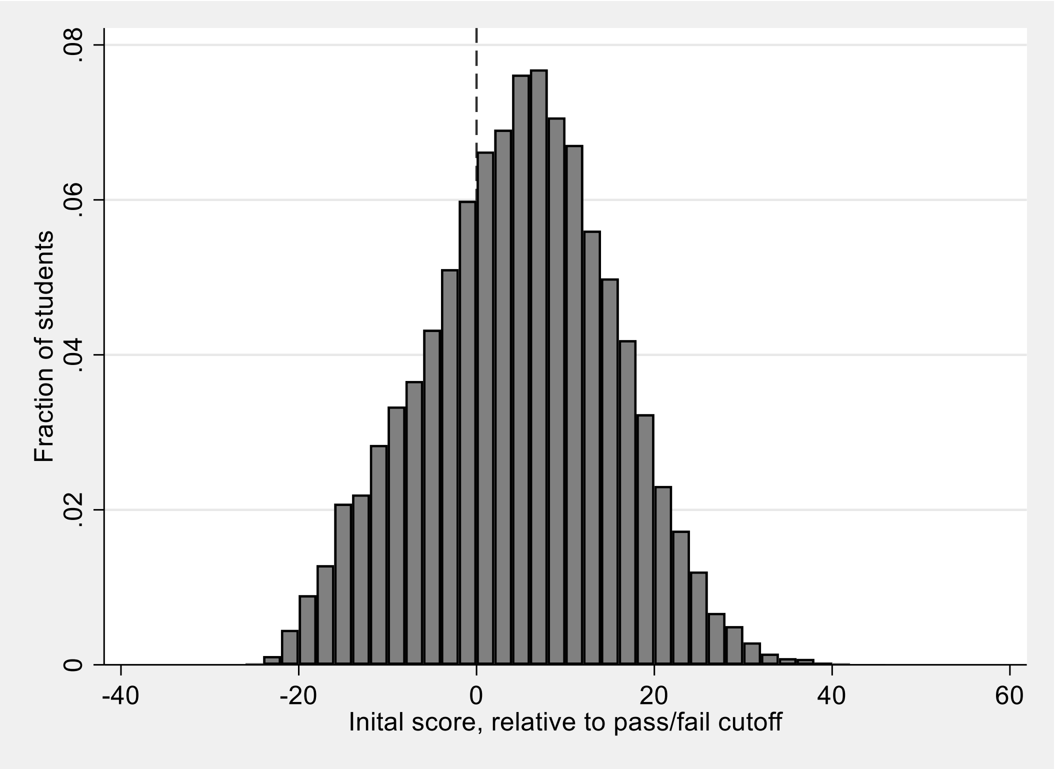
Does evaluation change teacher effort and performance? Quasi-experimental evidence from a policy of retesting students

