# Does evaluation change teacher effort and performance? 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

July 2020

We document measurable, lasting gains in student achievement caused by a change in teachers' evaluation incentives. A short-lived rule created a discontinuity in teachers' incentives when allocating effort across their assigned students: students who failed an initial end-of-year test were retested a few weeks later, and then only the higher of the two scores was used when calculating the teacher's evaluation score. One year later, long after the discontinuity in incentives had ended, retested students scored 0.03σ higher than non-retested students. Otherwise identical students were treated differently by teachers because of evaluation incentives, despite arguably equal returns to teacher effort.

(JEL No: I2, M5)

[†] Aucejo: Arizona State University; Romano: Oxford College of Emory University; Taylor: Harvard University. Generous financial support was provided by the Jacobs Foundation. We greatly appreciate the help of the North Carolina Education Research Data Center. We thank our anonymous reviewers and seminar participants at AEFP, APPAM, CESifo, BYU, Harvard, Stanford, and Virginia 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. The design of evaluation can be particularly challenging in multitask jobs where often the evaluation measure captures only a subset of (or differentially weights) the employee's many job tasks (Holmstrom and Milgrom 1991, Baker 1992). Improved performance on measured or incentivized tasks can come at the expense of poorer performance on un-measured or non-incentivized tasks. In this paper we provide an empirical example from schools in North Carolina. A teacher's different students are the different job tasks over which she must allocate effort, but her evaluation incentives differ sharply across students who are otherwise identical.

We document relatively large and long-lasting gains in students' math achievement caused by a change in teachers' evaluation incentives. The gains arose, most likely, during a brief 2-3 week period when the incentives were active, but the gains lasted a year or more beyond that period. However, these gains came at a cost. The new incentives caused new inefficiencies in how teachers allocated their effort across students, creating new inequities.

A short-lived feature of North Carolina evaluation rules created sharp differences across students (job tasks) in the returns to teacher effort—specifically, returns in the form of higher evaluation scores. During all years in our data (2003 to 2015), teachers and schools were evaluated based on end-of-school-year student test scores.[1] During the "retest policy years" (2009 to 2012), 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 new rule created strong and clear incentives to allocate more effort to students who

[1] For simplicity, throughout the paper we refer to school years by their spring number, thus the 2002-03 school year is $t = 2003$. In the education sector, the term "accountability" is often used to mean the same thing as performance evaluation with explicit (implicit) incentives.

failed the initial test and would soon be retested. Effort directed at retested students could only increase teacher evaluation scores, while effort directed at non-retested students could not change teacher evaluation scores. Critically, the incentives changed discontinuously at the pass/fail cutoff score. Students just above (not retested) and just below (retested) that cutoff were weighted very differently in the teacher evaluation score even though they were otherwise identical.

The retest policy—with its new incentives for teachers—made retested students better off, as measured by larger gains in their math achievement. We use regression discontinuity (RD) methods to compare students who barely fail (retested) and barely pass (not retested) the year $t$ math test. Our main outcome measure is the year $t + 1$ math test, which comes nearly one year after the teachers' retest incentives end. Our outcome is not the retest score itself. The difference in scores at $t + 1$ is approximately 0.03 student standard deviations (σ).

This 0.03σ gain is educationally and economically meaningful, even if a small share of the total variation in student scores. First, this gain in output is large relative to the change in inputs. The improvement (most likely) arose in just 2-3 weeks of work. For comparison, over an entire typical school year, the standard deviation of teacher contributions to student achievement is 0.10-0.20σ (Hanushek and Rivkin 2010, Jackson, Rockoff, and Staiger 2014). Second, the gains were long lasting. Our 0.03σ estimate is the effect measured nearly one year after the retest, and thus one year after the discontinuity in incentives across students had gone away. The effect at two years out is 0.023σ. Moreover, given the typical decay (or "fade out") of student test score effects, it is likely the effect at the end of the 2-3 weeks was as much as 0.06σ. Finally, a back-of-the-envelope application of estimates from Chetty, Friedman, and Rockoff (2014) suggests that the 0.03-0.06σ gain may be worth $1,500-3,000 per student in net present earnings.

Retested students were better off, but the gains came at a cost. The gains reveal a new allocative inefficiency created by teachers' response to evaluation

incentives. If our identifying assumptions hold, then the students who barely failed and those who barely passed are identical (at expectation), except that those who failed were favored by the retesting incentives for teachers. Retested and non-retested students would have had equal returns to teacher effort, but they were not given the same effort. This new allocative inefficiency is likely at odds with the objectives of the school system, and likely an unintended consequence of the change in evaluation rules. Moreover, our results are suggestive of the distortions that can arise in evaluation of multitask jobs (Holmstrom and Milgrom 1991). Otherwise identical students were treated differently by their teachers because of the different evaluation incentives attached to different students. However, a full welfare analysis would require a richer set of measures, and a different identification strategy.[2]

Our identification strategy uses two sources of variation: First, the discontinuity at the pass/fail cutoff in how students are treated. Second, variation over time in state evaluation rules. We combine these two in a difference-in-RD, or event study of RD estimates, strategy. The pass/fail cutoff exists in all thirteen years of our data, but the retest policy incentives at that cutoff exist only from 2009-2012.

The 0.03σ difference in achievement between barely failing and barely passing 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 new rules, and incentives, in the retest policy years.

Causal interpretation of our diff-in-RD estimates requires a weaker assumption than the simple RD or simple diff-in-diff. If we only had data from the

[2] For example, perhaps teachers (schools) took short run actions—i.e., giving more effort only to retested students—because they knew there would be long run benefits—e.g., avoiding the sanctions of NCLB would leave more resources to improve the achievement of all students.

four retest policy years, and thus only a conventional RD estimate, we would need the conventional RD assumption: that absent the retest policy's new discontinuity in teacher incentives, students' math scores would be continuous at the pass/fail cutoff. Using the diff-in-RD design we can relax the assumption by adding: however, if there are discontinuities in potential outcomes unrelated to the new incentives, those discontinuities are constant across the retest policy years (treatment) and comparison years. In the body of the paper we show evidence, and discuss institutional details, consistent with these assumptions. A key consideration here is that other relevant educational inputs may be discontinuously assigned at the pass/fail cutoff; we discuss this in detail in Section 4. Similarly, a conventional diff-in-diff would also require the stronger assumption that the appropriate counterfactual for North Carolina can be estimated with some (weighted) average of other states. More importantly, the conventional diff-in-diff would estimate a different parameter, some broad effect of the policy. The diff-in-RD focuses more sharply on the discontinuity in teacher evaluation incentives across students (tasks).

