NATIONAL CENTER ON
Performance Incentives

The Effects of No Child Left Behind on School Services and Student Outcomes

Randall Reback
Jonah E. Rockoff
Heather L. Schwartz

Paper presented at the *NCLB: Emerging Findings Research Conference* at the Urban Institute, Washington, D.C. on August 12, 2009.

Conference Paper
August 2009

LED BY
VANDERBILT
PEABODY COLLEGE

IN COOPERATION WITH:
CALDER

The National Center on Performance Incentives (NCPI) is charged by the federal government with exercising leadership on performance incentives in education. Established in 2006 through a major research and development grant from the United States Department of Education's Institute of Education Sciences (IES), NCPI conducts scientific, comprehensive, and independent studies on the individual and institutional effects of performance incentives in education. A signature activity of the center is the conduct of two randomized field trials offering student achievement-related bonuses to teachers. The Center is committed to air and rigorous research in an effort to provide the field of education with reliable knowledge to guide policy and practice.

The Center is housed in the Learning Sciences Institute on the campus of Vanderbilt University's Peabody College. The Center's management under the Learning Sciences Institute, along with the National Center on School Choice, makes Vanderbilt the only higher education institution to house two federal research and development centers supported by the Institute of Education Services.

The conference was supported, in part, by the National Center for the Analysis of Longitudinal Data in Education Research (CALDER), funded by Grant R305A060018 to the Urban Institute from the Institute of Education Sciences, U.S. Department of Education and the National Center for Performance Incentives (NCPI) at Vanderbilt University. This research project was made possible by funding from the Institute for Education Sciences and the Spencer Foundation, as well as a seed grant from the Columbia University Institute for Social and Economic Research and Policy. Ben Lockwood, Julia Zhou, and Molly Alter provided outstanding research assistance. The views expressed in the paper are solely those of the authors and may not reflect those of the funders or supporting organizations. Any errors are attributable to the authors.

Please visit **www.performanceincentives.org** to learn more about our program of research and recent publications.

The Effects of No Child Left Behind on School Services and Student Outcomes

RANDALL REBACK
Barnard College

JONAH E. ROCKOFF
Columbia Business School

HEATHER L. SCHWARTZ
Columbia University

Abstract

Under the *No Child Left Behind Act* (NCLB), in theory, schools on the margin for meeting AYP face strong short-term incentives to increase students' pass rates on specific exams, and may change their behavior accordingly. Using a comprehensive, national, school-level data set concerning schools' AYP status, student characteristics, and student test score performance, as well as NCLB test score performance and school characteristics, the authors predict which schools were near the margin for failing to meet their state's AYP standards in math and reading. Variance in state policies creates several cases where schools near the margin for satisfying their *own* state's AYP requirements would have almost certainly failed or almost certainly passed AYP if they were located in other states. Using the nationally representative Early Childhood Longitudinal Survey (ECLS), the authors examine how NCLB incentives affect students' academic achievement and non-academic outcomes, and school resource allocation. States vary widely in the percent of schools that fail or struggle to meet AYP, with cross-state variation in student academic aptitude or in exam difficulty explaining relatively little of this variation. Rather, cross-state variation in AYP failure rates is due to states' choice of policy parameters concerning AYP rules, such as the minimum student enrollment size required for student subgroups' pass rates to contribute to a school's AYP status, the size of confidence intervals and the "safe harbor" rules that effectively lower the minimum pass rates required for smaller subgroups to pass AYP. Our very preliminary results suggest that NCLB pressure influences student and staff attitudes and teachers' time use and instructional strategies but has little net effect on mean student test score growth on low-stakes exams.

1. Introduction

In 2002, President George W. Bush signed into law the No Child Left Behind Act (NCLB), which many consider to be the largest change to federal education policy since the authorization of the Elementary and Secondary Education Act in 1965. The most significant policy change was to require, under the compulsion of federal funding, that states adopt school accountability systems based on minimum competency testing. This accountability system requires that states determine whether public schools satisfy Adequate Yearly Progress (AYP) based on the fraction of students taking and demonstrating proficiency on statewide exams. Public schools have strong incentives to satisfy AYP: failure triggers sanctions that escalate over consecutive years, and states are required to publish annual school report cards, which can affect school prestige, local property values, and financial rewards to schools and teachers.

This paper examines the short term incentives that schools face to satisfy AYP during a particular year. We investigate whether schools on the cusp of failing AYP alter their behavior in ways that improve student test scores and other outcomes, affect teachers' priorities and time-use, and change the services offered to students. In addition, we plan to examine whether schools respond to incentives to focus on the test score performance of specific types of students depending on which student subgroups are at threat of failing AYP. NCLB requires states to base schools' AYP status not only on the overall pass rates in math and reading, but also on pass rates among student subgroups, separated by ethnicity, (e.g., White, Hispanic, African American), and special needs (e.g., students from low income families, special education students, limited English proficiency), so long as there are a sufficient number of students in those subgroups continuously enrolled in the school. Student performance is thus far more critical to schools' AYP status if the students are, first, part of a subgroup that is sufficiently large to count towards

AYP, and, second, the subgroup has a relatively high risk of missing the required pass rate in that subject.

Our analysis is based on two unusually comprehensive sets of data on schools, classrooms, and students. First, we assemble a national dataset of school-level AYP related outcomes, as well as states' AYP rules and regulations. These AYP data are then matched to panel data on a nationally representative sample of elementary school children from the Early Childhood Longitudinal Survey (ECLS), available from the National Center for Educational Statistics. The ECLS is an extremely detailed panel data set, and we can examine ECLS data for more than 9,000 students who were in first grade during the 1999-00 school year. These students were surveyed again in the school years 2001-02 and 2003-04, when most would have been in third and fifth grade, respectively. The timing of the ECLS is serendipitous, as this student cohort was among the first to be affected by the requirements and incentives of NCLB. Importantly for our research, ECLS students are administered independent assessments in math, reading, science, and social studies that are separate from states' high-stakes tests used to determine schools' AYP status.

We implement a "difference-in-differences" style approach to identify the effects of NCLB incentives on ECLS students' behavior, academic performance, and school resource allocation. First, we use schools' campus-wide and subgroup-specific test score performance and school characteristics from the school year 2001-02—which is after the passage of NCLB but prior to the measurement of AYP—to predict which schools were near the margin for failing to meet their states' various AYP standards for math and reading for the 2003-04 school year. Each state designates its own standardized tests and its own rules for satisfying AYP, so there are numerous cases where schools near the margin for satisfying their own state's AYP requirements would have almost certainly failed or almost certainly passed AYP if they were located in other states. We exploit this variation by comparing differences in outcomes for ECLS students in schools on the margin and in schools not on the margin within the same state with differences in outcomes for students in similar schools in other states where neither school is on the margin. Because ECLS tracks students over time, we can also control for students' initial achievement levels and a host of other student and school characteristics. We examine the impact of schools' AYP incentives on students' academic achievement, their social and emotional development, their teachers' survey responses about classroom activities, and the school resources directed towards them.

Our analyses confirm that states vary widely in the percent of schools that fail or struggle to meet AYP, with cross-state variation in student academic aptitude or exam difficulty explaining relatively little of this variation. Most of the cross-state variation in AYP failure rates is due to states' choice of policy parameters concerning issues such as the minimum student enrollment size required for student subgroups to count towards AYP, the degree of racial and economic heterogeneity within schools and thus the number of student subgroups for which schools are accountable, the number of high stakes grade levels that count toward AYP designations, and various confidence interval and "safe harbor" rules allowing lower pass rates for smaller groups.

Our **very preliminary** results suggest that broad NCLB pressure influences student and staff attitudes and teachers' instructional strategies but has little effect on mean student performance on low stakes exams. As expected, we find that teachers spend more time on test preparation in schools facing greater short term NCLB pressure. They are also less likely to report feeling that they have great control over classroom-level decisions.[1] These changes do not appear to have either positive or negative net effects on mean student achievement gains in math, reading, or science, though NCLB pressure does appear to bolster students' confidence and motivation in math and lower it in reading. In future drafts, we plan to examine the robustness of these findings to alternative specifications and to examine whether particular types of students perform better or worse due to within-school variation in schools' incentives to improve the performance of various types of students.

From a policy perspective, understanding the impact of NCLB on schools' behavior provides important feedback on both its intended and unintended consequences. It is important to determine whether schools' responses to strong accountability incentives lead to improvements or declines in academic and non-academic outcomes for different types of students. This information could guide future federal policies, including the potential re-authorization of NCLB, as well as states' decisions concerning the specific design of their school accountability systems.

We begin by framing our project within the larger body of research about the effects of NCLB on student outcomes. To motivate our research questions, we next set out our conceptual framework about how schools might respond to NCLB pressures. We then discuss the NCLB data we have collected as well as the ECLS survey data. We then discuss our methodology, report our findings, and briefly discuss their implications for policy.

