EdWorkingPaper No. 21-444

Cramming: Short- and Long-Run Effects

Michael Gilraine
New York University

Jeffrey Penney
University of Alberta

An administrative rule allowed students who failed an exam to retake it shortly after, triggering strong `teach to the test' incentives to raise these students' test scores for the retake. We develop a model that accounts for truncation and find that these students score 0.14 standard deviations higher on the retest. Using a regression discontinuity design, we estimate thirty percent of these gains persist to the following year. These results provide evidence that test-focused instruction or `cramming' raises contemporaneous performance, but a large portion of these gains fade-out. Our findings highlight that persistence should be accounted for when comparing educational interventions.

VERSION: August 2021

Suggested citation: Gilraine, Michael, and Jeffrey Penney. (2021). Cramming: Short- and Long-Run Effects. (EdWorkingPaper: 21-444). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/94pe-5j18

Cramming: Short- and Long-Run Effects[*]

Michael Gilraine
Department of Economics
New York University

Jeffrey Penney
Department of Economics
University of Alberta

August 4, 2021

ABSTRACT

An administrative rule allowed students who failed an exam to retake it shortly after, triggering strong 'teach to the test' incentives to raise these students' test scores for the retake. We develop a model that accounts for truncation and find that these students score 0.14σ higher on the retest. Using a regression discontinuity design, we estimate thirty percent of these gains persist to the following year. These results provide evidence that test-focused instruction or 'cramming' raises contemporaneous performance, but a large portion of these gains fade-out. Our findings highlight that persistence should be accounted for when comparing educational interventions.

Keywords: Fade-out; Teaching to the Test; Education Production; Regression Discontinuity.

JEL codes: I20, I21, I28, J24.

[*] Thanks to Randy Reback, Joshua Hyman, Barbara Biasi and Steven Lehrer for helpful discussions about this topic. Thanks also to seminar participants at the University of Rochester and participants at the Northeast Economics of Education Workshop for additional comments. Jeffrey Penney thanks SSHRC for independent research support. Gilraine: Department of Economics, New York University, 19 West 4th Street, New York, NY 10003. Penney: Department of Economics, University of Alberta, 7-19 Henry Marshall Tory Building, Edmonton, Alberta, Canada T6G 2H4. Contact emails: Gilraine, mike.gilraine@nyu.edu; Penney, penney1@ualberta.ca.

1 Introduction

"During remediation (the week before the retest) teachers usually focus on two or three skills from the test [...] so they can bump them up those few questions/points to try to help them pass." - Quote from a teacher in North Carolina.[1]

In response to chronic concerns about the quality of public education, test-based accountability systems have become a key element of education reform. These schemes incentivize educators by tying rewards and punishments to the test score performance of their students. Opponents to using test-based measures argue that they induce teachers to take actions that boost student test scores at the expense of imparting deep underlying knowledge, or 'teach to the test' in common parlance. Oftentimes these critics specifically point to educators conducting weeks- (or months) long test preparation or 'cramming' programs in advance of standardized tests as lost opportunities for deeper learning.[2] Proponents counter that students benefit by learning the tested material.

Key to the debate about test-based incentives is whether they lead to longer-term student improvements or whether initial gains subsequently fade-out. A large literature highlights that contemporaneous test score gains decrease over time, often at high rates. For example, the contemporaneous test scores gains induced by various interventions such as remedial math courses (Taylor, 2014), pre-kindergarten programs such as Head Start (Deming, 2009), class size reduction (Chetty et al., 2011), and teacher assignment (Jacob et al., 2010; Jackson et al., 2014) all fade-out by approximately half after a single year; this fade-out can alter the policy calculus of educational interventions significantly (Macartney et al., 2018). There is limited evidence, however, whether contemporaneous test score gains from test-focused learning fade-out faster relative to other educational interventions.

This paper investigates the short- and long-term effects on academic achievement from a brief yet intense preparation period for a specific test which we call 'cramming.' To do

[1] See: `http://www.proteacher.net/discussions/showthread.php?t=405354` (Accessed May 30th, 2021).

[2] See, for instance, Mehta and Fine (2019).

so, we take advantage of a policy in North Carolina whereby students who fail to reach the proficiency threshold on a standardized test in either mathematics or English are compelled to reattempt the failed test(s) within one- to three-weeks. Because student performance on these retakes directly contribute to accountability metrics that affect the educators themselves, there are strong incentives for educators' to raise these students' test scores above the proficiency standard on the retest. In response, schools assigned students who were retaking the test to a one- to three-week intensive study program for that subject which they called 'remediation' (see introductory quote). The explicit goal of remediation was to improve students' test-taking ability for the year's accountability metrics, providing a unique case to investigate an extreme 'teach to the test' incentive and whether this type of test-focused instruction raises contemporaneous test scores and, if so, to what extent these gains persist.

Our empirical goal is thus to find the causal impact of the cramming program on both contemporaneous and subsequent student outcomes. For the latter, we employ a regression discontinuity design that compares students who just failed to reach the proficiency threshold on the initial test to those who just exceeded it. Students who failed to reach the proficiency threshold received the one- to three-week cramming intervention, while those that achieved the proficiency standard on the initial test were offered 'enrichment activities' of varying quality during this time.[3] Since only those who failed to achieve proficiency were eligible to receive the test-focused instruction, this comparison yields a clean estimate of 'cramming' on subsequent outcomes.

Estimating the impact of 'cramming' on contemporaneous outcomes is more challenging. In particular, we are unable to apply our regression discontinuity approach as only students who crammed were retested, and so we lack retest scores for students who did not cram. Furthermore, we cannot simply calculate the gain from the intervention as being the difference between scores on the initial test and the retest since scores on the initial test are truncated

[3] Anecdotally, these 'enrichment activities' included being kept "busy with extra assignments" (Wagner, 2013), looking forward to material from the subsequent grade, outdoor activities, or watching movies. The exact enrichment activity was usually left up to the teacher, and so likely varies substantially across and within schools. Unfortunately, we do not observe these 'enrichment activities' in our administrative data.

as, by definition, these students' initial scores could not exceed the proficiency threshold.[4] Given this truncation, retested students should, on average, score higher on the retest even in the absence of cramming affecting test performance.

We therefore develop a structural approach to estimate the effect of cramming on the retest based on a model of education production whereby test scores are a function of student test-taking ability plus measurement error. Cramming can then improve a student's test-taking ability before the retest. We allow the measurement error to be 'U-shaped' with respect to underlying student ability, in line with psychometric estimates. Our model predicts that students who score lower on the initial test will have a larger test score increase for the retest as lower-scoring students are more likely to have received a worse draw from the error distribution during the initial test. We find this exact pattern in the raw data.

To estimate the model, we assume that both measurement error and student ability are normally distributed and that the variance of the ability distribution is known. The variance estimates we use are from psychometric estimates based on test-retest and parallel-forms reliability constructed specifically for the tests we study. We estimate our structural model via minimum-distance estimation. The model precisely replicates the pattern of test score improvements in the raw data, whereby lower-scoring students achieve higher test score gains on the retest.