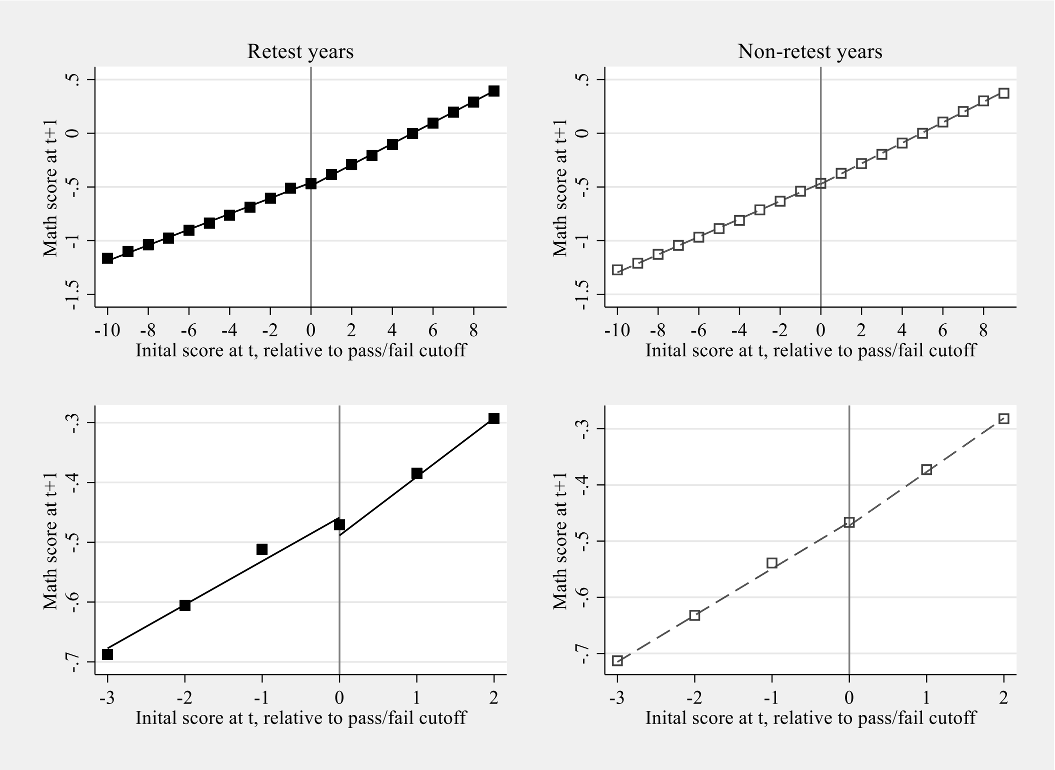
Appendix

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

July 2020



Appendix Figure A1—Histogram of end-of-year initial math test score (running variable)



Appendix Figure A2—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 comparison 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.

Appendix Table A1—Summary statistics

	All students	Level II & III students (within bandwidth for main estimate)
	(1)	(2)
Math score, $t-1$	0.015	-0.212
	(0.992)	(0.774)
Math score, t	0.000	-0.263
	(1.000)	(0.646)
Math score, $t+1$	0.024	-0.248
	(0.991)	(0.777)
Reading score, $t-1$	0.015	-0.172
	(0.990)	(0.854)
Reading score, t	0.002	-0.203
	(0.999)	(0.848)
Reading score, $t+1$	0.022	-0.194
	(0.986)	(0.859)
Retained in year t	0.009	0.010
Female	0.493	0.502
White	0.546	0.496
Days absent	6.422	6.483
	(6.344)	(6.222)
Free or reduced lunch	0.508	0.550
Special education	0.101	0.092
Limited English proficiency	0.082	0.089

Note: Means and (standard deviations) for grade 3-7 students in 2003-2015. Column 1 is based on 7,019,698 student-by-year observations in year t. Column 2 is restricted to student-by-year observations where the student's year t math test score was "Level II" or "Level III," and we observe the student's year $t+1$ math test score. Column 2 has 3,978,190 observations. Free or reduced lunch and special education data not available for 2003-2005.

Appendix B: Additional Details to Accompany Section 1

B.1 North Carolina state tests

Our study focuses on annual end-of-grade math tests for grades 3-8, and also uses parallel reading tests.[1] Individual student scores are reported in two ways. First, a "scale score" which is a weighted average of individual test items, with the weights determined by Item Response Theory (IRT). Second, the scale scores are divided into four mutually exclusive and exhaustive ordered categories called "Level I" through "Level IV". Our regression discontinuity analysis uses only "Level II" and "Level III" scores. "Level II" is failing and "Level III" is passing.

Test items change from year to year, but the IRT methods link item weights over time to keep scale scores and cutoffs constant across years. Thus, for a given grade level and subject, the scale score that separates pass/fail ("Level II"/"Level III") remains constant over time. However, the scaling and pass/fail cutoffs do change when the test design changes. This occurred twice during the 13 years of our study: the first ending in 2006, the second from 2007-2012, and the third from 2013 forward. We standardize all test scores (mean 0, standard deviation 1) within year-by-grade cells.

In 2014, the four "level" categories were expanded to five categories "new Level I" through "new Level V". The pre-2014 "old 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

[1] 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. High school students take end-of-course exams in select courses.