[1] Some of these findings, though substantial in size, are currently of modest statistical significance and we need to re-examine them in future drafts using larger sample sizes.

2. Related Literature

Most research on schools' responses to accountability programs examines state and local systems that preceded No Child Left Behind (e.g., Ladd and Zelli, 2002; Hanushek and Raymond, 2005; Reback, 2008; Chakrabarti, 2007; Rouse et al., 2007, Chiang, 2008, Rockoff and Turner, 2008). These studies find evidence that accountability pressure can cause schools to change resource allocation in ways that raise average student achievement. For example, in addition to finding improved achievement, Rouse et al. (2007) find that schools under accountability pressure in Florida devoted more time to instruction, increased resources available to teachers, and decreased principal control. Chiang (2008) finds increased spending on instructional equipment, curricular development, and teacher training using the same data.

However, research has also revealed that schools may shift their resources towards students and subjects that are most critical to the schools' accountability rating, to the detriment of lower stakes subjects and students whose outcomes carry less weight. Further, additional studies uncover ways in which accountability can produce unintended effects, such as teaching to the test (Jacob, 2005; Figlio and Rouse, 2006), eliminating low performing students from the testing pool (Figlio & Getzler, 2006; Figlio, 2006, Cullen & Reback, 2006), or outright cheating (Jacob and Levitt, 2003).

Knowledge about the impacts of NCLB (and not just state accountability systems more generally) on school practice and student performance is still nascent. Studies that have been national in scope generally fail to employ rigorous methods required to make causal claims (e.g., Center on Education Policy, 2007). To our knowledge, only one study examines the impact of NCLB incentives in multiple states using rigorous identification methods. Ballou and Springer (2008) identify responses to accountability incentives using variation in the timing in which

NCLB testing was administered at various grade levels across seven states. They find that students generally perform better on low-stakes exams during high-stakes years, particularly students near the margin for passing their states' high-stakes exam. The few additional studies of NCLB incentives that apply more rigorous methods use data from only one state or one city. These studies have found that below-proficient students enrolled in schools failing AYP tend to make greater than expected test score gains, but there is conflicting evidence concerning the effects on students scoring at either the highest or lowest end of the performance spectrum (see Springer, 2008; Krieg, 2008, and Neal and Whitmore Schanzenbach, forthcoming).

Our work here extends this prior literature in several ways. First, we examine NCLB incentives using a nationally representative sample of students and their schools. Second, we employ a rich panel dataset that includes a wide range of academic outcomes based on identical assessments across states, allowing us to measure how individual students' academic outcomes change over time. Third, we study how NCLB accountability incentives affect mediating and moderating variables such as student behavior, teacher behavior, and school resource allocation. Taken together, these aspects enable us to provide a much richer picture of the impacts of NCLB pressure on students and schools.

3. Conceptual Framework

Our empirical strategy is based on a simple but flexible framework for how schools respond to a system of accountability such as No Child Left Behind. Schools have various resources they can use to improve the skills of their students (e.g., school staff, curriculum, facilities, parental involvement, etc.), and all of these resources have associated costs. Subject to a budget constraint, schools choose an allocation of resources that determine students' skills.

Schools have preferences about the relative importance of helping students improve different types of skills and the relative importance of helping different types of students make improvements.

But there are also competing demands that constrain the amount and allocation of school resources. Community members and school staff care about other things in addition to students' skill acquisition. For example, teachers and school administrators care about their own leisure time and would perceive a cost to extra time spent with students, such as spending weekends in the classroom. Similarly, community members care about their consumption of other goods and services, and would perceive a cost to increasing property taxes in order to increase resources devoted to education.

More formally, suppose that schools' resources can be classified into four general types: the first type (denoted u) helps to improve all skills for all students (e.g., the overall effort level of teachers), the second (denoted a_s) is skill-specific and serves all students (e.g., math lessons that equally help all students learn math), the third type (denoted b_i) is student-specific and serves all skills (e.g., providing individual attention to students to improve study-skills or behavior), and the fourth type (denoted c_{is}) is skill-specific and student-specific (e.g., individual math tutoring). Suppose that there are three categories of student skills that schools aim to improve: reading (s=r), math (s=m), and all other academic and non-academic skills (s=z), and that schools place weights (denoted γ_{is}) on each type of skill for each student, due to the preferences of school staff and the community. Finally, denote by l the type of resources that improve consumption of goods and services that are valued by the community and school staff but unrelated to skill acquisition (e.g., teacher leisure time). Schools with N students and total resources equal to K will choose an allocation of resources to maximize:

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

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

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

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

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

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

In this equation, ε_i is idiosyncratic noise due to imperfect test measurement (with mean zero and known variance), the function g_i maps resources and test measurement error into whether student *i* passes the standardized tests, and the function *V* maps the school-wide pass rate into benefits or costs.[2] Note that resources which do not improve skills in math or reading (a_z and c_{iz}) do not enter in the function g_i.

[2] Alternatively, the function V could enter into the school's budget constraint, rather than the utility function, but the qualitative results from this alternative framework would be the same.

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

$$V = \begin{cases} \overline{V}\ if \left(\frac{1}{N} \sum_{i=1}^{N} g_{i} \left(u, a_{r}, a_{m}, b_{i}, c_{ir}, c_{im}, d_{i}, \varepsilon_{i} \right) \right) \geq P* \\ \underline{V}\ if \left(\frac{1}{N} \sum_{i=1}^{N} g_{i} \left(u, a_{r}, a_{m}, b_{i}, c_{ir}, c_{im}, d_{i}, \varepsilon_{i} \right) \right) < P* \end{cases} \quad where\ \overline{V} > \underline{V}$$

In other words, the school is worse off if the pass rate falls below some threshold *P**, but all other variation in the pass rate above or below that threshold does not have immediate consequences related to NCLB. The "all or nothing" structure has important implications for the accountability pressure faced by different schools. Because input allocations and the variance of test measurement error are known, schools will form expectations about their probabilities of making AYP. If a school has a very high probability of making AYP under its optimal pre-NCLB resource allocation**,** it will face very little pressure and, consequently, resource allocation under NCLB should resemble "business as usual." In contrast, if a school expects to be close to the margin of making AYP under their optimal pre-NCLB resource allocation, the school and its community will face considerable pressure to improve student pass rates. This is the key identifying assumption in our methodology.[3]

There are several ways in which an accountability system such as NCLB may change resource allocation decisions. It may induce schools and communities to reduce resources devoted to consumption of goods and services (*l*) and direct them towards improving student skills. Accountability pressure may also induce schools to spend fewer resources on non-tested

[3] Schools with a very low probability of making AYP in the current year will also face pressure to improve over a longer period of time. In our current analysis, we focus on student achievement and resource allocation in the current year, and group together schools with both very low and very high probabilities of making AYP. In future work, we will examine the importance of this specification restriction.

skills (a_z and c_{iz}), and more resources into the math and reading skills of students for whom extra resources will improve the probability of passing the standardized exams (b_i, c_{ir}, and c_{im}). Finally, schools may allocate resources to activities that improve pass rates (d_i) but not skill acquisition.

The extent to which the NCLB incentives change the allocation of resources in a school depends greatly on schools' preferences and the functions that determine how resources affect skill acquisition and exam pass rates. So the actual impact of NCLB on students is an empirical question. Schools may respond by having the school's staff perform more effectively and/or exert greater effort, raising student achievement for all students without negatively affecting any student. In less ideal circumstances, schools may shift resources in ways that improve their chances of making AYP but at the expense of the acquisition of skills by some students or the acquisition of non-tested skills.

4. Data and Descriptive Analysis

Our analysis requires a comprehensive database of NCLB-related outcomes across all states in the ECLS. We therefore compile data from a number of existing (but incomplete) sources covering multiple states, and add newly collected data from school report cards and other state-level sources. Specifically, we collect three categories of data that influence school AYP determinations: (i) the percentage of students within each numerically significant subgroup that passed the state math and reading tests, (ii) the state's determination whether each numerically-significant subgroup passed its proficiency rate targets in math and reading—a determination that accounts for confidence intervals, safe harbor provisions, and the appeals process, and (iii) the number of students within each numerically significant subgroup, which is a tally of tested and “continuously enrolled” students (according the state’s definition of that term). We also determine whether the school passed or failed AYP overall, which is based on not only proficiency targets but also on participation rate targets and the state’s designated “other” category, which is typically the attendance rate at the elementary and middle-school levels and the graduation rate for high schools. After compiling data that is already available from national datasets such as School Data Direct and the National AYP and Identification Database, we supplement missing information with data from state departments of education, including data from individual school report cards. For example, student subgroup size data was largely unavailable from existing national datasets, while other categories were frequently missing. Appendix 1 displays the categories of data collected and their sources.

