan to remain in teaching?" We coded their response as planning to teach until retirement if they responded either "Until I am eligible for retirement" or "As long as I am able." The other possible responses were "Will probably continue unless something better comes along," "Definitely plan to leave teaching as soon as I can," or "Undecided at this time."

[†††]We set teachers' work-related hours and instructional hours to missing if reported instructional hours were 60 hours or greater, a suspiciously high level of 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."

the accountability pressure that NCLB placed on schools and if this heterogeneity is related to the probability of attrition, though the direction in which this pushes the local effects we can identify is unclear.

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

Panel B of Table 1 provides descriptive statistics for our ECLS outcome measures. Since most students in this wave of the ECLS were fifth graders, we limit the sample of students to those attending regular public schools that served (at least five) fifth graders in the school year 2001–2002. We standardize students' scores within subject and year so that the national mean score equals zero and the national standard deviation equals one. In addition to standardized exams, we examine students' reported enjoyment of math and reading, as well as reported anxiety over standardized tests.[12] Tables 2A and 2B provide descriptive statistics on control variables used in our regression analyses. We show statistics separately for our samples of public school teachers from the SASS and for public school students from the ECLS.

In addition to our analysis of SASS and ECLS data, we examine a set of survey responses from the Implementing Standards-Based Accountability (ISBA) study, conducted by the RAND Corporation (Hamilton et al. 2007; Stecher et al. 2008). 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 a substantial number of principals and teachers were surveyed in each state, nonrandom participation means that these are unlikely to be representative samples. These data are not public, but 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 III.

B. *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

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

Table 2A—Descriptive Statistics for Teacher Control Variables, SASS Sample

Teacher characteristics	Mean	SD
Age	41.4	10
Total years of experience	13.1	9.6
Total years of experience at same school	6.9	7.5
Female	87%	
White, non-Hispanic	78%	
Black	10%	
Hispanic	9%	
Has full certification	88%	
Has Master's in education	41%	
Has Master's in other field	3%	
Completed an undergraduate certification program	46%	
Teaches grades four or five	61%	
Teaches grades six or higher	8%	

Table 2B—Descriptive Statistics for Control Variables, ECLS Sample

Variable	Mean	SD
Family characteristics		
Two parent household	67%	
Mother's education level unknown	9%	
Mother has at least a high school diploma	89%	
Mother possesses a BA	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		
Reading *Z*-score in spring 2000	0.017	0.950
Math *Z*-score in spring 2000	0.027	0.920
Reading *Z*-score in spring 2002	—	0.981
Math *Z*-score in spring 2002	0.028	0.971
Reading enjoyment *Z*-score in spring 2002	—	1.031
Math enjoyment *Z*-score in spring 2002	0.030	1.018
African American	18%	
Hispanic	20%	
Asian	3%	
Other	5%	
Female	48%	
Date of birth (measured in days)	3/18/93	140

test participation cutoff of 95 percent. Secondary schools must also meet thresholds for graduation rates, and primary schools must also perform sufficiently well on a state-selected "additional indicator," which is typically the attendance rate. Beyond this set of parameters, states have a great deal of flexibility in setting a number of other rules and regulations. Table 3 lists ten important factors states must determine, but even this multitude of choices does not fully capture all the minutia of NCLB rulemaking. For example, while most states consider the performance of five ethnic subgroups (Asian/Pacific Islander, black, Hispanic, Native American, and white) in their AYP determinations, California and Alaska added additional subgroups

Table 3—Ten Important Rules and Regulations Chosen by States under NCLB

1. Selection of standardized tests in math, reading, and (since 2007–2008) science
2. Selection of which grade levels to test (until 2005–2006)[a]
3. Establishment of proficiency rate thresholds, i.e., the percent of students that must score proficient or higher, which apply to the whole school as well as individual subgroups
4. Determination of whether to calculate proficiency rates using students across all tested grade levels within each school or within each tested grade level[b]
5. Determination of whether to calculate subgroup proficiency rates using multiple test years
6. Definition of continuous enrollment, where only continuously enrolled students count towards calculation of subgroup size as well as test participation and proficiency rates
7. Selection of 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
8. Determination of the size of confidence intervals applied to student subgroups' raw proficiency rates; larger intervals effectively lower proficiency rate needed to make AYP
9. Determination of 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
10. Design the process by which schools can appeal their AYP status from the state

[a] 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. Beginning in 2005–2006 states had to assess the math and reading proficiency of all third through eighth graders and at least one level for grades 10 to 12.

[b] Arizona, Colorado, Maine, New York, New Jersey, Rhode Island, Tennessee, and Washington disaggregate subgroup size and subgroup results to the grade or grade span level. All other states determine subgroup size using students across all tested grades within a school.

(Filipino and Alaskan Native, respectively) while Asian/Pacific Islander is not an AYP subgroup in Texas.[13]

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

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 twentieth percentile of baseline (spring 2002) school performance. More importantly, as shown in Figure 1, there is little relationship between the fraction of schools failing to make AYP in a state and the state's average student achievement as measured on the National Assessment of Educational Progress (NAEP), a federal exam that has been administered to nationally representative samples of students in grades four and eight for decades.[14] States with the highest NAEP proficiency rates have slightly lower AYP failure rates than other states, but

[13] Additional analysis of the vagaries of states' NCLB rules, including several illustrative case studies, can be found in Davidson et al. (2013).

[14] Note that we plot AYP failure rates for schools serving fifth grade students, which is the type of schools we analyze using the SASS and ECLS data. Across states, the correlation between elementary and high school AYP failure rates was about 0.7.

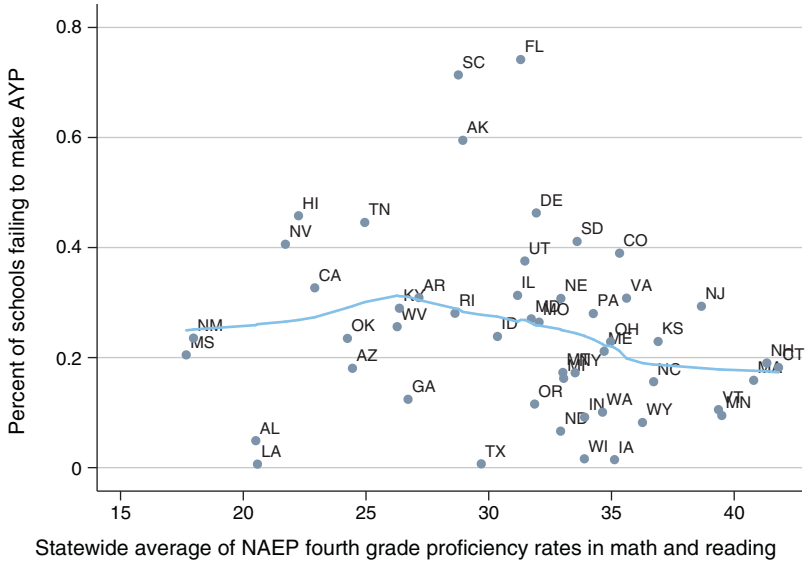

FIGURE 1. AYP FAILURE RATES VERSUS NAEP PROFICIENCY RATES BY STATE, 2003

Notes: Line represents a locally weighted regression. Failure rates are based on schools serving at least five fifth grade students.

this relationship is not statistically significant and NAEP scores explain almost none of the cross-state variation in AYP failure rates.

We have been unable to find any single aspect of NCLB design that can explain the wide variation in failure rates. However, by testing a number of factors we have come to the conclusion that the interaction of four features significantly influences the fraction of schools failing AYP: (i) state rules specifying how large subgroups must be to count towards AYP; (ii) diversity of student populations within schools, which influences how many student subgroups per school are accountable; (iii) the generosity of the state's confidence intervals; and (iv) the generosity of the state's safe harbor provisions. Differences in the leniency of various NCLB requirements across states allow us to identify the impact of accountability pressure.

II. Predicting the Probability of Failing AYP

In the first stage of our analysis, we use our assembled data on NCLB-related inputs and outcomes, along with data from the Common Core on school-level demographics (listed in Tables 2A and 2B) and test performance variables from the school year 2001–2002, after the passage of NCLB but prior to the first AYP determinations. [15] Our goal is to determine which student subgroups and, by extension, which schools were on the margin of failing to make AYP in the first two years that NCLB

[15] 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

was in effect. We begin by estimating state- and subject-specific probit regressions to generate predictions of the likelihood that each numerically significant student subgroup would pass AYP proficiency targets in the spring of both 2003 and 2004.