One key result emphasizes the distinction between the effects of retesting *per se* and the effect of the retest policy's new incentives. In select grades and school districts, there was some retesting prior to the 2009-2012 retest policy years. Importantly, however, the pre-2009 retest scores were not used in calculating teacher evaluation scores. 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 retest policy years. In fact, we do not see such effects. Effects only appear when the stakes change for the adults in the school.

The most likely mechanism for our main result, we argue, is that students were taught more math during the 2-3 weeks before the retest, and that extra learning persisted one year later. However, the $0.03\sigma$ difference at $t+1$ could arise from other types of mechanisms: students themselves behaved differently

independent of any prompting from their teachers,[3] or teachers (schools) treated students differently during the $t+1$ school year after the retest. Across several empirical tests we do not find evidence supporting these alternatives. For example, the pre-2009 retesting shows no evidence of changes in student behavior, and we find no effect on student absences during $t+1$ as one imperfect measure of effort. Consistent with increased effort in the 2-3 weeks before the retest, we find evidence that from 2009-2012 schools chose test dates, which are partly at their discretion, to increase the time between the initial test and retest. But before and after the retest policy years, schools chose dates to increase time before the initial test. Several additional tests are described in Section 4.

While the likely mechanism is changes in teacher effort, we do not have data to further describe those changes. Teachers may have shifted effort across students, from non-retested to retested; or may have shifted across subjects within retested students, for example from art to math. Teachers could have reduced their own leisure to give more effort to retested students. Additionally, the changes may have been individual teacher decisions, or may have been collective decisions taken by all teachers (and administrators) in the school. We cannot test these alternatives empirically. Nevertheless, all of these hypotheses require a change in teacher effort.

Our paper makes a novel contribution to the broad literature on differences in teacher job performance, and the more specific literature on how evaluation affects that performance. A well-established literature shows large differences between teachers in their causal contributions to student test scores, among other outcomes (see Jackson, Rockoff, and Staiger 2014 for a review). Existing estimates measure teacher contributions which accumulated over an entire school year. This

[3] Student and teacher effort are complementary inputs. One fundamental way that teacher effort creates student achievement is through inducing students to give greater effort. Here we have in mind mechanisms due to changes student effort (relatively) independent of changes teacher effort.

paper shows that, at least over short intervals (2-3 weeks), teachers can make even larger contributions to student learning.

The existing literature also includes many papers on intended and unintended responses to evaluation by teachers. On the intended side of the ledger, we have several (quasi-)experimental examples testing whether evaluation improves outcomes for students (see for example Neal and Schanzenbach 2010, Glewwe, Ilias, and Kremer 2010, Dee and Jacob 2011, Taylor and Tyler 2012, and Deming et al. 2016, among others).[4] One empirical challenge is that in practice incentives are often weak and difficult for teachers to understand. Our contribution is a case where the incentives are quite strong and obvious; for a similar example see Dee and Wyckoff (2015). Though in both this paper and Dee and Wyckoff (2015) the generalizability of the estimates are limited by the relatively narrow RD LATE. On the unintended side of the ledger, there are examples of teachers manipulating the evaluation measure; the clearest example being outright cheating on tests (Jacob and Levitt 2003).[5] Consistent with the Holmstrom and Milgrom (1991) theoretical example, teachers appear to shift effort away from untested subjects and toward tested subjects (Jacob 2005, Glewwe et al. 2010, Dee and Jacob 2011, though Cohodes 2016 is a counter example). Also, the results in Macartney (2016) suggest schools reallocate effort across years within students. This paper adds an example where evaluation incentives cause teachers to reallocate of effort across otherwise identical students.

[4] For reviews of the literature see Neal (2011), Jackson, Rockoff, and Staiger (2014), Deming and Figlio (2016), and Neal (2018). Additional examples of how school accountability programs effect student performance include Chiang (2009), Rockoff and Turner (2010), and Rouse et al. (2013). Duflo, Dupas, and Kremer (2011) is an example of how changes in other education management decisions can interact with existing teacher evaluation incentives to change student outcomes.

[5] Other examples include manipulating which students are tested (Cullen and Reback 2006), e.g., by differential suspension from school during the test period (Figlio 2006), special education designation (Jacob 2005, Figlio and Getzler 2006), and 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, Glewwe et al. 2010, Koretz 2017).

A small set of papers focus, as we do, on teachers' effort allocation across students, and how evaluation incentives can change that allocation: Burgess et al. (2005), Reback (2008), Springer (2008), and Neal and Schanzenbach (2010). The common theme of these papers is to show how teachers (schools) respond when the (new) evaluation measure is simply the percent of student who pass the end-of-year exam. The simple incentive in such systems is to give more effort to students at the margin of passing or failing—the "bubble kids"—and all four papers show evidence that those marginal students did make larger gains in achievement. The papers differ on the question of whether those gains for marginal students came at the expense of losses for their higher or lower performing peers.

Despite similarities, the retest policy incentives we study in this paper are quite different from the "bubble kids" incentives. The "bubble kids" incentives are a smooth function of students' prior scores, while the retest policy incentives have a sharp discontinuity. That discontinuity is econometrically advantageous, allowing us to compare students who are identical (in expectation) except for a large difference in evaluation incentives attached to them. Thus, we can make stronger claims about the causal effects of evaluation, including claims about teachers' willingness to treat students differently in response to evaluation incentives. The disadvantage is that our results, strictly speaking, apply to a relatively small population of students, while the four papers cited above characterize important changes in outcomes for a wide range of students.

In the next section we describe the evaluation program and incentives teachers (schools) faced across the time period we study, and 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. We conclude in Section 5.

## 1. Setting and data

We study students, teachers, and schools in North Carolina over a 13-year period from 2003-2015. (For simplicity, throughout the paper we refer to school years by their spring date, thus the 2002-03 school year is $t = 2003$.) Throughout this 13-year 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: 2009-2012. In this section we describe key details of the evaluation systems, and how the "retest policy" years differed. Figure 2 is a timeline of key changes. Additional details are provided in the online appendix.

All data used in this paper were provided by the North Carolina Education Research Data Center. The data are typical of school administrative data, including annual records for each student with school, test scores, demographic characteristics, and program participation variables.

### *1.1 School and teacher evaluation in North Carolina, 2003-2015*

In all years we study, North Carolina teachers and schools were subject to both federal and state evaluation systems. The federal system is commonly known as No Child Left Behind (NCLB) and began in 2003. Students in grades 3-8 were tested annually each spring in math and reading. The student scores were then converted into a performance measure used to evaluate teachers and schools, specifically, the percent of students who passed the exam. The measure was often referred to as "percent proficient."

