Does evaluation distort teacher effort and decisions? Quasi-experimental evidence from a policy of retesting students[†]

Esteban Aucejo, Arizona State University
Teresa Romano, Oxford College of Emory University
Eric S. Taylor, Harvard University

January 2019

Performance evaluation may change employee effort and decisions in unintended ways, for example, in multitask jobs where the evaluation measure captures only a subset of (differentially weights) the job tasks. We show evidence of this multitask distortion in schools, with teachers allocating effort across students (tasks). Teachers are evaluated based on student test scores; students who fail the test are retested 2-3 weeks later; and only the higher of the two scores is used in the teachers' evaluations. This retesting feature creates a sharp difference in the returns to teacher effort directed at failing versus passing students, even though both barely failing and barely passing students have arguably equal educational claim on (returns to) teacher effort. Using RD methods, we show that students who barely fail the end of school-year t math test, and are then retested, score higher one year later $(t+1)$ compared to those who barely pass. This difference in scores occurs during the four years of the retest policy, but not in the years before or after. We find no evidence that the results arise from retesting *per se*, or from changes in students' own behavior alone. The results suggest teachers give more effort to some students (tasks) simply because of the evaluation system incentives.

JEL No: I2, M5

[†] Aucejo: esteban.aucejo@asu.edu. Romano: teresa.romano@emory.edu. Taylor: eric_taylor@harvard.edu. Generous financial support was provided by the Jacobs Foundation. We greatly appreciate the help of the North Carolina Education Research Data Center at Duke University. We thank seminar participants at Brigham Young and Harvard for their comments and suggestions on earlier drafts.

Employers adopt employee evaluation systems—including performance measures and implicit and explicit incentives—to change employee effort, ideally to bring that effort in line with the employer's objectives. Evaluation can, however, change employee effort and decisions in unintended ways. This potential distortion of effort is of particular concern in multitask jobs if the evaluation measure captures only a subset of (differentially weights) the job tasks (Holmstrom and Milgrom 1991, Baker 1992). In this paper we provide an empirical example of this multitask distortion in schools. In their seminal paper, Holmstrom and Milgrom (1991) use the example of a teacher tasked both with teaching basic (testable) skills and with teaching higher-order thinking (non-testable) skills, but where only the basic skills contribute to her evaluation because they can be measured in student tests. In contrast, we study an example where a teacher's different students are the different tasks over which she must allocate effort.

The design of evaluation is a salient and timely topic in the education sector. In particular, the use of student test scores in school and individual teacher performance evaluation has become a central feature of education policy and management in recent decades.[1] The stated motivations for evaluation are, first, to inform personnel decisions about individual teachers, for example, retention, tenure, and hiring decisions. And, second, to induce increases in teacher (school) effort, but only occasionally by attaching explicit financial incentives.[2] Higher average performance of the teacher workforce, either through increased effort from

[1] The recent use of tests began first with state-level policies in the 1980s and 90s, then the federal No Child Left Behind (NCLB) regulations in the early 2000s, and more recently individual teacher evaluation prompted by the federal Race to the Top and NCLB waivers.

In US schools, at least in recent decades, the term "accountability" is often used to mean the same thing as performance evaluation with explicit (implicit) incentives. In this paper we use the terms performance evaluation and performance measure, but, as we describe in detail later, the evaluation systems in this setting include NCLB regulations and similar state programs.

[2] Incentives to increase effort need not be monetary, of course. Reback, Rockoff, and Schwartz (2014) report evidence that untenured teachers work longer hours when under pressure from test-based performance evaluation; teachers also feel less job security.

or better selection of teachers, would benefit students. A large literature now documents meaningful variation between teachers in their contributions to student learning (see Jackson, Rockoff, and Staiger 2014 for a recent review). Moreover, teachers who make larger contributions to learning, as measured by standardized tests in elementary and middle school, are also teachers who make larger contributions to their students' outcomes in adulthood, like labor market earnings and college attendance (Chetty, Friedman, and Rockoff 2014).

In the case we study, a short-lived feature of state evaluation rules created sharp differences across students (tasks) in the evaluation-score returns to teacher effort directed at different students. Briefly, in all years, teachers and schools were evaluated based on end-of-school-year student test scores. During the short-lived "retest policy" years, students who failed the test were retested 2-3 weeks later, and then only the higher of the two scores was used to calculate teacher and school performance evaluation measures. This retest policy created a strong incentive to allocate more effort to students who failed, (potentially) at the cost of less effort for students who passed. Critically, the incentives changed discontinuously at the pass/fail cutoff score, so that students near that cutoff who had arguably identical educational need for (returns to) teacher effort were weighted differently in the teacher evaluation because test measurement error had assigned their binary "pass" or "fail" status.

We find evidence consistent with teachers distorting their effort toward the retest students, as we would expect given the evaluation incentives. Using regression discontinuity methods, we show that students who barely fail the end of school-year t math test, and are then retested, score higher one year later (the $t+1$ test) than their classmates who barely pass at t. Note the outcome here is not the

retest score itself.[3] The difference in scores at $t + 1$ is approximately 0.03 student standard deviations (σ). The difference is small, but not inconsistent with a 2-3 week treatment, assuming the effects operate primarily (only) through changes during the period between the initial and retest.

This difference between passing and failing students occurs during the retest policy years, but not in the years before or after. Figure 1 shows year-by-year RD estimates for $t = 2003$ through 2015; the retest policy years are 2009 through 2012. The pattern of results in Figure 1 strengthens the case for interpreting the 0.03σ difference as occurring because of the retesting rules. A key category of threat to that interpretation is that other relevant educational inputs may be discontinuously assigned at the pass/fail cutoff. One such potential threat is the probability of repeating a grade, which we address in detail in the paper.

We combine the pass/fail cutoff (RD) variation and the over-time retest policy variation in a difference-in-RD identification strategy. Causal interpretation of our diff-in-RD estimates requires a weaker assumption than the simple RD. If we only had data from the four retest policy years, and thus a conventional RD estimate, we would need the conventional RD assumption: that potential outcomes (math scores) are continuous at the pass/fail cutoff both with the retest policy and without the policy. Using the diff-in-RD design we can relax the assumption by adding: however, if there are discontinuities in potential outcomes unrelated to the retest policy, those discontinuities are constant across the retest years and non-retest years. In the body of the paper we show evidence, and discuss institutional details, consistent with these assumptions.

One key result emphasizes the distinction between effects of retesting *per se* and effect of the retest policy. There was some retesting prior to the start of the

[3] The end of $t + 1$ test is the first time we observe math outcomes for both passing and failing students. These main results do not use the retest score, though we do make use of the retest scores when discussing mechanisms.

retesting policy in select grades and districts. Importantly, however, the pre-2009 retest scores were not used in calculating teacher evaluation scores; the pre-2009 retesting had no (little) effect on teachers' evaluation incentives. If retesting *per se* caused improvements in students' future math scores, we should see those effects in the select grades and districts even before the retesting policy. In fact, we do not see such effects. Our estimates are quite similar for grades and districts with and without pre-2009 retesting. Effects only appear when the stakes change for the adults in the school.

One mechanism for our main result is that (a) students were taught more math during the 2-3 weeks before the retest, and that extra learning persisted one year later. However, the 0.03σ difference at $t+1$ could arise from other types of mechanisms: (b) teachers and schools treated retest students differently during the subsequent school year after the retest, or (c) students themselves behaved differently. We do not find evidence consistent with (b) or (c). For example, relevant to mechanism type (c), treatment did not affect student absences during the subsequent year, nor did failing the math test affect reading scores at $t+1$; both suggesting students did not differentially change their effort. Also, relevant to mechanism type (b), we do not find differences in repeating grade or differences in the quality of teachers or peers to which students are assigned. Moreover, among students who barely failed the initial test nearly two-thirds passed the retest which would have removed, or at least blunted, the signal of "failed the test last year."

We do find evidence consistent with mechanism type (a): students were taught more math in the 2-3 weeks before the retest. For example, schools' test date decisions suggest they were trying to exploit this mechanism. In brief, schools have some discretion over test dates within a state-defined window. Before the retest policy schools facing pressure to improve test scores scheduled their tests relatively late in the window; after the retest policy began those schools moved their initial

test date up earlier. This switch is consistent with schools valuing time between the initial test and retest more than they value time before the initial test.

Our paper makes a novel contribution to the existing literature on unintended or unwanted responses to evaluation in schools. That literature already includes examples of manipulation of the performance measure: simple cheating *per se* (Jacob and Levitt 2003), but also manipulating which students are tested (Cullen and Reback 2006). Examples of the latter include differential suspension from school during the test period (Figlio 2006), special education designation (Jacob 2005, Figlio and Getzler 2006), grade retention in untested grades (Jacob 2005). Manipulation can also occur through test-taking skills and tricks which are orthogonal to the learning tests are intended to measure (Jacob 2005, Koretz 2017).