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

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

$$(1) \qquad \mathbf{AYP}_{jks03\text{–}04} = \begin{cases} 1 & \textit{if} \quad \alpha_q + \mathbf{X}_{jks02}\boldsymbol{\beta}_1 + \mathbf{N}_{jks04}\boldsymbol{\beta}_2 + \mathbf{XN}_{jks}\boldsymbol{\beta}_3 + \mathbf{W}_{j02}\boldsymbol{\beta}_4 \\ & \qquad + \mathbf{M}_{jks03\text{–}04}\boldsymbol{\beta}_5 + \zeta_{jks} > 0 \\ 0 & \textit{otherwise}, \end{cases}$$

where $\mathbf{AYP}_{jks03\text{–}04}$ denotes whether subgroup k at school j met its AYP proficiency rate targets in 2003 and 2004 in subject s. $\mathbf{X}_{jks02}$ is a vector containing a cubic polynomial for subgroup performance on statewide exams during the school year 2001–2002,[16] $\mathbf{N}_{jks04}$ is a vector of student subgroup size variables in 2004, $\mathbf{XN}_{jks}$ represents interactions of test score and subgroup size variables, $\mathbf{W}_{j02}$ is a vector of control variables for school-level demographics from the school year 2001–2002 (listed in Tables 2A and 2B), $\mathbf{M}_{jks03\text{–}04}$ is a vector of indicators for whether the subgroup was numerically significant in only 2002–2003 or only 2003–2004 (but not both), and ζ_{jks} is a normally distributed latent disturbance term. The subgroup size variables ($\mathbf{N}_{jks04}$) and interactions with test score measures ($\mathbf{XN}_{jks}$) are included to account for states' confidence interval adjustments and the mechanical decrease in the error variance of student pass rates as the number of tested students within subgroup k increases. In particular, the $\mathbf{N}_{jks04}$ vector contains cubic terms for the inverse of the square root of the number of accountable test-taking students in subject s in subgroup k in school j during the school year. The Appendix provides detailed descriptions of

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.

[16] 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 four or grade six; full details are provided in the Appendix. 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.

each predictor and its data source. We exclude subgroups from our estimation if they were too small to be accountable under AYP in *both* 2003 and 2004, since none of these subgroups would have failed state proficiency targets.

We restrict our sample to schools that were (i) operational from at least 2001–2002 through 2003–2004; (ii) neither technical/vocational nor only for special education students according to the school classifications in the Common Core of Data; and (iii) enrolled at least five students in the fifth grade as of the school year 2001–2002.[17] We are forced to omit nine states from the SASS sample and five states from the ECLS sample due to missing data (e.g., 2002 test scores or a state's AYP determinations for subgroups were missing). Our numerous attempts at gathering these data from state departments of education have either been unsuccessful or, in most cases, states claim that the data simply do not exist or are too unreliable to release. Fortunately, these states have relatively small populations; more than 92 percent of the US population resides in one of the 41 states with sufficient data for our analyses.

A. *Defining the AYP Margin*

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

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

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

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

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

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

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

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

In these 41 states, we classify 65.7 percent of schools above the AYP margin, 23.8 percent on the AYP margin, and 10.5 percent below the AYP margin. The actual rates with which schools made AYP in both 2003 and 2004 were 85.9 percent for schools above the margin, 37.0 percent for schools on the margin, and 6.8 percent for schools below the margin, demonstrating that our specification has sufficient power to identify substantial variation in which schools were at risk of failing to make AYP. However, our analyses below are predicated on the idea that the risks of AYP failure were foreseeable to school administrators and teachers. To the extent that measurement error causes us to misclassify which schools *believed* they were on the AYP margin, our estimated effects of NCLB pressure may be biased towards zero. This possibility motivates the need to examine whether our estimates are related to teachers' and administrators' reported sense of accountability pressure, which we do below.

Table 4 provides additional descriptive statistics for our measures of NCLB pressure. With the exception of white and economically disadvantaged students, most student subgroups were typically *not* numerically significant and thus did *not* count towards AYP. For example, more than 68 percent of schools did not have a sufficient number of disabled (special education) students in either 2003 or 2004 to be held accountable for that group's performance. This rate varied across states depending on minimum subgroup size requirements, again underscoring the importance of AYP formulas. For example, disabled subgroups were accountable under NCLB in either 2003 or 2004 in just 7 percent of Arizona schools, compared with 61 percent in Massachusetts.

Among subgroups that were numerically significant and thus accountable, the fraction we predict to have a moderate or low chance of making AYP varies considerably. The subgroups most frequently predicted to have a *moderate* chance of passing in reading were disabled and limited English proficient (each about 38 percent), and those predicted to have a moderate chance in math were disabled and black (33 and 28 percent, respectively). Disabled student subgroups also have relatively high fractions predicted to have *low* chances of passing proficiency targets (16 percent for math, 18 percent for reading). In contrast, white subgroups are nearly always predicted to have a high chance of passing proficiency targets.

B. *Variation in Predicted NCLB Pressure across States*

Similar schools faced different levels of NCLB pressure because of the state in which they were located, but 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 these points, we match each school with similar out-of-state schools to simulate the counterfactual NCLB pressure in math and reading that schools would face had they been located in other states. We find the most

Table 4—Predictions of AYP Outcomes for Subgroups

	Numerically significant subgroup	Conditional on numerical significance			
		Predicted moderate chance		Predicted low chance	
		Math	Reading	Math	Reading
Overall school population	92.6%	7.8%	9.5%	2.2%	2.6%
Actually made AYP in subject in '03 and '04		52.0%	52.8%	10.8%	8.5%
Economically disadvantaged	61.5%	15.6%	19.4%	3.7%	4.7%
Actually made AYP in subject in '03 and '04		54.3%	54.2%	12.6%	13.0%
Limited English proficient	23.6%	19.4%	37.6%	4.9%	10.6%
Actually made AYP in subject in '03 and '04		58.1%	49.8%	14.2%	19.4%
Disabled	31.5%	32.9%	38.0%	15.9%	17.9%
Actually made AYP in subject in '03 and '04		51.4%	52.1%	14.9%	12.8%
White	71.6%	1.4%	1.0%	0.1%	0.0%
Actually made AYP in subject in '03 and '04		56.0%	62.6%	15.8%	30.0%
Black	33.7%	28.2%	25.9%	10.0%	8.0%
Actually made AYP in subject in '03 and '04		51.7%	53.9%	16.8%	14.5%
Hispanic	29.9%	12.2%	20.9%	1.4%	2.7%
Actually made AYP in subject in '03 and '04		56.1%	54.9%	12.2%	15.9%
Asian/Pacific Islander/Filipino	11.2%	6.1%	11.5%	7.0%	9.5%
Actually made AYP in subject in '03 and '04		39.4%	46.9%	42.3%	11.2%
Native American/Alaskan Native	4.8%	12.5%	13.9%	10.8%	10.0%
Actually made AYP in subject in '03 and '04		53.5%	46.3%	6.8%	10.2%

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

similar school in each of the other states in our 41 state sample based on an index of school characteristics, which includes student math and reading test scores from the spring of 2002, the number of total enrolled students in that year, the percent of students who are from low-income households, and the racial composition of students.[18] We then calculate the fraction of matched schools that are on the AYP margin and the fraction below the AYP margin.