Drawing on our compiled NCLB data, Table 1 presents information on the schools that did and did not make AYP in the school year 2003-04.[4] This is, to the best of our knowledge, the first time such a comparison has been possible, for any school year since the passage of NCLB. Roughly one in five schools did not make AYP based on tests taken in 2004. Schools that failed to make AYP had higher enrollment and higher student/teacher ratios, were more likely to be eligible for Title I funding or be located in a city, and were less likely to serve students in primary grades.[5] They also had higher fractions of students eligible for free or reduced price lunch (a measure of poverty), and higher fractions of black and Hispanic students.

The fraction of schools in each state making AYP in 2004 varied considerably (Figure 1). Failure rates ranged from roughly five percent in Ohio to 77 percent in Florida, with about half

[4] Data on school characteristics is taken from the Common Core of Data (CCD), compiled by the National Center for Educational Statistics (NCES). The merging of CCD and AYP data, based on NCES identifiers, was imperfect. Only a handful of schools with AYP information did not merge with CCD data, but 13 percent of schools in the CCD for 2003-04 did not merge with AYP information. However, of these unmerged CCD schools, more than half were listed as having no positive enrollment or having opened after the previous CCD report was submitted to NCES. Tennessee stopped reporting school level demographic information to the federal government after 1998-99. Rather than drop Tennessee from our analysis, we use data from 1989-99.

[5] The definition of grade levels served is taken directly from the CCD data.

the states falling between 15 and 30 percent. Importantly for our study, the variation in the fraction of schools making AYP seems to be a somewhat complex function of state's policy choices surrounding proficiency measurement and AYP determination. For example, the fraction of schools failing to meet AYP by state bears no relation to the fraction of students statewide deemed proficient in math (Figure 2, top panel). While this may seem surprising, it has a simple explanation—there is a very strong positive relationship between the fraction of students proficient in math statewide and the required fraction needed to make AYP at the school and subgroup level (Figure 2, bottom panel). States' NCLB regulations "grade on a curve,"[6] so differences in exam difficulty, opinions of what constitutes proficiency, or true variation in 'proficiency' levels across states made little difference in schools' AYP outcomes.[7]

One factor we can examine which does bear a positive relation with AYP outcomes across states is the average number of significant subgroups per school (Figure 3). The variation in significant subgroups stems from cross-state variation in the number of students constituting significance and the size and heterogeneity of school enrollment. For example, in Louisiana, only 10 students are needed to constitute a significant subgroup, while most states set this number at 30 or 40. Yet South Dakota, where the number of subgroups is quite low on average, also places the cutoff for significance at 10 students. The difference in average number of significant subgroups per school across these two states is likely driven by heterogeneity and size—85 percent of students are white in South Dakota (48 percent in Louisiana), and average school enrollment is 170 (510 in Louisiana). Nevertheless, few schools failed to make AYP in 2004 in

[6] Most states' initial AYP standards for required pass rates were linked to some sort of statewide measure of school performance at the 20th percentile in baseline (pre-AYP) spring 2002 testing, but there was wide variation in how states calculated 20th percentile performance and how states applied required pass rate thresholds across subjects and grades. For example, some states based the 20th percentile measure on baseline school-wide pass rates and some used grade-specific and/or subject specific baseline pass rates.

[7] Figures for using reading proficiency (available upon request) are quite similar.

both South Dakota and Louisiana. While Louisiana had many significant subgroups per school, it also had very generous 99% confidence interval which built wide bands around the raw proficiency rates of subgroups, which proportionally lowers the adjusted proficiency bar for AYP in relation to the size of the subgroup.

We performed our analyses in two stages. The first stage requires the NCLB data we described to predict whether a school is on the margin of passing or failing AYP in 2004. In the second stage, we merge this information about schools' predicted probabilities with student-level outcomes and student characteristics available from the nationally-representative Early Childhood Longitudinal Survey-Kindergarten Class of 1998-99 (ECLS) dataset. The ECLS survey is sponsored and distributed by the National Center for Educational Statistics and follows the same children from kindergarten through the end of elementary school. Data collection took place in the fall and the spring of the school years 1998-99 and 1999-2000 (kindergarten and first grade), and in the spring of the school years 2001-02, 2003-04, and 2006-07 (third grade, fifth grade, and eighth grade).[8] We have gained access to the restricted-use version of the ECLS, which identifies individual schools and can therefore be linked with school accountability measures.

The ECLS sample was designed to be nationally representative of kindergartners, their classrooms, and their schools in the 1998-99 school year, (and also representative for first grade

[8] Due to student attrition and movement, the ECLS may not be perfectly representative in the later years of data collection. In the school year 1999-2000, the sample remained representative by surveying a random 50 percent sub-sample of students who transferred from their original school and adding another random sample of first graders in the same schools where transfer students were followed. However, this "freshening" of the sample was not repeated in the third, fifth, and eight grades. This might affect the validity of our empirical methodology to the extent that school accountability pressure from NCLB had an impact on student mobility, and we return to this issue below. The number of children per school decreases with each round of data collection as children change schools. Approximately, one-quarter of children changed schools between kindergarten and first grade, and half of the children had changed schools at least once between kindergarten and third grade. The ECLS continued to collect data from students who were retained within the same grade or skipped a grade level.

students in the 1999-00 school year), but the sample only includes students from 40 states.[9] While it would be preferable to have nationwide data, the ECLS has the widest coverage of any longitudinal dataset during the NCLB time period. It also includes a broad range of information collected from students, families, teachers, and schools. Of particular interest is student performance on a series of standardized tests. Reading and math tests were administered at every stage of data collection, tests of both science and social studies were administered in kindergarten and first grade, and science tests were administered in third, fifth, and eighth grade.

Unlike the tests that states administer for accountability purposes, the ECLS tests were un-timed and administered adaptively using Item Response Theory (i.e., questions are selected based on a student's performance on preceding questions), which prevents a floor or ceiling effect and increases reliability. For our primary test score outcome measures, we convert students' test scores to nationally standardized scores (Z-scores), so that a score of zero should be equivalent to performing at the national mean. Students also answered questionnaires regarding their perceptions of their own skills and knowledge, their emotional/behavioral/mental health, their experiences in school, and their other activities. The survey also collected practical measures of behaviors that are relevant to schools, including attendance and tardiness.

In addition to data collected on students, the ECLS contains information on schools' curricula and resources and teachers' time-use. These are the primary mediators that will allow us to examine how the accountability incentives created by NCLB may have changed the teaching and learning environment in schools. While it is impossible to establish a definitive, empirical link between changes in a moderating variable and changes in a specific type of

[9] It used a multistage probability sample design, first selecting broad geographic areas (e.g., a county), then selecting schools within that area, and finally selecting students within those schools. On average, 23 kindergarten students were sampled from each school. During 1999-00, ECLS refreshed the sample so that researchers can calculate nationally representative estimates for students who were first graders in 1999-00.

student outcome,[10] examining these moderating variables offers complementary evidence on the mechanisms by which incentives affect students.[11]

Table 2 provides the descriptive statistics for our various outcome measures for the sample of students surveyed at the end of 2003-2004 attending regular public schools. Table 3 provides the descriptive statistics for the measures we use as control variables in our regression analysis.

5. Methods and Empirical Results

We investigate whether the incentives built into NCLB cause changes in outcomes and resource allocation at schools that expect to be near the margin of making AYP. Substantial variation across states in the difficulty of making AYP allows us to compare schools with very similar student characteristics and prior achievement but very different accountability pressure. Our “difference-in-differences” style approach compares (a) differences in outcomes for students in the same state enrolled in schools that are and are not near the AYP margin to (b) differences in outcomes of students in other states enrolled in similar schools that are both not near the AYP margin. Additionally, we control for student characteristics (including prior achievement) and a host of other factors that might separately contribute to outcomes, (see Table 3 for a list of these control variables).[12]

[10] Rouse et. al (2008) discuss this same issue within the context of analyzing student outcomes and potential moderating variables in their Florida accountability data.

[11] It is worth noting two factors that help ensure the validity of the ECLS measures. First, students and schools became involved in the survey prior well before NCLB. This is likely to increase their level of comfort with the ECLS surveyors and their understanding that the ECLS survey is completely independent of NCLB policies. Additionally, there were no stakes attached to any of the outcomes we examine, which diminishes any concern that schools were “teaching to” these tests or that respondents would strategically misreport answers to survey questions.

[12] Note that, in addition to the advantages enumerated in section 4, the fact that we do not examine student achievement on tests directly related to NCLB ensures there will be no mechanical “reversion to the mean” for schools near the margin of making AYP due to idiosyncratic measurement error in test scores.

In the first stage of our analysis, we estimate which schools could have expected to be on the margin for satisfying AYP during the school year 2003-04. Specifically, we first estimate the likelihood that each numerically-significant student group within a school satisfied AYP requirements in 2003-04 using a probit model.[13] This means a single school will have as many AYP predictions as it has numerically significant student subgroups. Fortunately, we can use both school demographic characteristics (listed in Table 3) and actual 2001-02 test performance to predict the likelihood of satisfying AYP. The 2001-02 test performance variables serve as powerful, exogenous predictors, because AYP accountability began the following year and testing during 2001-02 was only a "dry run" for NCLB.