The existing literature also includes examples of unintended distortion of teacher and school effort. For example, teachers shifting effort away from untested subjects, like science and social studies, and toward tested subjects, like math and reading (Jacob 2005, Dee and Jacob 2011). The results in Macartney (2016) suggest schools distort effort across time within students; in short, schools with different grade spans (e.g., K-5, K-6, K-8) face different incentives in an evaluation based on student score growth over time.[4]

Our paper is most closely related to Neal and Schanzenbach (2010, "NS") which also focuses on teachers' effort allocation *across students*, and how evaluation based on student test scores changes that allocation. Studying the Chicago Public Schools, NS measures how NCLB affected student scores, and, importantly, how those effects differed (or not) across the distribution of students' pre-NCLB scores. The NCLB performance measure is the percent of students who pass the test, thus incentivizing teachers and schools to make effort allocation

[4] Adnot (2016) shows evidence that teachers distort effort across instructional tasks when evaluated by classroom observations; teachers respond to the incentives implied by the relative ease or difficulty of scoring well on a given task in the observation scoring system.

across students a function of the student's probability of passing. In slang terms, focus on the "bubble kids." Consistent with the incentives, NS finds NCLB-induced gains for students in the middle of the achievement distribution, but not for the low (high) achieving students with little chance of passing (failing). NS also finds this pattern of results for a district policy similar to NCLB but introduced in 1996, and evidence from other settings is consistent with the NS results (Burgess et al. 2005, Reback 2008, Springer 2008).

Our paper differs from Neal and Schanzenbach (2010, "NS"), and related work, in two ways both arising from the sharp discontinuity in incentives in the retest policy. First, the shift in teacher (school) effort across students implied by the NS results was partially intended by NCLB and similar policies, specifically the shift in effort from students higher in the distribution to students further down. Less effort for the lowest achieving students was presumably an unintended consequence. In the retesting policy case, students just above and below the pass/fail cutoff have arguably identical educational claim on (returns to) teacher effort; a difference in teacher effort across that pass/fail margin is clearly unintended. Second, the NS results clearly document different effects of NCLB across the distribution of achievement; students in the middle of the distribution benefited more. However, it may be that teachers and schools increased effort for the lowest achieving students as well, but those students' scores were not responsive to the increase in effort; this seems more plausible the further down the distribution one moves. By contrast, in the current paper, students just above and below the pass/fail cutoff should be equally responsive to any change in teacher effort, and indeed the students should be equal at expectation in other unobservable ways.

In summary, in this paper we show evidence that teacher effort can be unintentionally distorted by performance evaluation systems. Given two students with identical educational needs, teachers can be induced to give more effort to one

of the two through evaluation incentives. This result is contrary to the idea, claimed by some, that teachers are perfectly motivated agents whose effort is determined by professional judgement and whose effort is invariant to evaluation incentives. This result also emphasizes, alongside the other literature cited above, the importance of careful design of evaluation systems.

In the next section we describe the evaluation program and incentives teachers (schools) faced across the time period we study, and, critically, the changes that the retest policy made. Section 2 describes the econometric details. In section 3 we present our main results, and in section 4 we examine robustness and potential mechanisms. Then we conclude with some discussion.

1. Setting and data

We study students, teachers, and schools in North Carolina over a 14-year period from the 2002-03 through 2015-16. Throughout this period, North Carolina's schools and teachers were subject to state and federal systems of evaluation (accountability) based on student test scores. The retest policy was in place for four years: 2008-09 through 2011-12. In this section we describe the evaluation systems, retest policy, and other related details.

The specifics of the retest policy are the explicit mechanical features of the "treatment" we are studying. However, as is clear by the paper's introduction, we do not see these explicit mechanical specifics as the only mechanisms behind the results. We return to a discussion of the mechanisms in Section 4.

All data used in this paper were provided by the North Carolina Education Research Data Center at Duke University. The data are typical of school administrative data, including annual records for each student with school attended, test scores, demographic characteristics, and program participation variables. We also construct data linking students to teachers and classes. Throughout the paper

we refer to school years by their spring date—the 2002-03 school year is $t = 2003$ and so on.

1.1 School and teacher evaluation in North Carolina, 2003-2016

Throughout the period we study, 2003-2016, North Carolina teachers and schools were evaluated based on both student test score levels and test score growth. Our description of the details focuses on elementary and middle schools. Students in grades 3-8 were tested annually each spring in math and reading.[5] These test scores are the basis for the performance measures and incentives described in the next few paragraphs.

North Carolina's state evaluation system, known as the ABCs of Public Instruction (ABCs), began in 1996-97 several years before our study period. The ABCs performance measure was a school-level measure; it was a weighted average of a test-score *growth* measure and a test-score *levels* measure.[6] The levels measure was the percent of students who passed the end of year exam, much like the NCLB measure. Test scale scores were divided into four mutually-exclusive and exhaustive ordered categories, known as "Level I" through "Level IV," for each grade and subject. A student passed if they scored Level III or higher. This passing threshold is the same cutoff for the retesting policy we discuss later.

The ABCs growth measure was the school mean of the difference between a student's actual score and a state-specified expected score. The growth score was quite close in practice to the school mean of predicted student residuals, after a regression of standardized (mean 0, s.d. 1) year t score on standardized year $t-1$

[5] Tests in other subjects are more sporadic. In more recent years, for example, grade 5 and 8 students were tested in science. Earlier in our study period, grade 4 and 7 students were tested in writing.

[6] Our description of ABCs, READY, and other state programs in this section is drawn from historical documents found on the North Carolina Department of Public Instruction's website (www.ncpublicschools.org). The authors are happy to share those documents. In this paper we use the terms "growth" and "level." In North Carolina, the growth component was also sometimes called "gain" and the level component was known as "performance."

score for a given subject and grade. However, the standardization was based on a preset mean and standard deviation, and the coefficient on $t-1$ score was estimated in a regression using prior years' data. In short, the growth measure was responsive to any increase (decrease) in individual student test scores, not just changes in passing status.

Beginning in 2002-03 North Carolina's schools, along with all other schools receiving federal Title I funds, were also subject to the student test score based evaluation regulations known as No Child Left Behind (NCLB). Note that our data also begin in 2003. The NCLB performance measure was also a school-level measure, but it used only student test-score *levels*.[7] Just like ABCs, the levels measure was the percent of students who passed (scored Level III or higher) the end of year exam. The measure was often referred to as "percent proficient."

Varied consequences and incentives were attached to the ABCs and NCLB performance measures. As part of ABCs, the state gave schools labels with positive and negative connotations. Though the specific labels changed over time, examples include "Most Improved," "Expected Growth," "Honor Schools of Excellence," "Low-performing schools," and "No Recognition." The ABCs program also provided bonuses to teachers of between $750-1,500 per year based on the growth measure, but funding for the bonuses ended after the 2007-08 school year.

As is more well known, NCLB specified a set of escalating consequences for schools if the percent passing did not rise year after year. Schools were said to have made or met "adequate yearly progress" (AYP) if the percent of passing students was above the year's target numbers.[8] The ambitious yearly AYP targets were set under the legislated constraint that 100 percent of students pass by 2013-

[7] In the 2005-06 school year North Carolina was part of a pilot program which allowed the state to use a growth measure in their NCLB evaluations.

[8] Indeed, AYP target numbers were set, and had to be met, for several subgroups of students within each school. For example, racial and ethnic groups and different grade levels.

14, though different states set the slopes of the year AYP functions differently. Schools that missed AYP in two consecutive years had to write a plan for improvement, and students could transfer to another school. After a third consecutive year missing AYP schools had to provide students additional tutoring. Missing AYP in four or five consecutive years could lead to firing the school faculty and staff, or even simply closing the school.

Late in the period we study, the some details of the federal and state evaluations changed, but the core performance measures remained the same. First, starting with the 2011-12 school year, North Carolina received a "NCLB waiver."[9] Under the waiver, North Carolina's federal performance measure continued to be the percent of students passing, however, the annual targets for schools were reset.[10] Second, in 2012-13 North Carolina introduced READY, a bundle of features including revised content standards, tests to accompany the standards, and test-based evaluation (accountability).[11] While the standards and tests changed, the evaluation performance measures remained the same: the growth measure and the levels measure, now weighted 20/80 respectively. Replacing the descriptive school labels were school grades A-F.

1.2 The retest policy, 2009-2012

School years 2008-09 through 2011-12 were the "retest years" in North Carolina. This subsection describes how evaluation rules during the retest years

[9] Such waivers were granted liberally, in part because of failed attempts to reauthorize and update NCLB as the "100 passing" rule approached for 2013-14. Replacement legislation, known as the Every Student Succeeds Act, passed in December 2015. The period we study ends with the 2015-16 school year.

[10] The new targets were still ambitious. Schools would be expected to reduce the proportion of failing students by half over six years, relative to 2010-11 failure rates.

[11] The new tests were given five ordered "levels" instead of the prior four. We return to this change later in the paper and show that are results are robust to different approaches to dealing with this change. Additionally, the math test also changed substantially in 2005-06. Again, later we show that our results are robust to different approaches to dealing with test changes.

differed from the policies described in the previous subsection. First, as in all other years, all students in grades 3-8 were tested annually each spring in math and reading. We call this the "initial test score" or the time t score.

Second, any student who failed the initial test would be retested. Empirically, nearly all students (98 percent) who scored "Level II" (failing) on the initial test were retested, and zero students who scored "Level III" (passing) were retested.[12] The retest students took a different form from the same grade-and-subject level test; in other words, test items for the initial and retest were different but drawn from the same item bank.

Finally, only the higher of a student's two scores—initial score or retest score—would be used to calculate the school performance measures, that is, the growth and levels measures described in the prior subsection. Retesting could only increase the performance measure used to evaluate schools and teachers.[13]

Schools and teachers had approximately 2-3 weeks or more between the initial test and retest. During the retest years, the initial test could be given no earlier than 22 school days (4.4 five-day weeks) before the end of the school year. Schools had some discretion to give the tests later (closer to the end of the year), but not earlier. The retest had to occur before the end of the year.[14]

While the retest policy was only in place between 2009-2012, there was some retesting prior to 2009 but it did not change school and teacher performance evaluation measures. The pre-2009 retesting was much more limited: only about one-third of districts retested failing students, and then only in grades 3, 5, and 8.