Substantial cross-state variation in NCLB rules creates substantial cross-state variation in AYP pressure. For schools below, on, and above the margin in their actual state, the percentage of matched schools *below* the margin are 40 percent, 21 percent, and 7 percent, respectively. The percentage of matched schools *on* the margin are 35 percent, 32 percent, and 20 percent, respectively, and the percentage of matched schools *above* the margin are 26 percent, 46 percent, and 74 percent, respectively. Thus, while there is some correlation in AYP outcomes for schools with similar characteristics, this correspondence is quite weak, leaving us with considerable identifying variation. Figure 2, panels A and B illustrate how matched schools' AYP pressure

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

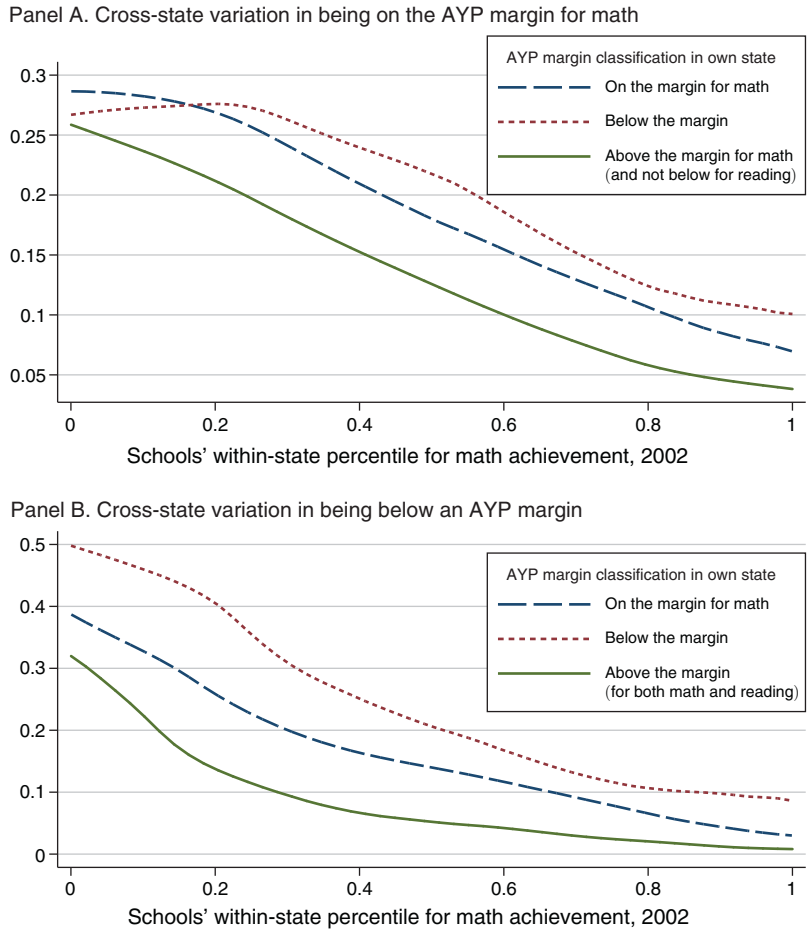

FIGURE 2

Notes: To produce these figures, we used baseline school characteristics to match every school in the country with the most similar school in each of the other states (see footnote 18 in the text). We then found the fraction of these paired schools that were either on or below the AYP margins based on their own states' rules. The downward slope of the curves in these figures implies that higher-achieving schools are generally less likely to face accountability pressure than lower-achieving schools. Yet the curves' substantial distances from 0 percent and from 100 percent on the *y*-axis imply that schools on the AYP margin would often have instead been above or below the AYP margin had they been subject to other states' rules. For example, if a school is on the AYP margin for math and has baseline student math test performance at the sixtieth percentile of schools in its state, then about 15 percent of its paired schools were on the AYP margin for Math in *their* states (panel A) and about 10 percent of its paired schools were below an AYP margin in *their* states (panel B). Based on these rates, we estimate that this type of school would actually have been above the AYP margin if it were subject to NCLB rules similar to those found in one of the other 75 percent of states.

varied with a school's own prior level of math achievement (reading results are quite similar) and AYP pressure. As expected, schools with lower prior achievement have peer schools that are more likely to be on or below the AYP margin. Yet the figures also reveal that schools frequently would face different pressure in other states than they do in their own state. For example, for schools we classify below the AYP margin in their own state and whose math scores are in the bottom decile, not even half of

the matches are classified below the AYP margin in their states. Moreover, less than a third of these matches are classified as on the AYP margin in their states, meaning that roughly a quarter of these matched schools—with very similar demographics and equally low test scores—are classified as above the AYP margin.

To isolate variation in schools' treatment status that depends on how their *own* states' rules differ from those in other states, we include these out-of-state counterfactual accountability pressure measures as control variables in our analysis below. However, their inclusion ultimately has very little impact on our findings, because their correlations with outcomes of interest are mostly captured by our other control variables such as schools' relative performance on statewide examinations prior to NCLB. Still, controlling for these counterfactual accountability pressure measures increases our confidence that identification is not based on any general tendency of schools with low-achieving or low socioeconomic status students to face more NCLB pressure.

Another potential concern with our approach is the endogeneity of accountability policy; states with tougher AYP standards may have also adopted other school accountability policies at the same time. To the extent that these policies affect all schools, they will be picked up by state fixed effects in our analysis, but it is reasonable to imagine such parallel policies were designed to improve outcomes in low-achieving schools. To test this possibility, we examine whether our estimates are affected by the addition of interactions between the state AYP failure rate and schools' pre-NCLB achievement. Controlling for these interactions changes our point estimates slightly but does not change our main findings.[19] Below, we also examine whether NCLB pressure has different impacts in states that had a test-based accountability policy prior to NCLB.

C. *Assessing Our Measure of NCLB Pressure in the ISBA Surveys*

To get an initial sense of the validity of our measures of NCLB pressure, we examine aggregate statistics from surveys of principals and math teachers in California, Pennsylvania, and Georgia by the RAND Corporation in the school year 2003–2004 (Hamilton et al. 2007; Stecher et al. 2008). While these cross-tabulations are only suggestive—the micro data from these surveys are not publicly available—these data are unique in that principals and teachers were asked specifically about NCLB pressure.

We examine principals' survey responses in 21 schools that we classified as on the AYP margin and 104 schools above the AYP margin; no principals were surveyed at any school that we predicted to be below the margin of making AYP. Among principals working in schools above the AYP margin, 96 percent felt they would make AYP in the school year 2003–2004. Only 71 percent of principals felt the same who worked in schools on the AYP margin. Indeed, among principals in schools above the AYP margin, 72 percent felt they would make AYP for *the next five years*, relative to only 48 percent in the marginal group (Table 5, panel A). Principals in schools on the AYP margin were also between 9 and 14 percentage points more likely to say that they had taken the following actions: encouraged teachers to focus

[19] For example, the estimated effect of a school being on the margin of AYP failure increases from 0.062 to 0.065 for reading test scores and falls from 0.054 to 0.052 for science test scores.

TABLE 5—EVIDENCE ON NCLB PRESSURE FROM THE ISBA SURVEY IN CALIFORNIA, GEORGIA, AND PENNSYLVANIA

	Above AYP margin ($N = 104$)	On AYP margin ($N = 21$)	
Panel A. Principals			
Do you agree with the following statement:			
My school can attain the AYP targets for 2003–2004	96.1%	71.4%	
My school can attain the AYP targets for the next five years	71.6%	47.6%	
Has your school and/or district done any of the following:			
Encouraged or required teachers to spend more time on tested subjects and less time on other subjects	49.0%	61.9%	
Distributed commercial test preparation materials	67.0%	81.0%	
Distributed released copies of the state test or test items	76.9%	85.7%	
	Above AYP margin ($N = 1{,}074$)	On AYP margin ($N = 224$)	Below AYP margin ($N = 19$)
Panel B. Math teachers			
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: 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 II of the paper for details. No principal surveyed was in a school classified by our analysis as below the AYP margin. All of the differences in rates between the groups above the AYP margin and either of the other two groups are statistically significant at approximately the 0.01 level or better. Differences in rates between teachers in schools above the AYP margin and those in schools on the AYP margin are statistically significant at the 0.05 level for "I focus more effort on students who are close to proficient," and at the 0.01 level for "I look for particular styles…" and "I search for more effective teaching methods."

more time on tested subjects; distributed commercial test preparation materials; or distributed copies of previous state tests or test items. All of these differences in responses across principals in the two groups are statistically significant at approximately the one percent level.