NCLB specified a set of escalating consequences for a school if its percent passing did not rise year after year. In the worst case, a school which failed to meet its annual targets for four or five consecutive years could simply be closed or have its entire faculty and staff fired. Intermediate consequences included paying for more tutoring, allowing students to transfer out, and writing an improvement plan.

Further, each school's annual targets were, in fact, a vector of "percent proficient" values, one for each of several subgroups defined by grade level, test subject, and student demographic categories.

North Carolina's state evaluation system, known as the ABCs of Public Instruction (ABCs), began in 1997. ABCs used the same annual student tests as NCLB, but converted those scores into two distinct performance measures to evaluate schools and teachers: (1) the percent of students who passed the exam or "percent proficient," just as under the NCLB system; and (2) a "test score growth" measure. The growth measure was the school mean of $\left(Y_{i,t} - \hat{\beta} Y_{i,t-1}\right)$, where $Y_{i,t}$ is the test score for student $i$ at the end of school year $t$. The $\hat{\beta}$ term remains (relatively) fixed over time and was estimated by regressing $Y_{i,t}$ on $Y_{i,t-1}$ using data from a period before $t-1$. The growth measure was responsive to any increase (decrease) in individual student test scores, not just changes in passing status.

The consequences and incentives under ABCs were quite different from NCLB. First, based on ABCs performance measures, North Carolina 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." From 2013 the labels were replaced with A-F letter grades. Second, initially 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 2008 school year.[6]

[6] The end of bonuses is concurrent with the beginning of the "retest policy" in 2009. As a first-order effect, the end of bonuses should weaken teachers' incentives to focus on student test scores generally, and thus (likely) bias against our finding an effect of the retest policy. Perhaps teachers reduced effort and attention on "test score growth" and increased effort on "percent proficient," and thus increased effort toward students near the margin of passing and failing. But such a shift would not create or explain a discontinuity at that passing cutoff.

Both NCLB and ABCs were team evaluation not individual evaluation. Neither NCLB nor ABCs calculated performance measures specific to individual teachers; neither included the individual "teacher value-added scores" which have become more common in recent years. Nevertheless, NCLB and ABCs were teacher evaluation systems, even if teachers were evaluated in grade-level teams or subject teams or teams called "schools." Student scores include many family and school inputs; but among school inputs, teacher labor is by far the most influential (Jackson, Rockoff, and Staiger 2014). Moreover, the teams could be quite small, for example, NCLB measures performance at the grade-by-subject level for each school.

North Carolina teachers were also subject to individual evaluation, but not during the entire the period we study. Beginning in 2008 and each year after, the state calculated individual teacher "value added" scores using the annual grade 3-8 math and reading tests. Beginning in 2010, a state policy change allowed (but did not require) districts to use value-added scores in formal evaluation. Finally, beginning in 2012 and through the end of our data, state policy required districts to use value-added scores as one component of formal individual evaluation, alongside classroom observation measures and subjective measures.[7]

*1.2 The retest policy years, 2009-2012*

School years 2009 through 2012 were the "retest policy years" or the "treatment years." This subsection describes how evaluation rules during the retest policy years differed (or did not differ) from the details described in the previous subsection. First, as in all other years, students in grades 3-8 were tested annually

[7] This section describes the details of NCLB, ABCs, and individual value-added scores relevant to grade 3-8 math and reading teachers. Additional details are provided in the appendix. In 2012 North Carolina, along with many other states, received an "NCLB waiver" which reset, but did not eliminate, the percent passing targets. From 2013 the name ABCs was replaced by READY Accountability, but the relevant features of the evaluation system did not change.

each spring in math and reading. We call this the "initial test score" or the score at time $t$. This initial score is observed for all students.

Second, students who failed the initial test were retested. Empirically, compliance with the retesting requirement was near perfect. Nearly all (98 percent) of students scoring "Level II"—failing but with scores relatively close to the pass/fail cutoff—were retested during these four years. The retested students took a different test form of the same grade-level math test; in other words, the specific test items on the initial test and retest were different but all items were drawn from the same item bank.[8]

Finally, only the higher of a student's two scores—initial score or retest score—would be used to calculate the performance measures used to evaluate schools and teachers. Thus retesting could only increase the "percent passing" and "test score growth" measures used to evaluate schools and teachers.

Schools and teachers had approximately 2-3 weeks between the initial test and retest, or about 6-8 percent of the school year. By rule both the initial test and retest had to occur during the last 22 school days of the school year. Each school had some discretion about when to give tests within that 22-day window, and we examine that choice empirically later in the paper.

*1.3 Retesting before 2009, and additional details*

While the teacher (school) evaluation "retest policy" described above was only in place between 2009-2012, there was some retesting of students occurring prior to 2009. Importantly, the pre-2009 retests were not used in the calculation of teacher (school) evaluation scores. From 2013 forward there was no retesting.

[8] Scores on the initial test were divided into four ordered levels: "Level I" and "Level II" were failing, and "Level III" and "Level IV" were passing. The state required that all students who scored "Level II" be retested, and 98 percent were retested. Students who scored "Level I" could be retested but retesting was not required by the state.

The pre-2009 retesting was mechanically similar but more limited in scope. Just as in the 2009-2012 years, students who failed the initial test would be retested. However, the time between the initial test and retest was somewhat shorter; before 2009 the test window was the last 15 days of the year. Different from the 2009-2012 years, the pre-2009 retesting was much more limited in scope: it only occurred in grades 3, 5, and 8, and only in about one-third of school districts which chose to retest. Thus for these select grades and districts the only change in 2009 was that retest scores would affect teacher (school) evaluation scores.

The pre-2009 retesting was linked to decisions about retaining students in grade. A North Carolina state policy, which ended after 2010, required that "one factor" schools must consider in retention decisions was the student's end-of-year test score, especially if the student failed the test. This state policy applied only to grades 3, 5, and 8. The one-third of district which retested before 2009 chose to retest (some of) their failing students in order to refine grade retention decisions.[9] While failing the test was nominally "one factor", later in the paper we show empirically that failing the math test had little effect on actual grade retention decisions.

Finally, over the 13 years in our study, North Carolina had three different math test designs: the first ending in 2005, the second from 2006-2012, and the third from 2013 forward. We standardize all test scores (mean 0, standard deviation 1) within year-by-grade cells, and later we show that our results are robust to using only data from the 2006-2012 test.

[9] The other two-thirds of districts adopted 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 was treated as having passed for the purposes of grade retention decisions.

