Published in final edited form as:

Sociol Educ. 2014 April ; 87(2): 125–141. doi:10.1177/0038040714525787.

Measure for Measure: How Proficiency-Based Accountability Systems Affect Inequality in Academic Achievement

Jennifer Jennings¹ and

New York University, Department of Sociology, New York, NY, USA

Heeju Sohn

University of Pennsylvania, Graduate Group of Demography, Department of Sociology,
Philadelphia, PA, USA

Abstract

How do proficiency-based accountability systems affect inequality in academic achievement? This paper reconciles mixed findings in the literature by demonstrating that three factors jointly determine accountability's impact. First, by analyzing student-level data from a large urban school district, we find that when educators face accountability pressure, they focus attention on students closest to proficiency. We refer to this practice as *educational triage*, and show that the difficulty of the proficiency standard affects whether lower or higher performing students gain most on high-stakes tests used to evaluate schools. Less difficult proficiency standards decrease inequality in high-stakes achievement, while more difficult ones increase it. Second, we show that educators emphasize test-specific skills with students near proficiency, a practice that we refer to as *instructional triage*. As a result, the effects of accountability pressure differ across high and low-stakes tests; we find no effects on inequality in low-stakes reading and math tests of similar skills. Finally, we provide suggestive evidence that instructional triage is most pronounced in the lowest performing schools. We conclude by discussing how these findings shape our understanding of accountability's impacts on educational inequality.

Most educational accountability systems evaluate schools based on the percentage of students performing above a score deemed “proficient.” This focus on proficiency potentially influences how educators distribute instructional resources, as schools get no credit for raising the achievement of students already above or far below proficiency. Teachers thus have a short-term incentive to focus on “bubble” students, those close to the proficiency cut score, especially when increasing a small number of students' scores can improve the school's accountability rating. This practice, which has been referred to as “educational triage,” may have important implications for educational stratification.

Though the short-term incentives created by proficiency-based accountability systems are clear, the literature on their distributional effects is decidedly mixed. By the distributional effects of accountability systems, we refer to effects that vary across the test score distribution. In cases where the proficiency standard is high relative to the achievement of

¹Corresponding author, Department of Sociology, New York University, 295 Lafayette Street, New York, NY 10003, USA. jj73@nyu.edu.

students, schools may neglect students far below proficiency. As a result, students at the 50th percentile of the test score distribution may benefit, while those at the 10th percentile do not. Alternatively, where the proficiency standard is set low, schools may focus less attention on higher-achieving students. In this case, students at the 10th percentile of the distribution may improve, while those at the 90th percentile do not. Qualitative and survey studies have found that educators focus more attention on students close to proficiency (Author; Gillborn and Youdell 2000; Hamilton et al. 2007; Weitz-White and Rosenbaum 2007), sometimes to the detriment of the lowest performing students, but quantitative studies have come to a variety of conclusions about accountability's consequences on inequality in academic achievement.

In this paper, we propose that these mixed results can be reconciled by understanding that three factors influence the effects of proficiency-based accountability systems on inequality in math and reading skills. The first is the percentage of the population that falls below the proficiency standard. That proficiency standards differ dramatically across states makes the difficulty of the cut score (i.e. the score above which students must score to be considered proficient) a prime suspect for producing variation across studies. If teachers do direct resources to students close to proficiency, we expect the lowest scoring students to benefit when proficiency standards are low and to fall behind when proficiency standards are high.

The second factor is the extent to which gains on high-stakes state tests used for accountability generalize to other assessments. There is a substantial literature suggesting that schools focus on test-specific skills in order to increase high-stakes test scores (for an excellent summary, see Koretz 2008). Paying more attention to content that predictably appears on high-stakes tests or coaching students to respond to predictable recurrences in question structure may bolster scores without improving other measures of achievement. We call this practice *instructional triage*. As a result of instructional triage, studies using high-stakes measures may tell a different story about the effects of accountability on inequality than those using other measures of achievement, and come to different conclusions about accountability's impacts on inequality. This is important not only because it affects the inferences we make about accountability's effects on math and reading skills, but because high-stakes tests have multiple users. Students', teachers', schools', and parents' future educational decisions may be affected by high-stakes test scores, and policymakers may use these data to shape and legitimate future educational policies.

The third factor is the concentration of low-performing students in the school. We expect that educators' focus on students near proficiency will be most acute in schools where the number of students in academic need exceeds the resources available to serve them, and also predict that high-stakes gains are least likely to generalize to other assessments in the lowest-performing schools. Because low-performing schools face heightened pressure to increase scores of near-proficient students, we expect to see larger gains for "bubble" students in schools with concentrated academic disadvantage that are struggling to make accountability targets. Since these schools often have limited organizational capacity to systemically improve instruction, however, we expect these gains to largely reflect instructional triage.

In what follows, we first review the literature on the distributional consequences of accountability and the impact of accountability pressure on instruction. We propose that mixed findings on distributional effects emerge both from differences in the outcomes and the contexts studied. We then describe our analytic strategy, which uses variation in two cohorts' exposure to the No Child Left Behind Act to identify its effects. We exploit differences in the difficulty of the reading and math proficiency standards relative to the achievement distribution to establish how this factor impacts student performance. Unlike existing studies, however, we are able to evaluate how proficiency-based accountability systems affect two outcomes: state tests used for accountability ("high-stakes tests") and a second measure of achievement in which the stakes are low for individual teachers and schools (hereafter, "low-stakes tests").

Our analysis addresses three questions:

- To what extent does the difficulty of the proficiency standard influence the distributional effects of accountability systems on academic achievement?
- Are these distributional effects similar on high- and low-stakes tests?
- Do distributional effects on high- and low-stakes tests vary across low-, mid-, and high-performing schools?

The Consequences of Educational Accountability for Inequality in Academic Achievement

How is the achievement of students at different parts of the score distribution affected by proficiency-based accountability systems? While a number of studies have considered the average effects of state accountability systems (Carnoy and Loeb 2002; Dee and Jacob 2011; Hanushek and Raymond 2004; Hout and Elliott 2011; Wong, Cook, and Steiner 2009), whether these systems consistently help lower-, average-, and high-performing students is less clear.

We begin by reviewing the qualitative and survey literature on how educators respond to accountability systems. Ethnographic studies of educators' responses to accountability pressure in Texas, Chicago, and the UK have all demonstrated that educators focus their attention on students close to passing when they are rewarded for moving students to proficiency, and refer to this process as educational triage (Author; Gillborn and Youdell 2000; Weitz-White and Rosenbaum 2007). Survey evidence corroborates these findings. In the early period of NCLB implementation, the RAND three state study found that 77 to 90 percent of elementary school principals reported encouraging teachers to "focus their efforts on students close to meeting the standards," and 29-37 percent of teachers reported doing so (Hamilton et al. 2007). Reback, Rockoff, and Schwartz (2010) later merged these survey data with state test score data, allowing them to estimate the amount of accountability pressure schools faced. They found that in schools that had low probabilities of failing NCLB's Adequate Yearly Progress requirement, 26 percent of teachers reported focusing on students close to proficient. In schools more likely to fail AYP, 53 percent of teachers did.