[12] The state required that all students who scored "Level II" be retested. Students who scored "Level I" could be retested, but retesting was not required by the state.

[13] The change was not misunderstood by the policymakers who made it. Indeed, the state Department of Public Instruction warned that ABCs and NCLB measures for 2009-08 and later should not be compared to nominally similar measures from before retesting started. The motivation for the change seems to have been concerns about test reliability.

[14] We return to schools' choices of test dates later in the paper. Before the retest policy in 2009, the spring tests would be given no earlier than 15 days before the end of the year.

Later we show that our results are robust using just the other two-thirds of districts, or just grades 4, 6, and 7. The purpose of the pre-2009 retesting was to inform decisions about retaining students in grade, not to inform school performance measures.[15] In section 4 we describe the grade retention policies in more detail, and show that are results are not driven by any (potential) grade retention differences at the pass/fail cutoff.

2. Identification strategy

Our empirical objective is to estimate the causal effect of the retest policy on student math achievement scores. Our approach is a difference-in-RD, or an event study of year-by-year RD estimates. Conceptually, we first obtain a separate RD estimate at the pass/fail margin for each school year. Our primary outcome of interest is student math test scores one year later; thus, under the conventional RD identification assumptions, each year's RD estimate is the effect of barely failing the year t math test on year $t+1$ math scores. The effect of *failing* is not necessarily the effect of being *retested,* if failing brings consequences or interventions orthogonal to the retesting policy. By applying a difference-in or event study logic to the yearly RD estimates our goal is to difference out any effects of these potential other unrelated treatments at the pass/fail cutoff.

We fit the following specification, and variations on it, by local linear regression:

$$Y_{i,t+1} = f(Y_{it}) + \gamma F_{it} + \delta F_{it} R_t + \pi_{s(it),g(it),t} + \varepsilon_{it} \quad (1)$$

where Y_{it} is the running variable: student i's score on the end-of-year t math test, the initial test not the retest. The indicator variable $F_{it} = 1$ if student i received a failing score on Y_{it}. The indicator $R_t = 1$ during the four retest policy years, 2009-

[15] Instead of retesting the other two-thirds of districts used just the initial test score and the standard error of measurement for that score to inform retention decisions in grades 3, 5, and 8.

2012. The term $\pi_{s(it),g(it),t}$ represents fixed effects for each school-by-grade-by-year cell, which aid in precision. Our primary outcome of interest, $Y_{i,t+1}$, is student i's math score one year later. Throughout the paper we cluster standard errors at the values of the running variable Y_{it}.

Notice that equation 1 is a more typical RD specification if the $F_{it}R_t$ is omitted. The $F_{it}R_t$ allows the (potential) discontinuity at the pass/fail cutoff to be different in the non-retest years, γ, and the retest years, $\gamma + \delta$. The main effect for R_t is subsumed by the -by-year fixed effects.

A key choice in any RD analysis is how to model $f(Y_{it})$, the relationship between the running variable and outcome. In the LLR style, we fit a linear relationship which is allowed to be different above and below the pass/fail cutoff, i.e., $f(Y_{it}) = \alpha_1 Y_{it} + \alpha_2 Y_{it} F_{it}$. Further, we allow the parameters of $f(Y_{it})$ to be different year by year. In practice, as we show below, our estimates are not very sensitive to making f more or less flexible.

The bandwidth for our LLR varies by grade and year, but on average is 9 scale score points above or below the pass/fail cutoff. Student scores, Y_{it}, are divided into four ordered categories known as "proficiency levels." Students scoring Level I or II fail, and students scoring Level III or IV pass. Thus the cutoff between Level II and Level III is the pass/fail cutoff. In our main estimates the LLR bandwidth is all scores in Level II or III. We exclude Level I and Level IV because the Level I/Level II and Level III/Level IV cutoff (may have) induced other discontinuities in potential outcomes. As we show later, our results are robust to using smaller bandwidths.

The key parameter in equation 1 is δ, the effect of being retested for students near the margin of failing the exam (LATE). Strictly speaking we report intent-to-treat (sharp RD) estimates, though during the retest policy 98 percent of Level II

students were retested, and only about 0.1 percent of Level III students were retested.

To interpret our diff-in-RD estimate of δ as the causal effect requires a weaker assumption than the simple RD. If we only had data from the four retest policy years, and thus a conventional RD estimate, we would need the conventional RD assumption: (a) that potential outcomes (math scores) are continuous at the pass/fail cutoff both with the retest policy and without the policy. Using the diff-in-RD design we can relax the assumption by adding: (b) however, if there are discontinuities in potential outcomes unrelated to the retest policy, those discontinuities are constant across the retest years and non-retest periods. Part (b) addresses the possibility that the pass/fail cutoff may be used by schools to determine other consequences or interventions for students orthogonal to any effect of retesting which then in turn may create a discontinuity in potential outcomes at the pass/fail cutoff. In section 4 return to the topic of "other interventions," like repeating a grade. In the remainder of this section we report the tests relevant to judging part (a), the continuity of potential outcomes at the cutoff.

In this setting there is little, if any, scope for manipulating running variable scores relative to the cutoff. The running variable is a weighted average of test items where the weights are unknown to the student or school; the weights are determined by an item response theory (IRT) procedure. In other words, students and schools cannot rely on a simple rule like: answer *n* out of *N* items correctly and you will pass. The pass/fail cutoff is set by the state and does not change from year to year.[16]

[16] A pass/fail cutoff is specific to grade level, subject, and test design. During the period we study there are three math test designs: one used up through 2005, one used from 2006 through 2012, and a third used from 2013 forward. So, for example, the pass/fail cutoff for the 2006-2012 test was determined in 2006 and remained fixed, notably constant over the years just before and during the retest policy. Test items change from year to year, but IRT methods link item weights over time to keep scale scores and cutoffs constant.

Consistent with the institutional details, empirically we find no evidence of manipulation. Appendix Figure 1 is a histogram of the forcing variable centered at the pass/fail cutoff; the distribution appears smooth with no visible extra (missing) density above (below) the cutoff. The McCrary test statistic is -0.008 (st.err. 0.002), quite a small difference in density but statistically significant at conventional levels partly given substantial power.

As complementary evidence, Table 1 reports covariate balance style tests for several student characteristics. For example, there is no discontinuity at the pass/fail cutoff for students prior $(t-1)$ math test scores. Column 1 shows estimated difference in the non-retest years, $\hat{\gamma} = -0.004\sigma$, and column 2 shows the diff-in-RD estimate for the retest years, $\hat{\delta} = 0.003\sigma$; both come from estimating equation 1 where the dependent variable is $Y_{i,t-1}$. Students are also balanced on prior reading scores, retention in grade, absences, etc. Of the nine student characteristics tested, two show statistically significant differences for δ: female and special education status.

3. Main estimates

The retest policy generates improvements in the future math scores of retested students. Students who barely fail the initial end-of-year test (time t)—and thus are retested 2-3 weeks later—score 0.03σ higher one year later $(t+1)$ compared to otherwise-identical students who barely passed at time t. These differences at the pass/fail cutoff occur during the four years of the retest policy, but not in the years before or after.

Figure 1 shows the event study of RD estimates. Each square is an RD point estimate for a given school year. These estimates are obtained by fitted equation 1, but interacting F_{it} with year indicators instead of R_t. (The omitted year is 2008.) There is no trend in pass/fail differences in the years leading up to the retest policy.

However, the retest years, 2009-2012, are a clear deviation from that prior trend. The difference between the retest years and prior years appears to be between 0.02-0.03σ. The retest years are also clearly different from the post years, though the post years also appear different from the pre years. The post years may be partly explained by a change in the math test, which discuss below with other robustness tests.

Figure 2 is a rough visual representation of specification 1. For each value of the running variable ($A_{i,t}$ in scale score units), square markers represent the mean outcome score ($A_{i,t+1}$ in student standard deviations) net of grade-by-year-by-school fixed effects. Fitted lines are estimated using the underlying student-by-year data. There is a visible discontinuity in the retest years (left column, solid squares). The discontinuity is easier to see in the lower row where we "zoom in" to a smaller x-axis range; the marker means and fitted slopes are identical in the upper and lower rows. By contrast, there is no apparent discontinuity at the pass/fail cutoff in the non-retest years (right column, hollow squares). Nevertheless, our diff-in-RD estimate take account of any discontinuity in the non-retest years even if it is difficult to see in the picture.

The main diff-in-RD estimates are shown in the top row of Table 2. In the non-retest years, there is little difference at the pass/fail cutoff. The RD estimate γ from equation 1 is -0.002 (st.err. 0.008). In the retest years, the pass/fail difference increases by 0.031 (st.err. 0.005), which is our estimate of δ in equation 1. Thurs, during the retest years students who barely failed scored $\gamma + \delta = 0.029\sigma$ higher than they would have if they had barely passed.[17]

[17] The estimates in Table 1 row 1 (and similarly other estimates in the paper) come from a single regression fitting specification 1. An alternative two-step approach is to, first, estimate 13 year-specific RD point estimates, like the 13 shown in Figure 1. Then, second, estimate a bivariate regression with 13 observations where the depended variable is the year-specific RD estimate from step one, and the dependent variable is an indicator = 1 for the 4 retest years. Using this alternative approach, our point estimate is 0.030 (st.err. 0.004), nearly identical to the one-regression approach.