Level III" versus "new Level IV" cutoff, as we assume in our analysis, or at the "new Level II" versus "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.

B.2 No Child Left Behind

The federal NCLB regulations required that a school's "percent passing" increase every year. But individual states set their own (stricter or more lenient) *standards* for "passing," and set their own targets for *how much* the "percent passing" should increase each year. Our analysis uses only within state (North Carolina) variation, and so these standards and targets were constant across schools.

Beginning in 2012, North Carolina was operating under an "NCLB waiver." Similar waivers were granted to most other states in 2012 or shortly after. The relevant details of the North Carolina waiver for this paper are straightforward. First, the waiver did not change the evaluation performance measure; schools and teachers were still evaluated based on the "percent passing." However, the annual targets were reset. The new targets were still ambitious. Schools would be expected to reduce the proportion of failing students by half over six years, relative to the failure rates in 2011. In the absence of the waivers, the existing NCLB statute required that 100 percent of student be passing the exams by 2014. If that target had persisted its impossible expectation would have likely weakened the incentives for teachers to worry about NCLB consequences.

Second, the NCLB waivers generally required states to evaluate teachers individually, based in part on student test scores. This coincides with North Carolina requiring that districts begin using "value added" scores in 2012, though districts could choose to use those scores in prior years.

In December of 2015, the middle of the 2016 school year, NCLB was replaced by legislation known as the Every Student Succeeds Act. The period we study ends with the 2016 school year.

B.3 ABCs / READY Accountability

Our description of ABCs, READY Accountability, and other state programs 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."

In 2013, North Carolina introduced "READY", a bundle of education policy features including revised content standards, new test designs to accompany the standards, and changes to test-based evaluation (accountability). While the standards and test designs changed, the evaluation performance measures remained largely the same: The state's composite measure—a weighted average of the "percent passing" and "test score growth" scores—was reweighted to favor the "percent passing" more. The descriptive school labels were replaced with letter grades A-F.

B.4 The retest policy

By using only the higher of a student's initial score and retest score, the retest policy fundamentally changed the "percent passing" and "test score growth" evaluation measures. This fact was understood by the policymakers who made the change. The state Department of Public Instruction warned that ABCs and NCLB measures for 2008-09 and later should not be compared to nominally similar measures from before retesting started. The change to the retest policy seems to have been motivated by considerations of test reliability.

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.

B.5 Grade retention policies

As explained in the main text, 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. Before 2009, one-third of districts chose to retest (some of) their failing students in order to refine grade retention decisions. The other two-thirds of districts adopted the "SEM rule." If a student failed, but their score was within one standard error of measurement (SEM), the student was treated as having passed for purposes of grade retention decisions.

The end of the retention policy was officially approved on October 7, 2010, approximately six weeks into the 2011 school year. The policy change explicitly allowed schools to reverse the retention decisions for 2011 they had made previously.

Appendix C: Additional Results Summarized in Sections 4.3 and 4.4

C.1 Additional results related to retention

As stated briefly in the main text, our effect estimates do not covary with how districts' retention behavior changed after the end of the state retention policy in 2010. The results are reported in Appendix Table C1. First, for each district in North Carolina we estimate the discontinuity in retention probability at the pass/fail cutoff both before and after the 2010 state policy change described in Section 1.3. Then we divide districts into terciles of the difference between the before and after point estimates. Roughly speaking we divide districts into those who retained barely failing students more after the 2010 change, districts with no change in retention behavior, and districts who retained barely failing students less. Second, we estimate out main diff-in-RD model separately for each of these three subsamples. We find no difference in effects across these three groups.

C.2 RD estimates using the retest pass/fail cutoff

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 (Appendix Table C2 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 Appendix Table C2 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.

C.3 Value-added to retest analysis

As mentioned briefly in the main text, one result 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). 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 a quadratic in student i's year $t - 1$ initial math score, student demographic characteristics, and teacher fixed effects. The estimated teacher fixed effects are our "value added to retest" measure.