These data suggest that accountability pressure is associated with how teachers allocate instructional resources between students.

A variety of scholars claim to test the “educational triage” hypothesis, but no agreement exists about the meaning of this concept. Disagreement partly arises from one group of scholars defining the term based on student outcomes, while others rely on inputs. In the qualitative studies from which this concept originally derives, scholars described educational triage based on observed educator behavior and organizational strategies (inputs), and referred to educators allocating resources in ways that maximize proficiency rates. For other scholars analyzing administrative data sets, educational triage implies a system of distributing educational resources within schools that benefits the test scores of students close to proficiency while harming those at the top and bottom of the performance distribution.

We believe that the operationalization that provides the most conceptual leverage uses the concept of educational triage to refer to schools and educators allocating resources to maximize proficiency rates. Consider two states in which all teachers adopt the same behavior of focusing resources on students closest to proficiency. In the first state, only 15 percent of all students are not proficient, so the lowest performing students in the state are the “bubble students.” In a second state, 85 percent of students are not proficient, so some of the state's highest performing students are the bubble students. The definition that we propose would call teachers' allocation process in both states educational triage, and is not dependent on the difficulty of the cut score relative to the achievement distribution of students in a state or in a school. Because our definition highlights both the allocation process *and* the outcomes of the allocation process, it can be used to describe behaviors observed in qualitative studies, and can also be detected in quantitative studies when we observe larger than expected gains near proficiency.

A large number of quantitative studies have addressed the issue of distributional effects. Figure 1 summarizes these studies by reporting the type of test analyzed in the study (i.e. low or high-stakes tests), the difficulty of the proficiency standard relative to the performance distribution, the time period studied, and key findings. In this table, the proficiency difficulty classification has three categories – high, mixed, and low. Low difficulty refers to studies in which more than 70 percent of students meet the proficiency standard across tested grades 3-8 in reading and math, on average, while studies in which fewer than 70 percent of students meet the proficiency standard are referred to as “high difficulty.” When studies include multiple states with a mix of high and low proficiency standards, we refer to these as “Mixed.”

A number of patterns emerge from this table. Low-performing students appear to consistently benefit when the proficiency standard is low (Clotfelter, Ladd, and Vigdor 2009; Ladd and Lauen 2009; Reback 2008), and lose ground when the proficiency standard is high or rising from low to high (Jacob 2005; Krieg 2008; Lauen and Gaddis 2012; Neal and Schanzenbach 2010). In addition, distributional effects on low-stakes tests, on average, appear to be more muted or inconsistent across grades and subjects (Dee and Jacob 2011;

Reback, Rockoff, and Schwartz 2008; Springer 2008), with the exception of Ballou and Springer (2011).

In the next section, we explore the mechanisms that may produce different distributional effects on high and low-stakes tests.

Instructional Responses to Accountability Pressure

Teachers can also change how they allocate instructional time to different parts of the state standards, in addition to changing how they distribute resources between students. This can be an effective strategy for raising scores because tests are based on a sampling principle. Students are administered a sample of items from a knowledge domain of interest (i.e. 8th grade mathematics), and a well-designed test allows us to make inferences from performance on this sample to the whole domain. When the stakes are high, as they have been for schools, teachers, and sometimes students in recent years, educators have an incentive to narrow instruction to focus on predictable material on state tests (Koretz 2008). The result is often score inflation, which occurs when test scores overstate students' skills in the tested area and do not allow us to make valid inferences of their skills in the larger domain that is measured.

Studies of standard coverage show that state tests do not sample the full state standards and are predictable in ways that facilitate test-specific instruction (Author et al.). Author et al. found that teachers respond to predictable test features by focusing on frequently-assessed standards, and this response is stronger in schools facing greater accountability pressure. Survey evidence largely confirms these findings. In the RAND study of NCLB implementation in three states, teachers reported that they identified “highly assessed standards” on which to focus their attention (Hamilton and Stecher 2007). Reback, Rockoff, and Schwartz’s study of the same data found that in schools facing substantial accountability pressure (those below the AYP margin), 84 percent of teachers reported focusing on topics emphasized on the state test, while 69 percent of those in schools at low-risk of failing AYP did.

A complementary group of studies on “teaching to the format” illustrates how the design of high-stakes tests may lead educators to focus on particular formats in their teaching that parallel those appearing on high-stakes tests (Pedulla et al. 2003; Shepard and Dougherty 1991). The most well-known example of this phenomenon comes from Shepard’s (1988) finding that students could effectively add and subtract decimals when they were presented in a vertical format, but struggled when decimals were presented in a horizontal format. While most “teaching to the format” studies pre-date the NCLB era, Reback, Rockoff, and Schwartz (2010) found that teachers in higher accountability pressure schools reported looking for “particular styles and formats of problems in the state test and emphasize“d” those in “their” instruction.” 100 percent of teachers in schools below the AYP margin reported doing so, while 67 percent of teachers at lower risk of failing AYP did.

Taken together, these studies suggest that predictable assessments create incentives for teachers to engage in *instructional triage*; that is, to focus on predictably-assessed content

and present questions as they will appear on the state test. We believe that this concept is a useful companion to educational triage, which instead describes *which students* are the focus of attention. Instead, instructional triage describes *what content and presentation of material* receives more attention. To the extent that students learn how to correctly answer questions when they are presented in a specific format, but struggle with the same skills when they are presented in a different format, instructional triage can create gains on the high-stakes test while leaving students' underlying skills largely unchanged.

In sum, though the literature on the distributional effects of accountability is mixed, we attribute these inconsistencies, in large part, to differences in proficiency standards across states and differences in the stakes associated with the outcome measured. Our study is the first to bring together these two issues and isolate the relevance of proficiency standard difficulty for inequality in academic achievement on both high and low-stakes tests.

Data and Methods

We analyze longitudinal student-level data from Houston Independent School District over the period 2001 through 2004, the period during which NCLB was first implemented. HISD is the seventh largest school district in the country and the largest in the state of Texas. Fifty-three percent of grades 3-8 students are Hispanic, while 32 percent are African-American, 10 percent are white, and 3 percent are Asian. Close to 80 percent of students are considered by the state to be economically disadvantaged, 27 percent have a Limited English Proficient classification, and 11 percent have a special education classification.

A unique feature of these data is the availability of high and low-stakes test scores for each student over the period before and after the assignment of the first “Adequate Yearly Progress” ratings under the No Child Left Behind Act. When we describe “stakes” in this paper, we are referring to the stakes for teachers and administrators, not the stakes for students. The Texas Assessment of Knowledge and Skills (TAKS), the district's “high-stakes test,” is administered to students in grades 3 to 11 in reading/English Language Arts, mathematics, writing, science, and social studies, although reading and math are the only subjects tested every year between grades 3 and 8. The Stanford Achievement Test is administered to all students in grades 1 to 11 in reading, math, language, science, and social science. HISD added the Stanford Achievement Test under pressure from a business task force that sought a nationally-normed benchmark test (McAdams 2000), and all students except for those with the most severe disabilities have been required to take the Stanford. Both the TAKS and the Stanford have flexible time limits. During this period, both tests were administered in the spring, from early March (Stanford) to mid- to late April (TAKS).