Since exams, rules governing AYP, and available data differ across states, we estimate a separate probit model for each state, allowing state-specific slopes for the same general model specification. For all students and all student subgroups (indexed by *k*) contributing to the AYP status of school *j* in state *q,* we specify the following model:

$$(1)\ \ AYP_{jks2004} = \begin{cases} 1\ if\ \ \alpha_q + X_{jks2002}\beta_q + \rho_q W_{j2002} + \zeta_{jks} > 0 \\ 0\ otherwise \end{cases},\ j \in q\,,$$

where $AYP_{jks2004}$ denotes whether the pass rate for group *k* at school *j* met state requirements for AYP in subject *s*. W_{j2002} is a vector of control variables for school-level demographics from the school year 2001-02 (listed in Table 3), and $X_{jks2002}$ is a vector of test score variables for group *k* based on performance on statewide exams in subject *s* during the school year 2001-02, and ζ_{jks} is

[13] NCLB requires that, in order to make AYP for the school year, a school must satisfy AYP for each of its "numerically significant" student subgroups, as well as at the school-wide level. The definition of numeric significance varies across states. For example, in California there are ten student subgroups that can be each assessed for AYP performance: (1) African American, (2) white, (3) Asian, (4) Filipino, (5) Hispanic, (6) socio-economically disadvantaged, (7) disabled, (8) limited English proficient, (9) American Indian, and (10) Pacific Islander. California deems a student subgroup numerically significant if it has either (a) 100 or more students, or (b) at least 50 students and the subgroup comprises at least 15% of the total enrollment in the school. This policy has the effect of making many student subgroups within schools exempt from AYP criteria. California is one of the only states that require such a large count for numeric significance; most states deem a subgroup numerically significant if it contains 30 or more students.

a normally distributed disturbance term. The $X_{jks2002}$ vector includes cubic terms for either the percent of students in group *k* passing the exam in subject *s* or, if this is unavailable in state *q*, the percent of students overall in the given grade level passing the exam in subject *s*.[14] The $X_{jks2002}$ vector also includes each of these cubic performance terms divided by the square root of the number of accountable test-taking students in group *k* at school *j* during the 2003-04 school year.[15] These additional terms serve a dual purpose: (i) they help to account for the mechanical decrease in the variance of student pass rates as the number of tested students in 2004 increases, and (ii) they also help to account for state-specific NCLB "confidence interval" rules which lower the passing requirements for smaller groups.

In a few states, the 2002 test score performance of group *k* in school *j* is unavailable*,* and we instead use the performance of all tested students in one elementary grade (either grade three, four, or five depending on which grade is tested in the state) in the school for that year. To account for variation in the power of school-wide test performance to predict future student subgroup specific outcomes, in these states we also include interactions between school-wide performance and the fraction of tested students in group *k* in 2002. Thus, while Equation 1 provides a succinct functional form for estimating AYP pass rate probabilities, it is sufficiently flexible to take into account variation in data availability and states' NCLB standards.

We use predicted group-level AYP pass probabilities to construct proxies for accountability pressure faced by schools under NCLB. Schools with high pass probabilities for all of their accountable groups will face little pressure. In contrast, a school with at least one group on the margin for satisfying AYP requirements is likely to face accountability pressure.

[14] In practice, we find that either measure of pre-NCLB test score performance works equally well in predicting the likelihood that the schools' pass rates will be near the NCLB required cutoff in 2003-04.

[15] Thus, technically, the X_{jk2002} vector includes some information from the school year 2003-04. We use only the 2002 subscript for simplicity.

Furthermore, this pressure will be greatest in the short run if no numerically significant group within the school has a very low probability of making AYP. For example, if group *k* in school *j* has only a one percent expected probability of satisfying state requirements, there is likely to be little the school can do in the short run to change their AYP outcome.

Define Mrg_AYP_{jks} as an indicator for whether a single group *k* at school *j* was on the margin for satisfying state requirements for subject *s*. We set Mrg_AYP_{jks} equal to one if the estimated probability of this group satisfying the requirements was between 25 and 75 percent. Define Low_AYP_{jks} as an indicator variable for whether the estimated probability of group *k* at school *j* passing state requirements for subject *s* during 2003-04 is less than 25 percent. Our simplest measure of whether school *j* faced strong short term incentives during 2003-04, is Mrg_AYP_j, where:

$$Mrg\_AYP_j = \max_{k,s}\left(Mrg\_AYP_{jks}\right) * \min_{k,s}\left(1 - Low\_AYP_{jks}\right)$$

In other words Mrg_AYP_j indicates whether any group within the school was near the margin of state requirements in either math or reading and no group within the school had a very low probability of passing state requirements in either tested subject.

In our second stage, we use the proxies for school accountability incentives (Mrg_AYP_j) to predict various outcomes for student *i* attending school *j* located in state *q*. Our basic regression specification is shown by Equation 2:

$$(2)\ Y_{ijs} = \delta_q + \lambda Mrg\_AYP_j + \phi Z_{ij} + \rho W_{j2002} + \zeta_{ijs}\,.$$

Y_{ijs} is a student-level outcome of interest (e.g., math achievement), δ_q is a state fixed effect, Z_{ij} is a set of student characteristics, including prior achievement and other lagged outcome measures, and the W_{j2002} vector of school-level demographic variables is the same as in equation 1. Table 3 sets out the student, family, and school characteristics we employ as controls in equation 2.

The coefficient of interest is λ, which measures the average impact of NCLB accountability pressure on the outcome variable (Y_{ijs}). The variation that identifies the coefficient λ is at the school level, so we adjust our standard errors for school-level clustering. The control variables for student and school characteristics and the state fixed effects are used to account for any confounding influences on the outcomes of interest and to isolate the effects of being enrolled in a school on the margin of satisfying AYP.

5.1 First Stage Estimates of Schools on the Margin of AYP

Our current analysis is based on estimates of the probability of making AYP for schools in 23 states.[16] A summary of the results from our first stage analyses is shown in Table 4. Among these schools, 16 percent are estimated to be on the margin of making AYP, as defined above, while 2 percent of schools are estimate to have a low probability of making AYP (i.e., at least one group has an estimated probability below 25 percent in either math or reading). Among schools we place on the AYP margin, 43 percent made AYP in 2004. Of those we estimate to have a low probability of making AYP, 8 percent made AYP in 2004. Of the remaining schools, for whom none of their subgroups had a predicted probability of passing under 75 percent in either math or reading, 94 percent made AYP. Thus, our first stage specification seems to work well in predicting which schools were at risk for failing to make AYP. Of course, whether this risk was foreseeable in advance is an open question. To the degree with which we misclassify the

[16] Since our second stage concerns the effects of NCLB on ECLS students' academic and behavioral outcomes in the fifth grade (2003-04), we restrict our sample of schools in the first stage to non-special education schools with at least five fifth graders in the school year 2003-04.

schools which felt they were likely to be on the margin of making AYP, our second stage estimates will suffer from bias due to measurement error. If these errors are uncorrelated with our outcome measures, as we believe is likely, then it will bias our estimates of the impact of NCLB pressure towards zero.

Our examination of schools in 23 states reveals that, with the exception of white and economically disadvantaged students, the majority of student subgroups are not numerically significant.[17] For example, only 15 percent of approximately 33,000 schools had a sufficient number of IEP (special education), students to be held accountable for that group's performance on state AYP tests. However, this rate varies across states depending on their minimum subgroup size requirements. For example, two percent of IEP subgroups in sample schools in Tennessee counted towards AYP, whereas 39 percent did in North Carolina, and 85 percent did in Florida. The two most common numerically significant subgroups are economically disadvantaged students and white students. Black and Hispanic subgroups were numerically significant in about one in four schools, while disabled and limited English proficient students were numerically significant in about one in six schools.

Also shown in Table 4 are the proportion of numerically significant subgroups with a predicted probability of making AYP in math and reading between 25 and 75 percent or less than 25 percent respectively. The probability of having a moderate or low chance of making AYP, conditional on numerical significance, accords largely with our priors. For example, disabled students have the highest predicted probabilities of falling into these categories in both subjects, while white students have extremely low predicted probabilities of being on the margin of failing AYP in either math or reading. Also, Limited English proficient, Hispanic, and Asian students,

[17] Schools were restricted to non-special education schools that had at least five fifth graders in the 2004 school year.

conditional on numerical significance, are more likely to have moderate or low probabilities of making AYP in reading than in math.

5.2 Baseline Estimates of NCLB on Student Achievement

Table 5 displays our very preliminary second stage results. Column 1 of Table 5 displays our estimates of equation 2, and the remaining columns displays estimates from analogous models which replace the Mrg_AYP_j indicator with a subject-specific indicator Mrg_AYP_{js}, (based on whether the schools has any accountable subgroups near the margin of the states' requirements for performance in a specific subject). As of this draft, we employ linear probability models to estimate effects on dummy dependent variables, but using mean marginal effects from probit model estimation would not lead to substantially different findings. Also, as of this draft, our second stage standard errors do not account for prediction error in the first stage equation.