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

III. Estimates of the Impact of Accountability Pressure under NCLB

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

$$(2) \qquad \mathbf{Y}_{ij} = \sum_{q} \delta_q \mathbf{D}_j^q + \mathbf{Q}_{ij}\boldsymbol{\rho}_1 + \mathbf{W}_{j02}\boldsymbol{\rho}_2 + \mathbf{X}_{j02}\boldsymbol{\rho}_3 + \lambda \mathbf{M}_j^{AYP} + \gamma \mathbf{B}_j^{AYP} + \boldsymbol{\rho}_4 \%\mathbf{M}_j^{AYP} + \boldsymbol{\rho}_5 \%\mathbf{B}_j^{AYP} + \zeta_{ij}.$$

$\mathbf{Y}_{ij}$ is an outcome of interest for an individual teacher or student i in school j, δ_q is a state fixed effect while $\mathbf{D}_j^q$ is an indicator for school j located in state q, $\mathbf{Q}_{ij}$ is a vector of (student- or teacher-level) control variables, and $\mathbf{W}_{j02}$ is a vector of school-level control variables (as in equation (1)). The $\mathbf{X}_{j02}$ vector is analogous to the $\mathbf{X}_{jks02}$ vector in equation (1), with schoolwide student achievement measures replacing subgroup-specific achievement measures. $\mathbf{M}_j^{AYP}$ and $\mathbf{B}_j^{AYP}$ are indicators for schools on or below the AYP margin, respectively, and $\%\mathbf{M}_j^{AYP}$ and $\%\mathbf{B}_j^{AYP}$ are controls for the simulated fraction of states where the school would be on or below the AYP margin.

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

A. *Impacts on Teachers*

We examine the effect of NCLB pressure on teachers using the SASS data. The $\mathbf{Q}_{ij}$ vector includes the teacher-level control variables listed in Table 2A and Appendix Table A3, with both linear and squared terms for teachers' age, years of teaching experience, and years of experience teaching at the same school. We present results for our main sample from the 2003–2004 SASS in the top panels of Tables 6 and 7. To help address the concern that differences in outcomes for the kinds of schools that faced NCLB pressure are driven by (unobservable) school characteristics and not accountability pressure, we also compare our estimates with a falsification sample based on the 1999–2000 wave of the Schools and Staffing Survey, which predates NCLB.

We first examine teachers' views on job security and satisfaction, limiting our sample to those teaching in high-stakes grades and subjects tested under NCLB and

using linear probability models.[20] The first column of Table 6 displays estimated effects of NCLB pressure on whether teachers agreed with the statement: "I worry about the security of my job because of the performance of my students on state and/or local tests." Compared to teachers of high-stakes grades/subjects at schools above the AYP margin, those in schools on the AYP margin or below the AYP margin are, respectively, 3.8 percentage points and 9.7 percentage points more likely to report concern over their job security related to student test performance. The latter difference is of moderate statistical significance ($p = 0.11$) but fairly large considering that 38 percent of teachers reported this concern overall. Point estimates for the pre-NCLB SASS sample are of the opposite sign, supporting the notion that our measure of NCLB pressure is valid and captures significant variation in school staff members' perceptions of pressure.

Less-experienced teachers should be most sensitive to issues of job security, given the prevalence of seniority-based layoff and transfer rules (see Boyd et al. 2011). We therefore estimate regressions separately based on whether a teacher had more or less than ten years of experience—roughly the sample median. The effects of accountability on perceived job security are much stronger for teachers with fewer than ten years of experience; relatively inexperienced teachers in schools on the AYP margin or below the AYP margin are, respectively, 9.8 percentage points and 19.0 percentage points more likely to report concern over their job security related to student test performance (Table 6, column 2). These effects are statistically significant at the 0.10 and 0.05 levels respectively; they are also significantly different from the falsification estimates based on the pre-NCLB SASS sample.

Roughly three quarters of teachers in our sample indicate that they plan to (or hope to) teach until retirement (see Table 1), which we view as a measure of job satisfaction. We find that teachers in schools on or below the AYP margin are less likely to plan to teach until retirement than their counterparts teaching in schools with high probabilities of making AYP (Table 6, column 4), and these effects are especially strong for veteran teachers. Again, estimates from the falsification estimates based on the pre-NCLB sample wave are of the opposite sign and significantly different from the main estimates, supporting the idea that NCLB pressure had an impact on teachers' career plans. These results comport with findings by Feng, Figlio, and Sass (2010) that schools in Florida that received low ratings under the state's accountability system subsequently experienced higher rates of teacher turnover.

While the SASS does not ask about plans to teach next year, we are able to examine short-run turnover for approximately 540 teachers from the NCES Teacher Follow-up Survey, which tracks a subsample of teachers from the prior year's SASS wave. Using the same specification as above, we estimated a linear probability model predicting whether teachers applied to nonteaching jobs or left the profession the next year. We estimate coefficients of 0.15 and 0.10 for being on and below the

[20] One concern in this analysis is that we limit the sample to high-stakes teachers, and principals at schools facing NCLB pressure might strategically place teachers into high-stakes grades and subjects. While we cannot test this hypothesis, we believe such behavior would most likely create a bias against our findings. If principals facing strong pressure wish to boost high-stakes test performance, then they should assign their most talented teachers to high-stakes areas. Yet we find these teachers are more concerned about their job security, are less likely to plan to teach until retirement, and work fewer hours.

Table 6—Effects of NCLB Pressure on Teacher Attitudes

Dependent variable:	Concerned about job security due to student test score performance			Plan to teach until retirement		
Teachers:	All (1)	Less experienced (<10 years) (2)	More experienced (>= 10 years) (3)	All (4)	Less experienced (<10 years) (5)	More experienced (>= 10 years) (6)
Panel A. Main sample: NCLB sample wave						
On AYP margin	0.038	0.098*	−0.016	−0.064**	−0.011	−0.085**
	(0.038)	(0.059)	(0.050)	(0.032)	(0.056)	(0.036)
Below AYP margin	0.097	0.190**	−0.009	−0.153***	−0.112	−0.151**
	(0.054)	(0.076)	(0.075)	(0.048)	(0.084)	(0.059)
Panel B. Falsification sample: Pre-NCLB sample wave						
On AYP margin	−0.010	−0.064	0.049	0.036	0.007	0.050
	(0.041)	(0.066)	(0.051)	(0.035)	(0.063)	(0.039)
Below AYP margin	−0.020	−0.019	−0.003	0.065	0.062	0.066
	(0.058)	(0.091)	(0.071)	(0.051)	(0.090)	(0.061)
Panel C. Differences between actual and falsification estimates						
On AYP margin	0.048	0.162*	−0.065	−0.100**	−0.018	−0.135***
	(0.056)	(0.089)	(0.071)	(0.047)	(0.084)	(0.053)
Below AYP margin	0.117	0.208*	−0.005	−0.217***	−0.173	−0.217***
	(0.079)	(0.119)	(0.103)	(0.074)	(0.123)	(0.085)
N, main sample	2,800	1,270	1,520	2,800	1,270	1,520
N, falsification sample	3,200	1,310	1,890	3,200	1,310	1,890

Notes: The main sample includes teachers sampled in the 2003–2004 wave of the Schools and Staffing Survey (SASS) who were working in high-stakes grades/subjects. The falsification samples includes teachers sampled in the 1999–2000 wave of the SASS who were working in grades/subjects that later became high stakes for NCLB. For the falsification regressions, schools are classified as "On AYP Margin" or "Below AYP Margin" if they later had that status during the main sample period. To facilitate comparisons with the ECLS analysis in the remainder of the paper, we limit both samples to teachers in public schools that serve fifth grade students. All models control for the independent variables listed in Table 2A and Appendix Table A3, and also control for state fixed effects, a squared term for the number of Limited English proficient students in the grade, a squared term for the teacher's years of experience, and squared and cubic terms for schools' within-state standardized 2001–2002 test score performance in both math and reading. All models use the SASS cross-sectional sample weights to make the estimates nationally representative. Bootstrapped standard errors, adjusted for school-level clustering using 1,000 Monte Carlo simulations of both the first-stage and second-stage models, are displayed in parentheses below each estimate. Sample sizes are rounded to the nearest 10 to comply with restricted-data reporting requirements.

***Significant at the 1 percent level.

**Significant at the 5 percent level.

*Significant at the 10 percent level.

AYP margin, respectively, with the former coefficient statistically significant at the five percent level. This provides additional support to the notion that NCLB pressure affected teachers' job security and job satisfaction.