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. In other words, we test whether teacher contributions to retest scores persist and predict student scores one year later. This second step specification also includes fixed effects for year $t + 1$ teacher, and controls for student demographics and student i's initial test score from year t.

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. In other words, we test whether the nature of teacher's contributions to retest scores changed when the policy changed. Data on retest scores prior to the retest policy are available for only one year, 2008; and, as described in Section 1, 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.

The results of this test are consistent with two conclusions. Detailed results are shown in Appendix Table C3. 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 0.02σ 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 one-third in the retest policy years.

C.4 School choices of test dates

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

We estimate the following specification:

$$D_{st} = \alpha \bar{F}_{s,t-1} + \beta R_t + \delta R_t \bar{F}_{s,t-1} + \pi_t + \nu_{st} \tag{C.1}$$

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 Appendix Table C4 column 1 we show estimates fitting specification C.1 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. 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 C.1 using placebo policy changes in columns 4 and 5. We find no changes in test date decisions at these placebo years.

C.5 Effects of the retest policy more broadly

Here we provide some initial evidence that benefits for retested students did not necessarily come at the expense of other students. First, in Appendix Figure C1, we simply plot trends in math test scores over time, but separately for students who failed (“Level II”) and students who passed (“Level III”) the prior year’s exam. This figure is limited to 2006-2011 when there is a consistent math test to examine levels over time. For failing students (solid line), average scores improve between 2008 and 2009 when the retest policy starts and remain higher. Failing students score 0.074σ higher on average in the retest policy years than in the pre-policy years. For passing students (dashed line), the trend also rises in 2009 and afterward, but the improvement is smaller (0.044σ compared to 0.074σ). Passing students do not appear worse off under the policy.

We caution against making causal interpretations of the trends in Appendix Figure C1. We have no novel identification strategy to offer here. Perhaps, for example, students who passed were in fact worse off initially because teachers shifted effort away from non-retested students and towards retested students. But then, in the time before the $t+1$ test, teachers and schools took additional actions to bring non-retested students back up to their previous level. Or more simply, perhaps students who passed would have been even better off absent the retest policy, lost some of their teachers’ effort to their retested classmates, but nevertheless still did not experience a decline in their score levels.

The average trends in Appendix Figure C1 could mask important heterogeneity. In Appendix Table C5 we test whether passing students’ future math scores (in year $t+1$) are correlated with the proportion of their classmates (in year t) who failed and were retested. If teachers did shift their effort away from non-retested students to retested students, then non-retested students should be worse off the more retested classmates they had. We test this in Appendix Table C5 column 1 by fitting:

$$Y_{i,t+1} = g(Y_{it}) + \alpha \bar{F}^{t}_{c(it)} + \beta \bar{F}^{t}_{c(it)} R_t + \pi_{g(it),t} + \mu_j + u_{i,t+1}, \tag{C.2}$$