The TAKS represents the district's “high-stakes” test for several reasons. First, TAKS test scores are used to assess schools' success in meeting Adequate Yearly Progress (AYP) under NCLB. Second, proficiency rates on these tests have been an integral part of Texas' accountability system since 1994 (Reback 2008). Under this system—which served as the model for No Child Left Behind—schools and districts are labeled “exemplary,” “recognized,” “acceptable,” or “low-performing” based on their proficiency rates in each subject area. In most years, monetary rewards were available for high-performing or

improving schools, while low-performers were subject to sanctions, including school closure or reconstitution. Third, HISD operated a performance pay plan during this period that provided monetary rewards to schools and teachers for exemplary TAKS results.

The Stanford can be considered HISD's "lower-stakes" test, in that it is not tied to the state accountability system or teachers' performance pay during this period. However, this test plays several important roles in the district and is not a "no stakes" test for students. For example, it is used as one criterion for grade promotion in grades 1-8. HISD students are expected to perform above a minimum standard on the Stanford (e.g. one grade level below average or above) as well as the TAKS. In addition, the Stanford is used in decisions to place students in gifted and talented programs and special education. For our purposes, it is ideal that students have good reason to exert effort on both tests, but the significance of the tests for teachers and administrators varies.

The TAKS and Stanford tests are intended to test similar grade-level domains, but these are not identical. Because the Stanford is a proprietary test, we were not able to examine items directly to compare them with the TAKS items.² Both test contractors, however, report that their tests are aligned with the NAEP frameworks. We have only been able to identify one analysis (Hoey, Campbell, and Perlman 2001) that mapped the standards on the Texas Assessment of Academic Skills (TAAS) math test (in grade 4 only) to those covered on the Stanford and found considerable overlap, with 83 percent of the Texas standards represented on the Stanford. (The Stanford was a bit more inclusive, with 74 percent of Stanford standards represented on the TAAS).

Given our focus on the distributional effects on accountability, our study requires that students take the English version of both the Stanford and TAKS so that they can be located in one test score distribution. This requirement complicates the study of grades 1-5 because both the Stanford and the TAKS tests administer Spanish language versions in these grades. Approximately a third of students in grade 3 took the Spanish language version of these tests. The proportion of students decreases to 20 percent in fourth grade and to 6 percent in fifth grade. By grade 6 less than 1 percent of students take the Spanish versions. For this reason, we study student performance during the period between grade 6 and 8 where the vast majority of students are all taking the same English language version of the Stanford and TAKS, and do not include the small number of students taking the Spanish language version in our sample.

We estimate the distributional effects of NCLB on both high and low-stakes tests. Two main approaches have been used to estimate accountability's impact on student outcomes. The first approach has attempted to establish the *total* effect of having a stronger accountability system in place across all schools, where the counterfactual is either no accountability system or a weaker accountability system. The second approach involves answering a fundamentally different question. That is, given that an accountability system is already in place, how does facing heightened school accountability pressure within that system affect

²We asked Pearson Learning for permission to view these items so that we could map TAKS and Stanford items and compare their respective domains, but this request was not granted.

student outcomes? In these studies, schools facing less or no accountability pressure serve as the counterfactual for students in high pressure schools. Our study is in the tradition of the former, and attempts to identify the effect of having NCLB in place relative to a world where NCLB had not existed. We note, however, that because Texas implemented educational accountability in the 1990s, the pre-NCLB period already represents an era of heightened accountability, so the contrast we provide is not with a “no accountability” world.

Our strategy for identifying the effects of NCLB follows previous work by Neal and Schanzenbach (2010). We rely on a cohort design and study two cohorts who took both the same high and low-stakes tests in grades 6 and 8. Schools attended by the 2001 6th grade cohort were not yet rated under NCLB between 6th and 8th grade. However, the 2002 cohort entered 6th grade before the first AYP ratings were assigned to schools and completed 8th grade in the year after these ratings were assigned. In short, the difference between these cohorts is that the 2002 cohort was treated for one year after AYP ratings were first released, while the 2001 cohort was not. (See Appendix Table A1 for a figure outlining this strategy.) Both cohorts took all 6th grade tests under a pre-NCLB regime, which is crucial to the analytic strategy that we outline below.

We first assign students in the 2001 6th grade cohort to quintiles based on their 6th grade high-stakes test score, and assign the 2002 6th grade cohort to these quintiles based on the 2001 distribution. Using test scores from the untreated 2001 cohort, we regress the *8th grade high-stakes score* on the *6th grade low-stakes score* for each quintile in five separate regressions. These models also include students' gender, race, and free and reduced-price lunch status. We then estimate five additional regressions where the outcome is instead the *8th grade low-stakes score*, but the predictor remains the *6th grade low-stakes score*, and the controls remain the same.³ We apply these respective prediction equations to the 2002 cohort to estimate their 8th grade test scores on both high and low-stakes tests, which provides an estimate of how students in each quintile would have performed had NCLB not taken effect. We then compare the predicted 8th grade high and low-stakes scores with the observed 8th grade scores of the 2002 cohort, and infer that the difference can be attributed to the 2002 cohort's exposure to NCLB in their 8th grade year. Hereafter, we refer to the 2001 cohort as the “pre-NCLB cohort” and the 2002 cohort as the “post-NCLB cohort.”

This approach is not without limitations. It requires us to assume that no other unobserved factors correlated with time produced this result (for example, changes in the high-stakes test), and that these changes are not simply a function of cohort-to-cohort differences. While we believe this is the best available analytical strategy to estimate the impact of NCLB with these data and have done our best to rule out alternate explanations, we stop short of calling the estimated differences between the pre- and post-NCLB cohorts a causal effect.

To determine how school context affects the process we observe, we examine whether being enrolled in schools at different levels of the school performance distribution affects students' test performance after the implementation of NCLB. We assign each school in our analysis

³6th grade low-stakes test scores are correlated with 8th grade high-stakes test scores at 0.678 for reading and 0.745 for math, while they are correlated with 8th grade low-stakes test scores at 0.772 and 0.775 for reading and math, respectively.

to a performance tercile—low, mid, and high—based on each school's pre-NCLB (2001) 8th grade high-stakes math proficiency rate. Math pass rates in the three terciles are 51, 67, and 83 percent, respectively. We use this assignment for both math and reading analyses as reading proficiency rates are consistently high and there is substantially less variation across schools. We then separately report the mean difference between predicted and observed scores for students attending schools in each tercile.

Our sample includes students enrolled in grade 8 two years after they enrolled in grade 6 and who took high and low-stakes tests in both grades in a given subject. Our math analysis sample includes 8,659 students in the pre-NCLB cohort and 8,870 students in the post-NCLB cohort, while our reading analysis sample includes 8,632 students in the pre-NCLB cohort and 8,811 students in the post-NCLB cohort.