Our structural estimates indicate that the cramming program raises students' test scores on the retest by 0.140σ and 0.083σ for mathematics and English, respectively. These are large magnitudes for such a short intervention; as means of comparison, the highly-studied Project STAR small class size experiment yielded an effect of about 0.15 standard deviations for those students assigned to a small class for the duration of a year (Mueller, 2013). However, much of the benefits of cramming fades out over several years, with estimates from the regression discontinuity design finding that students who crammed score 0.041σ and 0.032σ on the mathematics and English test in the following year, respectively. These benefits

[4] Specifically, we are referring to the truncation of the dependent variable.

further decline to 0.025σ and 0.013σ two years into the future for mathematics and English, respectively, and the effect three years in the future is about 0.014σ for both subjects and remains statistically significant. We find precisely estimated zero effects of the cramming program on later life outcomes, including high school graduation, PSAT scores, SAT-taking, and SAT scores.[5]

We find that cramming does raise contemporaneous test performance substantially, but its impact fades out by about two-thirds the following year and by a further half the year after. Compared to fade-out estimates from other academic interventions, this is high. In comparison, Bailey et al. (2017) conduct a meta-analysis of 67 high-quality early childhood education interventions and find an average fade-out rate of fifty-five percent. A few interventions do find fade-out rates that compare to our estimates. (Appendix C provides a brief (non-exhaustive) summary of the estimated fade-out from some high-profile interventions). Notably, estimates suggest test score gains from the Project STAR class size experiment fade-out by three-quarters the first year after the program, although these gains did not fade-out further in the subsequent period (Krueger and Whitmore, 2001; Chetty et al., 2011). Furthermore, researchers found that Project STAR caused large improvements in long-run outcomes, such as earnings. In contrast, we find that cramming features high fade-out and does not yield any benefits to long-term outcomes such as high school graduation. Our results therefore lend some credence to opponents of test-focused preparation as its benefits do appear to rapidly fade-out and do not improve longer term outcomes. That said, our results indicate that some knowledge from cramming does persist over several years.

A closely related paper to ours is Aucejo et al. (forthcoming), who examine the same retesting policy. Those authors also examine the effect of the retest on test scores in the following year using a regression discontinuity design, finding near-identical results.[6] As we

[5] Our statement on finding a precisely estimated null effect is based on our point estimate being close to zero and the upper and lower 95% confidence interval bounds not being of practical significance. See Penney (2013) for further discussion on testing for a precisely estimated zero.

[6] Slight differences in the estimates are attributable to the different bandwidths used in the regression discontinuity design.

do, they contend that the gains arose during the remediation period. The authors make that contention by hypothesizing that the increase in subsequent test scores can come from three sources: changes in student treatment, changes in student behaviour, or changes in teacher behaviour. As they do not find changes in student absences, teacher assignments, or peer composition in the year following the retest, they rule out changes in student treatment and behavior as explanations for the test score gains. By elimination, the authors conclude that the change in student performance must therefore come from teachers helping students prepare for the retest.

In contrast, our paper directly estimates the knowledge gain students receive between the initial test and the retest. Estimating this knowledge gain then allows us to consider whether test-focused interventions fade-out quicker than other interventions. Highlighting the importance of differential fade-out, those authors conclude that "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" (Aucejo et al., forthcoming, p. 3). We find substantially higher gains for the retest (as high as 0.14σ), highlighting that the fade-out of test-focused interventions are by no means typical.

More generally, our work speaks to the importance of taking fade-out into account when comparing the impacts of various educational interventions. In particular, the level of fade-out is likely to be intervention-specific, and so researchers cannot simply use contemporaneous test score gains when comparing different policies. These concerns are especially salient when comparing interventions targeting foundational knowledge to those that focus on raising short-term test scores.

This paper is organized as follows. The next section details the institutional background along with the related literature. Section 3 develops our theoretical model, details our estimation strategies, and describes the data we use. Our empirical results are contained in Section 4. Section 5 concludes.

2 Background and Literature

The No Child Left Behind Act of 2001 (NCLB) in the United States mandated that annual assessments in mathematics and English were required in grades 3 through 8 and that there were consequences for schools that failed to achieve “adequate yearly progress” (AYP). The key determinant of AYP was whether the proportion of a school’s students scoring at or above the proficiency level (hereafter “proficient”) in each subject exceeded a state-determined threshold.[7] Schools that failed to achieve AYP faced escalating punishments, up to school closure or the firing of all staff.[8]

In the 2008-09 school year, the federal Department of Education granted North Carolina a waiver that allowed students who failed to reach the proficiency threshold on the end of year exam to be retested; for retested students, only the maximum score over the two tests would then be counted for accountability purposes. Since NCLB incentivized schools to maximize the proportion of proficient students, schools could improve their probability of achieving AYP by re-testing all students scoring below the proficiency threshold on the initial test.

In response to the retest policy, the state mandated that students scoring below the proficiency threshold in either the mathematics or English test had to retake the failed exam(s). Accordingly, schools notified parents after the initial test if their child failed to achieve the proficiency threshold in one (or both) tests and told them that their child had to retake the test(s).[9] Both the initial test and the retest for both subjects had to occur within the last 22 school days of the school year, although schools had discretion about when

[7] Schools had to reach this state-determined percent proficient threshold for both the school overall and in each of nine subgroups. Schools also had to meet several additional criteria, with some leeway also provided through safe harbor and confidence interval provisions. See Ahn and Vigdor (2014) for a detailed description of AYP determination for schools in North Carolina.

[8] The firing of all staff or school closure punishment was imposed when the school failed to reach AYP for six consecutive years. See Ahn and Vigdor (2014) for a detailed description of punishments for failing AYP in North Carolina.

[9] Given that the test was multiple choice, the state was able to quickly inform schools of their students’ performance. For instance, some teachers report that they were provided their students’ test scores on the same day that the test was taken.

to administer the tests within the 22-day window. Indeed, Aucejo et al. (forthcoming) find that schools moved up the dates of their initial test when the retest policy was in place to maximize the time between the two tests; intuitively, this indicates that schools were trying to maximize the time they could prepare their students for the retest. On average, the retest was taken approximately two weeks after the initial test. The rule was active through the 2011-12 school year,[10] with over ninety-eight percent of students who failed to reach the proficiency threshold in mathematics and English being retested in that subject.[11]

Students who failed to meet the proficiency threshold on one or more tests were subjected to a program which consisted of teachers providing focused instruction on test materials and additional tutoring in the failed subject(s).[12] This was a short-duration intervention focused on preparing students for the retest. In contrast, students that passed the proficiency standard on the initial test were given 'enrichment activities' during this time. These activities were at the teacher's discretion and so the quality of education during this time likely varied substantially both across and within schools. Anecdotally, some teachers used this time to teach "some items I want to cover to prep them for next year (as well as some fun things we didn't get time to do)!", while others used the time less productively, leading to some parents complaining that students would do "the EOG's and then they spend the rest of the year watching movies."[13] As the end-of-grade tests occurred in May, there was a sense that, for the students that passed, the post-test period was an opportunity to unwind during the final weeks of school. Indeed, one teacher who taught the remediation program for the students

[10] After 2011-12, North Carolina was no longer subject to NCLB requirements as they received a statewide NLCB waiver to participate in the Race to the Top initiative. The retest rule was therefore halted.