All standard errors reported in the outcomes table (Table 5) adjust the standard errors for clustering at the school level, except for the school-level outcomes models which instead adjust for state level clustering. The student-level results also use the appropriate panel sampling weights which will ultimately make the estimates nationally representative for students in public schools in the 2003-04 school year who attended first grade during the 1999-2000 school year.

The first three rows of Panel A of Table 5 reveal that AYP pressure does not have a strong effect in either direction on mean student achievement growth on low-stakes exams in reading, math, or science. None of the estimates is statistically significant at the .10 level. It will be important to determine whether these estimates become larger or more precise when we use the complete ECLS sample and test more refined ways of measuring NCLB pressure.

Even without a strong mean impact on low-stakes exam performance, it is possible that accountability pressure influences the acquisition of specific types of skills. There could be disproportionate growth in relatively basic skills that would be important for students to pass their states' NCLB exams, at the expense of growth in more advanced skills. The next rows of Panel A display results of models estimating the impact of AYP pressure on the likelihood that students have mastered specific skills in reading and math, where mastery is defined as a greater than 90% probability that the student has mastered the skill according to ECLS test diagnostics. The results do not suggest strong effects on mean rates of student skill acquisition. For the effects of same-subject incentives (the last two columns), only one of the six coefficients is statistically significant at the .10 level (reading extrapolation skills), though it does suggest a rather large 4.9 percentage point decline in acquisition rates. In future drafts, we plan to examine the academic progress of students with various levels of lagged achievement and to test whether there are substantial effects in either subject when students are near their states' passing threshold for a particular subject.[18]

5.3 Effects of NCLB on Attitudes, Behavior, and Resources

The remainder of Panel A in Table 5 examines student-level outcomes concerning students' behavior, attitudes, and services received. We examine three indicators of whether students are having difficulties with behavior and attitudes: students' confidence and interest in reading, their confidence and interest in math, and whether they are having externalized behavior problems (e.g., anger or distractibility). Each of these three outcomes is based on student

[18] For most of the ECLS sample, will be able to identify which students are near their states' cutoffs by linking the ECLS students' 3rd grade national percentile equivalency for ECLS test scores to estimates of the national percentile equivalence of their states' passing threshold. A very helpful report published by the Institute for Education Sciences (2007) provides a crosswalk between most states' passing cutoff standards and 4th grade 2004 NAEP reading and math test scores, and we have obtained centile equilvalent estimates for the distribution of NAEP scores.

responses to a battery of related questions on these topics, and we classified students as having difficulties in these areas if the ECLS-published index based on these student responses placed them in roughly the "worst" quintile nationally. Interestingly, our measure of AYP pressure in reading is associated with a 7.5 percentage point increase (*p*-value=.03) in the likelihood that students struggle with their confidence/interest in reading, but AYP pressure in math leads to a 5.3 percentage point decrease in the likelihood that students struggle with their confidence/interest in math (*p*=.12). These results are in the same direction as the estimated effects of AYP pressure on the low-stakes exam scores in these subjects, (though the exam results were not as statistically significant). School-wide NCLB incentives do not appear to influence the likelihood that students have behavioral problems.

The results for student-level services outcomes are shown at the bottom of Panel A in Table 5. Attending a school that is on the margin for making AYP is associated with a reduction in the rate with which students receive small group instruction or individual tutoring in reading, even if the school is on the AYP margin for reading performance in particular. While this result is somewhat surprising, it is possible that schools become more selective in terms of which students they select for extra help in reading and reduce overall reading resources in order to best cater to the students who are on the margin for passing their states' NCLB reading exams. Being on the margin for AYP also substantially reduces the likelihood that students meet with school counselors, though this estimate has modest statistical significance.

Panels B and C in Table 5 examine the attitudes and instructional strategies of reading and math teachers. The same teacher often instructs ECLS students in both subjects, so there may be some overlap across the sample of teachers used in these panels. However, the outcome measures are based on subject-specific questions, so that the same teacher may provide different information about different subjects. Consistent with the idea that our measure of schools being on the AYP margin should induce principals and teachers to more heavily emphasize students' test performance, both reading and math teachers spend a greater amount of time preparing students to take tests in those subjects when the school is on the AYP margin in those subjects. Reading teachers spend more than three additional hours on test preparation when their school is on the AYP margin in reading (p=.09), while math teachers spend 2.7 additional hours when their school is on the AYP margin in math (p=.21). Consistent with the literature on NCLB pressure leading to prescriptive classroom practices, being on the AYP margin makes math teachers feel that they have less control over classroom decisions concerning curriculum, pedagogy, and discipline (p<.10).

There are numerous additional teacher survey questions in the ECLS that short term school accountability incentives could theoretically influence. In the interest of parsimony and of avoiding finding incidental, spurious effects, we focus on a few responses that might be particularly relevant in an environment in which high-stakes test outcomes are increasingly important. Interestingly, neither reading nor math teachers report a substantial change in the rate of grouping students by their ability in class. It is possible that accountability incentives lead to greater levels of ability tracking across (rather than within) classrooms, but the ECLS data are ill-equipped to explore that issue. There is also not any substantial change in the likelihood that either reading or math teachers report feeling that test scores help to guide their instruction. Reading teachers are not much more likely to focus more on non-fiction reading than on literature when their school is on the AYP margin, loosely suggesting that added concern over the high-stakes exams does not correspond with greater concern over students' abilities to read particular types of material.

Panel D in Table 5 displays the estimated effects of being on the AYP margin on various school-level services and attitudes. The first three outcomes are based on how the majority of surveyed teachers at the school responded, while the final two outcomes are based on surveys of one administrator per school. The most striking result in Panel E is that, when their school is on the AYP margin, teachers are more likely to express that their school administrators are not very well adept at handling outside pressure. While our current estimate is only marginally statistically significant (*p*=.155), it is very large, representing a more than five percentage point increase for an outcome with a frequency of only 19 percent. This again verifies that the schools that we have classified as on the margin for AYP are feeling the heat created by greater short term accountability incentives. The remainder of Panel E suggests that these schools do not substantially alter their provision of physical education classes, recess, or gifted and talented programs, and are not changing the length of the school year.

6. Next Steps

As we continue our analysis, we will augment our basic specification in a number of ways to further investigate the manner in which accountability incentives impact student outcomes. First, we will examine whether the impact of NCLB depends on the grade levels tested to determine AYP. Until the school year 2005-2006, when states were required to implement testing in grades three through eight, there was considerable variation across states in which grades were tested in accordance with NCLB. Ballou and Springer (2008) use this type of variation to identify the impact of high-stakes NCLB testing on students' low-stakes test performance in seven states. Of the 40 states in the ECLS data, our preliminary counts suggest that 25 states tested students in fourth grade for AYP in 2002-03, and 23 states tested students in fifth grade in 2003-04.[19] In future drafts, we can therefore test whether ECLS students experience different outcomes when they are exposed to either zero, one, or two years of high-stakes testing during the 2002-03 and 2003-04 school years.

We will also employ more nuanced definitions of AYP pressure to test whether attending a school where the *focal student's* subgroup(s) (i.e., not just the school-wide definition we employ in this paper) are on the margin of passing or failing AYP affects those students' outcomes. As described in our conceptual framework, we hypothesize that students who belong to numerically significant student subgroups that are at risk of failing AYP will receive the strongest pressure and school responses.

We will explore additional forms of within-school variation in accountability incentives, similar to those examined in Reback (2008). These issues include whether the impact of NCLB differed for students who were in the terminal grade of their school, or whether there are cross-subject effects of NCLB incentives.

Finally, another important test we will perform is whether the impact of accountability incentives differs if a school receives federal funding through the Title I program. One mechanism by which failing to make AYP can reduce school resources is the redirection of Title I funding to Supplemental Education Services from outside vendors.[20] Thus, there may be larger effects on school resource allocation and student outcomes for schools near the margin of making AYP if these schools receive Title I funding.

[19] Note that most states tested the same grades in both math and reading, but not always. For example, AYP determinations in Kentucky were based on reading tests in grades 4 and 7 and math tests in grades 5 and 8.

[20] For academic impacts of Supplement Education Services (SES), see Heinrich, Meyer, and Whitten (2008) and Springer, Pepper, and Ghosh-Dastigar (2009). Heinrich, Meyer, and Whitten (2008) first review findings from their parent and students surveys in Milwaukee about awareness and take up of supplemental education services. They also examine the impacts of supplemental education services on student achievement, finding no effects of any participation in supplemental education services on middle and high school students' achievement. However, Springer, Pepper, and Ghosh-Dastigar (2009) found positive effects on elementary and middle school students' test score gains in math, using longitudinal data from one large urban school district.