Table 7 presents results concerning how NCLB pressure affects teachers' self-reported total weekly work hours, whole-class instructional hours, and coverage of science and social studies.[21] Because teachers were surveyed in the fall, well ahead of NCLB spring testing, their responses should reflect general shifts in instruction rather than last-minute preparation for high-stakes tests. We examine

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

TABLE 7—EFFECTS OF NCLB PRESSURE ON TEACHERS' WORK HOURS AND INSTRUCTION

Dependent variable:	Work hours in a typical week		Whole-class instructional hours in a typical week		Taught at least one lesson during the last week in … Science	Taught at least one lesson during the last week in … Social studies
Teachers:	Generalists	Specialists	Generalists	Specialists	All	All
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A. Main sample: NCLB sample wave						
On AYP margin	−1.64**	4.07**	−0.90**	−1.39	−0.061*	−0.034
	(0.86)	(1.83)	(0.43)	(1.12)	(0.037)	(0.035)
Below AYP margin	−2.64**	4.32	−1.63**	−2.11	−0.150***	−0.147***
	(1.35)	(2.76)	(0.74)	(1.63)	(0.054)	(0.049)
Panel B. Falsification sample: Pre-NCLB sample wave						
On AYP margin	0.352	1.35	N/A	N/A	0.006	0.022
	(0.677)	(2.24)			(0.036)	(0.034)
Below AYP margin	0.850	0.421			−0.034	−0.005
	(0.876)	(3.09)			(0.048)	(0.051)
Panel C. Differences between actual and falsification estimates						
On AYP margin	−1.99*	2.72			−0.068	−0.056
	(1.09)	(2.89)			(0.052)	(0.049)
Below AYP margin	−3.49**	3.89			−0.116	−0.142**
	(1.61)	(4.14)			(0.073)	(0.071)
N, main sample	2,300	500	2,300	500	3,370	3,370
N, falsification sample	2,710	480	N/A	N/A	3,870	3,870

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

***Significant at the 1 percent level.

**Significant at the 5 percent level.

*Significant at the 10 percent level.

whole-class instructional hours in addition to total work hours to assess whether teachers engage in more targeted instruction. Weekly reported whole-class instructional hours, which average 29 hours, are a subset of total reported weekly work hours (including preparatory time in school and work done at home), which average 53. Most teachers in these elementary and middle schools are generalists, meaning they cover multiple subjects and teach in self-contained classrooms where the same students remain for most of the day. Other teachers are specialists, such as math instructors who see several different groups of students during the same day. As of the school year 2007–2008, almost two-thirds (63 percent) of all US public schools had staff with a specialist or coaching assignment (ED 2007–2008). Of these schools, 80 percent have reading specialists and 32 percent have math specialists. Given the different roles of these teachers, we divide our sample between generalist or specialist teachers when examining work time.

We find that greater accountability pressure is associated with a decrease in work hours among generalists but an increase among specialists, while whole-class instructional hours fall for both types of teachers. More than half of the generalists' decrease in work hours is accounted for by fewer hours per week devoted to

whole-class instruction. We cannot estimate impacts on instructional hours for the pre-NCLB wave of the SASS because this question was not asked, but point estimates for changes in total work hours in the pre-NCLB sample are insignificant, small, and significantly different from the main estimates for generalist teachers. Thus, in schools facing NCLB pressure, the overall trend is to have generalist teachers working fewer hours while their specialist colleagues work longer hours on activities other than whole-class instruction. We can only speculate on the mechanisms for these time reallocations; one possibility is that specialists spend more time working with small groups of students (for example by pushing into the classroom or pulling students out) and teachers spend more time on test preparation, student assessments, and tutoring. Regular teachers may tend to work slightly fewer hours due to the increased role of specialists and other factors, such as decreased teacher autonomy. There is insufficient research to indicate whether there are relative advantages of specialists versus generalists on student achievement, but specialists' content knowledge might help to boost student achievement. The literature on effective math instruction has underscored the importance of both content knowledge and pedagogical knowledge (Hill, Rowan, and Ball 2005).

Finally, we use the SASS data to explore shifts in instructional time across subject areas, using self-reports of whether the teacher taught a science lesson or a social studies lesson during the prior week. Unlike the estimates discussed above, we now include teachers in the sample regardless of whether they taught a high-stakes subject, which, due to the random sampling of teachers in the SASS, allows us to capture both shifts in the subject composition taught by generalists and shifts in the employment of specialists teaching low-stakes subjects.[22] The estimates suggest that schools facing accountability pressure reduce the frequency of science and social studies lessons (Table 7, columns 5 and 6). Compared to teachers at schools above the margin, teachers are 6.1 percentage points less likely to have taught a science lesson in the last week and 3.4 percentage points less likely to have taught a social studies lesson. The former estimate is statistically significant at the 0.10 level but the latter estimate is not statistically significant. The effects on science and social studies offerings in schools below the AYP margin are even larger: a roughly 15 percentage point reduction in the likelihood of either type of lesson. These effects are substantial, considering that 59 percent and 62 percent of teachers in this sample taught science and social studies lessons, respectively, in the previous week. Schools well under the margin of passing AYP in the short run may be shifting greater amounts of time to tested subjects in order to be able to reach this threshold several years in the future. Our falsification results based on the pre-NCLB produce very small and statistically insignificant estimates for these models, suggesting that our main estimates are capturing responses to NCLB pressure. Schools may shift instructional time from science and social studies to reading and math in order to increase their chances of making AYP.

[22] The results are similar if we instead limit the sample to generalist teachers: an estimated coefficient of being on the AYP margin of −0.09 (0.04 bootstrapped standard error) for science lessons and −0.05 (0.04 bootstrapped standard error) for social studies. This suggests that schools do not ramp up their use of social studies and science specialist teachers to compensate for the lower frequency of social studies and science lessons taught by generalists.

B. *Impacts on Students*

Using the ECLS data, we now turn to whether NCLB pressure affected student achievement, enjoyment of material, and test anxiety. Recall that the ECLS sampled geographic areas, then schools, then kindergarten students within schools, and then follows students if they transfer; our sample of roughly 7,000 students are spread among approximately 1,450 schools. Our specification is again shown by equation (2), and student-level controls are listed in Table 2B and Appendix Table A3. Importantly, we include third-degree polynomials of the student's prior math and reading performance, measured in the first and third grade waves of the ECLS. Thus, in addition to controls at the school level, our identification comes only from comparisons of students with very similar prior learning trajectories. In these regressions, we focus on the AYP margin for the most relevant subject(s): math for math test performance or enjoyment, reading for reading test performance or enjoyment, and "either math or reading" for science test performance and for anxiety about standardized tests. We lack power to separate relevant-subject and cross-subject effects using the ECLS.

Our estimates, displayed in Table 8, reveal that NCLB pressure has either neutral or positive effects on student achievement growth in both low- and high-stakes subjects. When schools are on the AYP margin and thus have strong short-term incentives to raise high-stakes test performance, their students perform better on low-stakes reading tests and perform at least as well on low-stakes math and science tests. Students' reading scores are 0.062 of a standard deviation greater on average when schools are on the AYP margin. This estimate is statistically significant at the 0.10 level and its bootstrapped 90 percent confidence interval ranges from 0.025 to 0.138. This estimate is meaningfully large since previous estimates of the impact of accountability pressure on *high-stakes* tests are typically between 0.1 and 0.2 standard deviations (e.g., Rouse et al. 2007; Rockoff and Turner 2010).[23] Although we are examining results for multiple dependent variables, a power test suggests that these three estimated effects for schools on the AYP margin are too large to likely have occurred by chance.[24] Students also perform no worse on low-stakes exams when their schools are below the AYP margin rather than above this margin, suggesting the instructional shifts observed in Table 7 may not be harmful for general learning, at least not in the short term.

We conducted falsification tests (not reported here) using ECLS data to examine the effects of NCLB pressure on students' achievement growth between the fall and spring of kindergarten. We used spring kindergarten *Z*-scores for students' math or reading performance as the dependent variable, changed the prior student-level performance controls to third-order polynomial measures of fall kindergarten *Z*-scores, and included the other controls from our main specification. The coefficient on

[23] Smaller effects of accountability pressure on low-stakes exams are also in line with Corcoran, Jennings, and Beveridge's (2011) findings that teacher effects tend to be smaller for low-stakes exams than for high-stakes exams.

[24] To test the joint significance of these test score estimates, we simulated estimations of these three models after randomly reassigning schools to different AYP status. Only 0.5 percent (5 out of 1,000) of these simulations produced a set of counterfactual estimates that had values, from greatest to least, greater than the values of the actual highest, second highest, and third highest estimate reported in the first three columns of the first row of Table 8. Only 6.7 percent of these simulations produced *any* estimates that were greater than the actual estimated effect for reading test scores in the first row of panel I of Table 8.