Results

Table 1 displays the demographic characteristics of our two cohorts and illustrates the incentives that schools face to focus on different parts of the achievement distribution. The proficiency standards for math and reading are set at very different levels relative to the Houston distribution. Students in the pre-NCLB cohort are much less likely to pass the math exam (67 percent) than the reading exam (87 percent), so schools have incentives to focus on different parts of the distribution on the two tests. Table 1 reports that in the pre-NCLB cohort, only 11.2 percent of quintile 1 students passed the math test. This increases to 24.1 percent for quintile 2 students and 49.1 percent for quintile 3 students. The overwhelming majority of quintile 4 and 5 students are proficient in math. In reading, most students above the 2nd quintile achieve proficiency on the reading test (79 percent and above), while 47 percent of students in quintile 1 do. In the third through fifth quintiles, almost all students score at proficient on the reading test.

These facts lead to two general predictions. First, we expect teachers to focus efforts towards the middle of the distribution in math and towards the lower part of the distribution in reading. Second, we expect that schools will put more emphasis on the math test rather than the reading test, since it is generally the math test that puts them at risk of failing to make Adequate Yearly Progress.

We now turn to our main results, which provide our estimates of how the implementation of NCLB affected the 8th grade high and low-stakes scores of students with different initial levels of achievement. Figure 2 provides kernel densities of the predicted and observed outcomes for the post-NCLB cohort's high-stakes test scores. The left hand panel shows that the lowest performing students in the post-NCLB cohort performed worse than predicted on the 8th grade high-stakes test, which is what we would expect if resources were reallocated from the bottom of the distribution towards the middle to top. In the right hand panel, observed scores for the lowest performing students shift to the right relative to their predicted scores. This is also what we would expect to see if schools allocated resources to the lowest performing students in reading. In sum, the difference between the observed and predicted distributions plotted in Figure 2 provides descriptive evidence that inequality

between lower and higher performing students in math grew, while it declined marginally in reading.

Table 2 reports the mean difference between observed and predicted outcomes in each quintile for both math and reading tests. Our estimates are consistent with teachers focusing on students who are closer to the cut score for proficiency, with some important exceptions that we discuss below. Students in the bottom quintile scored lower than predicted on their eighth grade high-stakes math test, while students in quintiles 3-5 performed higher than what they would have had they not been exposed to NCLB. These effects are not trivial in size. Students in the first quintile lost .11 standard deviations on the high-stakes test, while students in the top quintile gained .25 standard deviations.

However, we do not observe similar distributional effects on the low-stakes math test, suggesting that the relatively better and worse performance of the post-NCLB cohort at different parts of the distribution may be a function of acquiring test-specific skills that do not generalize to the low-stakes test – what we have referred to as instructional triage. We observe that the post-NCLB cohort overall performed better than expected on their low-stakes eighth grade Stanford tests, on the order of .064-.099 SDs, but the positive effects are similar across quintiles. This result is more consistent with the interpretation that increased attention to the subject of mathematics equally benefitted students across the distribution. That is, if schools systematically reallocated time from reading to math for all students since math proficiency rates were much lower, we might expect to see relatively even gains across quintiles.

Counter to our expectations, students in the fourth and fifth quintiles also experienced significant gains as a result of NCLB. While there is still considerable room for increasing proficiency in the fourth quintile – with 75.5 percent proficient in the pre-NCLB cohort– this is not the case for the 5th quintile (91.8 percent proficient). Why might we nonetheless observe gains in the upper quintiles? The most likely explanation, in our view, is that students in these quintiles are not sorted into separate classes. Especially in middle school, we might expect relatively higher achieving students to have math classes together, and interventions intended to help bubble students may also spillover to higher-achieving students. It may also be the case that teachers do not target quite as precisely as the educational triage hypothesis suggests, whether because they actively reject this approach or because they underpredict math proficiency for high-achieving students.

We find substantially different distributional effects for reading, which we attribute to the lower difficulty of the reading proficiency standard. Lower quintile students performed better than expected on high-stakes tests, while the upper quintile students performed worse than their predicted high-stakes scores. First quintile students saw the most benefit from NCLB on their high-stakes tests, scoring 0.141 standard deviations above their mean predicted scores. They nonetheless experienced significant losses on their low-stakes tests. Unlike for math, the post-NCLB cohort performed worse than predicted on their eighth grade low-stakes reading test, with losses ranging from .203-.372 standard deviations.

We believe that part of the negative effect in reading, which is surprisingly large by education policy standards, is attributable to an enlarged emphasis on math over reading more generally. The binding constraint faced by schools in their efforts to make NCLB Adequate Yearly Progress targets was clearly the math test, which could lead to a greater emphasis on math at the expense of reading.

What impact, then, did these gains and losses have on changes in proficiency rates? Despite average losses in achievement gains for students in the first and second math quintiles, the bottom panel of Table 1 shows that proficiency rates in those quintiles substantially increased relative to our predictions, with proficiency gains of 13.3 and 23.9 percentage points, respectively. This suggests that small gains for some students that moved them over the proficiency cut score were offset by larger losses for other students in those quintiles. (See Appendix Figure A2 for plots of the observed and predicted distributions relative to the cut score.) Table 1 also shows that the greatest gains in the reading proficiency rate were for the lowest performing students. Students in the first quintile of the reading distribution performed substantially better in the post-NCLB cohort in terms of proficiency (58.6 versus the predicted 33.9 percent proficiency rate).

Table 3 decomposes these results by comparing students' actual 8th grade math scores with their predicted scores by school performance level (low, mid, and high performing). Two major findings emerge from Table 3. First, we observe the steepest distributional gradient on the high-stakes test in the lowest performing schools. The difference between the losses of the bottom quintile and the gains at the top are .483 SDs, while they are .11 SDs in mid-performing schools. Because of the imprecision with which these effects are measured, we cannot conclude that the high-stakes distributional gradient varies across school types. However, students in the bottom performing third of schools did not display any gains on their low-stakes tests as a result of NCLB, while we do observe low-stakes gains in the middle and top school performance terciles. Moreover, if we consider the point estimates, the low-stakes gains are generally greatest in the highest performing schools. While we are careful not to over-interpret these results given their imprecision, our results provide suggestive evidence that practices of instructional triage are most prevalent in the lowest performing schools, which almost half of the lowest-performing students (those in the bottom quintile) attend. (See Table 5.)

Table 4 shows the effect of NCLB on reading scores by school performance tercile. Unlike for math, we do not see evidence that schools facing more pressure respond differently. There are at least two possible explanations for this finding. The reading proficiency rate is already relatively high across the entire district, which means that schools were under less pressure to raise proficiency rates through targeting high-return students and content. In addition, multiple studies have demonstrated that math scores have been more responsive to test-specific coaching than have reading scores. Similarities in the presentation of math problems on high-stakes tests enable these practices, while fewer test-specific coaching practices have been identified for reading.

To summarize our results, the proficiency standards for math and reading vary in their difficulty. The average student in the first two quintiles is expected to fail the eighth grade

math high-stakes tests while the average student in quintiles 2-5 is expected to score at proficient on the reading high-stakes test. Our findings support the hypothesis that schools facing accountability pressure focus attention on students who are in the vicinity of the proficiency standard. We observe high-stakes test benefits for students in math quintile 3 and above, and for reading, we observe these benefits in quintile 1. The difference between the effects on high and low-stakes tests is most pronounced in the lowest performing schools, which we cited as suggestive evidence that school context plays an important role in shaping responses to accountability pressure. We attribute this finding to a process of instructional triage through which teachers imparted test-specific skills to students near proficiency.