Is the difference of 0.03σ large or small? It is small as a share of the total variation in student math scores, but larger in context. First, consider 0.03 in the context of a teacher's total contribution to student test scores, assuming the gain comes through an increase in teacher effort. One standard deviation of the teacher effort ("value-added") distribution is typically estimated to be between 0.10-0.20σ (Hanushek and Rivkin 2010, Jackson, Rockoff, and Staiger 2014), making our estimated effect equivalent to 15-30 percent of the between-teacher variation.

A second relevant comparison is other estimates of the returns to quantity of instruction. If our estimated effect is due to teaching during the 2-3 weeks between the initial and retest, then it is relatively large compared to other estimates. Sims (2008) and Aucejo and Romano (2016) estimate the benefit of adding 1 week before the end-of-year test, finding 0.03σ and 0.02σ respectively. Raudenbush, Reardon, and Nomi (2012) and Taylor (2014) estimate the benefit of doubling a student's math class time for an entire year, finding 0.21σ for grade 9 students and 0.17σ for grade 6-8 students respectively. On the school calendar, 2-3 weeks represents approximately 6-8 percent of the year.

Finally, before turning to mechanisms, we briefly show that our results are robust to a number of critical estimation choices. The robustness test results are shown in the remaining rows of Table 2. Our main specification allows the slope parameters of $f(Y_{it})$ to differ year by year. The treatment effect estimate is essentially unchanged if we make $f(Y_{it})$ more or less flexible, i.e., allowing the parameters to differ for each grade-by-year cell, or only allowing them to differ for the binary retest and non-retest periods. Our estimates are also robust to using all of the within grade-by-year variation, instead of the within school-by-grade-by-year variation in our preferred specification.

Critically given the RD LLR feature of our approach, our effect estimates are not sensitive to bandwidth choice. For example, using just one-quarter of our

preferred bandwidth, the diff-in-RD estimate is 0.028σ; one-quarter is approximately 2-3 scale score points on either side of the cutoff.

The bottom two rows of Table 2 show estimates restricting the years used in estimation. Our results are essentially unchanged if we us only data from $t =$ 2006 through 2012. This is the period over which there was no change in the math test design; a new test was introduced in 2013 and an older test was used until 2005. Similarly, our results are unchanged if we use only the retest years and pre-retest years, $t =$ 2003 through 2012. Figure 1 suggests some additional change may have occurred post the retest years.[18]

4. Mechanisms

In this section we discuss evidence (in)consistent with different potential mechanisms for the retest policy's effects we documented above. Together the evidence suggests the effects arose because teachers (schools) changed their effort in response to the new, sharp incentives in their evaluation measure. We show evidence that the effects are not the result of retesting *per se*, nor the result of changes in students' own effort or behavior. It appears most likely that teacher (school) effort changed primarily during the 2-3 weeks before to the retest, at least more likely than changes in the subsequent school year.

[18] One possibility is the following: Beginning with 2014, the four proficiency levels were expanded to five levels. The old pre-2014 Level II was subdivided into two levels, and the ordered level numbers were adjusted accordingly. Thus, the new pass/fail cutoff was between new "Level III" and new "Level IV." For our analysis, we convert the new five levels back to the old four levels simply by collapsing new Level II and new Level III back into a single old Level II.

There was no retest policy in 2014, and so we do not know for certain whether the correct counterfactual would have been to apply retesting at the new Level III/new Level IV cutoff, as we assume in our analysis, or at the new Level II/new Level III cutoff. We have replicated our results using the latter assumption. In Figure 1, only the 2014 and 2015 points change, of course, and in this alternative those two points are closer to zero and their 95 percent confidence intervals include zero.

4.1 Retesting per se versus the retest policy

We begin with a result that emphasizes the difference between retesting *per se* and the retesting policy. There was some retesting in North Carolina prior to the start of the retest policy in 2009. Importantly, however, the pre-2009 retest scores were not used in calculating performance measures for teacher and school evaluation. Recall that under the retest policy, from 2009-2012, only the higher of a student's initial score and retest score would be used in teacher (school) evaluation measures. The prior retesting, but with quite different stakes for teachers, provides a convenient empirical test which we construct in this subsection.

First we need to explain a few institutional details. Prior to 2009 retesting was mechanically similar but more limited in scope. Similar to the 2009-2012 policy, students who failed the initial test would be retested a few weeks later, though there were fewer school days between the initial and retest pre-2009.[19] However, pre-2009 retesting was limited to grades 3, 5, and 8 only. The retesting was intended to help schools make decisions about which students to retain in grade, and the state policies about grade retention applied only to grade 3, 5, and 8. Moreover, districts could choose one of two rules about whom to retest: (1) Districts could simply retest failing students. This would become the rule for all districts in the retest policy years. Or (2) districts could adopt the "SEM rule." If a student failed, but their failing score was within one standard error of measurement (SEM) below the pass/fail cutoff, then the student did not need to be retested. Such students could be treated as effectively having passed for purposes of grade retention decisions.

If retesting *per se* drove the effects, then the start of the retesting policy in 2009 would no change student outcomes in grades 3, 5, and 8 in districts which were retesting prior to 2009. Here "no change" is equivalent to null hypothesis of

[19] During the retest policy years schools had to conduct both the initial test and retest in the final 22 school days of the year. Prior to 2009 the testing window was the final 15 school days.

$\delta = 0$ for the diff-in-RD. We test this prediction in Table 3 row 1.[20] The diff-in-RD estimate is 0.034 (st.err. 0.010), quite similar to our main estimate. Additionally, we should also expect a positive γ if retesting *per se* had benefits, but the estimate of γ in row 1 is -0.007 (st.err. 0.013). In short, for this sample retesting was a constant across all the years, and what changed in 2009 was how the retest scores affected teachers' (schools') evaluation measure; that change in 2009 induced changes in student outcomes.

For completeness, rows 2-4 of Table 3 are results for the other three possible combinations of grade level group and district retesting rule. Each of these three is also similar to the main effect estimate.

This pattern of results strongly suggests teachers (schools) changed their effort in response to the retest policy, specifically features of the retest policy other than the act of retesting failing students. One relevant feature—perhaps the only relevant feature—was the change in how retest scores were used to calculate teacher (school) evaluation measures.

4.2 Student effort or behavior

We now shift focus to the students themselves. One category of potential mechanism is changes in student effort or behavior. Student and teacher effort changes are not mutually exclusive mechanisms. Indeed, teacher contributions to student outcomes include teachers inducing their students to give more effort.

One specific hypothesis is that students may find being retested distasteful and give more effort in the future to avoid being retested again. This hypothesis could be true even if there is no change in teacher effort. We do not find any evidence consistent with this hypothesis. First, the results using pre-2009 retesting

[20] In Table 3 rows 1-4 we limit the sample to $t = 2003$ through 2012. We exclude the post period because there was no retesting. The results are substantively no different if we use all years in the data.

differences (Table 3 rows 1-4) are largely inconsistent with this hypothesis. If students feared being retested, they should fear it before and after 2009. It is possible that before 2009 students retested in grade 3 or 5 knew that there was no chance of being retested in grade 4 or 6 and thus they did not need to give extra effort on the $t + 1$ test.

We have one more-direct but imperfect measure of student effort: absences from school. Table 4 row 1 shows results where the outcome measure is changed from test score at the end of year $t + 1$ to absences during year $t + 1$. The diff-in-RD point estimate is a reduction in absences, but only 0.013 days, and far from statistically significant.

If students did change their effort to avoid being retested in the future, we might expect effects to be correlated with grade level. Older students may be more aware of retesting rules, or younger students may be less likely to shirk in the first place. We find no evidence effects are correlated with grade. Table 3 rows 5-9 provide estimates by grade. We can reject equality in some pairwise grade comparisons, but there is no monotonic relationship between grade and estimated effect.

Additionally, if students did fear being retested the following year, we might expect their increased effort to show up in other tests. Table 4 row 2 shows estimates where the outcome variable is reading/language arts score at $t + 1$, but the RD remains the math pass/fail cutoff. The point estimate is positive and marginally statistically significant, but an order of magnitude smaller than the main effect for math score.

One final hypothesis is about students learning from the experience of sitting for an additional test. Suppose simply by taking the retest students strengthened their general test-taking skills, knowledge of specific test item types, etc. Retested students would thus have an advantage over their peers who passed

and were not retested. This hypothesis is also contradicted by the results using pre-2009 retesting.

4.3 The weeks before the retest versus the subsequent school year

Incentives were clear and strong during the weeks between the initial and retest: effort directed at to-be-retested students could only improve the teacher (school) evaluation measure. However, our outcome, $A_{i,t+1}$, is only measured at the end of the subsequent school year, some 11 months after the retest. Thus, it is possible that our estimated effects, 0.03σ, arose (partly) because students who barely failed or barely passed the initial end-of-year t test were treated differently the following school year.

To explain the discontinuous jump in outcomes at the time t cutoff, any other candidate explanation—like a treatment during year $t+1$—would also need to change discontinuously at the t cutoff. For example, studying data from Miami schools, Taylor (2014) finds that middle-school students who barely fail the year t math test are assigned to lower level math courses the for year $t+1$ and have lower-achieving classmates; these "tracking" patterns are not required by any policy. In the current setting, we also see evidence that the pass/fail cutoff affects future peer, and perhaps teacher, assignments. Importantly, however, these discontinuities in peer (teacher) assignment existed absent the retest policy and do not change under the policy. In Table 4 row 3 the outcome measure is the mean prior math achievement, A_{it}, of student i's math classmates in year $t+1$; row 4 is the proportion of $t+1$ classmates who failed the t test; and row 5 is the value-added score of student i's year $t+1$ math teacher. There are some (marginally) significant differences in column 1, but the diff-in-RD estimates are zero.