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

	Reading score	Math score	Science score	Enjoyment of reading	Enjoyment of math	Anxious about standardized tests
On the AYP margin	0.062*	0.032	0.054	0.031	0.112	−0.090***
	(0.035)	(0.038)	(0.034)	(0.070)	(0.077)	(0.037)
Below the AYP Margin						
	0.022	−0.005	0.032	0.052	−0.027	−0.159***
	(0.054)	(0.049)	(0.049)	(0.094)	(0.090)	(0.051)

Notes: Each column displays estimates from one regression model using data from the Early Childhood Longitudinal Survey-Kindergarten Cohort (ECLS). These estimates describe effects based on whether the school was on the AYP margin or below the AYP margin in the relevant subject: math for math test performance or enjoyment, reading for reading test performance or enjoyment, and *either* math or reading for science test performance or anxiety about standardized tests. All models control for the variables listed in Table 2B and Appendix Table A3, 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 6,870 (rounded to the nearest 10 to comply with data reporting requirements). All models weight observations using the student-level longitudinal sample weights provided in the ECLS data. Bootstrapped standard errors, adjusted for school-level clustering using 1,000 Monte Carlo simulations of both the first-stage and second-stage models, are displayed in parentheses below each estimate.

***Significant at the 1 percent level.

**Significant at the 5 percent level.

*Significant at the 10 percent level.

having a fifth grade school on the margin for making AYP is small, statistically insignificant, and *negative*, for both reading (−0.007) and math (−0.021) performance. These falsification tests provide reassuring evidence that our main results are not driven by spurious correlation with an unobserved factor generating positive trends in student achievement.

Importantly, our results also suggest that when schools face NCLB pressure, gains in achievement do not decrease students' enjoyment of reading or math and are likely to decrease anxiety over testing. While it is hard to establish causality, anxiety is generally thought to impede learning while enjoyment increases it (OECD 2004). Respective point estimates for the impact of NCLB pressure on students' enjoyment of reading and math are 0.03 standard deviations and 0.11 standard deviations, though neither estimate is statistically significant at the 0.10 level. We find statistically significant decreases of 9 and 16 percentage points in rates of students' reported anxiety over testing for schools on and below the AYP margin, respectively. Student anxiety might decrease as students feel more comfortable with the tested subject material. While it is possible that schools under pressure focus more on testing—using practice exams, motivational techniques, etc.—these actions might also alleviate student anxiety rather than exacerbate it.

The previous school accountability literature motivates the idea that the impacts of NCLB may differ across students within a school. In a companion set of specifications (not shown here), we examine whether our estimates depend on whether schools faced strong pressure to raise proficiency rates for the overall student population or for the focal student's own subgroup(s). We replace the single "on the AYP margin" variable with three mutually exclusive indicators for whether the school was on the AYP margin due to: (i) the overall student group, (ii) any one of the

student's own subgroups (and *not* the overall student group as well), and (iii) other subgroups (*not* the student's own subgroup or the overall student group). The point estimates for reading performance are always positive, regardless of whether the students are members of subgroups whose performance is most critical to the schools' AYP ratings, and the estimates for math and science performance are all statistically insignificant. We cannot reject that gains on low-stakes reading exams are equal, though the largest point estimate is for students whose own performance will likely not affect their schools' chances of making AYP, suggesting that all students may improve their reading skills when their schools spend more time on reading instruction due to accountability pressure.

In summary, using reliable, low-stakes examinations, we find no evidence that NCLB pressure systematically leads to adverse achievement outcomes. We find some evidence that this pressure improves both achievement and students' outlook towards academics and testing.

C. *Heterogeneous Effects on Students*

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

Previous work suggests schools might direct resources to students who are likely to score close to the threshold of passing the exam (Reback 2008; Neal and Schanzenbach 2010). We classify students as "on the bubble" for passing their state exam if their third grade test scores were estimated to be within 15 percentiles below or 5 percentiles above their states' NCLB exam passing threshold.[25] We then reestimate the specification above adding an indicator for "on the bubble" and an interaction term between this indicator and whether the school is on the AYP margin. For reading, math, and science performance, these estimates are −0.006 (0.065 standard error), 0.047 (0.094 standard error), and 0.074 (0.054 standard error) respectively. Thus, while students on the bubble of passing high-stakes exams do not appear to perform very differently on *low-stakes* exams when their schools face strong NCLB pressure, the imprecision of these point estimates only allows us to rule out large differences.[26]

One concern with accountability pressure is that schools might alter instruction for traditionally underperforming students, such as students from relatively poor households, in ways that raise test scores in the short-run but are less valuable in the

[25] 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.

[26] We can also rule out large differences in test anxiety for students "on the bubble"; the estimated coefficient of the interaction term between a school being on the AYP margin and a student being "on the bubble" for either math or reading performance suggests a statistically insignificant decrease in anxiety of less than one percentage point.

long-run (e.g., shift more time to test preparation). NCLB's use of subgroup-specific pass rates is intended to mitigate these effects, but they might still occur—especially for low-stakes rather than high-stakes test performance. We examine this hypothesis by adding to our main specification an indicator for a student having family income of $35,000 or less and an interaction of this indicator with whether the student's school is on the AYP margin. The estimated coefficients for this interaction terms are all statistically insignificant, though again imprecise.[27]

In many cases, NCLB was layered on top of states' preexisting school accountability systems, and the impact of NCLB pressure might be dampened in states that already put schools under pressure based on student test performance according to the state's own rating system. On the other hand, schools in states with pre-NCLB accountability systems might have more experience quickly mobilizing their resources to meet performance targets. Dee and Jacob (2011) find evidence that supports the former hypothesis: states lacking strong accountability systems prior to NCLB had stronger upward trends in fourth grade students' math performance comparing pre- and post-NCLB cohorts. They do not find a similar trend for reading performance.

Using Dee and Jacob's (2011) classifications, we add to our main specification an interaction term between the state having a strong pre-NCLB accountability system and the school being on the margin for making AYP. Our estimates are somewhat consistent with Dee and Jacob's (2011) findings—math AYP pressure had a more positive effect on students in states lacking strong accountability systems prior to NCLB, though this difference is not significant ($p = 0.16$), and we find even less significant differences in effects of pressure on reading or science performance across states with or without strong pre-NCLB accountability. While our estimates are too imprecise to determine the extent to which these schools contributed to their states' upward math performance trajectories, these results suggest that the upward trends in math performance observed by Dee and Jacob (2011) might have been driven by schools facing the greatest short-term pressure to increase math pass rates.

IV. Conclusion

As a result of the No Child Left Behind Act, virtually every public school in the United States is now accountable for meeting targets for student test performance. To further our understanding of the incentives created by NCLB on students and teachers nationwide, we assemble an extensive national dataset of school and student subgroup performance on the examinations required under NCLB and exploit extensive cross-state variation in states' rules and standards to examine how the threat of failing under NCLB affects school resource allocation and student achievement. We find that teachers report greater concern over how student test performance will affect their job security and they expect to leave the profession sooner. We find changes in work hours suggesting a shift towards teachers who specialize in single subjects and away from instruction of low-stakes subjects like science and social studies.

[27] For models with reading, math, or science test performance as the dependent variable, the estimated coefficients of the interaction terms are 0.025, 0.046, and −0.059, respectively. These estimates continue to be statistically insignificant if one instead uses a smaller income cutoff of $20,000 instead of $35,000.

Nevertheless, we find that students perform at least as well academically in schools facing strong short-term pressure from NCLB as those in comparable schools that do not face such pressure. In schools facing stronger short-term incentives to improve student proficiency rates on high-stakes exams, students raise their achievement by 0.06 standard deviations on low-stakes reading exams, do similarly on low-stakes math and science tests, report more enjoyment of math (and no less enjoyment of reading), and report less test anxiety. We do not find significant heterogeneity in these effects across students in/out of subgroups that have greater influence on whether their schools meet NCLB requirements, students who will likely score close/far from the passing score on their states' high-stakes exam, or students who are from poorer/wealthier families. However, our sample size limits the power of these tests for heterogeneous effects.

Our findings are important given widely held concerns that test-based accountability systems crowd out learning material outside of the high-stakes test. On the other hand, our results also suggest that NCLB pressure may discourage the work effort and career length of teachers working in schools with little chance of meeting student performance standards. This result echoes findings by Li (2012) that relatively effective principals are more likely to exit from schools facing NCLB pressure in North Carolina, and could undermine schools' ability to improve student performance in response to accountability pressure. In addition, our finding that schools respond to accountability pressure by reducing instruction in low-stakes subjects may have negative consequences in the longer term.