[11] The ninety-eight percent retest rate is among students scoring in the 'below proficient' category who are the focus of our analysis. There is also a 'well-below proficient' category for very low-performing students. For these students, parents had to request that their child retake the test rather than be automatically enrolled. Even among these students, however, around ninety percent were retested.

[12] For example, see http://www1.gcsnc.com/boe/2010/5_11/procedure_ikea_p.pdf (accessed May 30th, 2021), which details the retest policies of Guilford County Schools. In the document, the school district details that students who fail to reach the proficiency threshold will be retested after a focused intervention involving extended instructional opportunities such as: smaller class sizes, modified instructional programs, extended school or tutorial sessions, Saturday school, special homework, or parental involvement.

[13] These quotes are taken from https://www.city-data.com/forum/charlotte/654710-eog-burn-out-how-your-students.html (accessed May 30th, 2021).

that failed the initial test commented that "the hardest part is keeping them working while everyone else is shutting down."[14]

Teachers and schools were highly incentivized to raise the performance of these students in light of the possible consequences to the school of failing AYP. In contrast, students faced no explicit incentives to perform well on the retest, although very low performance could affect whether educators believed the student should be held back. Retention decisions, however, were never solely based on a student's performance on the test.[15]

2.1 Related Literature

This paper seeks to identify the impact of short-term test-focused preparation using a rule whereby schools were incentivized to invest substantially in students whose test scores are below a certain threshold.

Of particular interest for the policy examined in this paper is the extent to which the benefits persist. It is well known that the effects of educational interventions decrease over time; this is known in the literature as "fade-out". For example, a review of the effects of teacher quality on student achievement found that approximately half the benefit from being assigned to a teacher dissipates after two years (Jackson et al., 2014). (See Appendix C for a brief summary of the estimated fade-out from some high-profile interventions). A recent comprehensive study of fade-out in academic ability and other psychological processes found that there are many potential reasons for its occurrence (Bailey et al., 2020a). The most important of the reasons explored for purposes of our paper is what is known as overalignment between treatments and outcomes; 'teaching to the test' is a specific example of this. For example, consider a math test designed to evaluate mathematical ability. Since the test

[14] See: `http://www.proteacher.net/discussions/showthread.php?t=405354` (accessed May 30th, 2021).

[15] North Carolina experimented with a policy meant to end social promotion by requiring students to perform at a certain level in 3rd, 5th, and 8th grade; this policy was scrapped in 2010 because it was found to not have affected the rate at which students failed to promote to the next school grade, ostensibly due to principals using exceptions to advance students. For more information, see `https://www.edweek.org/ew/articles/2010/10/20/08brief-3.h30.html` (accessed October 5th, 2018).

does not measure underlying ability perfectly, the teacher can focus on tested material (or 'teach to the test') to raise test scores even though teaching a more balanced curriculum may increase unobserved mathematical ability by more.[16] The worry is that teaching to the test creates a sort of hollow knowledge that is more easily forgotten (Jensen, 1999). Therefore, we may expect larger than usual fade-out from this cramming intervention.

A second consideration is the influence of the intervention on long-term outcomes, such as high school graduation rates. Many educational interventions fade-out over the following years, only to have effects occur much later on in life. For example, Chetty et al. (2011) and Chetty et al. (2014) find benefits in adulthood from lowered class sizes and assignment to higher quality teachers respectively, even though the benefits to test scores have substantially dissipated. The shallow learning afforded by teaching to the test for a brief period may have less potential for long-term benefits.

Our intervention has a similar flavor to those in other contexts. In particular, it is somewhat comparable to a large literature on programs that provide additional instruction to low-performing students in K-12 education; we briefly overview these interventions in Appendix Section B.[17] In general, the literature has found that these types of programs generate a moderate increase to standardized test scores.

Additionally, papers have found that mentoring students – which provides a form of additional instruction – can generate large positive effects on student achievement (Rodriguez-Planas, 2012; Oreopoulos et al., 2017). Researchers have also used exogeneous variation in the length of school year (Pischke, 2007), teacher strikes (Baker, 2013), and snow days (Goodman, 2014) to show that additional instructional days increase student achievement. Finally,

[16] The challenge in test design is the creation of a set of problems that can best reflect the latent level of ability of interest. In theory, if a test measured latent ability in a complete fashion, it would not be possible to teach to the test because that would be equivalent to increasing ability. Teaching to the test is taking advantage of the misalignment between measured ability using a testing instrument and actual ability.

[17] Grade retention is a type of remediation program which generally finds negative effects on student achievement (Roderick and Nagaoka, 2005; Manacorda, 2012), although this finding is disputed (e.g., see Fruehwirth et al., 2016). However, this policy may have very different effects compared to other remediation programs that keep children in the same grade but which provide additional instruction. Grade retention is also often combined with summer school; Jacob and Lefgren (2004) suggests that the zero effect in these cases might be the result of a positive summer school effect being masked by a negative retention effect.

outside of K-12 education, college-level remediation courses have been demonstrated to often generate positive effects (Bettinger and Long, 2009; Calcagno and Long, 2008; De Paola and Scoppa, 2014), although some papers find no effect (Lagerlöf and Seltzer, 2009; Martorell and McFarlin Jr, 2011).

3 Empirical Framework

The goal of our empirical analysis is to estimate the impact of the cramming program on student knowledge in both the retest and later academic outcomes, as well as its influence on later life outcomes such as SAT test-taking and high school graduation rates. To estimate the effect of the cramming program on future outcomes, we compare students that just exceeded the proficiency threshold (and thus did not 'cram') to those that just failed to achieve proficiency (and therefore 'crammed') using a regression discontinuity design. However, this research design cannot be employed to estimate the impact of the cramming program on the retest as we do not observe retest scores for those who exceeded the proficiency threshold on the initial test; that is, we are working with a truncated sample for the retest. To overcome this issue, we develop a theoretical model of educational production to estimate test score gains for these students from the initial test to the retest.

3.1 Model of Retest Knowledge Gain

This subsection introduces our model of knowledge acquisition to estimate the increase in learning students obtain during the cramming program using a minimum of assumptions. The *initial* test score of student i in grade g at time t is given by y_{igt} and consists of both the knowledge that the student has, α_{igt}, and an additively separable error term representing measurement error of the student's level of knowledge, ϵ_{igt}, which is mean zero, symmetric, single-peaked, and whose c.d.f. is given by $F(\cdot)$. Therefore, a student's initial test score is:

$$y_{igt} = \alpha_{igt} + \epsilon_{igt}\,. \tag{3.1}$$

Students who score below some threshold are then compelled to take a retest and receive the 'cramming' intervention. Formally, let the retest threshold in grade g at time t be given by r_{gt}. We assume throughout that the retest threshold is set below the mean initial test score (i.e., $r_{gt} < \bar{y}_{gt}$), in line with our empirical application. Further, denote the score the student achieves on the retest by y^*_{igt} and assume that student knowledge from the initial test fully persists to the retest (sensible, given that they are only one- to three-weeks apart and the retest is designed to test the same knowledge as the initial test).