The retest policy effects are not explained by grade retention differences. In Table 4 row 6 the outcome is an indicator $=1$ if student i is retained in grade, that

is, student i is assigned the same grade level in year t and $t+1$. As with the teachers and peer variables, we find no difference in discontinuity in the probability of being retained at the cutoff.

Differences in grade retention are, perhaps, a more plausible explanation than differences in tracking or teacher effectiveness because there is a relevant explicit policy. In 2009 and the years before, the state of North Carolina required that a student failing the end-of-year test must be "a factor" in the school's decision about whether to retain the student in grade. While failing was nominally a factor, as shown in Table 4 row 5 the empirical discontinuity is small. Two additional results are evidence against retention as a mechanism for our estimated effects. First, the state retention policy only applied to grades 3, 5 and 8. As shown in Table 3 the effects are not limited to those retention policy grades, nor limited to districts who used the SEM rule in pre-2009 retention decisions. Second, the state's retention policy ended after 2010. As shown in Figure 1, the retest policy effects continue in 2011 and 2012.[21] Additionally, our effect estimates do not covary with how districts' retention behavior changed after the end of the state policy (Appendix Table 1).[22]

Students may have been treated differently by their teachers and schools in ways we cannot observe in the data. Whether observed or unobserved, however, this type of mechanism requires that barely failing students be treated differently than barely passing students simply because of their "passed" or "failed" status or label.

[21] The end of the policy was officially approved on October 7, 2010, approximately six weeks into the 2010-11 school year. The policy change explicitly allowed schools to reverse the retention decisions for 2010-11 they had made previously.

[22] For each district we estimate the discontinuity in retention probability at the pass/fail cutoff before and after the 2010 state policy change. Then we divide districts into terciles of the change in the discontinuity point estimate; roughly districts who retained barely failing students more, districts with no change, and districts who retained barely failing students less. We find no different in effects across these three groups.

Hypotheses which rely on a student's pass/fail status *per se* changing her treatment in the subsequent school year are unlikely to explain our estimated retest policy effect. First, among students who barely failed the initial test, nearly two thirds (62 percent) passed the retest (Table 5 column 3). In other words, in the population to whom our LATE RD estimate applies, most students ended the school year having a label of "passed." This would substantially weaken any differences in future decisions by teachers or schools made on the basis of pass/fail status.

Nevertheless, many students failed again on the retest. Perhaps 0.03σ is simply a weighted average $= (0.62)0 + (0.38)0.08$, where failing the retest is the critical variable which triggers different treatment of the student in the subsequent school year. We test this hypothesis using RD methods to estimate the effect of barely failing the retest, compared to barely passing the retest, among the students who barely failed the initial test.

As shown in Table 5 column 1, we find no difference in the year $t + 1$ test scores of students who barely failed or barely passed the *retest*. In other words, our key outcome is not influenced by pass/fail status of the *retest*, but is influenced by pass/fail status of the *initial test*. This is consistent with teachers (schools) who react to the retest cutoff—and its strong evaluation incentives—but not other seemingly similar cutoffs. This result holds for students near the cutoff on the initial test, who are in the LATE population our main RD estimates apply to. Moving further away from the initial test cutoff, there is some evidence that students who fail the retest may be worse off, if anything.

A final test uses value-added-style estimates of teachers' contributions to *retest* scores. We first produce a "value added to retest" estimate for a given teacher j using conventional value-added methods, except that the dependent variable is the difference between student i's retest and initial scores (both from the same year

t).[23] In the second step we regress student i's test score from the following year, $t+1$, on the "value added to retest" score of student i's year t teacher.[24] In other words, we test whether teacher contributions to retest scores persist and predict student scores one year later. The final feature of this test uses the fact that retesting occurred before the retest policy began in 2009, as detailed above. In the second step regression we interact the "value added to retest" score with an indicator for the retest policy years.[25] In other words, we test whether the nature of teacher's contributions to retest scores changed when the policy changed.

The results of this test are consistent with two conclusions. Detailed results are shown in Appendix Table 2. First, teachers do make contributions to their student's retest scores beyond their contributions to initial scores; estimates of those contributions predict students' future test scores. If a teacher induces a 0.10σ improvement between initial and retest, her students are predicted to score approximately 0.026-0.035σ higher the following year. Second, the retest policy changed the nature of teachers' contributions to retest scores. The coefficient on prior teacher's "value added to retest" increases by about 15-25 percent in the retest policy years.

The evidence presented thus far suggests the retest policy effects were the result of teachers' (schools') actions during the weeks leading up to the retest.[26]

[23] We fit a specification where the dependent variable is retest minus initial score for student i in year t, and the right hand side has: student i's lagged math score, student demographic characteristics, student absences, and teacher fixed effects. The estimated teacher fixed effects are our "value added to retest" measure.

[24] This second step specification also includes fixed effects for year $t+1$ teacher, and controls for student demographics and absences, and student i's initial test score from year t.

[25] Data on retest scores prior to the retest policy are available for only one year, 2008; and, as described above, for only grades 3 and 5 given the retesting rules before the retest policy. Thus our "pre" period is limited to $t = 2008$. We limit the post period to 2009 for balance.

[26] Changes in teacher effort or decisions need not be changes by individual teachers independent of other teachers. For example, perhaps the teachers in a school decided to place all to-be-retested students in an ad-hoc remedial class, and then assigned the "best" math teacher to teach that ad-hoc class. This is still a change in effort, though a reallocation across teachers rather than simply within a teacher.

Whatever those actions were their effects persisted for at least a year and showed up on the $t+1$ test score, and so students presumably learned something during those 2-3 weeks. An additional piece of evidence consistent some learning is that students' retest scores are better predictors of their $t+1$ score than are their initial scores.[27]

What did retested students learn in those weeks? We cannot say for certain given the data available. The answer could be math skills, or test taking skills, or both; as long as those skills were valuable to next year's math test.[28] Moreover, math skills and test taking skills are not disjoint sets.[29] Whatever students learned, however, the resulting discontinuity in students' $t+1$ scores is consistent with teachers distorting their effort in response to the incentives of their evaluation measure.

4.4 School choices of test dates

One final complementary piece of evidence comes from the test dates chosen by schools. As we show in this subsection, before the retest policy schools with a higher proportion of failing students set their test date later in the window. But during the retest years the pattern goes away with such schools setting their test dates just as early as higher performing schools. Moving up the test date is consistent with schools which (plan to) make good use of the time before the retest.

[27] We estimate two simple bivariate regressions using only observations on Level II retested students from 2009-2012. For the regression of $A_{i,t+1}$ on the initial score (A_{it}) the *R*-squared is 0.11. For $A_{i,t+1}$ on the retest score the *R*-squared is 0.21.

[28] Our understanding is that teachers and schools did not have information about which specific test items students missed on the initial test, and thus could not use that information in their teaching before the retest.

[29] Koretz (2017) provides examples of test taking skills which overlap with math skills. Students can be coached to score higher based on knowing things like: (a) 3:4:5 and 5:12:13 are the most common Pythagorean triples used in test items, and (b) $y = mx + b$ is the most common representation of a simple linear equation used in test items. Knowing only 3:4:5 is not the same as fully understanding the Pythagorean Theorem, but knowing that 3:4:5 characterizes a right angle is a useful math skill, ask any carpenter.

After the retest policy ended, schools switched back. These choices are further evidence that teachers and schools understood and acted on the incentives created by the retest policy.

The relevant institutional details for this empirical test are as follows: The state of North Carolina sets a testing "window"; schools then choose which day, within that window, they will have their students take the test. Both the initial test and the retest had to occur during the window. During the retest policy years, 2009-2012, the window was the last 22 school days of the year. In the years before 2009, the window was the last 15 days.[30]

We estimate the following specification:

$$D_{st} = \alpha \bar{F}_{s,t-1} + \beta R_t + \delta R_t \bar{F}_{s,t-1} + \pi_t + \nu_{st} \quad (2)$$

where $\bar{F}_{s,t-1}$ is the school proportion of students failing the math exam the prior year, and $R_t = 1$ for the retest years. Recall that $\bar{F}_{s,t-1}$ is quite similar to the performance evaluation measures schools faced under NCLB and ABCs (see Section 1). The outcome variable, $D_{st} \in [0,1]$, is constructed so that if school s chooses to test on the first day of the window $D_{st} = 0$, and D_{st} is the proportion of the test window elapsed before the chosen test date. We also include non-parametric year controls, π_t, and cluster standard errors at the school level.

In Table 6 column 1 we show estimates fitting specification 2 but limiting the data to two years before and after the change (2007-2010). First, prior to the retest policy, a school with more failing students set their test dates later in the window (row 1). For example, a school with 10 points more failing in $t-1$ set its test date 1.3 percentage points further into the test window in year t. This average difference is not large, perhaps one-fifth of a day, but is statistically significant.

[30] For 2013 the window reverted to 15 days, and beginning in 2014 the window was 10 days.

Second, after the retest policy begins in 2009 the correlation between the proportion of failing students and test date essentially goes to zero.

When the retest policy ended after 2012, the patterns of test date choices reverted to what we observed in the pre-policy period, though perhaps slightly weaker. In column 2 we fit specification 2 limiting to two years before and after the end of the policy (2010-2014).

We test the robustness of these results in two ways. First, we simply pool all of the pre, during, and post retest years in column 3. The pattern of results is unchanged. Second, we fit specification 2 using placebo policy changes in columns 4 and 5. We find no changes in test date decisions at these placebo years.