7. Conclusion

Policymakers are currently considering reforming *No Child Left Behind,* and there has been a recent movement toward aligning state education standards with national standards. As of June, 2009, 46 states had agreed to work toward setting common academic standards. One of the intended goals of the U.S. Department of Education's new "Race to the Top" funding program is to entice states to alter their standardized testing practices.[21] These reform efforts are partly motivated by concerns about variability in the difficulty of states' NCLB exams, with the result that AYP fails to reflect real differences in student achievement across states. However, our analyses suggest that the difficulty of state tests bears little relation to the percent of schools passing AYP. Instead, cross-state differences in the fraction of schools making AYP are largely driven by the minutiae of states' AYP rules rather than by differences in exam difficulty or required pass rates. If policymakers would like establish uniformity across states' respective AYP standards, then NCLB reforms must address not only the nature and difficulty of the high-stakes exams but also the details of AYP rating formulae including minimum student subgroup size, confidence intervals applied to groups' raw proficiency rates, and school safe harbor provisions.

In our study, we compile an extensive national dataset concerning NCLB performance and standards and we exploit the extensive cross-state variation in the nature of AYP ratings to examine the impact of incentives under NCLB. To our knowledge, this is the first nationally representative, longitudinal study of the impact NCLB incentives on school services and student outcomes. Based on our preliminary analysis using a sample from 23 states, we find changes in resource allocation and in teacher attitudes and behavior when schools face relatively strong

[21] See, for example, Sawchuk's recent article in *Education Week* (8/5/09) for a discussion of Race to the Top and the debate over testing reform.

short term NCLB incentives. In particular, we find that teachers devote more time preparing students to take standardized tests and feel that they have less control over classroom level decisions. Students are also less likely to receive tutoring or small group instruction in reading, and students' confidence and interest increases for math but not for reading.

Our preliminary findings suggest that these changes in practices and attitudes do not translate into meaningful average changes in students' test score growth on low-stakes exams in reading, math, or science. As discussed in the previous section, we plan to explore whether responses to NCLB incentives influence the academic growth of particular types of students. A solid understanding of the impact of school accountability pressure under No Child Left Behind should be central to guiding accountability reforms designed to better align incentives with desired outcomes.

References

Ballou, D. and Springer, M.G., Vanderbilt University, “Achievement Trade-Offs and No Child Left Behind.”

Center on Education Policy. (2008). Instructional time in elementary schools: A closer look at changes for specific subjects. Retrieved May 18, 2008, from Center on Education Policy’s web site: http://www.cep-dc.org/

Center on Education Policy. (2007). Answering the question that matters most: Has student achievement increased since No Child Left Behind? Retrieved April 4, 2008, from Center on Education Policy’s web site: http://www.cep-dc.org/

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

Chiang, H. (2008). How accountability pressure on failing schools affects student achievement. Harvard University mimeo. Retrieved June 4, 2008, from Hanley Chiang’s website: http://www.people.fas.harvard.edu/~hchiang/

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

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

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

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

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

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

Heinrich, C., Meyer, & Whitten. (2008). University of Wisconsin-Madison – “Supplemental Educational Services under No child Left Behind: Who Signs Up, and What do They Gain?”, LaFollette School Working paper No 2008-005.

Jacob, B. (2005). Accountability, incentives and behavior: the impact of high-stakes testing in the Chicago Public Schools. *Journal of Public Economics 89*(5-6), 761-796.

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

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

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

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

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

Rouse, C., Hannaway, J., Goldhaber, D., & Figlio, D. (2007). Feeling the Florida heat? How low-performing schools respond to voucher and accountability pressure. *National Bureau of Economic Research*, working paper 13681. Retrieved May 18, 2008 from NBER's web site: http://www.nber.org/papers/w13681

Sawchuk, S. (8/5/09). Experts Hope Federal Funds Lead to Better Tests. *Education Week.*

Springer, M.G. (2008). The influence of an NCLB accountability plan on the distribution of student test score gains. *Economics of Education Review* 27(5), 556-563.

Springer, M.G., Pepper, M.J., and Ghosh-Dastidar , B. (2009). "The Effect of Supplemental Educational Services (SES) on Student Test Score Gains."

Figure 1: Distribution of AYP Failure Rates Across States, 2004

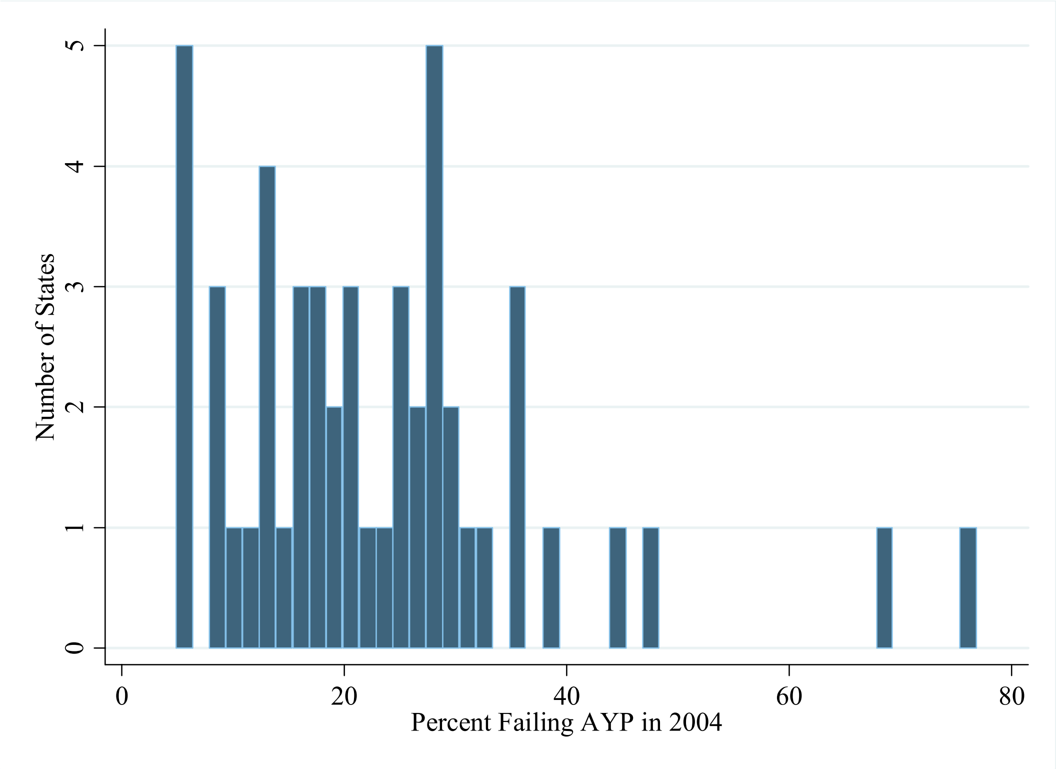
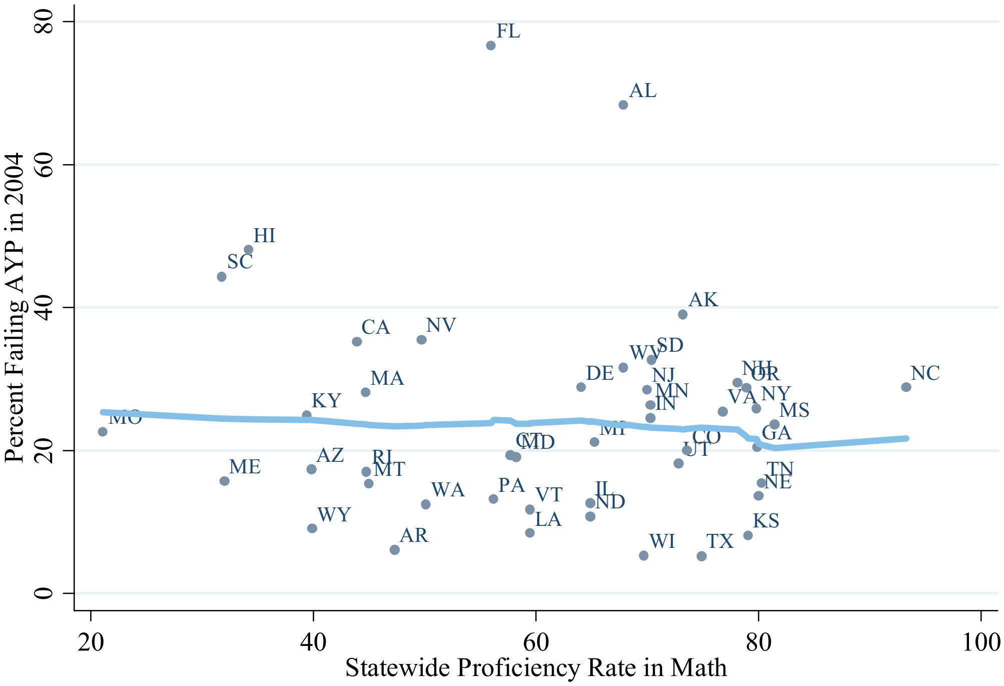
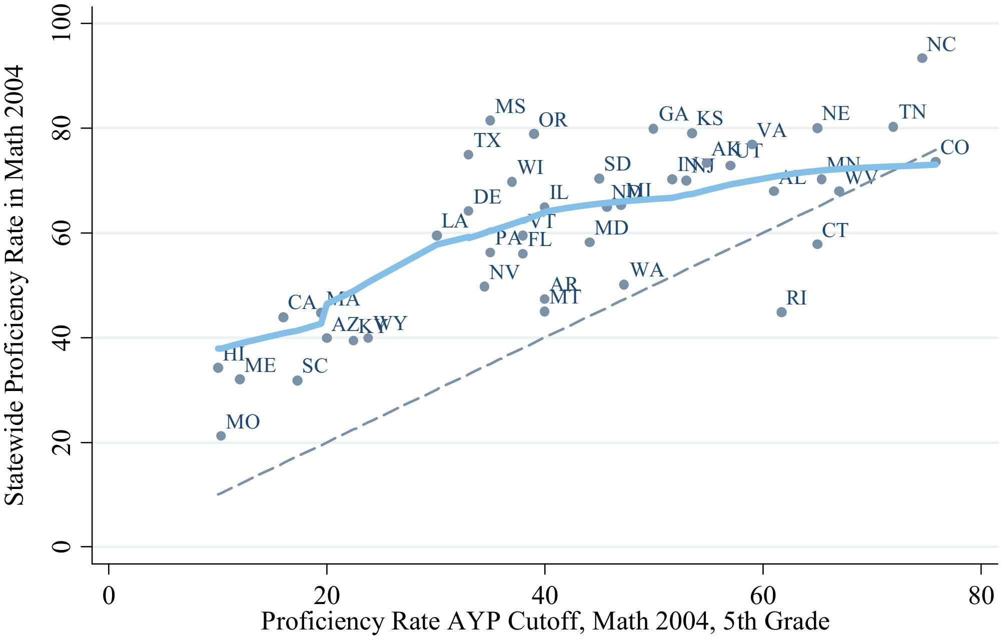