The 'cramming' intervention then raises student knowledge by z_{igt}, which we assume to be additively separable.[18] The retest score, y^*_{igt}, can therefore be expressed as:

$$y^*_{igt} = \alpha_{igt} + z_{igt} + \epsilon^*_{igt}\,, \tag{3.2}$$

where ϵ^*_{igt} is the error term on the retest which is drawn from the same distribution as the initial test. Our goal is to estimate the average increase in student knowledge from cramming, $\bar{z}_{gt}$.

A student's test score gain between the two tests is:

$$Gain_{igt} \equiv y^*_{igt} - y_{igt} = z_{igt} + (\epsilon^*_{igt} - \epsilon_{igt})\,. \tag{3.3}$$

The two test scores y^*_{igt} and y_{igt} are observable, and so one may be tempted to just treat the increase in test scores from the initial test to the retest as the knowledge gain. Unfortunately,

[18] The additive separability assumption ensures that there is no interaction between a student's initial ability, α_{igt}, and their retest knowledge gain, z_{igt}. This assumption is required for estimation to ensure that there is no heterogeneity in retest knowledge gain by initial ability. During estimation, however, we only use students scoring up to 1.4 standard deviations below retest threshold and so additive separability must only hold among these students who are of relatively similar ability.

such an estimate would be biased upward because ϵ_{igt} is no longer mean zero since the sample of students taking the retest is truncated based on their performance on the initial test: students who score below the retesting threshold will, on average, have a lower value of ϵ_{igt} compared to those who pass. Since the mean of ϵ^*_{igt} is zero, estimating the average knowledge gain, $\overline{z}_{gt}$, by summing the differences between the retest and initial test scores of all students falling below the threshold r_{gt} will thereby *overestimate* the true knowledge gain students receive from the cramming intervention.

Specifically, we have that student i is retested if $y_{igt} < r_{gt}$ or $\epsilon_{igt} < r_{gt} - \alpha_{igt}$. In expectation, the probability that student i is retested, P_{igt}, is then given by:

$$P_{igt} = F\left(\alpha_{igt} - r_{gt}\right). \tag{3.4}$$

Proposition 1 *Let students who score below the test score threshold r_{gt} be indexed as $j = 1, ..., J$. For students scoring below the threshold, the average difference between the retest and initial test scores, $\sum_{j=1}^{J}(y^*_{igt} - y_{igt})/J$, will overestimate the students' average knowledge gain between those two tests, $\overline{z}_{gt}$.*

Proof. See Appendix D. ■

We highlight the intuition underlying Proposition 1 through a simple example. Suppose that the student knowledge distribution for the initial test is standard normal, the error term is distributed $\epsilon_{igt} \sim \mathcal{N}(0, 0.25)$, the retest threshold is $r_{gt} = -0.5$, and the average treatment effect of cramming, $\bar{z}_{gt}$, is 0.1. Given the assumed distributions, a student who scores right at the retest threshold, $y_{igt} = -0.5$, would receive, on average, a draw of ϵ_{igt} of -0.10. Since the error on the retest (ϵ^*_{igt}) is mean zero, the student's expected increase in test score from the initial test to the retest is:

$$\mathbb{E}[y^*_{igt} - y_{igt} | y_{igt} = -0.5] = \mathbb{E}[z_{igt} + \epsilon^*_{igt}] - \mathbb{E}[\epsilon_{igt} | y_{igt} = -0.5] = 0.1 + 0 - (-0.1) = 0.2\,. \tag{3.5}$$

Therefore, this student's retest score will, on average, be $y_{igt}^* = -0.3$, a 0.2 increase relative to the initial test score of -0.5. This represents twice the actual knowledge gain of 0.1.

Corollary 1 *The overestimation of students' average knowledge gain between the initial test and retest is decreasing with respect to a student's initial test score,* y_{igt}^*.

Proof. See Appendix D. ■

Corollary 1 states that students who score lower on the initial test will have, on average, worse draws of the error distribution ϵ and so will expect a higher test score gain from the initial test to the retake.[19] Using the same example as above, we see the overestimation becomes more pronounced for students scoring lower on the initial test: for instance, the retest gain is 0.3 for those with an initial test score of $y_{igt} = -1.5$ (in contrast to the same 0.1 actual knowledge gain for these students).

In the raw data, we observe exactly this type of relationship. Figure 1 shows the mathematics and English retest gain, $y_{it}^* - y_{it}$, for students who score below the retest threshold. As expected, students scoring lower on the initial test achieve a higher retest gain, indicating that lower performing students on the initial test have, on average, lower draws of the error term, ϵ_{igt}. The relationship is highly nonlinear. While the c.d.f. introduces some non-linearity into the relationship, the high degree of non-linearity suggests that another factor is at play: in particular, a relationship between the variance of the measurement error, σ_ϵ^2 and underlying test-taking ability, α_{igt}, could introduce this high degree of non-linearity. Given that psychometricians find that error distributions are more disperse for students whose knowledge is further away from the mean, we introduce a possible relationship between the variance of the error term and test-taking ability in the estimation procedure that follows.

[19] If one wishes to assume non-constant treatment effects, monotonicity is a sufficient condition in order for this statement to be true. We feel this is a reasonable assumption to make in this context.

3.2 Estimation of Retest Knowledge Gain

To estimate the knowledge increase from the initial test to the retest, we impose some structure: in particular, we assume that test-taking knowledge and test scores are normally distributed and allow the variance of the error term and students' test-taking knowledge to be correlated. With these assumptions in hand, we then match the variance of the underlying ability distribution to the reliability of the test that has been estimated by psychometricians during test design using test-retest reliability and parallel-forms reliability.[20] Using this variance, we then minimize the mean squared error of our simulated retest gain relative to the actual retest gain observed in the data. We accomplish this using three parameters which we estimate: two that govern the variance of the error term and its relationship to students' underlying knowledge, and one that captures students' average knowledge gain from the initial test to the retest. This knowledge gain is our parameter of interest.

The first assumption that we impose on the data is that student test scores follow a standard normal distribution. Formally,

Assumption 1 *Test scores y_{igt} are distributed $y_{igt} \sim \mathcal{N}(0, 1)$.*

This assumption is relatively harmless since we follow much of the literature and standardize test scores to have mean zero and variance one at the school-grade level, and so the first two moments of the standard normal distribution are automatically imposed through test score normalization. For higher order moments, we do not observe large deviations from normality.[21]

Next, we need to make an assumption on the distribution of the measurement error, ϵ_{igt}. As previously stated, psychometricians find that measurement error is more dispersed for students whose ability is further away from the mean. For example, Figure A.1 reports

[20] Test-retest reliability refers to how closely repeated measurements using the testing instrument (e.g., an exam) produce the similar results. Parallel-forms reliability measures how different forms of the testing instrument produce the same results (e.g., two different versions of a test that are designed to test the same underlying ability).