5. Conclusion

Evaluation can change employee effort and decisions in both intended and unintended ways. In this paper we describe an empirical example of an unintended change; a distortion in teacher effort. The distortion occurs because teachers must allocate effort across many students (tasks), but the evaluation measure weights those students differently. An example of the multitask evaluation problem described by Holmstrom and Milgrom (1991).

Our paper is subtly but meaningfully different from prior papers which also document changes in teacher effort allocations across students. In Neal and Schanzenbach (2010), for example, a new evaluation system caused teachers to give more attention to students in the middle of the achievement distribution, and less attention to students higher (and lower) in the distribution. That change, however, was at least partly intended by the evaluation system, and depended on students' prior achievement.

In contrast, we document a case where students who are educationally identical (at expectation) ended up with different outcomes because of the evaluation system's incentives. Students who barely failed the end of year test, and

were later retested, were better off than their classmates who barely passed. Here "better off" is scoring higher on the state math test one year later. The barely failing students were no more or less deserving of their teachers' effort than were the barely passing, and no more or less likely to benefit from that effort. Barely failing students would have no higher or lower demand for their teachers' effort, at least absent being labeled "failing" by the test. In other words, the regression discontinuity difference we estimate is inconsistent with the notion that teachers are perfectly motivated agents whose decisions are invariant to performance evaluation.

The estimated effects on student test scores are most likely the result of teachers (schools) response to the retest policy's evaluation incentives. To reiterate those incentives: Effort directed at students who failed the initial test would unambiguously increase the teacher's (school's) performance measure, because only the higher of a student's initial and retest score would be used to calculate the measure. Effort directed at otherwise-identical passing students, however, would not change the performance measure. Our regression discontinuity estimate of the resulting benefit to barely failing students is approximately 0.03σ. That difference occurs only in the retest policy years, but not before or after. Additionally, the difference is not the result of retesting *per se* the data suggest. In the years before the retest policy began in 2009 failing students in select grades and districts were retested, but those pre-2009 retest scores were not used in the teacher and school evaluation measures. For those select grades and districts the retest policy only changed how retest scores would be used to judge the adults, and we find the same effects of the retest policy. Several other changes, or lack of changes, we observe in the student and school behavior are also consistent with a distortion in teacher and school decisions.

These results emphasize the importance of careful design in employee performance evaluation systems. The particulars of this paper's setting—especially

the student-level discontinuity and the variation in retest policy over time—are empirically advantageous for demonstrating distortion from evaluation. But the broader concern about distortionary incentives is more general. Evaluation designers—like legislators, superintendents, principals, and unions in the schools setting—would do well to think explicitly about the implicit incentives.

References

Adnot, Melinda. (2016). Effects of incentives and feedback on instructional practice: Evidence from the District of Columbia Public Schools' IMPACT teacher evaluation system.

Aucejo, Esteban M., and Teresa Foy Romano. (2016). Assessing the effect of school days and absences on test score performance. *Economics of Education Review, 55* : 70-87.

Baker, George P. (1992). Incentive contracts and performance measurement. *Journal of Political Economy, 100* (3): 598-614.

Burgess, Simon M., Carol Propper, Helen Slater, and Deborah Wilson. (2005). Who wins and who loses from school accountability? The distribution of educational gain in English secondary schools. CMPO Working Paper.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. (2014). Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood. *American Economic Review, 104* (9): 2633-79.

Cullen, Julie Berry, and Randall Reback. (2006) Tinkering toward accolades: School gaming under a performance accountability system. In Timothy J. Gronberg and Dennis W. Jansen (eds.) *Improving School Accountability (Advances in Applied Microeconomics, Volume 14)*. Emerald Group Publishing Limited.

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

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

Figlio, David N., and Lawrence S. Getzler. (2006). Accountability, ability and disability: Gaming the system? In Timothy J. Gronberg and Dennis W. Jansen (eds.) *Improving School Accountability (Advances in Applied Microeconomics, Volume 14)*. Emerald Group Publishing Limited.

Hanushek, Eric A., and Steven G. Rivkin. (2010). Generalizations about using value-added measures of teacher quality. *American Economic Review, 100* (2): 267-71.

Holmstrom, Bengt, and Paul Milgrom. (1991). Multitask principal-agent analyses: Incentive contracts, asset ownership, and job design. *Journal of Law, Economics, and Organization, 7* (special issue): 24-52.

Jackson, C. Kirabo, Jonah E. Rockoff, and Douglas O. Staiger. (2014). Teacher effects and teacher-related policies. *Annual Review of Economics, 6* (1): 801-825.

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

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

Koretz, Daniel. (2017). *The testing charade: Pretending to make schools better.* University of Chicago Press.

Macartney, Hugh. (2016). The dynamic effects of educational accountability. *Journal of Labor Economics, 34* (1): 1-28.

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

Raudenbush, Stephen W., Sean F. Reardon, and Takako Nomi. (2012). Statistical analysis for multisite trials using instrumental variables with random coefficients. *Journal of Research on Educational Effectiveness, 5* (3): 303-332.

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

Reback, Randall, Jonah Rockoff, and Heather L. Schwartz. (2014). Under pressure: Job security, resource allocation, and productivity in schools under No Child Left Behind. *American Economic Journal: Economic Policy, 6* (3): 207-41.

Sims, David P. (2008). Strategic responses to school accountability measures: It's all in the timing. *Economics of Education Review, 27* (1): 58-68.

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

Taylor, Eric. (2014). Spending more of the school day in math class: Evidence from a regression discontinuity in middle school. *Journal of Public Economics, 117* : 162-181.

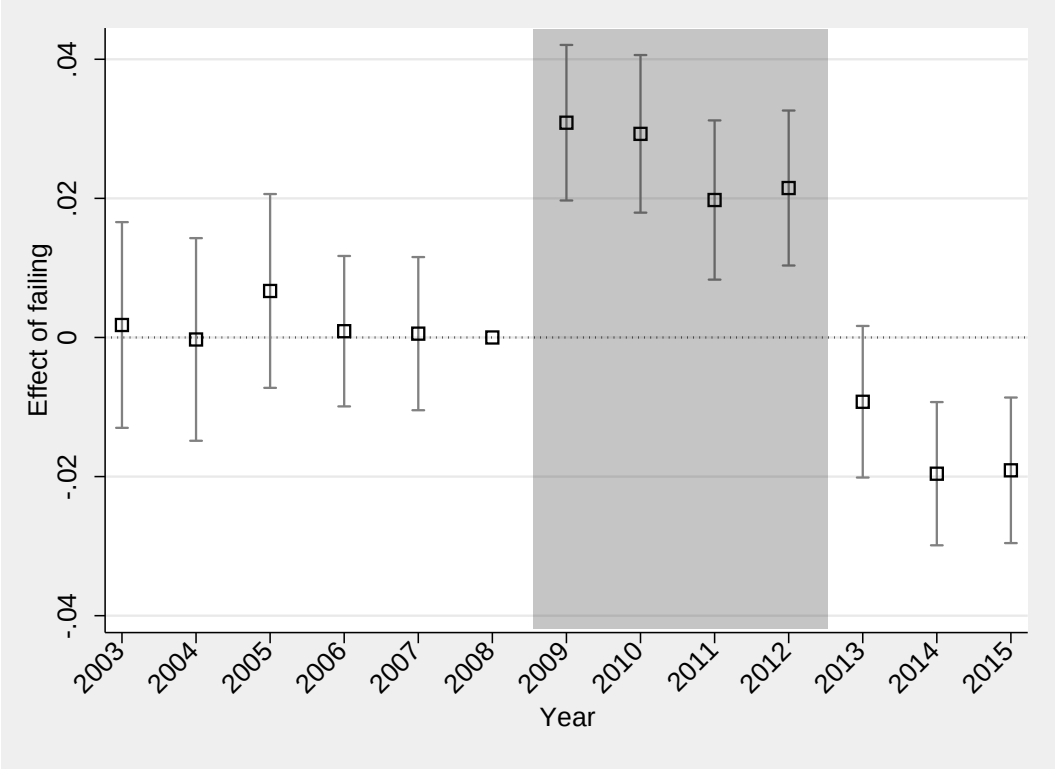


Figure 1—Estimated (RD) effect of failing math on next year's math score; before, during (2009-2012), and after the retest policy

Note: Each hollow square represents, for a given school year (x-axis = t), the estimated effect of failing the end-of-year t math test on math test score at $t + 1$ for students near the pass/fail cutoff, measured in student standard deviation units (y-axis). Each hollow square is a regression discontinuity (RD) estimate using the local linear regression methods described in the text. Vertical lines represent the 95 percent confidence interval for each RD estimate. All estimates are relative to 2008.