Figure 2a: AYP Failure Rate vs. Fraction Proficient by State, Math, 2004



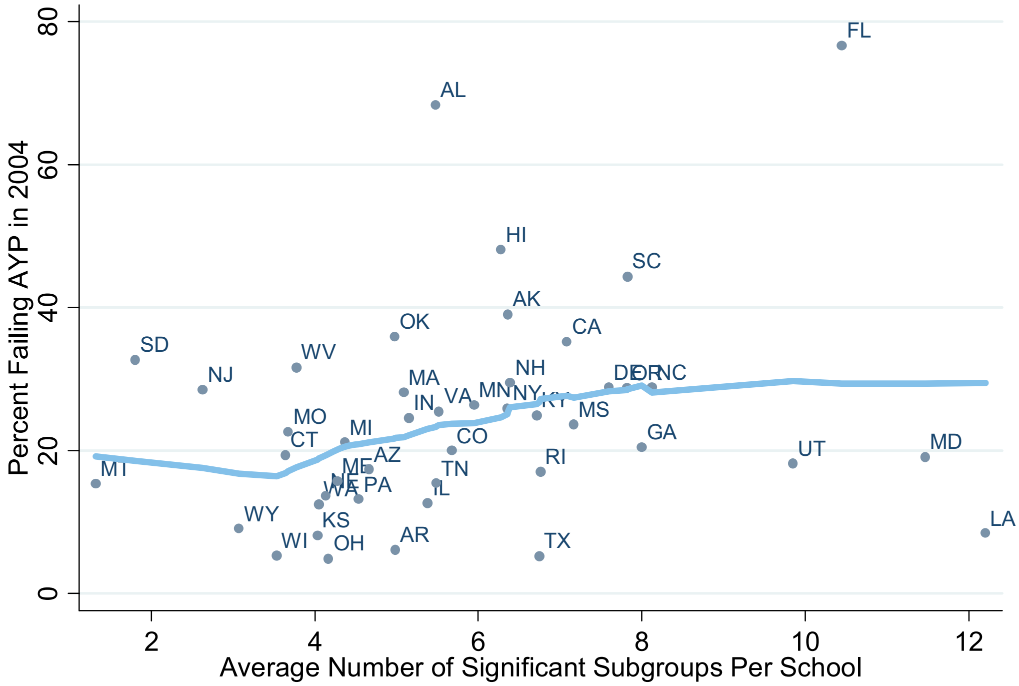
Note: Data from five states are unavailable. Line represents a locally weighted regression.

Figure 2b: Fraction Proficient by State vs. Required Proficiency Rate, Math, 2004



Note: Data on statewide proficiency missing for two states. Three states use index points instead of proficiency. Line represents a locally weighted regression.

Figure 3: AYP Failure Rates and State Variation in Accountable Student Subgroups



Note: Data from six states are unavailable. Line represents a locally weighted regression.

Table 1: Characteristics of Schools Nationwide by AYP Status, 2003-04

	Unweighted		*Weighted by Enrollment*	
	Failed AYP	Made AYP	Failed AYP	Made AYP
Number of Schools	19,483	65,332	19,483	65,332
Average Enrollment	727	500	1,216	800
Student/Teacher Ratio	17.4	16.1	18.4	17.2
Percent of Schools…				
School Title I Eligible	39.9%	29.4%	34.8%	27.6%
Located in City	36.5%	23.1%	39.2%	27.7%
Located in Urban Fringe	31.9%	33.3%	37.8%	41.1%
Located in Town or Rural Area	31.5%	43.5%	23.0%	31.2%
Serving Primary Grades	36.0%	64.1%	27.1%	56.3%
Serving Middle Grades	27.4%	15.5%	29.4%	17.0%
Serving High School Grades	27.9%	17.1%	39.1%	24.5%
Ungraded/Other	8.5%	3.2%	4.4%	2.2%
Percent of Students…				
Eligible for Free/Reduced Lunch	54.2%	39.9%	49.3%	37.0%
White	46.5%	67.7%	46.2%	63.6%
Black	25.6%	13.2%	24.1%	14.1%
Hispanic	20.8%	13.6%	23.7%	16.3%
Asian	3.4%	3.4%	4.3%	4.5%
Native American	3.3%	1.7%	1.5%	1.1%

Note: Data on school characteristics taken from the Common Core of Data, 2003-04. All U.S. public schools are included, regardless of their grade span. For schools in Tennessee, data on student ethnicity taken from 1998-99 and data on free/reduced price lunch eligibility is unavailable.

Table 2: Descriptive Statistics for Current ECLS 2004 Sample

Panel A: Student-Level Outcomes (N=5650)	Mean	SD
Reading Z-score	0.17	0.90
Math Z-score	0.16	0.93
Science Z-score	0.15	0.91
Student has Mastery of… *(0/1 Indicators)*		
Math: Place value	0.63	
Math: Rates and Measurement	0.24	
Math: Fractions	0.06	
Reading: Literal Inference	0.72	
Reading: Extrapolation	0.54	
Reading: Evaluation	0.05	
Student has difficulties with…		
Confidence/Interest in Reading	0.22	
Confidence/Interest in Math	0.20	
Behavior	0.23	
Student receives the following services…		
Meets with a Counselor	0.12	
Small Group/Individual Reading Tutoring	0.19	
Panel B: Reading Teacher-Level Outcomes (N=2050)		
Hours Spent on Test Preparation	13.48	16.93
Control over Class Curriculum, Pedagogy, Discipline	0.49	
Uses Ability Grouping at Least Once/Week	0.56	
Test Scores Help Guide Instruction	0.36	
More Time on Non-fiction than Literature	0.44	
Panel C: Math Teacher-Level Outcomes (N=1380)		
Hours Spent on Test Preparation	13.08	16.35
Control over Class Curriculum, Pedagogy, Discipline	0.49	
Uses Ability Grouping at Least Once/Week	0.42	
Test Scores Help Guide Instruction	0.36	
Panel D: School-Level Outcomes (N=1380)		
≥ 50% of Teachers Sampled in the School Say That…		
School Administrator Handles Outside Pressure Well	0.19	
Students Have Physical Education < 3 Days/Week	0.67	
School Administrator Responses		
# of School Days in the year	279	2.7
School has gifted/talented program	0.76	

Note: In seven schools, administrator questionnaires but no teacher questionnaires were completed. Means and standard deviations for the student-level variables are nationally representative, computed using the same panel sampling weights as we use in our main analyses. To comply with restricted-use data guidelines, sample sizes are rounded to the nearest 10.