[21] This aligns with Goldhaber and Startz (2017) and Gilraine et al. (2020) who investigate the test score distribution in North Carolina and do not find large deviations from normality.

the estimated standard error of measurement for the fourth grade mathematics and English tests in North Carolina by underlying student test-taking ability.[22] These estimates were constructed by the psychometricians who designed the test and rely on test-retest and parallel-forms reliability by student performance levels to estimate the standard errors of the test by test-taking ability. The estimated error is 'U-shaped' with respect to ability with the minimum variance occurring for mean ability students. We note that this 'U-shaped' variance figure is common to nearly every item response theory test analyzed by psychometricians.

While we could directly take these psychometric estimates, we instead estimate the underlying relationship by choosing the error distribution that will most closely match the retest gains we observe, although we note that our estimated error distributions are similar to the psychometric estimates. To do so, we approximate the true relationship between test score dispersion and underlying student knowledge using the following functional form:

Assumption 2 *The error term, ϵ_{igt}, is distributed $\epsilon_{igt} \sim \mathcal{N}(0, \sigma_\epsilon^2)$, where $\sigma_\epsilon \equiv \theta_0 + \theta_1 |\alpha_{igt}|$ and θ_0 and θ_1 are unknown parameters to be estimated.*

The approximation captures the fact that test scores become noisier measures of ability for the most and least able students; our assumed functional form also imposes symmetry in the error distribution with respect to ability. Though the true distribution may not be perfectly symmetric, the assumption of symmetry is harmless in our empirical implementation because the retest threshold is set well-below the mean (by approximately 0.8 standard deviations). Therefore, effectively no students with above-mean knowledge levels are retested, and so above-mean ability students do not contribute to the estimation of the retest gain (that is, we effectively only use the left-side of the 'U' in our estimation procedure).

To solve for the average test score increase of students from the intervention, we must also know the underlying ability distribution of the students. Fortunately, the variance of this distribution has already been estimated. The psychometricians who constructed the North Carolina end-of-grade tests calculate the ability distribution by assuming normality

[22] The figures for other grades are similar.

of both the ability and error distributions and then, using the test-retest and parallel-forms reliability statistics, calculate the proportion of variance coming from variation in student ability and the error term. We take these test-specific reliability measures and treat them as known:

Assumption 3 *Ability is distributed $\alpha_{igt} \sim \mathcal{N}(0, \sigma^2_{\alpha_{gt}})$, where $\sigma^2_{\alpha_{gt}}$ is derived from psychometric estimates based on test-retest and parallel-forms reliability.*

For instance, the fourth grade mathematics test in North Carolina has a reliability measure of 0.915. This implies that fourth grade ability is distributed $\alpha_{i4t} \sim \mathcal{N}(0, 0.915)$.

To estimate our model, we first standardize test scores for both the initial test and the retest by grade-year using the mean and standard deviation from the initial test. This allows us to estimate our model using data from all grades jointly.[23] We then define our estimation sample as students who took the retest and scored within 1.4 standard deviations of the retest threshold, which covers students scoring one to fifteen developmental scale points below the retest threshold.

With our sample in hand, we estimate our model through the following algorithm. First, we simulate student knowledge and guess a value of θ_0.[24] Second, a value of θ_1 is chosen so that the variance of our simulated initial test scores equals one, ensuring that our simulated test score distribution is standard normal (i.e., Assumption 1 holds). Third, we calculate the expected gain in the retest for each value of the initial simulated test score ($= \alpha^{Sim}_{igt} - y^{Sim}_{igt}$); this third step therefore finds the expected test score gain between the initial test and the

[23] Alternatively, we could estimate our model grade-by-grade. Our estimate is near-identical when we do the estimation grade-by-grade; given that, we estimate our model jointly as it substantially speeds up estimation. For example, our estimate of the retest gain in mathematics is 0.140σ. When doing the estimation grade-by-grade, the average retest gain is 0.141σ, with a range 0.13-0.16σ across the five grades used for estimation.

[24] We simulate student knowledge by imposing the reliability of the test, which is grade-specific. To estimate our model using all data jointly, we average these grade-specific test reliability measures and so our assumed ability distributions are $\alpha \sim \mathcal{N}(0, 0.914)$ for mathematics and $\alpha \sim \mathcal{N}(0, 0.912)$ for English. (Simulated test scores are based on 2.5 million simulated scores.) These grade-specific ability distributions tend to be very similar across grades (e.g., ranges from 0.905 to 0.922 for mathematics), and so this averaging is relatively harmless which is confirmed by the near-identical mathematics retest gain estimates of 0.140σ versus 0.141σ when using joint versus grade-by-grade estimation.

retest if there were no knowledge gain. Fourth, we find the average knowledge gain, $\bar{z}_{gt}$, that minimizes the mean squared error between the observed and simulated retest gain: $\bar{z}_{gt} = \underset{\bar{z}_{gt}}{argmin}(y^{*,Observed}_{igt} - y^{Observed}_{igt} - (y^{*,Sim}_{igt} - y^{Sim}_{igt}) - \bar{z}_{gt})^2$. We iterate until we find the value of θ_0 that minimizes the mean squared error between the observed and simulated retest gain. Since our observed initial test scores are somewhat discrete, we discretize the observed and simulated initial test scores throughout into 0.1σ bins to map our continuous simulated initial test scores to the observed discrete test scores.

3.3 Regression Discontinuity Design

We obtain reduced form estimates of the effect of the one- to three-week cramming program on student achievement in future years using a (sharp) regression discontinuity design. The identification strategy takes advantage of the administrative rule that students are only subject to the cramming program if their test score on the initial test falls below the proficiency standard. In contrast, students who exceeded the proficiency threshold were not availed this opportunity.

To illustrate the empirical approach, consider two students: one who scores just below the proficiency threshold and another that meets or just exceeds the proficiency threshold. Due to the policy rule, the student scoring just below the threshold is provided with the cramming program and is retested, while other is not. Since students scoring just below the proficiency threshold are unlikely to differ much from students scoring just above, we can compare the future outcomes across these two students to examine the effect of the cramming program.

As we incorporate students further away from the proficiency threshold into the analysis, possible confounding relationships between the student's score on the initial test and their performance in future periods may appear, making it necessary to control for current test scores through some function, which we assume to be linear in the main analysis.[25] To keep

[25] Table A.1 examines robustness to alternative control functions.

with the spirit of comparing students close to the cutoff, we restrict attention to students within a bandwidth of five scaled test score units of the proficiency cutoff – Appendix Figure A.5 shows robustness to alternative bandwidths.[26]

Formally, we estimate the effect of the retest on knowledge in period $t+s$ through the following equation:

$$y_{ig,t+s} = \zeta + \beta_{RD}^{t+s} D_{igt} + \theta X_{igt} + \phi D_{igt} \cdot X_{igt} + \eta Z_{igt} + \lambda_{gt} + \epsilon_{ig,t+s} \,,$$
$$for - b_{gt} \leq X_{igt} \leq b_{gt} \,, \tag{3.6}$$

where $y_{ig,t+s}$ is student i's test score in their first sitting of the end-of-grade test in period $t+s$ (where $s > 0$), $D_{igt} \equiv \mathbb{1}\{X_{igt} \leq 0\}$ is an indicator variable set to one if the student was below the retest threshold in period t (and thus had to take the retest), X_{igt} is student i's initial test score in period t normalized by its distance to the proficiency cutoff, Z_{igt} is a set of student-level controls (e.g., ethnicity, gender, etc.), while λ_{gt} are grade-by-year fixed effects. The regression is restricted to observations whose test score is within b_{gt} of the retest threshold threshold, and so b_{gt} is the chosen bandwidth. The coefficient of interest in equation (3.6) is β_{RD}^{t+s} which, under assumptions that are tested in Section 4.3, represents the causal effects of the cramming program on student achievement in period $t+s$. Our dependent variable in equation (3.6) has a correlated error structure since students within a school may face common shocks and because our data contain repeat observations on students in different grades, and so we therefore two-way cluster our standard errors by both school and student.[27]

[26] Our chosen bandwidth of five is very close to the optimal data-driven bandwidth found using the methodology in Calonico et al. (2014), which is usually around four (depending on the regression).