## 2. Identification strategy

Our empirical objective is to estimate the causal effect of the 2009-2012 retest policy, with its new incentives for teachers, 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 identifying 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 (LLR):

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

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, 2009-2012, the "treatment years." $R_t = 0$ for the "comparison years," 2003-2008 and 2013-2015. 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 comparison years, $\gamma$, and the retest years, $\gamma + \delta$. The main effect for $R_t$ is subsumed by the 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. North Carolina divides student scores, $Y_{it}$, into four ordered categories known as "proficiency levels." Students scoring "Level I" or "Level II" failed the exam, and students scoring "Level III" or "Level IV" passed. 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 $\delta$, 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 $\delta$ 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 policy years and comparison 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 we 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.

Consistent with the institutional details, empirically we find no evidence of manipulation. Appendix Figure A1 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 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 comparison years, $\hat{\gamma} = -0.004\sigma$, and column 2 shows the diff-in-RD estimate for the retest policy 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 $\delta$: female and special education status. Two significant differences is more than we would expect by chance, but the test scoring procedures described above make the typical concerns about manipulating assignment in an RD setting very unlikely in this case.[10] Still, relative to the comparison years, in the treatment years students who barely failed, and were retested, were slightly more likely to be male and special education than those who barely passed. These small discontinuities are unlikely to explain our results.[11]

## 3. Main estimates

The retest policy generates lasting improvements in the 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 when tested 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 fitting 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.

[10] Perhaps girls, but not boys, would give some extra effort in order to pass, if during the test they knew they were just below the passing threshold. But during the test students cannot know their place relative to the cutoff, nor can their teachers know. Girls might give more effort generally but that would not create a discontinuity. Jacob (2005) and Figlio and Getzler (2006) find that students were pre-emptively classified as special education so that they would not count toward the evaluation score. That manipulation would not create the discontinuity in Table 1; schools might use failing status to inform future special education classification but not prior classification.

[11] At this age boys do score higher in math than girls. Fryer and Levitt (2010) estimate the gap to be 0.20σ. Even with that large gap, the discontinuity for female in Table 1 would have to be at least an order of magnitude larger to explain our 0.03σ estimate. Moreover, the 0.20σ gap is driven largely by differences in the tails of the distribution. Among students in "Level II" or "Level III" in our data, girls slightly outperform boys on average. In general, special education students score lower than their classmates.

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 are also potentially different from the pre years. The post years may be partly explained by a change in the math test, which we discuss below with other robustness tests.

The main diff-in-RD estimates are shown in the top row of Table 2. In the comparison years, there is little difference at the pass/fail cutoff. The RD estimate $\gamma$ 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 $\delta$ in equation 1. Thus, during the retest years students who barely failed scored $\gamma + \delta = 0.029\sigma$ higher than they would have if they had barely passed. This 0.03σ difference is also visible in the conditional expectation graphs common in RD work (Appendix Figure A2).

*3.1 Magnitude and persistence of the effects*

Is the effect long lasting? Recall, importantly, that 0.03σ is not the effect on retest scores. (We do not use the retest scores as outcomes in the paper, except briefly in Appendix Table C2.) The 0.03σ is the effect on scores (nearly) one year later, which is the next time both retest and non-retested students are measured. Put differently, 0.03σ is the effect which is still measurable one year after the discontinuity in teachers' incentives has gone away. The incentives go away after students take the retest. Two years out the effect is 0.023 (st.err. 0.008).

Is the difference of 0.03σ large or small? It is small as a share of the total variation in student math scores but large in context. First, consider 0.03σ compared to a teacher's total contribution to student test scores, assuming the gain comes through a reallocation of teacher effort. One standard deviation of the teacher effect ("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. Assume our estimated effect is due to additional teaching during the 2-3 weeks between the initial test and retest. Existing estimates suggest adding 2.5 weeks of instruction would add 0.05-0.075σ to student test scores (Sims 2008, Fitzpatrick, Grissmer, and Hastedt 2011, Aucejo and Romano 2016). Our estimate of 0.03σ is smaller, but the simple comparison ignores potential effect decay, sometimes called "fadeout."

Student test score gains, induced by some added treatment, often decay by half one year after the treatment ends. Relevant to this paper, fadeout has been documented for both the treatment of assigning students to more effective teachers (Kane and Staiger 2008, Jacob, Lefgren, and Sims 2010, Chetty, Friedman, and Rockoff 2014) and of increasing instructional time (Taylor 2014) as well as other educational treatments. Assume 0.03σ is due to additional teaching during the 2-3 weeks, but then note 0.03σ is measured one year after treatment ends. Scaling up by the typical fadeout would suggest a treatment effect of about 0.06σ at the time of the retest.[12] We cannot measure decay over the first year (from $t$ to $t+1$) in our data, but we can measure decay over the second year (from $t+1$ to $t+2$). There are fewer estimates of second-year fadeout, but they range from one-third to half. The treatment effect on math scores at the end of year $t+2$ is 0.023, suggesting decay of about one-quarter (though we cannot reject decay of half).

*3.2 Robustness*

Our results are robust to a number of estimation choices. The robustness test results are shown in the remaining rows of Table 2. Our main specification allows

[12] The nature of the 2-3 week period and the retest may have incentivized teacher to use more short-run strategies, and thus the fadeout would be greater.

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 grouping: retest policy years versus comparison years. 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 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 use only data from $t =$ 2007 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 up through 2006. 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.[13]

*3.3 Reading test scores*

Our focus in this paper is math tests, but the motivation for our study applies to reading tests as well. The bottom row of Table 2 reports diff-in-RD estimates for reading tests. The pattern of results matches math, but with smaller differences. During the retest years students who barely failed reading scored 0.013σ higher one year later in reading than they would have if they had barely passed.

[13] One possibility is the following: Beginning with 2014, the four proficiency levels were expanded to five levels (see Appendix B). What had been the "level II/III" cutoff became the "level III/IV" cutoff. Figure 1, and all of our analysis, uses the new "level III/IV" cutoff as the counterfactual cutoff for retesting in 2014-2015. If instead we use the new "level II/III" cutoff the 2014 and 2015 point estimates are closer to zero and their 95 percent confidence intervals include zero.

Smaller effects for reading is a common pattern in the economics of education literature (e.g., Abdulkadiroglu et al. 2011, Fryer 2014), including the study of teacher performance specifically (e.g., Hanushek and Rivkin 2010, Taylor and Tyler 2012). Sims (2008) reports extra instruction time benefits math but not reading. One common hypothesis is that, in contrast to reading and language, math is learned primarily (entirely) in school classrooms (Jackson, Rockoff, and Staiger 2014), thus an increase in teacher effort of a given size will be a larger percentage increase in math instruction. A second, but related, hypothesis is that typical state reading tests are not as sensitive to teacher effort (Kane and Staiger 2011). Teachers' incentive in reading—the incentive to give greater effort before the retest—would be attenuated under these hypotheses.