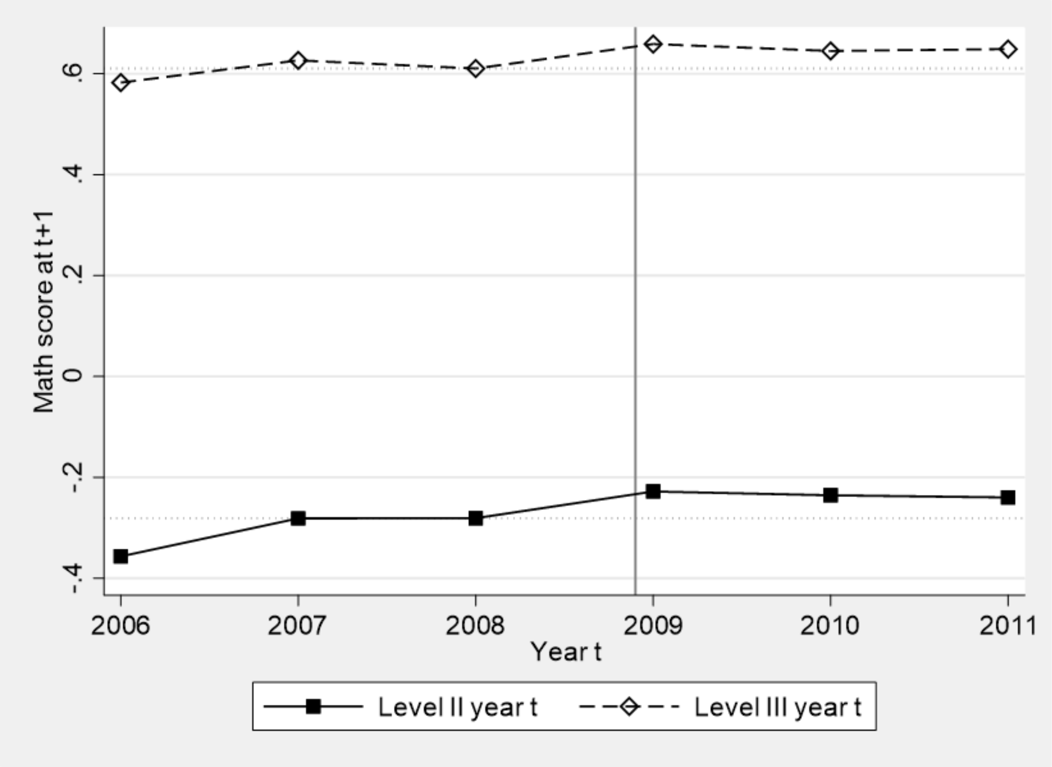
limiting the estimation sample to students who passed the year t initial exam. The new index $c(it)$ is the class c which student i was a member of in year t, and the new variable $\bar{F}_{c(it)}^{t}$ is the proportion of student i's classmates in year t who failed the initial exam. As above $R_t = 1$ during the retest policy years. Specification 3 also includes grade-by-year fixed effects, $\pi_{g(it),t}$; teacher fixed effects, μ_j; and flexible controls for student i's own prior year score, $g(Y_{it})$.[2]

Passing students' test scores next year $(t + 1)$ are higher when they have more failing classmates this year (t). However, this relationship existed in the years before the retest policy, and the relationship did not change once the policy started.

Column 2 is a different test. Students' test scores in year $t + 1$ are also likely affected by the classmates they have in year $t + 1$, and some of those $t + 1$ classmates will have been retested in year t. In Appendix Table C5 column 2 we modify specification 3, replacing $\bar{F}_{c(it)}^{t}$ with $\bar{F}_{c(i,t+1)}^{t}$ which is the share of student i's year $t + 1$ classmates who failed the initial exam the prior year (year t). In the years before the retest policy, passing students' $t + 1$ test scores were lower when they have more $t + 1$ classmates who failed in year t. However, that negative correlation is reduced by about one-third under the retest policy. If failing classmates made passing students worse off, that penalty appears to have been reduced by the retest policy.

To be clear, we view these tests as suggestive; we do not have a strong identification strategy to warrant causal interpretation. The proportion of failing (or passing) classmates may well be correlated with omitted variables, and the proportion may be endogenous to the retest policy.

[2] In Appendix Table C5 the estimation sample is "Level III" students, and $\bar{F}_{c(it)}$ is the proportion of "Level II" students. This matches the general pattern of the paper. However, the results in Appendix Table C5 are quite robust to using "Level III or IV" as the estimation sample, and "Level II or I" to define $\bar{F}_{c(it)}^{t}$. The controls $g(Y_{it})$ are a quadratic in Y_{it} where the parameters of the quadratic are allowed to differ in each grade-by-year cell. The results are robust to simpler versions of $g(Y_{it})$.



Appendix Figure C1—Math test score trends over time

Note: Each marker represents mean math score, $A_{i,t+1}$, in student standard deviation units (y-axis). Scores are standardized (i) separately by grade (ii) using the mean and standard deviation from 2006, the first year of the test, applied to all years. The solid line is students who scored at "Level II" (failing) on the year t test. The dashed line is students who scored at "Level III" (passing) on the year t test. The dotted horizontal lines mark the series value in 2008 which is the year immediately prior to the start of the retest policy.