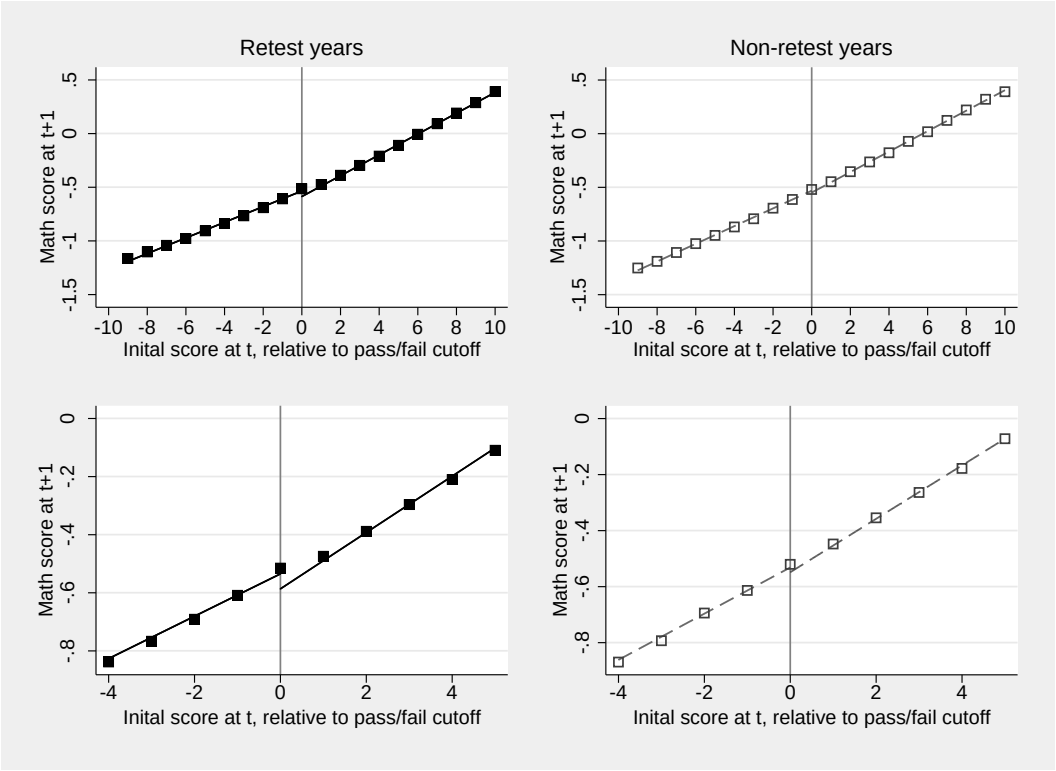


Figure 2—Initial math test scores and scores one year later

Note: Each square represents the mean outcome math score, $A_{i,t+1}$, in student standard deviation units (y-axis) for students with a given initial test scale score (x-axis), net of grade-by-year-by-school fixed effects. Filled squares are pooling retest policy years. Hollow squares are pooling non-retest policy years. Fitted lines are by OLS using data at the student-year observation level. The first and second rows are identical, except that in the second row the x-axis range is smaller to aid in visibility of the discontinuity; the square values and fitted line slopes are identical.

Table 1—Covariate balance

Pre-treatment covariate	RD estimate of difference at pass/fail cutoff: Non-retest years (1)	Difference-in-RD estimate: Retest years − Non-retest years (2)	Observations (3)
Math score, t-1	-0.004	0.003	3,695,517
	(0.004)	(0.009)	
Reading score, t-1	-0.002	-0.001	3,661,258
	(0.004)	(0.005)	
Retained in year t	0.000	0.000	3,674,709
	(0.000)	(0.001)	
Female	0.002	-0.008**	3,691,893
	(0.002)	(0.003)	
White	-0.001	0.001	3,691,893
	(0.001)	(0.002)	
Days absent	0.043	0.072	3,686,276
	(0.041)	(0.049)	
Free or reduced lunch	0.005*	-0.004	3,276,286
	(0.002)	(0.004)	
Special education	-0.009**	0.012**	3,276,286
	(0.003)	(0.003)	
Limited English proficiency	0.001+	0.001	3,684,301
	(0.001)	(0.002)	

Note: Each row reports estimates from a separate local linear regression with student-by-year observations. Each dependent variable is a pre-treatment student characteristic, and is described in the row label. The specification is the difference-in-RD described in detail in the text. The right-hand-side has separate linear terms for the running variable (initial math test score in year t) above and below the cutoff, and the slopes are allowed to differ in the retest policy years versus non-retest years. The specification also includes school-by-grade-by-year fixed effects. Standard errors in parentheses are clustered at the values of the running variable. Free or reduced lunch and special education data not available for 2004 and 2005.
+ indicates p<0.10, * 0.05, and ** 0.01

Table 2—Difference-in-RD estimates and robustness tests

	RD estimate of difference at pass/fail cutoff	Difference-in-RD estimate	
	Non-retest years	Retest years − Non-retest years	Observations
	(1)	(2)	(3)
Main estimate	-0.002	0.031**	3,978,190
	(0.008)	(0.005)	
Alternative specifications			
$f(Y_{it})$ parameters by retest years v. non-retest years	0.006	0.024**	3,978,190
	(0.007)	(0.007)	
$f(Y_{it})$ parameters by grade-by-year	0.002	0.030**	3,978,190
	(0.009)	(0.005)	
Grade-by-year FE	-0.002	0.033**	3,978,190
	(0.009)	(0.005)	
Alternative bandwidths			
1/4	0.024**	0.028**	1,028,410
	(0.001)	(0.002)	
1/2	0.011+	0.033**	2,085,368
	(0.006)	(0.006)	
3/4	0.006	0.031**	3,009,167
	(0.006)	(0.005)	
Only 2006-2012, no change in math test	0.005	0.025**	2,430,907
	(0.010)	(0.005)	

Note: Each row reports estimates from a separate local linear regression with student-by-year observations. For the main estimate in row 1: The dependent variable is the student's standardized math test score in year $t+1$. The specification is a difference-in-RD. The right-hand-side has separate linear terms for the running variable (initial math test score in year t) above and below the cutoff, and the slopes are allowed to differ in each year as in Figure 1. The specification also includes school-by-grade-by-year fixed effects. For rows 2 through the end: The estimation details are identical to row 1, except for the variation(s) describe in the row headers. Standard errors in parentheses are clustered at the values of the running variable.
+ indicates p<0.10, * 0.05, and ** 0.01

Table 3—Retesting before 2009 and grade level estimates

	RD estimate of difference at pass/fail cutoff	Difference-in-RD estimate	
	Non-retest years	Retest years − Non-retest years	Observations
	(1)	(2)	(3)
Grades 3 and 5; retesting before 2009			
District retested at pass/fail cutoff in 3, 5, and 8	-0.007	0.034**	459,461
	(0.013)	(0.010)	
District used SEM rule	-0.009	0.029**	798,773
	(0.009)	(0.008)	
Grades 4, 6, and 7; no retesting before 2009			
District retested at pass/fail cutoff in 3, 5, and 8	0.022*	0.018**	657,417
	(0.010)	(0.006)	
District used SEM rule	0.015	0.018**	1,156,740
	(0.010)	(0.006)	
Grade level			
3	-0.008	0.033**	859,919
	(0.013)	(0.007)	
4	0.014	0.016**	782,929
	(0.013)	(0.005)	
5	-0.013+	0.036**	785,195
	(0.007)	(0.005)	
6	0.002	0.025*	778,896
	(0.011)	(0.012)	
7	0.016*	0.040**	771,251
	(0.006)	(0.010)	

Note: Each row reports estimates from a separate local linear regression with student-by-year observations. The estimation details are identical to Table 2 row 1 (the main estimate), except the estimation sample is restricted to subsamples as defined in the row headers. In all rows the specification is a difference-in-RD. The right-hand-side has separate linear terms for the running variable (initial math test score in year t) above and below the cutoff, and the slopes are allowed to differ in each year as in Figure 1. Standard errors in parentheses are clustered at the values of the running variable.

+ indicates p<0.10, * 0.05, and ** 0.01

Table 4—Alternative dependent variables

	RD estimate of difference at pass/fail cutoff	Difference-in-RD estimate	
	Non-retest years	Retest years − Non-retest years	Observations
	(1)	(2)	(3)
Absences in year t+1	0.010	-0.013	3,928,054
	(0.018)	(0.032)	
Reading test score t+1	0.001	0.006+	3,933,249
	(0.004)	(0.003)	
Mean year t score of t+1 peers	0.010*	-0.001	1,941,699
	(0.005)	(0.003)	
Proportion t+1 peers failed t test	0.026**	-0.001	1,941,699
	(0.001)	(0.001)	
Value-added score of t+1 teacher	0.002+	-0.001	1,676,408
	(0.001)	(0.001)	
Retained (same grade t and t+1)	0.003**	0.000	3,212,750
	(0.001)	(0.000)	

Note: Each row reports estimates from a separate local linear regression with student-by-year observations. The estimation details are identical to Table 2 row 1 (the main estimate), except with an alternative dependent variable described in the row headers. In all rows the specification is a difference-in-RD. The right-hand-side has separate linear terms for the running variable (initial math test score in year t) above and below the cutoff, and the slopes are allowed to differ in each year as in Figure 1. Standard errors in parentheses are clustered at the values of the running variable.
+ indicates p<0.10, * 0.05, and ** 0.01

Table 5—Retest scores

Scale score points below pass/fail cutoff	RD est. diff. at pass/fail cutoff Retest years	Observations	Proportion passing on retest	Mean improvement retest-initial
	(1)	(2)	(3)	(4)
1	-0.001	60,377	0.616	0.164
	(0.009)			
2	0.005	58,079	0.521	0.157
	(0.009)			
3	0.006	48,690	0.435	0.159
	(0.010)			
4	-0.013	45,465	0.361	0.171
	(0.011)			
5	-0.023+	40,144	0.289	0.180
	(0.012)			
6	-0.022	34,668	0.233	0.199
	(0.015)			
7	-0.009	32,260	0.183	0.216
	(0.021)			
8	-0.036+	26,732	0.144	0.240
	(0.021)			
9	-0.040+	26,096	0.105	0.250
	(0.024)			
10	-0.061+	16,763	0.081	0.274
	(0.034)			