We focus on math, however, mainly because the institutional details and available data favor identification for math. First, for reading there are no data on retesting that occurred in years prior to the retest policy, while there are such data for math. As elaborated next in Section 4, retesting prior to the retest policy is important for ruling out mechanisms unrelated to teacher effort and incentives. Second, the state introduced a new reading test in 2009 the same year that the retest policy began, which complicates the core identifying assumptions for reading.

## 4. Mechanisms

In this section we discuss evidence (in)consistent with different potential mechanisms for the retest policy's effects we documented above. On balance the evidence suggests the effects arose because teachers (schools) changed their effort in response to the new, sharp incentives in their evaluation measure. We do not have tests to rule out all possible alternative mechanisms. Nevertheless, even if the effects arose from some combination of teacher effort and other mechanisms, our estimates still document relatively large and lasting benefits to students caused by a change in teacher evaluation rules.

*4.1 Retesting per se versus the retest policy*

We begin with an important result that emphasizes the difference between retesting *per se* and the 2009-2012 retesting policy. There was some retesting in North Carolina prior to the start of the retest policy in 2009, as detailed in Section 1.3. Critically, 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. Additionally, the pre-2009 retesting occurred only in select grades (3, 5, and 8) and only in one-third of school districts.

The pre-2009 retesting allows for the following test: If retesting *per se* drove the effects, then the start of the retesting policy in 2009 would not change student outcomes in grades 3, 5, and 8 in districts which were already retesting prior to 2009. Here "no change" is equivalent to null hypothesis of $\delta = 0$ for the diff-in-RD. We test this prediction in Table 3 row 1. The diff-in-RD estimate is 0.034 (st.err. 0.010), quite similar to our main estimate.[14] Additionally, we should also expect a positive $\gamma$ if retesting *per se* had benefits, but the estimate of $\gamma$ in row 1 is -0.007 (st.err. 0.013). In short, for the row 1 subsample 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 did cause an improvement in student outcomes.

The 2009-2012 retest policy had two main features: retesting failing students and using the retest scores to calculate teachers' (schools') evaluation scores. The lack of effects from pre-2009 retesting strongly suggests the mechanism

[14] From 2009-2012 the test window as 22 days, while prior to 2009 the window was 15 days. Thus, before 2009 teachers would have had about two-thirds (15/22) as much time to influence test scores before the retest. Using the results from Table 3 row 1, our estimate of $(\gamma + \delta) \times (15/22)$ is 0.0186, and the confidence interval for $\gamma$ is -0.033 to 0.0188, which only just includes 0.0186. Additionally, in Table 3 rows 1-4 we limit the sample to $t = 2003$ through 2012, because after 2012 there was no retesting. Results are not substantively different if we use all years in the data.

behind the 0.03σ effect was not the retesting itself. The lack of pre-2009 effects is also evidence against student effort mechanisms, as we discuss next. That leaves mechanisms arising from the change in how teachers were evaluated.

*4.2 Student effort or behavior*

We now shift focus to the students themselves. One category of potential mechanisms is changes in student effort or behavior. Student and teacher effort changes are not mutually exclusive mechanisms. Indeed, one important way that teachers contribute to student outcomes is by inducing their students to give greater effort. Choosing to give a student more math instruction would be useless if the student did not participate. Thus, teacher-induced student effort is always one component of "teacher value added." However, in this subsection we focus on changes in student effort or behavior (relatively) independent of their teachers.

One specific hypothesis is that students may find being retested distasteful and give greater 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. The results using pre-2009 retesting (Table 3 rows 1-4) are inconsistent with this hypothesis. If students feared being retested, they should fear it before and after 2009.

It is possible that students feared being retested, but also knew when retesting was occurring and when it was not. For example, perhaps before 2008 students retested in grade 3 or 5 did fear being retested in the future, but nevertheless did not give greater effort on the $t+1$ test because they knew there was no retesting in grades 4 and 6. We can test this more-specific hypothesis using observations from 2008 for students subject to retesting; we re-estimate Table 3 row 1 column 1 using 2008 data only. This sample was "treated" with retesting in $t = 2008$, they faced the threat of retesting in $t + 1 = 2009$, but their retest scores in 2008 did not count for teacher evaluation. We find no effect of retesting for this

specific subsample ($\gamma = 0.007$, st.err. 0.012).[15] By similar logic, students who were retested in 2012 should not fear being retested in 2013 because there was no retesting in 2013. If the student fear hypothesis explained our results, there should be no effect in 2012. As shown in Figure 1 the effect in 2012 is similar to 2009-2011.

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 total days absent during the entire $t + 1$ school year. The diff-in-RD point estimate is a reduction in absences but only 0.013 days (the 95 percent confidence interval is -0.077 to 0.051 days).

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. Though, alternatively, younger students may be more responsive to encouragement from parents (or teachers) to give greater effort on the test. 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 again, we might expect student effort effects to carry across subjects. First, we test for effects of being retested in math in year $t$ on reading test scores in $t + 1$. The point estimate is positive and marginally statistically significant, but an order of magnitude smaller than the main effect for math score (Table 4 row 2). Though perhaps student fears are subject specific or spill over only weakly. Second, the motivation effect from being retested in math at $t$ might be attenuated or amplified if the student had also

[15] Another test in the same spirit is to re-estimate Table 3 row 1 where the outcome is math scores in year $t + 2$. Students were "treated" with retesting in year $t$ at grade 3 or 5, and they faced the threat of retesting in $t + 2$ at grade 5 or 7, but their retest scores did not count for teacher evaluation. We find no effect of retesting in this test ($\gamma = -0.011$, st.err. 0.010).

been retested in reading at $t$. Effect estimates for math are, in fact, quite similar for students who did or did not fail reading in the same year: 0.034 (st.err. 0.005) and 0.039 (st.err. 0.005) respectively.

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 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 $t+1$.

To explain our results, any candidate explanation—like a differential treatment during year $t+1$—would need to change discontinuously at the time $t$ pass/fail 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, even though these "tracking" patterns were 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 in the policy

years 2009-2012. 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 significant differences in column 1, but the diff-in-RD estimates are close to zero and not statistically significant.

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 teacher 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. As explained in Section 1.3, in 2010 and the years before, the state required that failing the end-of-year test must be "one 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 6 column 1 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. Additionally, our effect estimates do not covary with how districts' retention behavior changed after the end of the state policy in 2010 (see Appendix C including Table C1).