Table 3. Descriptive statistics for control variables

Student characteristics	**Mean**	**SD**
Reading IRT scale score in spring 2000	69.94	21.47
Math IRT scale score in spring 2000	56.63	15.98
Reading IRT scale score in spring 2002	115.86	25.28
Math IRT scale score in spring 2002	91.33	21.82
African American	18%	
Hispanic	21%	
Asian	3%	
Other	3%	
Female	50%	
Age (in days)		135.84
Family characteristics		
Two parent household	66%	
Mother's education level unknown	9%	
Mother has at least a high school diploma	89%	
Mother possesses a B.A.	25%	
Family income missing	16%	
Family income under $20,000	20%	
Family income $20,000 -$35,000	21%	
Family income $35,000 - $50,000	17%	
Family income $50,000 - $75,000	17%	
Family income $75,000 - $100,000	12%	
School characteristics		
Percent LEP students	6%	13%
Number of enrolled students	583.96	254.50
Missing Title I eligibility data	8%	
Eligible for Title I	62%	
Missing economically disadvantaged data	2%	
Percent economically disadvantaged students	45%	30%
Percent Asian students	4%	8%
Percent Hispanic students	17%	25%
Percent African American students	19%	26%

N = 5,340

Note: Although the table currently presents summary statistics for IRT scale scores, we convert these to nationally representative Z-scores before including them as control variables in the analyses. The models with science test scores as dependent variables also control for cubic terms for student-level Z-scores for the 2002 science test administered as part of the ECLS-K. To comply with restricted-use data guidelines, the sample size is rounded to the nearest 10.

Table 4: First Stage Predictions of Being on the Margin of Failing AYP

Panel A: School-wide Outcomes	Predicted Marginal	Predicted Low Probability	Predicted High Probability
Proportion of Schools	16.1%	2.0%	81.9%
Actually Passed AYP 2004	43.2%	7.8%	93.5%

Panel B: Subgroup Outcomes		Conditional on Numerical Significance			
	Numerically Significant Subgroup	Predicted Marginal		Predicted Low Probability	
		Math	Reading	Math	Reading
Overall School Population	98.4%	3.3%	5.1%	0.3%	0.4%
Economically Disadvantaged	60.4%	7.0%	9.9%	0.2%	0.4%
Limited English Proficient	18.0%	9.9%	37.5%	0.1%	0.7%
Disabled	15.1%	38.0%	39.4%	10.0%	7.4%
White	67.4%	0.0%	0.0%	0.0%	0.0%
Black	24.1%	25.2%	16.7%	1.4%	0.7%
Hispanic	27.8%	4.1%	13.3%	0.1%	0.2%
Asian/Pacific Islander/Filipino	4.1%	0.4%	11.3%	0.0%	0.1%
Native American	0.6%	33.3%	29.0%	0.0%	4.8%

Note: Current analysis limited to 23 states.

Table 5: Preliminary Second-stage Regression Results

	School is on the Margin for AYP in…				
	Either Subject		Reading		Math
Panel A: Student-Level Outcomes (N=5,650)					
Reading Z-score	-0.004		-0.032		
	(0.035)		(0.036)		
Math Z-score	-0.023				0.036
	(0.039)				(0.044)
Science Z-score	0.019				0.022
	(0.041)				(0.051)
Student has Mastery of… *(0/1 Indicators)*					
Reading: Literal Inference	-0.005		-0.008		
	(0.028)		(0.030)		
Reading: Extrapolation	-0.040		-0.049	*	
	(0.026)		(0.027)		
Reading: Evaluation	-0.007		-0.006		
	(0.010)		(0.010)		
Math: Place value	-0.031				0.008
	(0.026)				(0.029)
Math: Rates and Measurement	-0.015				0.005
	(0.023)				(0.025)
Math: Fractions	-0.012				-0.021
	(0.012)				(0.017)
Student has difficulties with…					
Confidence/Interest in Reading	0.049		0.075	**	
	(0.032)		(0.034)		
Confidence/Interest in Math	-0.030				-0.053
	(0.030)				(0.034)
Behavior	0.005				-0.009
	(0.034)				(0.041)
Student receives the following services…					
Meets with a Counselor	-0.045				-0.039
	(0.028)				(0.036)
Small Group/Individual Reading Tutoring	-0.087	**	-0.077	*	
	(0.038)		(0.040)		
Panel B: Reading Teacher-Level Outcomes (N=2050)					
Hours Spent on Test Preparation	2.525		3.024	*	
	(1.550)		(1.763)		
Limited Control over Classroom Decisions (Curriculum, etc.)	0.057		0.036		
	(0.035)		(0.036)		
Uses Ability Grouping at Least Once/Week	0.005		-0.005		
	(0.035)		(0.035)		
Test Scores Help Guide Instruction	0.026		0.013		
	(0.032)		(0.034)		
More Time on Non-fiction than Literature	0.013		0.014		
	(0.033)		(0.034)		

Table 5 (continued)

	School is on the Margin for AYP in…		
	Either Subject	Reading	Math
Panel C: Math Teacher-Level Outcomes (N=1380)			
Hours Spent on Test Preparation	2.406		2.713
	(1.728)		(2.139)
Limited Control over Classroom Decisions (Curriculum, etc.)	0.064 *		0.031
	(0.039)		(0.049)
Uses Ability Grouping at Least Once/Week	-0.026		-0.100 **
	(0.044)		(0.048)
Test Scores Help Guide Instruction	0.009		0.029
	(0.036)		(0.046)
Panel D: School-Level Outcomes			
> 50% of Teachers Sampled in the School Say That… (N=1380)			
School Administrators Do Not Handle Outside Pressure Well	0.052		
	(0.036)		
Students Have Physical Education < 3 Days/Week	-0.028		
	(0.040)		
Students Do Not Have Daily Recess	-0.033		
	(0.034)		
Reported by School Administrator… (N=1240)			
School Days in the Year	-0.058		
	(0.131)		
School Has Gifted/Talented Program	-0.033		
	(0.029)		

Note: To comply with restricted-use data guidelines, the sample sizes are rounded to the nearest 10.

Appendix 1: Data Sources for First Stage Analyses

	Available from Existing Database	We Have Collected	Not Available
School Made AYP in 2003-04	39	1	0
Subgroup Made AYP in 2003-04	31	7	2
Percent Proficient by Subgroup in 2003-04	14	25	1
Subgroup Size 2003-04	6	27	7

Note: Based on the 40 states sampled in the ECLS database. Existing databases refer to Standard & Poor's School Data Direct and the National AYP and Identification Database.

Faculty and Research Affiliates

Matthew G. Springer
Director
National Center on Performance Incentives

Assistant Professor of Public Policy
and Education
Vanderbilt University's Peabody College

Dale Ballou
Associate Professor of Public Policy
and Education
Vanderbilt University's Peabody College

Leonard Bradley
Lecturer in Education
Vanderbilt University's Peabody College

Timothy C. Caboni
Associate Dean for Professional Education
and External Relations
Associate Professor of the Practice in
Public Policy and Higher Education
Vanderbilt University's Peabody College

Mark Ehlert
Research Assistant Professor
University of Missouri – Columbia

Bonnie Ghosh-Dastidar
Statistician
The RAND Corporation

Timothy J. Gronberg
Professor of Economics
Texas A&M University

James W. Guthrie
Senior Fellow
George W. Bush Institute

Professor
Southern Methodist University

Laura Hamilton
Senior Behavioral Scientist
RAND Corporation

Janet S. Hansen
Vice President and Director of
Education Studies
Committee for Economic Development

Chris Hulleman
Assistant Professor
James Madison University

Brian A. Jacob
Walter H. Annenberg Professor of
Education Policy
Gerald R. Ford School of Public Policy
University of Michigan

Dennis W. Jansen
Professor of Economics
Texas A&M University

Cory Koedel
Assistant Professor of Economics
University of Missouri-Columbia

Vi-Nhuan Le
Behavioral Scientist
RAND Corporation

Jessica L. Lewis
Research Associate
National Center on Performance Incentives

J.R. Lockwood
Senior Statistician
RAND Corporation

Daniel F. McCaffrey
Senior Statistician
PNC Chair in Policy Analysis
RAND Corporation

Patrick J. McEwan
Associate Professor of Economics
Whitehead Associate Professor
of Critical Thought
Wellesley College

Shawn Ni
Professor of Economics and Adjunct
Professor of Statistics
University of Missouri-Columbia

Michael J. Podgursky
Professor of Economics
University of Missouri-Columbia

Brian M. Stecher
Senior Social Scientist
RAND Corporation

Lori L. Taylor
Associate Professor
Texas A&M University

NATIONAL CENTER ON

Performance Incentives

EXAMINING PERFORMANCE INCENTIVES
IN EDUCATION

National Center on Performance Incentives
Vanderbilt University Peabody College

Peabody #43
230 Appleton Place
Nashville, TN 37203

(615) 322-5538
www.performanceincentives.org

VANDERBILT
PEABODY COLLEGE