These issues loom larger every year as NCLB standards become more stringent and more schools fail to meet those standards. Congress will likely revisit the design of school accountability systems when they enact revisions to NCLB. In view of ballooning AYP school failure rates, the US Department of Education has been granting waivers to states so that schools can avoid AYP failure designations in spite of less than perfect proficiency rates (United States Department of Education 2012). These waivers are conditional on broad education policy reforms that some states view as too costly to implement.[28]

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

[28] See Howard Blume, "No Child Left Behind Waiver Could Cost $2 Billion, Report Says," *Los Angeles Times*, November 12, 2011, accessed November 14, 2011 http://articles.latimes.com/2011/nov/12/local/la-me-no-child-20111112.

Appendix

A. *Sources of Collected AYP Data*

Table A1

	Available in existing databases	We have collected	Not available	State abbreviations where data are not available
Panel A. States in 2002–2003				
School made AYP	24	44	0	—
Subgroup made AYP	5	38	9[a]	AL[b], 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
Panel B. 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: Existing databases refer to the NLSLSAS and NAYPI databases. Number of states per row can exceed 50 because we collected data in states included in existing databases.

[a]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.

[b]Although Alabama did not publish whether student subgroups made AYP in 2002–2003, we can include Alabama schools in our analyses because Alabama (incorrectly) did not base schools' AYP status in 2002–2003 on student subgroup performance.

TABLE A2—MODEL SPECIFICATION AND DATA CONSTRUCTION FOR STATE PROBITS ESTIMATING LIKELIHOOD OF SUBGROUPS MAKING AYP IN 2003 AND 2004

Variable description	Data sources	Variable coding
Panel A. 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 are from NLSLSAS Database or the NAYPI 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 percent accuracy when tested it in two populous states with complete data.	Equals 0 if the subgroup failed its AYP subject-specific proficiency target in either 2002–2003 or 2003–2004. Equals 1 if the subgroup (i) passed its AYP proficiency target in the given subject in 2002–2003 and 2003–2004, or (ii) 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.

(*Continued*)

Table A2—Model Specification and Data Construction for State Probits Estimating Likelihood of Subgroups Making AYP in 2003 and 2004 (*Continued*)

Variable description	Data sources	Variable coding
Panel B. 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 fifth grade proficiency rate on the statewide test administered in 2001–2002 for the focal subject. (We selected grade five 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 six states where proficiency rates are unavailable, we instead use the reported percentile rank scores or scale scores. For states that did not test grade five in 2001–2002, we use the next closest lower tested grade level (i.e., grade four, grade three) or, if that is unavailable, the next closest higher tested grade (i.e., grade six, grade seven). 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., fourth and sixth), 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 fifth 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.

(*Continued*)

TABLE A2—MODEL SPECIFICATION AND DATA CONSTRUCTION FOR STATE PROBITS ESTIMATING LIKELIHOOD OF SUBGROUPS MAKING AYP IN 2003 AND 2004 (*Continued*)

Variable description	Data sources	Variable coding
Percent 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 is missing within the National Longitudinal School-Level State Assessment Score Database, data is from the 2001–2002 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 is 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 $\times$ 2002 percent group interaction terms)	Wherever available, school report card data from state departments of education that list student subgroup size (using AYP definitions). Where not available from state sources, then drawn from 2003–2004 data in the National Longitudinal School-Level State Assessment Score Database or the 2003–2004 Common Core of Data.	This variable is derived from the state's count of continuously enrolled students per student subgroup accountable under NCLB (note that states' definitions of "continuous enrollment" for the purposes of AYP accountability differ somewhat from state definitions for state accountability systems or just cross-sectional enrollment counts as of the fall in the school year). Where state sources are not available, size is estimated using 2004 State Assessment Score data about number of students tested per subgroup. If this source is not available for the state, we used grade-specific CCD enrollment data and district-level LEP and disabled ratios and applied them to school-by-grade-level membership.
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 in 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: Percent of students who are black Percent of students who are Hispanic Percent of students who are Asian Percent of students who qualify for a free or reduced-price meal Whether the school is Title I eligible 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.).

Table A3—School Control Variables, SASS, and ECLS Samples

Variable	Mean	SD
Panel A. School characteristics, SASS sample		
% of states where schools would be on AYP margin (for math or reading)	23%	14%
% of states where schools would be below the AYP margin	17%	20%
Within-state *Z*-score for 2001–2002 reading	0.007	0.949
Within-state *Z*-score for 2001–2002 math	0.043	0.925
Eligible for Title I (from the CCD)	69%	
Number of enrolled students (from the CCD)	587	258
Percent Asian students (from the CCD)	4%	9%
Percent Hispanic students (from the CCD)	19%	28%
Percent African American students (from the CCD)	18%	26%
Percent economically disadvantaged students (from the CCD)	47%	30%
Panel B. School characteristics, ECLS sample		
% of states where schools would be on AYP margin, reading	20%	13%
% of states where schools would be on AYP margin, math	16%	11%
% of states where schools would be on AYP margin, either subject	25%	13%
% of states where schools would be below the AYP margin	14%	18%
Within-state *Z*-score for 2001–2002 reading	0.125	0.957
Within-state *Z*-score for 2001–2002 math	0.100	0.960
Eligible for Title I (from the CCD)	60%	
Number of enrolled students (from the CCD)	586	252
Percent Asian students (from the CCD)	5%	10%
Percent Hispanic students (from the CCD)	16%	24%
Percent African American students (from the CCD)	19%	26%
Percent economically disadvantaged students (from the CCD)	44%	30%
Number LEP students in the grade	5	13
Missing number of LEP students in the grade	14%	

Note: Construction for variables regarding the percentage of states where a school would be on or below the AYP margin is explained in Section IIB.

B. *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 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 schoolwide 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 schoolwide 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 in 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).

REFERENCES

American Institutes for Research. 2001–2002. *National Longitudinal School-Level State Assessment Score (NLSLSAS) Database. State Study of the Implementation of Accountability and Teacher Quality Under No Child Left Behind (SSI-NCLB)*. U.S. Department of Education.

American Institutes for Research. 2003–2004a. *National Adequate Yearly Progress (AYP) Database. State Study of the Implementation of Accountability and Teacher Quality under No Child Left Behind (SSINCLB)*. U.S. Department of Education.

American Institutes for Research. 2003–2004b. *National Longitudinal School-Level State Assessment Score (NLSLSAS) Database. National Longitudinal Study of No Child Left Behind (NLS-NCLB)*. U.S. Department of Education.

Ballou, Dale, and Matthew G. Springer. 2008. "Achievement Trade-Offs and No Child Left Behind." http://www.caldercenter.org/upload/NCLB-paper-Ballou-and-Springer.pdf.

Booher-Jennings, Jennifer. 2005. "Below the Bubble: 'Educational Triage' and the Texas Accountability System." *American Educational Research Journal* 42 (2): 231–68.

Boyd, Donald, Hamilton Lankford, Susanna Loeb, and James Wyckoff. 2011. "Teacher Layoffs: An Empirical Illustration of Seniority versus Measures of Effectiveness." *Education Finance and Policy* 6 (3): 439–54.

Brunner, Eric J., and Jennifer Imazeki. 2010. "Probation Length and Teacher Salaries: Does Waiting Pay Off?" *Industrial Relations and Labor Review* 64 (1): 164–80.

Chakrabarti, Rajashri. 2013. "Vouchers, Public School Response, and the Role of Incentives: Evidence from Florida." *Economic Inquiry* 51 (1): 500–526.

Chakrabarti, Rajashri. 2014. "Incentives and Responses under No Child Left Behind: Credible Threats and the Role of Competition." *Journal of Public Economics* 110: 124–46.

Chiang, Hanley. 2009. "How Accountability Pressure on Failing Schools Affects Student Achievement." *Journal of Public Economics* 93 (9–10): 1045–57.

Cooley Fruehwirth, Jane, and Jeffrey Traczynski. 2012. "Spare the Rod? The Dynamic Effects of No Child Left Behind on Failing Schools." Unpublished.