Appendix Table C1—District retention after 2010 and treatment effects

	RD estimate of difference at pass/fail cutoff	Difference-in-RD estimate	
	Comparison years	Treatment years − Comparison years	Observations
	(1)	(2)	(3)
District change in retention at pass/fail after policy ended in 2010			
Bottom tercile of change	-0.001	0.034	531,111
(less retention)	(0.012)	(0.005)	
Middle tercile	-0.001	0.029	2,513,917
(roughly no change)	(0.009)	(0.005)	
Top tercile (more retention)	-0.004	0.036	846,998
	(0.007)	(0.009)	

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.

Appendix Table C2—Retest scores

Scale score points below pass/fail cutoff	RD est. diff. at pass/fail cutoff			
	Treatment years	Observations	Proportion passing on retest	Mean improvement retest-initial
	(1)	(2)	(3)	(4)
1	0.002	59,084	0.632	0.167
	(0.009)			
2	0.008	56,241	0.544	0.163
	(0.009)			
3	0.015	46,342	0.464	0.166
	(0.012)			
4	-0.002	42,328	0.393	0.175
	(0.012)			
5	-0.014	36,364	0.321	0.182
	(0.014)			
6	-0.014	30,302	0.273	0.206
	(0.017)			
7	0.013	26,816	0.222	0.215
	(0.025)			
8	-0.018	21,248	0.180	0.236
	(0.026)			
9	-0.016	19,716	0.143	0.262
	(0.028)			
10	-0.017	12,269	0.115	0.289
	(0.038)			

Note: “Treatment years” are the years of the retest policy, 2009-2012. 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.

Appendix Table C3—Teacher value added to retest scores as a predictor for future scores
Dep. var. = year $t + 1$ math test

	(1)	(2)
Year t teacher's value-added to retest score	0.196	0.188
	(0.056)	(0.053)
Treatment year	0.028	0.017
	(0.014)	(0.018)
Treatment year *	0.066	0.057
Year t teacher's value-added to retest score	(0.070)	(0.070)
Year $t + 1$ teacher fixed effects		√

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 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 a quadratic in year t initial test score and student demographics. 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 year t teacher level.

Appendix Table C4—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 $t-1$	0.121	0.083	0.105	0.130	-0.004
	(0.027)	(0.025)	(0.020)	(0.054)	(0.024)
Proportion failing $t-1$	-0.125	-0.084	-0.108		
* Treatment years 2009-2012	(0.028)	(0.028)	(0.021)		
Proportion failing $t-1$				-0.009	
* Comparison years 2007-2008				(0.053)	
Proportion failing $t-1$					0.002
* Comparison years 2011-2012					(0.024)
School-by-year observations	6,671	6,792	17,863	6,275	6,868

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 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.

Appendix Table C5—Potential effects on passing (non-retested) students
Dep. var. = year $t + 1$ math test

	(1)	(2)	(3)
Proportion of year t classmates	0.254		0.267
who failed year t test	(0.009)		(0.009)
Proportion * Treatment years	-0.010		-0.018
	(0.010)		(0.010)
Proportion of year $t + 1$ classmates		-0.016	-0.102
who failed year t test		(0.016)	(0.016)
Proportion * Treatment years		0.041	0.049
		(0.017)	(0.017)
Observations	888,627	888,627	888,627

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 column reports estimates from a separate least squares regression, using only observations on students who scored Level III (passing) on the year t exam. The dependent variable is the student's standardized math test score in year $t + 1$. The key right-hand-side variables, shown in the table, are calculations of the proportion of the student's classmates who scored Level II (failing) on the year t exam. The additional covariates, not shown in the table, are a quadratic in the student's own year t test score, where the parameters of the quadratic are allowed to vary for each grade-by-year cell; grade-by-year fixed effects, and teacher fixed effects. Robust standard errors in parentheses. Standard errors in parentheses are clustered at the teacher level.