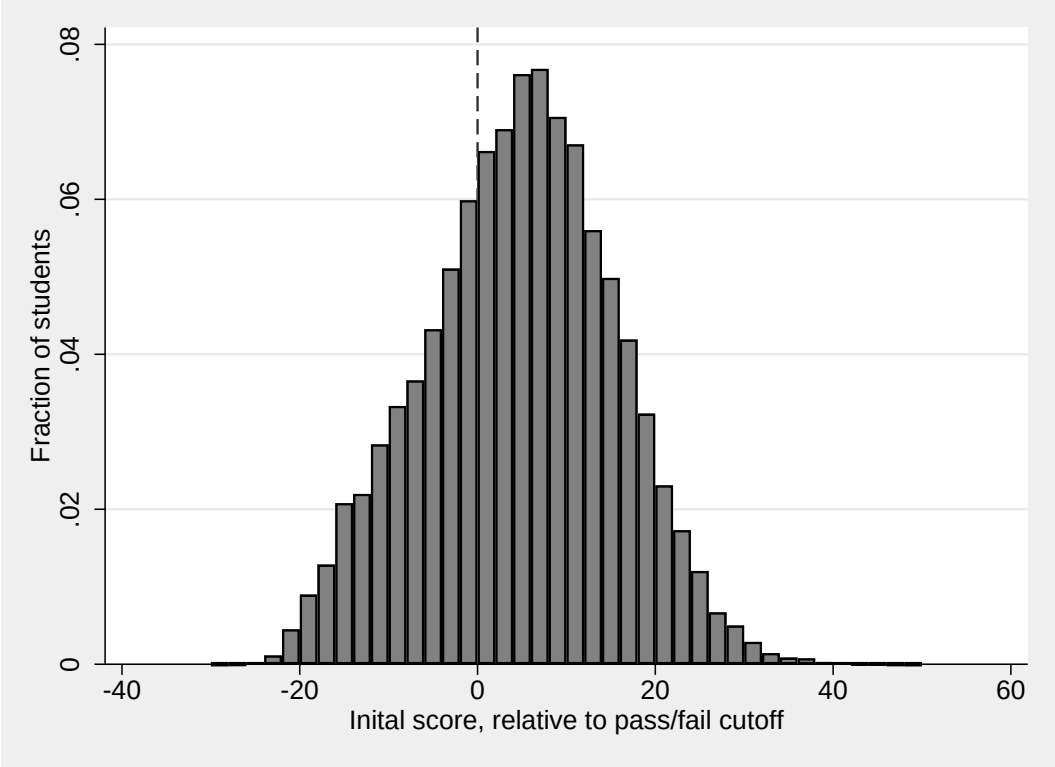
Note: In each row, column 1 reports a regression discontinuity estimate from a separate local linear regression with student-by-year observations. The dependent variable is the student's standardized math test score in year $t + 1$. In contrast to other estimates in the paper, the running variable is the student's retest score. The right-hand-side has separate linear terms for the running variable above and below the cutoff. Standard errors in parentheses are clustered at the values of the running variable. Row 1 is estimated using the sample of students who scored 1 scale score point below the pass/fail cutoff on the initial t test, row 2 for the sample 2 points below, and so on. Column 3 reports the proportion of students who passed on the retest, i.e., scored above the pass/fail cutoff on the retest. Column 4 reports the mean difference between retest score and initial score.
+ indicates p<0.10, * 0.05, and ** 0.01

Table 6—School initial test date choice
Dep. var. = proportion of test window elapsed before test date (0 on first day, 1 on last)

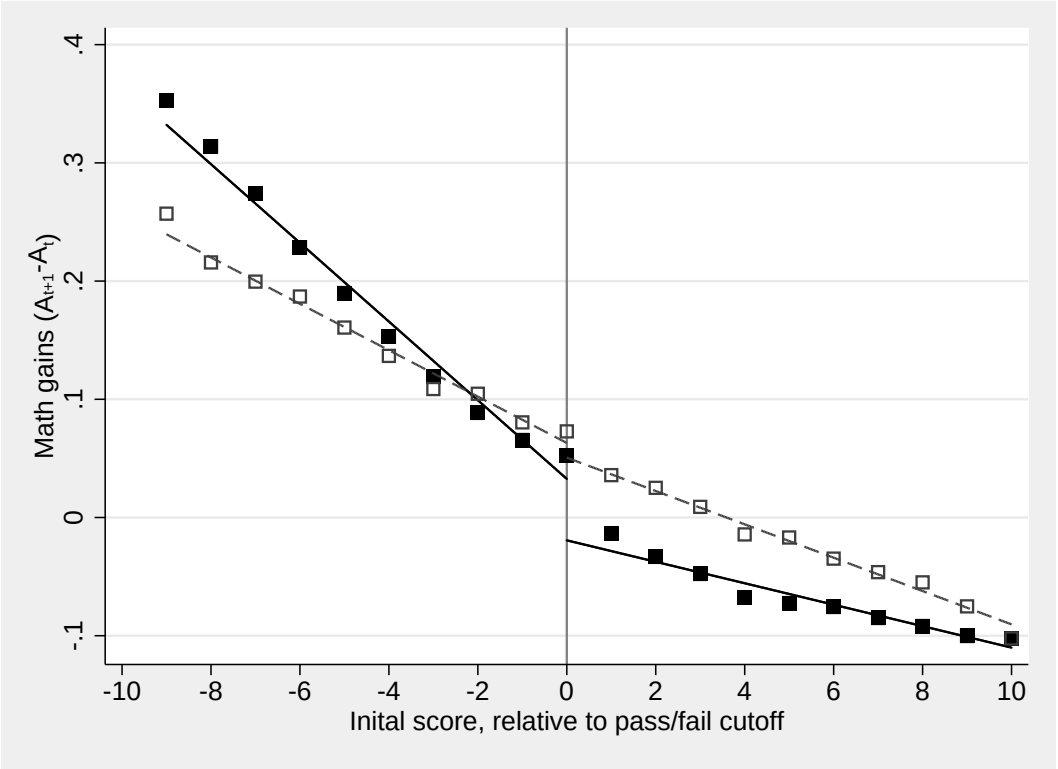
	Years used in estimation				
	2007-2010	2011-2014	All years	2005-2008	2009-2012
	(1)	(2)	(3)	(4)	(5)
Proportion failing t-1	0.133**	0.084**	0.108**	0.109+	-0.005
	(0.028)	(0.025)	(0.020)	(0.056)	(0.024)
Proportion failing t-1	-0.138**	-0.072*	-0.105**		
* Retest years 2009-2012	(0.029)	(0.029)	(0.022)		
Proportion failing t-1				0.024	
* Placebo years 2007-2008				(0.054)	
Proportion failing t-1					0.017
* Placebo years 2011-2012					(0.023)
School-by-year observations	6,444	6,612	17,050	5,720	6,801

Note: Each column reports estimates from a separate least squares regression. The dependent variable is the proportion of the test window elapsed before the date on which the test was given by school s in year t. The right-hand-side includes the regressors show above and year fixed effects. Standard errors in parentheses are clustered at the school level.
+ indicates p<0.10, * 0.05, and ** 0.01

Online Appendix



Appendix Figure 1—Histogram of end-of-year initial math test score (running variable)



Appendix Figure 2—Initial math test scores and *gain* scores one year later

Note: This figure is constructed just as Figure 2 is, except that the y-axis is gain scores. Each square represents the mean score gain ($gain = A_{i,t+1} - A_{it}$) in student standard deviation units (y-axis) for students with a given initial test scale score (x-axis), net of grade-by-year-by-school fixed effects. Filled squares are pooling retest policy years. Hollow squares are pooling non-retest policy years. Fitted lines are by OLS using data at the student-year observation level.

Appendix Table 1—Additional results

	RD estimate of difference at pass/fail cutoff	Difference-in-RD estimate	
	Non-retest years	Retest years − Non-retest years	Observations
	(1)	(2)	(3)
District change in retention at pass/fail after policy ended in 2010			
Bottom tercile of change (less retention)	-0.000	0.032**	654,634
	(0.012)	(0.005)	
Middle tercile (roughly no change)	-0.001	0.030**	2,104,152
	(0.009)	(0.005)	
Top tercile (more retention)	-0.003	0.034**	1,187,244
	(0.007)	(0.010)	

Note: Each row reports estimates from a separate local linear regression with student-by-year observations. The estimation details are identical to Table 2 row 1 (the main estimate), except the estimation sample is restricted to subsamples as defined in the row headers. In all rows the specification is a difference-in-RD. The right-hand-side has separate linear terms for the running variable (initial math test score in year t) above and below the cutoff, and the slopes are allowed to differ in each year as in Figure 1. Standard errors in parentheses are clustered at the values of the running variable.

+ indicates p<0.10, * 0.05, and ** 0.01

Appendix Table 2—Teacher value added to retest scores as a predictor for future scores

	All retested students		"Level II" students
	All LEAs	LEAs which used retest rule in 2008	
	(1)	(2)	(3)
Year t teacher's value-added to retest score	0.283**	0.278**	0.261**
	(0.017)	(0.025)	(0.03)
Retest policy year	0.052**	0.026+	0.009
	(0.009)	(0.013)	(0.016)
Retest policy year *	0.047*	0.069*	0.074+
Year t teacher's value-added to retest score	(0.023)	(0.034)	(0.041)

Note: Each column reports estimates from a separate least squares regression. The dependent variable is student i's initial test score in year $t+1$. The key dependent variable is the "value added to retest" score for student i's year t teacher. See text for the description of this value added score. The specification also includes fixed effects for year $t+1$ teacher. Additional covariates are year t initial test score; indicators for gender, race/ethnicity, limited English proficient; and days absent. The estimation sample is limited to $t = 2008$ and 2009, before and after the retest policy began respectively. Standard errors in parentheses are clustered at the teacher level.
+ indicates p<0.10, * 0.05, and ** 0.01