*4.4 Additional results*

Finally, we provide a brief summary of some additional tests and results. Full details and discussion are provided in online Appendix C. First, we find no effect of failing the *retest* on future test scores. Our simple RD estimate at the year $t$ retest pass/fail cutoff is 0.002 (st.err. 0.009) on $t+1$ scores, for students who barely failed the year $t$ initial test. Several hypothesized mechanisms, like those in Section 4.3 and others we cannot test, imagine teachers reacting to students' pass/fail status for reasons other than the evaluation incentives. But teachers do not respond to the pass/fail status of the retest, and that is difficult to reconcile with those hypotheses. Additionally, nearly two-thirds (63 percent) of students who barely failed the initial test passed on the retest, thus their final "fail" status changed to "pass."

Second, we estimate teacher contributions to the retest score—valued added to the retest—net of contributions measured on the initial test. These "value added to retest" scores do predict students' scores at $t+1$ in both 2008 and 2009. However, consistent with the retest policy changing teacher behavior, the predictive power increases during the 2009 retest year.

Third, a complementary piece of evidence comes from the test dates chosen by schools. Schools have some discretion over actual testing dates within a state-defined window. Before the retest policy started in 2009, schools with more failing students set their test date late in the window, much later than schools with fewer failing students. But then during the retest years, 2009-2012, the struggling schools moved their test date up, to a point much earlier in the window and similar to schools with few failing students. After the retest policy ended, the struggling schools switched back to a late test date. These choices are further evidence that teachers and schools understood and acted on the incentives created by the retest policy. Moving up the test date is consistent with schools which (plan to) make good use of the time before the retest.

## 5. Conclusion

In the design of employee evaluation, and related incentives, employers often face tradeoffs between intended and unintended changes in employee effort. This paper documents an empirical example from public schools. First, consistent with the goals of schools, we find that teacher effort responded to new evaluation incentives and, as a result, students made larger gains in math achievement. These were real gains lasting for a year or more beyond when the evaluation incentives had ended. However, likely inconsistent with the goals of schools, those added and lasting achievement gains were only among a select group of students—the students favored by the evaluation system incentives for teachers.

The first conclusion—larger and lasting gains for students—demonstrates the important role of teachers and teacher effort in educational production. On this point there is a substantial literature (see Jackson, Rockoff, and Staiger 2014 for a review). The evidence in this paper also demonstrates that teacher effort responds to evaluation incentives, and teacher contributions to student achievement can be larger, at least over short periods. On that point the literature is much thinner (see our introduction for a review). Broadly speaking, these conclusions lend support to the efforts of school policymakers and managers working to improve public education through teacher (school) evaluation.

The second conclusion—gains only for select students—demonstrates the challenge of designing evaluation incentives, especially in multitask jobs like teaching. The large and lasting gains are encouraging, but the cost of those gains was new inequity between students. Teachers responded to the incentives: students not favored by the evaluation system incentives had lower math scores than their favored classmates. This difference in math scores arose even though the students were identical (at expectation); in other words, even though non-favored students would have had equal returns to teacher effort. This allocative inefficiency is evidence inconsistent with the notion that teachers are perfectly motivated agents.

Our contributions arise in part from the distinctive features of the setting we study. Other papers also document changes in teacher effort allocations across students. What distinguishes our setting and contribution is the sharp discontinuity in incentives across students at the pass/fail threshold: Effort directed at students who failed the initial test would unambiguously increase the teacher's performance measure, but effort directed at otherwise-identical passing students could not change the measure. This sharp difference allows for sharper identification of the causal effects of evaluation incentives and enables us to make the claims about inequity and allocative inefficiency in the previous paragraph. However, the cost of our regression-discontinuity-based strategy is that, strictly speaking, we can make inferences only to students near the pass/fail cutoff. By contrast, Neal and Schanzenbach (2010) and papers with similar identification strategies, describe effects for a much wider range of students.

In the preceding paragraphs we have described the underlying mechanisms as teacher effort changing in response to new evaluation incentives. We believe this characterization is reasonable given the accumulated evidence. The paper documents several relevant empirical tests, including testing for changes in student effort or behavior (orthogonal to teacher effort). While we present several relevant tests, we cannot test all potential alternative mechanisms. Nevertheless, even if the effects arose from some combination of teacher effort and student effort or other mechanisms, our estimates still document relatively large and lasting benefits to students caused by a change in teacher evaluation rules.

We conclude with three additional caveats or limitations. First, while student math test scores rose, we cannot say for certain what the students learned. 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. Moreover, math skills and test taking skills are not disjoint sets. Whatever students learned, however, the resulting

discontinuity in students' $t+1$ scores is consistent with teachers reallocating their effort in response to the incentives of their evaluation measure.

Second, 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. As a second example, perhaps the school decided to hire temporary tutors for retested students or make greater use of tutoring software programs. This is also a change in the allocation of teacher effort or at least the teacher labor budget.

Third, this paper is not an evaluation of whether the retest policy was, summing up all its effects, good or bad for North Carolina schools. Our focus is students near the pass/fail cutoff where the teachers' incentives changed sharply. It is certainly possible that greater teacher effort directed at failing retested students came at the expense of less teacher effort for their passing classmates. Thus, passing students might well be worse off under the retest policy. However, effort reallocation across students is not necessary to explain our main results. Teachers' could give up their own leisure; or teachers could shift effort across subjects within individual students, for example, more math instruction and less arts instruction for a retested student.[16]

Teacher performance can be changed by evaluation, though intended changes can come with the costs of unintended changes. Evaluation incentives can induce teachers to give greater effort, or at least to reallocate their effort, and as a result make larger contributions to student learning. However, in the case documented in this paper, the larger gains came only for some students—students

[16] In online Appendix C we provide some descriptive evidence on how the retest policy affected students who barely passed the initial test.

favored by the teacher evaluation incentives. All of which emphasizes, alongside the other literature cited above, the importance of teacher effort in student outcomes and the importance of careful design of evaluation systems.

**References**


Abdulkadiroğlu, Atila, Joshua D. Angrist, Susan M. Dynarski, Thomas J. Kane, and Parag A. Pathak, "Accountability and flexibility in public schools: Evidence from Boston's charters and pilots," *Quarterly Journal of Economics* 126 (2011), 699-748.

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

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

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

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff, "Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood," *American Economic Review* 104 (2014), 2633-2679.

Chiang, Hanley, "How accountability pressure on failing schools affects student achievement," *Journal of Public Economics* 93 (2009), 1045-1057.