Discussion

This study demonstrated that the conclusions we draw about the effects of accountability pressure on inequality in reading and math achievement are sensitive to the choice of outcome and are affected by the difficulty of the proficiency standard. Overall, our results show that accountability pressure increases inequality between low and high-performing students on high-stakes math tests, decreases inequality on high-stakes reading tests, and has no effect on inequality in low-stakes math and reading test performance. This is, to be sure, a complex set of results. Helping sociologists of education understand their implications requires that we broaden our lens to consider three mechanisms through which the effects of accountability pressure on test scores may exacerbate or attenuate educational inequality beyond their effects on test scores alone. We refer to these as *skill consequences* (which derive from the math and reading skills that students gain or lose), *signaling and labeling consequences* (which occur because test scores serve as a signal that influences students', teachers', schools', and parents' future educational decisions), and *policy feedback consequences* (which result from policymakers' use of test results to shape future decisions).

Skill consequences

The motivating question of this study was how proficiency-based accountability systems affect inequality in math and reading achievement. Our results demonstrate that the answer depends on which outcomes you care about. If we rely on our high-stakes outcome, we come to the conclusion that proficiency-based accountability systems increased inequality in math and decreased it in reading. In math, these increases were the result of losses for the lowest performing students and gains for higher performing students. For reading, they were the result of gains for lower performing students and losses for higher performing students. If we rely on our low-stakes results, we come to the conclusion that all students benefitted and lost equally on the math and reading tests, respectively. In other words, there is no impact on inequality in low-stakes skills, even as the level of overall performance increased for math and decreased for reading.

Our results suggest that both sides of the debate on the distributional consequences of accountability systems have been, at least in some sense, correct. Educational triage – that is, focusing on students close to proficiency – appears to be a widespread response to proficiency-based accountability systems, and “bubble” students gain more on high-stakes tests. In the past, both scholars observing this behavior directly as well as those finding

particularly negative consequences for the lowest performing students have decried this practice and argued that it may play a central role in reproducing inequality. This prediction, however, is partially dependent on these skill advantages transferring to other educational outcomes in both the short and longer-term. We uncover the concomitant practice of instructional triage by evaluating the impact of proficiency-based accountability on a low-stakes test. Those who have argued that NCLB is not harming the lowest performing students' math and reading skills also have evidence to support their argument based on our results, although not for reasons they might have expected.

Our findings suggest that policymakers face a series of difficult normative questions when they decide where to set the cut score for proficiency. In contrast to a long history of failed reforms, our results provide suggestive evidence that education policy has direct effects on how resources are allocated within schools, though we cannot explicitly measure the instructional processes that produced these results. This also means that policy decisions about proficiency standard difficulty potentially affect school-level decisions about resource allocation, and thus inequality in math and reading skills. While our results may represent the short-term effects of accountability systems rather than their longer-run consequences, we cautiously conclude that higher cut scores are likely to increase skill inequality, while lower cut scores are likely to decrease it.

Even if the public could agree on whether we should privilege gains at the bottom of the distribution or gains at the top, the uneven distribution of higher and lower achieving students from different social backgrounds further complicates the question of policy design. Because of patterns of economic and racial segregation, high-achieving students from historically disadvantaged groups attend schools with higher fractions of low-achieving peers than do their high-achieving counterparts from historically advantaged groups. As a result, it may be the case that lower cut scores hurt the highest-achieving poor and black and Hispanic students, even as lower cut scores would benefit the performance of these groups on average. This concern appears to be born out in multiple papers tracking the performance of high-achieving black students (Clotfelter, Ladd, and Vigdor 2009; Hanushek and Rivkin 2009).

Signaling and labeling consequences

Test scores have multiple functions in schooling contexts that extend their effects beyond the measurement of cognitive skills. Because high-stakes test score data in particular are utilized in other ways by students, teachers, the school, and parents, distributional effects on high-stakes tests may be particularly important to understanding accountability's impact on educational inequality.

Unlike the low-stakes tests that we study, high-stakes tests divide students into multiple levels of proficiency based on their score. Students can be labeled as commended, meeting the standard, or not meeting the standard. Since the Pygmalion study found that providing randomly assigned performance labels to teachers could affect students' subsequent performance on standardized tests (Rosenthal and Jacobsen 1968), it has been clear that performance labels can have significant educational consequences. Papay, Murnane, and Willett's (2011) study provided a powerful demonstration of this concept by showing that

students earning an “advanced” label on Massachusetts' exit exam were more likely to attend college than those who scored just below this cut score. In this case, providing new information about student performance in the form of a performance label influences students' decisions to continue their formal education.

By shaping not only how students see themselves, but also how peers, parents, and teachers perceive students, high-stakes test scores plausibly affect academic identities, engagement in school, peer dynamics, teacher and parental expectations (and thus ability group and course placement – see Kelly and Carbonaro 2012), and future educational decisions. Considered from this angle, that the inequality effects we observe on high-stakes tests do not transfer to low-stakes tests is irrelevant because the effects of accountability on high-stakes tests go on to shape academic inequality through other mechanisms.

Policy feedback consequences

In addition to these micro-level mechanisms, the effects of accountability systems on inequality may result from how test score data shape policymakers' and the public's understanding of educational inequality. Measures themselves play a central role in constructing social problems. For example, poverty rates and unemployment rates are social metrics that affect how we understand the impacts of economic and social policy. Because poverty and unemployment measures have important implications for the policies that we adopt and for the success or failure that we ascribe to our public officials, how they are measured has been contested since they were first reported. Multiple versions of these measures exist, and policymakers and researchers advocate fiercely for the use of one measure versus another precisely because they are so important to framing policy debates (and thus the allocation of scarce resources). Similarly, the portrait of schools' performance and of educational inequality provided by high-stakes test score data matters because it shapes the actions taken to address these inequalities in the future.

It is from the “policy feedback loop” perspective that the implications of our results are the least clear. High-stakes scores typically receive the most attention in the news media and are used by policymakers as official measures of educational progress. We suggest that at least two different narratives about educational inequality and the role of schools in addressing it can arise from apparent increases in proficiency rates and declines in the size of proficiency-based achievement gaps. The first narrative is a “beating the odds” narrative. The fact that test scores tell a story of decreasing gaps may spur and legitimate investment in disadvantaged children, given that such gaps have been difficult to close historically. The second narrative is a “mission accomplished” narrative. From this perspective, our current policies have been successful at reducing inequality, and we need not look for or invest in new education or social policy solutions. Each of these narratives has powerful proponents, and the way in which actors deploy existing test score data can lend support to either narrative.

Low-stakes test scores, however, also play a role in constructing a counter-narrative about the gains produced by accountability systems. When former Houston Superintendent Rod Paige was appointed Secretary of Education, the *New York Times* analyzed low-stakes test data from Houston to argue that his accomplishments were inflated (Schemo and Fessenden

2003). The same occurred with the National Assessment of Educational Progress, another low-stakes test; when Arne Duncan was appointed Secretary, opponents pointed to low NAEP scores as evidence of his failure in spite of large gains on high-stakes tests (Anderson 2009). Each of these examples illustrates how the deployment of test scores in policy debates has the potential to amplify the micro-level consequences of test scores.