[27] These standard errors are more conservative than the heteroskedastic-robust standard errors suggested by Lee and Lemieux (2010). We do not cluster by the value of the running variable (Lee and Card, 2008) as Kolesár and Rothe (2018) show that standard errors clustered by the discrete running variable have poor coverage properties. Our confidence intervals end up being similar to the "honest CIs" from the bounded misspecification error model class in Kolesár and Rothe (2018).

3.4 Data

This paper employs detailed administrative data from the North Carolina Education Research Data Center (NCERDC) (2008-2019). These include information about all public school students in North Carolina for the 2008-09 to 2018-19 school years. The data set contains test scores for each student in mathematics and English for grades three through eight from standardized tests that are administered at the end of each school year in the state. In addition, the NCERDC also details whether the student retook either the mathematics or English test in any given year along with the test score on the retake. Since the retest policy was only in place from 2008-09 through 2011-12, our sample is given by students in grades 3-7 in those years, although we use the later years of data to obtain future outcomes for each student.[28]

Test scores are reported on a developmental Rasch scale that is designed such that each additional point represents the same gain in the level of knowledge, regardless of the student's grade or baseline ability. To facilitate the comparability of test scores across grades and years, we standardize this scale at the student level to have a mean of zero and a variance of one for each grade-year. The NCERDC data also contain unique student identifiers, which enable us to follow the same student over time and observe their subsequent test scores. A variety of student-level covariates are available, among which include race, gender, economically disadvantaged status, English learner status, disability status, and gifted status.

We are also able to link students to long-run outcomes that occur at the end of high school. We focus on several important high school outcomes including PSAT scores, SAT-taking, SAT scores, and high school graduation. These data cover any student that remains in North Carolina's public school system throughout high school. We link students to these later life outcomes, but restrict our sample to ensure that students have reached the required age to have realized these outcomes since our long-run outcome data only cover up to the

[28] While our data contain students' eighth grade scores, we do not use these scores in our analysis (except as future outcomes) as we do not have a subsequent test score for these students.

2018-19 school year (e.g., will have reached twelfth grade for the SAT outcomes).[29]

Table 1 displays the summary statistics of the data. Column (1) shows student characteristics for all students in the sample. North Carolina has a white student plurality and a substantial black minority population (twenty-six percent), with Hispanic and Asian students making up a further twelve and three percent of the student body, respectively. Column (2) restricts the sample to students scoring within five scale points of the proficiency standard, which constitutes our RD sample for mathematics. Since the proficiency standard in North Carolina is set well-below the mean, the RD sample contains students whose mathematics scores are, on average, a full 0.60σ below the mean. These students are also more likely to be Black and Hispanic relative to the overall sample; economically disadvantaged and English Learners are also over-represented. Similarly, column (3) restricts the data to students in our 'retest sample' that will be used in our structural estimation for mathematics, which consists of even lower-performing students given that those in this sample must have scored below the proficiency standard.

4 Results

This section presents our empirical results. First, we use our structural model to estimate the gains for the initial test to the retest. We then use our regression discontinuity design to calculate how these contemporaneous test score gains persist into future periods and what effect the intervention has on outcomes at the end of high school.

4.1 Initial Gains on the Retest

Figure 1 displays the relationship that we observe in our data between test score gains from the initial test to the retest as a function of a student's initial test score. On average,

[29] Specifically, we require that a student's cohort has reached grade 12 for the SAT outcomes, one year beyond grade 12 for high school graduation (to allow one year of grade repetition), and grade 11 for the PSAT outcomes. Since our PSAT data only begin in 2012-13, we also drop the 2008-09 seventh grade cohort as this cohort would have reached grade 10 before 2012-13 (as students take the PSAT in grade 10 or 11).

students in our 'retest sample' score a full 0.236σ and 0.175σ higher on their mathematics and English retake, respectively. We also observe that students with lower initial test scores experience higher test score gains between the two tests, as predicted by our model presented in Section 3.1 (see Corollary 1). Given this relationship, it is likely that a large portion of the retest gains are driven by the fact that retested students are more likely to have received a low draw from the error distribution (ϵ) on the initial test.

We therefore estimate the knowledge gain from the initial test to the retest using the structural estimation procedure outlined in Section 3.2. Model estimates are presented in Table 2 and indicate that, on average, the cramming intervention increased students' test-taking knowledge by 0.140σ in mathematics (s.e. = 0.002) and 0.083σ in English (s.e. = 0.003).[30] These represent gains that are quite large.

Figure 2 displays the fit of our structural model. To do so, it shows the average test score increase that retested students achieve on the retest in comparison to the initial test by distance to the retest threshold in both the raw and simulated data. The raw test score gains in Figure 2 are identical to those from Figure 1. The simulated test score gains are then superimposed using a dashed line. For both mathematics and English, the simulated test score gains lie nearly directly upon the actual test score gains. Our numerical model therefore fits the raw data near-perfectly, lending credence to the ability of our structural approach to capture the knowledge gains from cramming that we seek to estimate.

4.2 Gains on Future Outcomes

To estimate the impact of cramming on future outcomes, we turn to the regression discontinuity (RD) design introduced in Section 3.3.

RD Results for Subsequent Test Scores: We next estimate the effect of the short-

[30] Bootstrapped standard errors are employed. The estimates for the other two structural parameters that govern the variance of the error term and its relationship to students' underlying knowledge for mathematics are: $\theta_0 = 0.154$ (s.e. = 0.009) and $\theta_1 = 0.160$ (s.e. = 0.009). For English, the estimated structural parameters are $\theta_0 = 0.257$ (s.e. = 0.008), $\theta_1 = 0.041$ (s.e. = 0.009).

term intervention on test scores in the following grades. Recall that the intervention is subject-specific (e.g., failing the mathematics test will result in a mathematics-specific cramming program) and so we start by reporting the impact of the intervention on the subject that was its focus, the results of which are displayed in Table 3. The mathematics-focused intervention raises mathematics scores in the following year by approximately 0.04 standard deviations, while English-focused cramming increases reading scores by about 0.03 standard deviations. The results are similar regardless of whether demographic covariates are included in the specification. The magnitude of these coefficients is small, but not economically insignificant: the effects of the intervention are approximately one-third to one-sixth of small class size effects (Mueller, 2013).