Cohodes, Sarah R., "Teaching to the student: Charter school effectiveness in spite of perverse incentives," *Education Finance and Policy* 11 (2016), 1-42.

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

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

Dee, Thomas S., and James Wyckoff, "Incentives, selection, and teacher performance: Evidence from IMPACT," *Journal of Policy Analysis and Management* 34 (2015), 267-297.

Deming, David J., Sarah Cohodes, Jennifer Jennings, and Christopher Jencks, "School accountability, postsecondary attainment, and earnings," *Review of Economics and Statistics* 98 (2016), 848-862.

Deming, David J., and David Figlio, "Accountability in US education: Applying lessons from K-12 experience to higher education," *Journal of Economic Perspectives* 30 (2016), 33-56.

Duflo, Esther, Pascaline Dupas, and Michael Kremer, "Peer effects, teacher incentives, and the impact of tracking: Evidence from a randomized evaluation in Kenya," *American Economic Review* 101 (2011), 1739-1774.

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

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

Fitzpatrick, Maria D., David Grissmer, and Sarah Hastedt, "What a difference a day makes: Estimating daily learning gains during kindergarten and first grade using a natural experiment," *Economics of Education Review* 30 (2011), 269-279.

Fryer, Roland G., "Injecting charter school best practices into traditional public schools: Evidence from field experiments," *Quarterly Journal of Economics* 129 (2014), 1355-1407.

Fryer, Roland G. and Steven D. Levitt, "An empirical analysis of the gender gap in mathematics." *American Economic Journal: Applied Economics* 2 (2010), 210-240.

Glewwe, Paul, Nauman Ilias, and Michael Kremer, "Teacher Incentives," *American Economic Journal: Applied Economics* 2 (2010), 205-27.

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

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

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

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

Jacob, Brian A., Lars Lefgren, and David P. Sims, "The persistence of teacher-induced learning," *Journal of Human Resources* 45 (2010), 915-943.

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

Kane, Thomas J., and Douglas O. Staiger, "Estimating teacher impacts on student achievement: An experimental evaluation," NBER Working Paper 14607 (2008).

Kane, Thomas J., and Douglas O. Staiger, *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains.* (Seattle, WA: Bill & Melinda Gates Foundation, 2011).

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

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

Neal, Derek, “The design of performance pay in education,” in *Handbook of the Economics of Education* Volume 4, Eric A. Hanushek, Stephen Machin, and Ludger Woessmann, eds., 495–550. (Amsterdam: North-Holland, Elsevier, 2011).

Neal, Derek, *Information, incentives, and education policy.* (Harvard University Press, 2018).

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

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

Rockoff, Jonah, and Lesley J. Turner, “Short-run impacts of accountability on school quality,” *American Economic Journal: Economic Policy* 2 (2010), 119-47.

Rouse, Cecilia Elena, Jane Hannaway, Dan Goldhaber, and David Figlio, “Feeling the Florida heat? How low-performing schools respond to voucher and accountability pressure,” *American Economic Journal: Economic Policy* 5 (2013), 251-81.

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

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

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

Taylor, Eric S., and John H. Tyler, “The effect of evaluation on teacher performance,” *American Economic Review* 102 (2012), 3628-51.

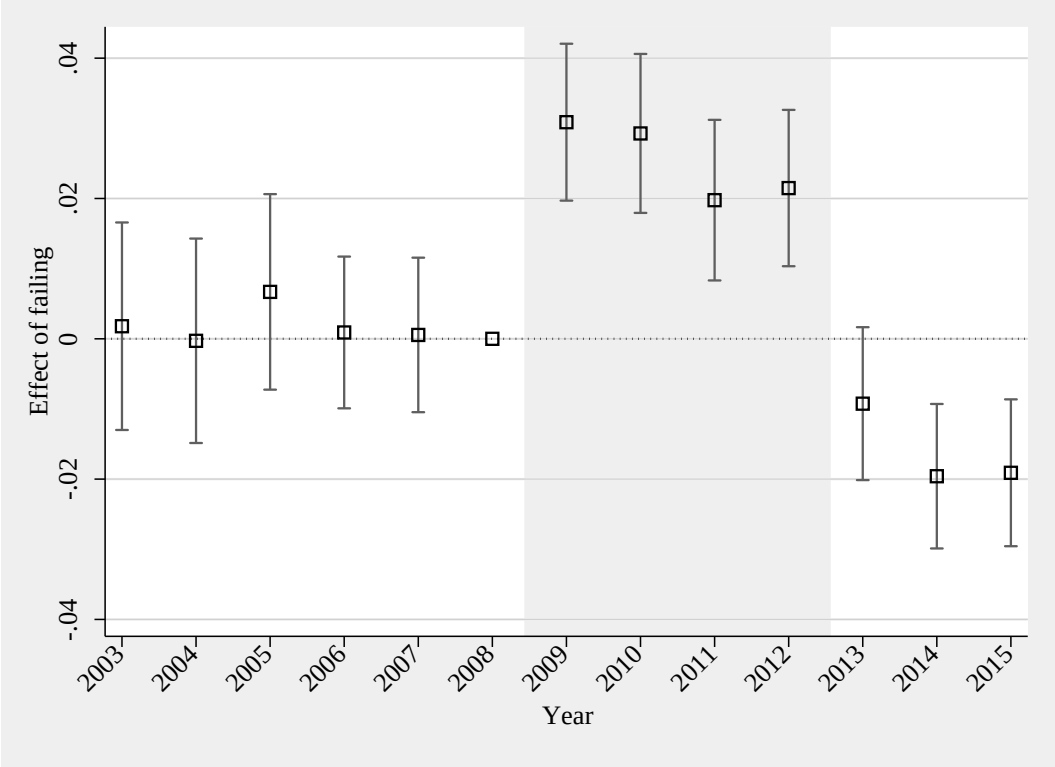


Figure 1—Estimated (RD) effect of failing math on next year's math score; before, during, and after the retest policy years

Note: "Treatment years" are the years of the retest policy, 2009-2012, shaded in gray. "Comparison years" are the years before and after the retest policy, 2003-2008 and 2013-2015. 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.