Our study is not without limitations. While we were able to strategically use the different proficiency standards for math and reading to identify their impact, math and reading scores are not necessarily independent, though they may be more so in a junior high school context than in a self-contained elementary classroom where one teacher is responsible for both subjects. The difficulty of the proficiency standard in one subject could certainly affect how resources are expended on the other, and indeed we see evidence of this pattern in our data. If both proficiency standards were set equally high relative to the Houston distribution, we might observe more muted distributional effects for the math test. Moreover, our study's analytic strategy relies on observing NCLB in the early years of implementation because this allows us to generate the cleanest short-term estimate of its impacts. It is possible that the effects vary over time as pressure becomes more intense. We suspect that this would exacerbate, rather than attenuate, the distributional effects that we uncover. On the other hand, it may be the case that over time schools develop new organizational procedures and instructional strategies that “lift all boats” or that disproportionately benefit students at the low end of the distribution. This study does not allow us to make inferences about the longer-term impacts of accountability pressure on the distribution of achievement, and this remains an important area for future research. Finally, we note that we study a disadvantaged urban school district, and that the patterns that we observe in this district may not generalize to schools of higher socioeconomic status.

Appendix Figure 1

Timeline

Appendix Figure 2

a. Distribution of Predicted and Observed Math 8th Grade High-Stakes Scores for the Post-NCLB Cohort

Note: This figure plots the predicted and observed scores for the post-NCLB cohort by quintile. The vertical line marks the proficiency cut score on each test.

b. Distribution of Predicted and Observed Reading 8th Grade High-Stakes Scores for the Post-NCLB Cohort

Note: This figure plots the predicted and observed scores for the post-NCLB cohort by quintile. The vertical line marks the proficiency cut score on each test.

References

- Andersen, Nick. Education Secretary Arne Duncan's Legacy as Chicago Schools Chief Questioned. *Washington Post*. 2009 Dec 29.
- Ballou, Dale; Springer, Matthew. Working Paper. Vanderbilt University; 2011. Achievement Trade-Offs and No Child Left Behind.
- Carnoy, Martin; Loeb, Susanna. Does External Accountability Affect Student Outcomes? A Cross-State Analysis. *Educational Evaluation and Policy Analysis*. 2002; 24:305–331.
- Clotfelter, Charles C.; Ladd, Helen F.; Vigdor, Jacob L. The Academic Achievement Gap in Grades 3–8. *The Review of Economics and Statistics*. 2009; 91(2):398–419.
- Dee, Thomas S.; Jacob, Brian. The Impact of No Child Left Behind on Student Achievement. *Journal of Policy Analysis and Management*. 2011; 30:418–446.
- Gillborn, David; Youdell, Deborah. *Rationing Education: Policy, Practice, Reform and Equity*. Buckingham: Open University Press; 2000.

- Hamilton, Laura S.; Stecher, Brian M. Measuring Instructional Responses to Standards-Based Accountability. Santa Monica, CA: RAND Corp; 2007.
- Hamilton, Laura S.; Stecher, Brian M.; Marsh, Julie A.; McCombs, Jennifer Sloan; Robyn, Abby; Russell, Jennifer Lin; Naftel, Scott; Barney, Heather. Implementing Standards-Based Accountability Under No Child Left Behind: Responses of Superintendents, Principals, and Teachers in Three States. Santa Monica, CA: RAND Corporation; 2007.
- Hanushek, Eric A.; Raymond, Margaret E. The Effect of School Accountability Systems on the Level and Distribution of Student Achievement. *Journal of the European Economic Association*. 2004; 2:406–415.
- Hanushek, Eric A.; Rivkin, Steven G. Harming the Best: How Schools Affect the Black-White Achievement Gap. *Journal of Policy Analysis and Management*. 2009; 28:366–393.
- Hoey, Lesli; Campbell, Patricia B.; Perlman, Lesley. Where's the Overlap? Mapping the SAT-9 and TAAS 4th Grade Test Objectives. Groton, MA: Campbell-Kibbler Associates; 2001.
- Hout, Michael; Elliott, Stuart W., editors. Incentives and Test-based Accountability in Education. Washington, DC: National Academies Press; 2011.
- Jacob, Brian A. Accountability, Incentives, and Behavior: Evidence from School Reform in Chicago. *Journal of Public Economics*. 2005; 89:761–796.
- Kelly, Sean; Carbonaro, William. Curriculum Tracking and Teacher Expectations: Evidence From Discrepant Course Taking Models. *Social Psychology of Education*. 2012; 15:271–294.
- Koretz, Daniel. Measuring Up: What Educational Testing Really Tells Us. Cambridge: Harvard University Press; 2008.
- Krieg, Jonathan. Are students left behind? The Distributional Effects of No Child Left Behind. *Education Finance and Policy*. 2008; 3:250–281.
- Ladd, Helen F.; Lauen, Douglas L. Status versus Growth: The Distributional Effects of Accountability Policies. *Journal of Policy Analysis and Management*. 2010; 29:426–450.
- Lauen, Douglas L.; Gaddis, S Michael. Accountability Pressure, Academic Standards, and Educational Triage. Paper presented at the annual meetings of the Society for Research on Educational Effectiveness; Washington, DC. 2012.
- McAdams, Donald R. Fighting to Save Our Urban Schools...and Winning!. New York: Teachers College Press; 2000.
- Neal, Derek; Schanzenbach, Diane Whitmore. Left Behind By Design: Proficiency Counts and Test-Based Accountability. *Review of Economics and Statistics*. 2010; 92:263–283.
- Papay, John P.; Murnane, Richard J.; Willett, John B. How Performance Information Affects Human-Capital Investment Decisions: The Impact of Test-Score Labels on Educational Outcomes. *National Bureau of Economic Research*, No. 2011:w17120.
- Pedulla, Joseph J.; Abrams, Lisa M.; Madaus, George F.; Russell, Michael K.; Ramos, Miguel A.; Miao, Jing. Perceived Effects of State-Mandated Testing Programs on Teaching and Learning: Findings from a National Survey of Teachers. Boston: National Board on Educational Testing and Public Policy; 2003.
- Reback, Randall. Teaching to the Rating: School Accountability and the Distribution of Student Achievement. *Journal of Public Economics*. 2008; 92:1394–1415.
- Reback, Randall; Rockoff, Jonah; Schwartz, Heather. Under Pressure: Job Security, Resource Allocation, and Productivity in Schools Under NCLB. Working Paper, Barnard College. 2010
- Rosenthal, Robert; Jacobson, Lenore. Pygmalion in the Classroom. New York: Holt, Rinehart & Winston; 1968.
- Schemo, Diana J.; Fessenden, Ford. A Miracle Revisited: Measuring Success, Gains in Houston Schools: How Real Are They? *New York Times*. 2003 Dec 3.
- Shepard, Lorrie A. The Harm of Measurement-Driven Instruction. Paper presented at the annual meeting of the American Educational Research Association; Washington, DC. 1988.
- Shepard, Lorrie A.; Dougherty, KD. Effects of High-Stakes Testing on Instruction. In: Linn, Robert L., editor. The Effects of High Stakes Testing; Symposium presented at the annual meetings of the American Education Research Association and the National Council of Measurement in Education; Chicago, IL. 1991.