The effects of the intervention on test scores two and three years into the future are as follows. Examining the results with covariates, there continue to be precisely estimated positive effects on student achievement in the $t + 2$ period for mathematics and English, although the impact has declined by roughly one half compared to the benefits experienced in the $t+1$ period. In the $t+3$ period, the effect falls to 0.015 standard deviations for math and 0.013 standard deviations for English, and the estimates remain statistically significant.

The fade-out of the estimates appears to be quite large for mathematics in the first year after the retest, falling from 0.140σ to 0.041σ, implying a fade-out of 71 percent. For English, the estimated fade-out the first year after the retest is 62 percent. This is a more rapid rate of decrease in comparison to other educational interventions (see Appendix C), which usually experience rates near 50 percent. The larger than normal level of fade-out lends credence to the theory that contemporaneous gains from test-focused interventions do not lead to commensurate improvements in long-term term outcomes.

Interestingly, the intervention also has an indirect spillover effect on the test scores of the other subject in the following year. These results are reported in Table 4. In particular, we find that receiving the mathematics-focused cramming program increases *English* scores in the following year by 0.013 standard deviations; we find a similar result (0.012σ) for the impact of the the English-focused intervention on mathematics scores. Because the intervention focuses only on the test that is failed, this increase in ability in the other subject is likely a behavioural response to having failed the exam in the previous year. For instance, students may have increased the amount of time they spent studying for these tests because their self-concept of ability was affected by failing the test in the previous year,[31] or because the cramming program improved study skills, some of which was applicable to other subjects. The magnitudes are about one-third of the effect on the own-subject retest for the same period.

We also perform our analysis on subsamples of students who are economically disadvantaged and by gender (results not displayed). We find minimal differences (if any) here, suggesting little heterogeneity in the treatment effect by observable student characteristics. Similarly, performing the analysis by grade indicates little heterogeneity in the treatment effect by the grade that the student was treated (results not displayed). Further, we have also investigated whether the impact of cramming was larger among schools that were more in danger of failing to meet AYP. Once again, we find little heterogeneity among this dimension (results not shown).

In addition, we also looked at whether students that failed *both* tests were affected differently by being subject to two cramming programs rather than just one using a multidimensional RD design.[32] We find that the math and English cramming programs are substitutes in the education production technology, with students subject to both cramming programs scoring roughly 0.01σ lower in mathematics compared to students who only failed mathematics on the initial test (the results for English are very similar). This result may indicate that cramming faces decreasing marginal returns, as test score gains from two cramming programs are less than twice the gains from a single program.

[31] Self-concept with regards to education is a student's perception of their own academic ability. Ding and Lehrer (2011) find that small class interventions raise both academic ability and self-concept in the Project STAR data, which suggests a possible link between the two.

[32] Following Dell (2010), our multidimensional RD design only include students near the retest threshold in both subjects and control for distance to both the math and English retest thresholds (along with an interaction between them) as forcing variables.

RD Results for Outcomes in High School: A natural question that arises is whether the high fade-out is permanent or if the contemporaneous gains from cramming fade-out but then resurface later in life. For instance, Chetty et al. (2011) find that the gains from smaller classes in Kindergarten have largely faded out by eighth grade, but these gains reappear for later life outcomes such as earnings. We investigate the impact of cramming on outcomes in high school in Table 5 which are PSAT and SAT-taking, PSAT and SAT scores, and high school graduation. Panels A and B report the results for the mathematics and English-focused cramming program, while Panel C stacks the data and so indicates the impact of a cramming program in either subject.

We find that the effect of cramming on SAT-taking, SAT scores, and high school graduation consist of precisely estimated zeros.[33] A small statistically significant effect of cramming on PSAT scores is found, although this impact is small: attending a cramming program raises PSAT scores by 2.7 points on the 1520 point scale. We do not interpret this as an economically meaningful effect, as this corresponds to approximately 0.5 percentile points at the median score for 10th graders.[34] It therefore appears that the massive 0.14σ mathematics test score improvement imparted by cramming largely fades out in subsequent periods, and does not meaningfully affect later life outcomes.

4.3 Robustness

To ensure the validity of our estimates, we conduct a series of robustness checks to verify that the regression discontinuity design is credible in our context.

RD Validity: For the RD design to be credible, we require that the administrative retesting rule is enforced. Figure 3 plots the proportion of students who retook the math or reading test by their (normalized) test score in the first sitting. We see that once a student's

[33] The ability to make this claim hinges on the fact that the standard errors are very small in absolute value (in the sense of economically significant magnitudes) and that the point estimate is close to zero. One cannot credibly make such a claim without both of these being true.

[34] See `https://research.collegeboard.org/programs/sat/data/data-tables-technical-resources/psat-nmsqt-score-information` (Accessed May 30th, 2021).

test score dips below the proficiency threshold, students almost always retake the relevant test, while they nearly never retake test if their test score is above the threshold. In fact, adherence to the retesting rule is nearly perfect: for mathematics, there is a 99 percentage point drop in the probability a student retakes the math test once the threshold is crossed, and the discontinuity for reading is even higher. Given the near-perfect adherence to the retesting rule, we do not adjust our point estimates for non-compliance, although the results are nearly identical if we do.[35]

A key identifying assumption behind the RD design is that students on either side of the retest cutoff are similar in both observable and unobservable dimensions. It seems unlikely in our context that test scores could be manipulated given that tests are scored centrally by the North Carolina Department of Public Instruction, and students are not perfectly aware of their score while taking the test. Regardless, we check for potential manipulation of test scores by seeing whether there is any bunching of test scores on either side of the proficiency threshold which would suggest that teachers or test markers are manipulating test scores so that students end up on one side of the proficiency threshold.

Figure A.2 plots the distribution of student test scores (normalized by the proficiency threshold) for both mathematics and English. If test scores are being manipulated, we would expect there to be a large number of test scores on one side of the test score cut-off for passing. Visually, there does not appear to be any excess density around the threshold. A formal test of continuity in the density around the threshold (Frandsen, 2017)[36] confirms the visual analysis: the null hypothesis of continuity at the cut-off is not rejected for mathematics. Continuity is rejected for English, although we note this is entirely driven by the distribution of seventh grade English scores and is likely just an artifact of the conversion of raw test scores to the developmental scale sometimes bunching at certain developmental scale units. Once seventh grade is excluded, continuity in the density at the threshold is no longer rejected;

[35] One could account for non-compliance by scaling up our reduced-form estimates by a factor of 1.01.

[36] We use the Frandsen (2017) test over the more commonly-used McCrary (2008) test as our running variable is discrete.

our results are near-identical if seventh grade English scores are excluded from all of our analysis.