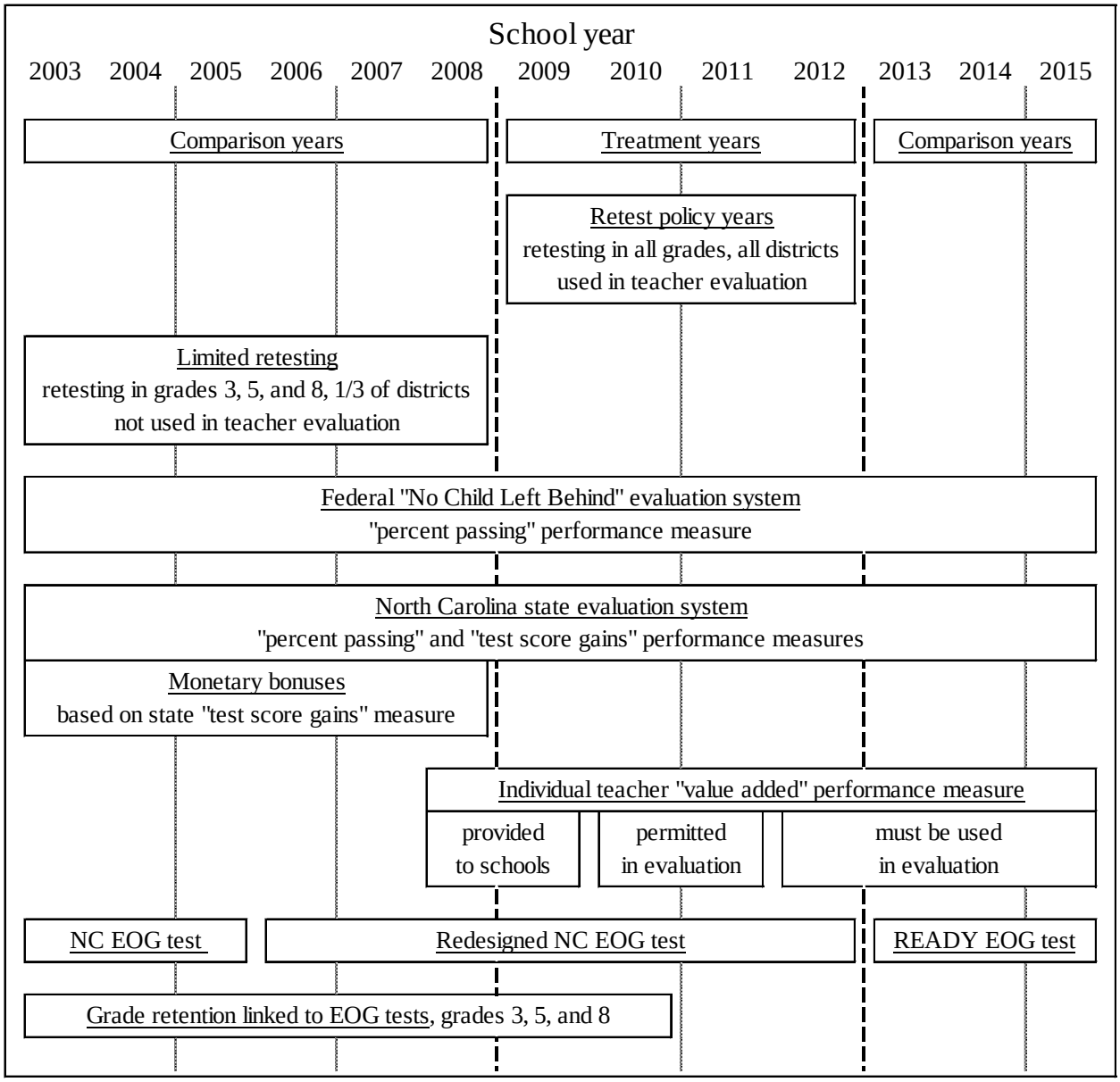


Figure 2—Timeline of relevant policies and tests

Note: School years are named by the spring year, e.g., the school year 2002-03 is 2003. See Section 1 for detailed description of the policies named in the figure.

Table 1—Covariate balance

| Covariate | RD estimate of difference at pass/fail cutoff: Comparison years | Difference-in-RD estimates: Treatment years − Comparison years | Observations |
|---|---|---|---|
| | (1) | (2) | (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: "Treatment years" are the years of the retest policy, 2009-2012. "Comparison years" are the years before and after the retest policy, 2003-2008 and 2013-2015. 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 comparison 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.

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

| | RD estimate of difference at pass/fail cutoff | Difference-in-RD estimates | |
|---|---|---|---|
| | Comparison years | Treatment years − Comparison 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 policy years v. comparison 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) | |
| Reading estimate | 0.006 | 0.013 | 4,015,113 |
| | (0.008) | (0.010) | |

Note: "Treatment years" are the years of the retest policy, 2009-2012. "Comparison years" are the years before and after the retest policy, 2003-2008 and 2013-2015. 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. The reading estimate simply replaces math outcome and running variables with their reading equivalents. Standard errors in parentheses are clustered at the values of the running variable.

Table 3—Retesting before 2009 and grade level estimates

| | RD estimate of difference at pass/fail cutoff | Difference-in-RD estimates | |
|---|---|---|---|
| | Comparison years | Treatment years − Comparison years | Observations |
| | (1) | (2) | (3) |
| Retesting prior to 2009 | | | |
| Grades 3 and 5 in districts which retested failing students in 3 and 5 | -0.007 | 0.034 | 459,461 |
| | (0.013) | (0.010) | |
| No retesting prior to 2009 | | | |
| Grades 4, 6, and 7 in districts which retested failing students in 3 and 5 | 0.022 | 0.018 | 657,417 |
| | (0.010) | (0.006) | |
| Grades 3 and 5 in districts which used SEM rule in 3 and 5 | -0.009 | 0.029 | 798,773 |
| | (0.009) | (0.008) | |
| Grades 4, 6, and 7 in districts which used SEM rule in 3 and 5 | 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: "Treatment years" are the years of the retest policy, 2009-2012. "Comparison years" are the years before and after the retest policy, 2003-2008 and 2013-2015. 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.

Table 4—Alternative outcomes

| | RD estimate of difference at pass/fail cutoff | Difference-in-RD estimates | |
|---|---|---|---|
| | Comparison years | Treatment years − Comparison 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 | 0.012 | -0.003 | 2,127,030 |
| | (0.005) | (0.003) | |
| Proportion $t+1$ peers failed $t$ test | 0.028 | -0.002 | 2,127,030 |
| | (0.002) | (0.001) | |
| Value-added score of $t+1$ teacher | 0.001 | 0.000 | 1,943,383 |
| | (0.000) | (0.000) | |
| Retained (same grade $t$ and $t+1$) | 0.003 | 0.000 | 3,212,750 |
| | (0.001) | (0.000) | |

Note: "Treatment years" are the years of the retest policy, 2009-2012. "Comparison years" are the years before and after the retest policy, 2003-2008 and 2013-2015. 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.