Corcoran, Sean, Jennifer L. Jennings, and Andrew A. Beveridge. 2011. "Teacher Effectiveness on High-and-Low-Stakes Tests." https://files.nyu.edu/sc129/public/papers/corcoran_jennings_beveridge_2011_wkg_teacher_effects.pdf.

Cullen, Julie Berry, and Randall Reback. 2006. "Tinkering Toward Accolades: School Gaming under a Performance Accountability System." In *Advances in Applied Microeconomics*, Vol. 14, edited by Timothy J. Gronberg and Dennis W. Jansen, 1–34. Bingley: Emerald Group Publishing Limited.

Davidson, Elizabeth, Randall Reback, Jonah E. Rockoff, and Heather L. Schwartz. 2013. "Fifty Ways to Leave a Child Behind: Idiosyncrasies and Discrepancies in States' Implementation of NCLB." National Bureau of Economic Research (NBER) Working Paper 18988.

Dee, Thomas S., and Brian Jacob. 2011. "The Impact of No Child Left Behind on Student Achievement." *Journal of Policy Analysis and Management* 30 (3): 418–46.

Feng, Li, David N. Figlio, and Tim Sass. 2010. "School Accountability and Teacher Mobility." National Bureau of Economic Research (NBER) Working Paper 16070.

Figlio, David N. 2006. "Testing, Crime, and Punishment." *Journal of Public Economics* 90: 837–51.

Figlio, David N., and Lawrence S. Getzler. 2006. "Accountability, Ability, and Disability: Gaming the System?" In *Advances in Applied Microeconomics*, Vol. 14, edited by Timothy J. Gronberg and Dennis W. Jansen, 35–49. Bingley: Emerald Group Publishing Limited.

Figlio, David N., and Maurice E. Lucas. 2004. "What's in a Grade? School Report Cards and the Housing Market." *American Economic Review* 94 (3): 591–604.

Figlio, David N., and Cecilia Elena Rouse. 2006. "Do Accountability and Voucher Threats Improve Low-Performing Schools?" *Journal of Public Economics* 90 (1–2): 239–55.

Figlio, David N., and Joshua Winicki. 2005. "Food for Thought? The Effects of School Accountability Plans on School Nutrition." *Journal of Public Economics* 89 (2–3): 381–94.

Hamilton, Laura S., Brian M. Stecher, Julie A. Marsh, Jennifer Sloan McCombs, Abby Robyn, Jennifer Russell, Scott Naftel, and Heather Barney. 2007. *Standards-Based Accountability under No Child Left Behind: Experiences of Teachers and Administrators in Three States*. Santa Monica, CA: RAND Corporation.

Hanushek, Eric A., and Margaret E. Raymond. 2005. "Does School Accountability Lead to Improved School Performance?" *Journal of Policy Analysis and Management* 24 (2): 297–327.

Hemelt, Steven W. 2011. "Performance Effects of Failure to Make Adequate Yearly Progress (AYP): Evidence from a Regression Discontinuity Framework." *Economics of Education Review* 30 (4): 702–23.

Hill, Heather C., Brian Rowan, and Deborah Lowenberg Ball. 2005. "Effects of Teachers' Mathematical Knowledge for Teaching on Student Achievement." *American Educational Research Journal* 42 (2): 371–406.

Hussain, Iftikhar. 2012. "Subjective Performance Evaluation in the Public Sector: Evidence From School Inspections." http://cee.lse.ac.uk/ceedps/ceedp135.pdf.

Jacob, Brian A. 2005. "Accountability, Incentives and Behavior: The Impact of High-Stakes Testing in the Chicago Public Schools." *Journal of Public Economics* 89 (5–6): 761–96.

Jacob, Brian, and Steven D. Levitt. 2003. "Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating." *Quarterly Journal of Economics* 118 (3): 843–77.

Krieg, John M. 2008. "Are Students Left Behind? The Distributional Effects of No Child Left Behind." *Education, Finance and Policy* 3 (2): 250–81.

Ladd, Helen, and Douglas Lauen. 2010. "Status versus Growth: The Distributional Effects of School Accountability Policies." *Journal of Policy Analysis and Management* 29 (3): 426–50.

Ladd, Helen F., and Arnaldo Zelli. 2002. "School-Based Accountability in North Carolina: The Responses of School Principals." *Educational Administration Quarterly* 38 (4): 494–529.

Li, Danielle. 2012. "School Accountability and Principal Mobility: How No Child Left Behind Affects the Allocation of School Leaders." Unpublished.

Organisation for Economic Co-operation and Development (**OECD**)**.** 2004. "Student Learning: Attitudes, Engagement, and Strategies." In *Learning for Tomorrow's World – First Results from PISA 2003*. Programme international pour le suivi des acquis des élèves (PISA). Paris.(accessed June 13, 2013).

National Center for Education Statistics. 2000. *Common Core of Data (CCD) Data File: Public Elementary/Secondary School Universe Survey Data: School Year 1998-99.* U.S. Department of Education. Washington, DC: U.S. Department of Education.

National Center for Education Statistics. 2003. *Common Core of Data (CCD) Data File: Public Elementary/Secondary School Universe Survey Data: School Year 2001-02.* U.S. Department of Education. Washington, DC: U.S. Department of Education.

National Center for Education Statistics. 2004. *Common Core of Data (CCD) Data File: Public Elementary/Secondary School Universe Survey Data: School Year 2002-03.* U.S. Department of Education. Washington, DC: U.S. Department of Education.

National Center for Education Statistics. 2005. *1999-2000 Schools and Staffing Survey (SASS) and 2000-01 Teacher Follow-up Survey (TFS) Restricted-Use Data with Codebook.* CD-ROM. Washington, DC: U.S. Department of Education.

National Center for Education Statistics. 2005. *Common Core of Data (CCD) Data File: Public Elementary/Secondary School Universe Survey Data: School Year 2003-04.* U.S. Department of Education. Washington, DC: U.S. Department of Education.

National Center for Education Statistics. 2006. *Early Childhood Longitudinal Study-Kindergarten Class of 1998-99 (ECLS-K) Fifth Grade Restricted-Use Data File and Electronic Codebook.* CD-ROM. Washington, DC: U.S. Department of Education.

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

National Center for Education Statistics. 2007. *2003–04 Schools and Staffing Survey* (*SASS*) *and 2004–05 Teacher Follow-up Survey* (*TFS*) *Restricted-Use Data with Codebook*. CD-ROM. Washington, DC: U.S. Department of Education.

Neal, Derek, and Diane Whitmore Schanzenbach. 2010. "Left Behind by Design: Proficiency Counts and Test-Based Accountability." *Review of Economics and Statistics* 92 (2): 263–83.

Reback, Randall. 2008. "Teaching to the Rating: School Accountability and the Distribution of Student Achievement." *Journal of Public Economics* 92 (5–6): 1394–1415.

Reback, Randall, Jonah Rockoff, and Heather L. Schwartz. 2014. "Under Pressure: Job Security, Resource Allocation, and Productivity in Schools under No Child Left Behind: Dataset." *American Economic Journal: Economic Policy*. http://dx.doi.org/10.1257/pol.6.3.207.

Rockoff, Jonah, and Lesley J. Turner. 2010. "Short-Run Impacts of Accountability on School Quality." *American Economic Journal: Economic Policy* 2 (4): 119–47.

Rouse, Cecilia Elena, Jane Hannaway, Dan Goldhaber, and David Figlio. 2007. "Feeling the Florida Heat? How Low-Performing Schools Respond to Voucher and Accountability Pressure." National Bureau of Economic Research (NBER) Working Paper 13681.

Springer, Matthew G. 2008. "The Influence of an NCLB Accountability Plan on the Distribution of Student Test Score Gains." *Economics of Education Review* 27 (5): 556–63.

Stecher, Brian M., Scott Epstein, Laura S. Hamilton, Julie A. Marsh, Abby Robyn, Jennifer Sloan McCombs, Jennifer Russell, and Scott Naftel. 2008. *Pain and Gain: Implementing No Child Left Behind in Three States, 2004–2006.* Santa Monica, CA: RAND Corporation.

Tyler, John H., Richard J. Murnane, and John B. Willett. 2000. "Estimating the Labor Market Signaling Value of the GED." *Quarterly Journal of Economics* 115 (2): 431–68.

United States Department of Education. 2012. Elementary and Secondary Education Act (ESEA) Flexibility. http://www.ed.gov/esea/flexibility (accessed July 24, 2012).