- Springer, Matthew. The Influence of an NCLB Accountability Plan on the Distribution of Student Test Score Gains. *Economics of Education Review*. 2008; 27:556–63.
- Weitz-White, Katie; Rosenbaum, James E. Inside The Black Box Of Accountabilty: How High-Stakes Accountability Alters School Culture and the Classification and Treatment of Students and Teachers. In: Sadvonik, A.; O'Day, J.; Bohrnstedt, G.; Borman, K., editors. *No Child Left Behind And The Reduction of the Achievement Gap: Sociological Perspectives on Federal Education Policy*. New York: Routledge; 2007. p. 97-116.
- Wong, Manyee; Cook, Thomas D.; Steiner, Peter M. Working Paper. Northwestern University; 2009. *No Child Left Behind: An Interim Evaluation of Its Effects on Learning Using Two Interrupted Time Series Each With Its Own Non-Equivalent Comparison Series*.

Studies	Dependent Variable and Context	Time Period	Test Type (HST or LST)	Proficiency Standard Difficulty	Findings
Ballou and Springer (2011)	Northwest Evaluation Association Scores (NWEA) test scores from CO, ID, IN, MN	2002-2006	LST: math only	Mixed	Effects of NCLB implementation on the distribution of math achievement vary across grades and states. Low-performing students' scores increased, while higher-performing students' scores fell.
Clotfelter, Ladd, and Vigdor (2009)	North Carolina state test scores	1995-2004	HST	Low	Black-white gaps in math among low-performing students shrink as students progress through grades 3-8, while gaps among for high-performing students widen. Authors suggest that accountability pressure may play a role in producing this pattern.
Dee and Jacob (2011)	National Assessment of Educational Progress test scores (state-level)	1992-2007	LST	Mixed	NCLB implementation is associated with increases in math achievement across the performance distribution in 4th grade math, though point estimates are larger for low-performing students (10th percentile) than for high-performing students (90th percentile). Low-performing students' 8th grade math scores increased, but high-performing students' did not. Low-performing students' 4th grade reading scores did not change, but high-performing students' scores increased.
Jacob (2005)	Iowa Test of Basic Skills test scores from Chicago Public Schools	1993-2000	HST	High	Chicago's district accountability system led to larger test score gains for students near the proficiency cut score (10-50th percentile) than those far below (<10 percentile) or above it, but differential effects of prior school performance were greater than differential effects of prior individual student performance.
Krieg (2008)	Washington State test scores	2001-2005	HST: math only	High	High and low-performing students score worse than expected if their school is in danger of failing to make Adequate Yearly Progress (AYP) under NCLB, while students close to proficiency score better. These effects accumulate over time in schools repeatedly failing to make AYP.
Lauen and Gaddis (2010)	North Carolina state test scores	2004-2009	HST	Low→High	Increases in the proficiency standard from a low standard to a more challenging one led to lower gains for the lowest performing students, suggesting that the difficulty of the proficiency standard may affect schools' responses to accountability pressure.
Ladd and Lauen (2009)	North Carolina state test scores	1996-2007	HST	Low	NCLB accountability pressure increases achievement gains for low-performing students. Negative effects of status-based accountability pressure (i.e. systems that do not take into account student growth) for higher-performing students in reading.
Neal and Schanzenbach (2010)	Iowa Test of Basic Skills and Illinois State Achievement Test scores from Chicago Public Schools	1994-2003	HST	High	Implementation of Chicago and NCLB accountability systems benefitted students close to achieving proficiency and had negative effects on the lowest performing students.
Reback (2008)	Texas state test scores	1992-1998	HST	Low	Under Texas' accountability system, students close to proficiency perform better than expected when their math test score will impact their school's accountability rating. Low-achieving students gain in both reading and math when there is a chance their schools may achieve a higher rating, even if they themselves will not reach proficiency. The effects on higher-achieving students vary by subject, and negative effects are greater in reading than in math.
Reback, Rockoff, and Schwartz (2010)	Early Childhood Longitudinal Study - Kindergarten Cohort test scores	1999-2004	LST	Mixed	NCLB accountability pressure has no impact on the scores of students close to the proficiency cut score.
Springer (2008)	NWEA test scores from one anonymous state	2002-2005	LST: math only	Unknown	Low-performing students in schools that failed to meet NCLB's AYP target made more progress than students in schools facing no accountability pressure, and no negative effects on high-performing students.

Note: The difficulty of the proficiency standard is characterized relative to the population being studied; the possible categories are high, low, and mixed. Texas and North Carolina are classified as low difficulty relative to the population being studied, with more than 70% of students proficient, on average, in grades 3-8 in the subject studied (Reback; Ladd and Lauen 2009; Clotfelter, Ladd and Vigdor 2009); Chicago and Washington State are classified as high. Studies that include multiple states that have both low and high proficiency standards are referred to as "Mixed."

Figure 1. Summary of Previous Studies on Distributional Effects of Accountability Systems

Figure 2. Distribution of Predicted and Observed 8th Grade High-Stakes Scores for the Post-NCLB Cohort

Note: This figure plots the predicted and observed scores for the post-NCLB cohort. The vertical line marks the proficiency cut score on each test.

Table 1
Descriptive Statistics by Cohort (%)

Panel A - Demographic Characteristics by Cohort (%)			
	Pre-NCLB Cohort	Post-NCLB Cohort	
Asian	3.5	3.0	
Black	30.9	29.7	
Hispanic	55.1	56.5	
White	10.6	10.8	
Limited English Proficiency	15.6	13.2	
Special Education	5.3	4.5	
Economically Disadvantaged	78.6	78.6	
Female	52.7	53.1	

Panel B – Predicted and Observed Proficiency Rates by Cohort (%)			
	Pre-NCLB Cohort (Observed)	Post-NCLB Cohort (Observed)	Post-NCLB Cohort (Predicted)
<i>8th grade Math proficiency rates (by quintile)</i>			
1	11.2	13.7	0.4
2	24.1	24.6	0.7
3	49.1	51.4	41.0
4	75.5	78.1	84.5
5	91.8	93.4	97.6
<i>8th grade Reading proficiency rates (by quintile)</i>			
1	47.0	58.6	33.9
2	79.0	85.7	99.5
3	91.9	94.2	99.8
4	96.4	98.3	100
5	99.5	99.3	100

Note: Pre-NCLB cohort (2001) n = 8,659 for math and n=8,632 for reading; Post-NCLB cohort (2002) n = 8,870 for math and n=8,811 for reading.

Table 2
Effects of Accountability on Grade 8 Low and High Stakes Scores by Performance Quintile

MATH	Quintile (1 lowest performing to 5 highest performing)				
	1	2	3	4	5
Low Stakes Test	0.099 (0.022)***	0.068 (0.018)***	0.064 (0.016)***	0.099 (0.02)***	0.080 (0.021)***
High Stakes Test	-0.108 (0.027)***	-0.102 (0.023)***	0.058 (0.018)***	0.195 (0.023)***	0.254 (0.025)***
READING					
	Quintile (1 lowest performing to 5 highest performing)				
	1	2	3	4	5
Low Stakes Test	-0.203 (0.023)***	-0.311 (0.016)***	-0.372 (0.019)***	-0.332 (0.015)***	-0.247 (0.019)***
High Stakes Test	0.141 (0.025)**	0.038 (0.017)*	-0.039 (0.02)**	-0.082 (0.018)**	-0.206 (0.023)**

Note:

* $p < 0.05$,

** $p < 0.01$,

*** $p < 0.001$.

This table reports the difference between predicted and observed high and low-stakes standardized 8th grade scores of the post-NCLB cohort based on prediction equations generated for each quintile from the pre-NCLB cohort. Analysis sample includes students enrolled in grade 8 two years after they enrolled in grade 6 and took high and low-stakes tests in both grades in a given subject. Pre-NCLB cohort (2001 6th grade students) $n = 8,659$ for math and $n = 8,632$ for reading; Post-NCLB cohort (2002 6th grade students) $n = 8,870$ for math and $n = 8,811$ for reading. The pre-NCLB cohort completed 8th grade before the implementation of NCLB, while the post-NCLB cohort completed 8th grade during the first year of NCLB implementation.

Table 3
Effects of Accountability on Grade 8 Low and High-Stakes Tests by School Performance, Math

<i>Students attending Low Performing Schools (1st tercile)</i>					
	1	2	3	4	5
Low Stakes Test	0.024 (0.033)	0.045 (0.028)	0.013 (0.027)	0.020 (0.036)	-0.028 (0.044)
High Stakes Test	-0.151 (0.039)***	-0.149 (0.033)***	0.059 (0.029)*	0.135 (0.041)***	0.332 (0.05)***
<i>Students attending Mid Performing Schools (2nd tercile)</i>					
	1	2	3	4	5
Low Stakes Test	0.137 (0.036)***	0.044 (0.03)	0.096 (0.026)***	0.113 (0.035)***	0.035 (0.038)
High Stakes Test	-0.073 (0.046)	-0.131 (0.038)***	0.072 (0.031)**	0.180 (0.039)***	0.037 (0.043)
<i>Students attending High Performing Schools (3rd tercile)</i>					
	1	2	3	4	5
Low Stakes Test	0.205 (0.057)***	0.134 (0.042)***	0.083 (0.029)***	0.135 (0.033)***	0.180 (0.03)***
High Stakes Test	-0.082 (0.062)	0.012 (0.054)	0.033 (0.034)	0.243 (0.039)***	0.328 (0.035)***

Note:

* $p < 0.05$,

** $p < 0.01$,

*** $p < 0.001$.

Schools were divided into three performance categories based on pre-NCLB proficiency rates. This table reports the difference between predicted and observed high and low-stakes standardized 8th grade scores of the post-NCLB cohort based on prediction equations generated for each quintile from the pre-NCLB cohort. Analysis sample includes students enrolled in grade 8 two years after they enrolled in grade 6 and took high and low-stakes tests in both grades in a given subject. Pre-NCLB cohort (2001 6th grade students) $n = 8,659$ for math and $n = 8,632$ for reading; Post-NCLB cohort (2002 6th grade

students) $n = 8,870$ for math and $n = 8,811$ for reading. The pre-NCLB cohort completed 8th grade before the implementation of NCLB, while the post-NCLB cohort completed 8th grade during the first year of NCLB implementation.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 4
Effects of Accountability on Grade 8 Low and High-Stakes Tests by School Performance, Reading

Students attending Low Performing Schools (1st tercile)					
	1	2	3	4	5
Low Stakes Test	-0.202 (0.033)***	-0.257 (0.027)***	-0.447 (0.035)***	-0.383 (0.032)***	-0.404 (0.046)***
High Stakes Test	0.169 (0.035)***	0.061 (0.026)**	0.010 (0.035)	-0.045 (0.035)	-0.157 (0.052)***
Students attending Mid Performing Schools (2nd tercile)					
	1	2	3	4	5
Low Stakes Test	-0.187 (0.04)***	-0.349 (0.025)***	-0.366 (0.028)***	-0.362 (0.025)***	-0.263 (0.039)***
High Stakes Test	0.137 (0.043)***	-0.003 (0.03)	-0.071 (0.033)**	-0.140 (0.032)***	-0.246 (0.05)***
Students attending High Performing Schools (3rd tercile)					
	1	2	3	4	5
Low Stakes Test	-0.226 (0.063)***	-0.361 (0.038)***	-0.276 (0.034)***	-0.271 (0.023)***	-0.167 (0.024)***
High Stakes Test	0.063 (0.069)	0.053 (0.038)	-0.078 (0.037)**	-0.073 (0.03)**	-0.194 (0.03)***

Note:

* $p < 0.05$,

** $p < 0.01$,

*** $p < 0.001$.

Schools were divided into three performance categories based on pre-NCLB 8th grade proficiency rates. This table reports the difference between predicted and observed high and low-stakes standardized 8th grade scores of the post-NCLB cohort based on prediction equations generated for each quintile from the pre-NCLB cohort. Analysis sample includes students enrolled in grade 8 two years after they enrolled in grade 6 and took high and low-stakes tests in both grades in a given subject. Pre-NCLB cohort (2001 6th grade students) $n = 8,659$ for math and $n = 8,632$ for reading; Post-NCLB cohort (2002

6th grade students) $n = 8,870$ for math and $n = 8,811$ for reading. The pre-NCLB cohort completed 8th grade before the implementation of NCLB, while the post-NCLB cohort completed 8th grade during the first year of NCLB implementation.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 5
Distribution of Student Achievement on Baseline High-Stakes Tests by School High-Stakes Performance Quintile

	1	2	3	4	5	Total
<i>Number of Students in Each Performance Quintile (Math, High Stakes Test)</i>						
Low-performing schools	992	1,273	1,652	1,013	1,113	6,043
(%)	16.4	21.1	27.3	16.8	18.4	100.0
Mid-performing schools	765	1,040	1,631	1,081	1,127	5,644
(%)	13.6	18.4	28.9	19.2	20.0	100.0
High-performing schools	449	680	1,328	1,265	2,296	6,018
(%)	7.5	11.3	22.1	21.0	38.2	100.0
<i>Number of Students in Each Performance Quintile (Reading, High Stakes Test)</i>						
Low-performing schools	1,148	1,520	1,121	1,469	726	5,984
(%)	19.2	25.4	18.7	24.5	12.1	100.0
Mid-performing schools	864	1,384	1,095	1,557	710	5,610
(%)	15.4	24.7	19.5	27.8	12.7	100.0
High-performing schools	399	770	879	1,848	2,114	6,010
(%)	6.6	12.8	14.6	30.7	35.2	100.0

Note: Analysis sample includes students enrolled in grade 8 two years after they enrolled in grade 6 (our baseline tests) and took high and low-stakes tests in both grades in a given subject. Pre-NCLB cohort (2001 6th grade students) n = 8,659 for math and n=8,632 for reading; Post-NCLB cohort (2002 6th grade students) n = 8,870 for math and n=8,811 for reading. The pre-NCLB cohort completed 8th grade before the implementation of NCLB, while the post-NCLB cohort completed 8th grade during the first year of NCLB implementation.