Next, we estimate reduced form econometric models to check whether predetermined covariates are balanced at the threshold. These covariates include other-subject test scores,[37] race, socioeconomic disadvantaged status, limited English-proficient status, disability status, and whether the student is repeating the grade. While observable characteristics are controlled for, discontinuities in observable characteristics may be suggestive of changes in unobservable characteristics around the cutoff. Figures A.3 and A.4 graph mean covariates by normalized student test scores for the mathematics and English tests, respectively. In addition, the discontinuity in the covariate at the threshold is estimated and reported above each figure. We find that the discontinuity is only statistically significant at the five percent level for one covariate, in line with expectations given that we check for covariate balance for eighteen covariates.[38]

RD Bandwidth and Control Function: We examine the robustness of our RD results by altering the chosen bandwidth and varying the function we use for the distance of a student's initial test score to the retest threshold. Figure A.5 plots our RD estimates using various bandwidths from 2 to 10 test score units (a bandwidth of 5 is chosen for our main specifications). For both mathematics and English, magnitudes are consistent across the selected bandwidths, although they decline somewhat as larger bandwidths are chosen. Table A.1 then demonstrates the robustness of our RD results to various functional forms to control for the distance of a student's initial test score to the retest threshold. To start, columns (1) and (4) show results using a linear functional form, which is what we use for our main specifications. Columns (2) and (5) then report estimates using a quadratic functional form, while columns (3) and (6) do so for a triangular kernel. The coefficient estimates are very

[37] Since own-subject test scores are the running variable, they are perfectly balanced around the threshold by definition.

[38] The probability that at least one of the covariates will be significant at the 5% level in a placebo test with 18 variables is 60.3%.

similar regardless of the chosen functional form.

5 Conclusion

This paper investigates the short- and long-term impacts of an approximately two weeks-long test focused intervention which we refer to as ‘cramming.’ The intervention we study embeds an extreme ‘teach to the test’ incentive for educators to raise contemporaneous test scores, providing a unique case study to investigate the initial impact and persistence of interventions focused on short-term academic performance. Our estimates indicate that cramming substantially raises students’ contemporaneous test scores, but much of this increase fades out the following year. We also find no evidence that cramming improved long-term outcomes as measured by PSAT scores, SAT scores, and high school graduation.

Our findings provide some support for both the proponents and opponents of test-based incentive schemes. On one hand, we do find that test-focused instruction substantially raises contemporaneous test scores and some of these gains persist over time. On the other hand, we find that the test score gain induced by test-focused instruction fade-out at faster rates relative to other educational interventions and do not reappear later in life.

More generally, our work speaks to the importance of taking fade-out into account when comparing the impacts of various educational interventions. In particular, fade-out is likely to be intervention-specific and so researchers cannot simply use contemporaneous test score gains to compare policies. These concerns are especially salient when comparing interventions targeting foundational knowledge to those that focus on raising short-term test scores. However, it is important to note that, given the short-term nature of the intervention, it appears to likely be highly cost-efficient relative to other policies that produce comparable gains in contemporaneous test scores.

References

Ahn, Thomas and Jacob Vigdor (2014), "The impact of No Child Left Behind's accountability sanctions on school performance: Regression discontinuity evidence from North Carolina." Working Paper 20511, National Bureau of Economic Research, URL `http://www.nber.org/papers/w20511`.

Aucejo, Esteban, Teresa Romano, and Eric S. Taylor (forthcoming), "Does evaluation change teacher effort and performance? Quasi-experimental evidence from a policy of retesting students." *Review of Economics and Statistics*.

Bailey, Drew, Greg J. Duncan, Candice L. Odgers, and Winnie Yu (2017), "Persistence and fadeout in the impacts of child and adolescent interventions." *Journal of Research on Educational Effectiveness*, 10, 7–39.

Bailey, Drew H., Greg J. Duncan, Flávio Cunha, Barbara R. Foorman, and David S. Yeager (2020a), "Persistence and fade-out of educational-intervention effects: Mechanisms and potential solutions." *Psychological Science in the Public Interest*, 21, 55–97.

Bailey, Drew H., Lynn S. Fuchs, Jennifer K. Gilbert, David C. Geary, and Douglas Fuchs (2020b), "Prevention: Necessary but insufficient? A 2-year follow-up of an effective first-grade mathematics intervention." *Child Development*, 91, 382–400.

Baker, Michael (2013), "Industrial actions in schools: Strikes and student achievement." *Canadian Journal of Economics/Revue canadienne d'économique*, 46, 1014–1036.

Banerjee, Abhijit V., Shawn Cole, Esther Duflo, and Leigh Linden (2007), "Remedying education: Evidence from two randomized experiments in India." *Quarterly Journal of Economics*, 122, 1235–1264.

Battistin, Erich and Elena Claudia Meroni (2016), "Should we increase instruction time in low achieving schools? Evidence from southern Italy." *Economics of Education Review*, 55, 39–56.

Bettinger, Eric P. and Bridget Terry Long (2009), "Addressing the needs of underprepared students in higher education: Does college remediation work?" *Journal of Human Resources*, 44, 736–771.

Black, Alison Rebeck, Marie-Andree Somers, Fred Doolittle, Rebecca Unterman, and Jean Baldwin Grossman (2009), "The evaluation of enhanced academic instruction in after-school programs: Final report. NCEE 2009-4077." *National Center for Education Evaluation and Regional Assistance.*

Calcagno, Juan Carlos and Bridget Terry Long (2008), "The impact of postsecondary remediation using a regression discontinuity approach: Addressing endogenous sorting and noncompliance." Working Paper 14194, National Bureau of Economic Research, URL `http://www.nber.org/papers/w14194`.

Calonico, Sebastian, Matias D. Cattaneo, and Rocio Titiunik (2014), "Robust nonparametric confidence intervals for regression-discontinuity designs." *Econometrica*, 82, 2295–2326.

Chetty, Raj, John N. Friedman, Nathaniel Hilger, Emmanuel Saez, Diane Whitmore Schanzenbach, and Danny Yagan (2011), "How does your kindergarten classroom affect your earnings? Evidence from Project STAR." *Quarterly Journal of Economics*, 126, 1593–1660.

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

Clements, Douglas H., Julie Sarama, Christopher B. Wolfe, and Mary Elaine Spitler (2013), "Longitudinal evaluation of a scale-up model for teaching mathematics with trajectories and technologies: Persistence of effects in the third year." *American Educational Research Journal*, 50, 812–850.

Cortes, Kalena E., Joshua S. Goodman, and Takako Nomi (2015), "Intensive math instruction and educational attainment long-run impacts of double-dose algebra." *Journal of Human Resources*, 50, 108–158.

De Paola, Maria and Vincenzo Scoppa (2014), "The effectiveness of remedial courses in Italy: A fuzzy regression discontinuity design." *Journal of Population Economics*, 27, 365–386.

Dell, Melissa (2010), "The persistent effects of Peru's mining mita." *Econometrica*, 78, 1863–1903.

Deming, David (2009), "Early childhood intervention and life-cycle skill development: Evidence from Head Start." *American Economic Journal: Applied Economics*, 1, 111–34.

Ding, Weili and Steven F. Lehrer (2011), "Experimental estimates of the impacts of class size on test scores: Robustness and heterogeneity." *Education Economics*, 19, 229–252.

Dougherty, Shaun M (2015), "Bridging the discontinuity in adolescent literacy? Mixed evidence from a middle grades intervention." *Education Finance and Policy*, 10, 157–192.

Dynarski, Mark, Susanne James-Burdumy, Mary Moore, Linda Rosenberg, John Deke, and Wendy Mansfield (2004), "When schools stay open late: The national evaluation of the 21st Century Community Learning Centers Program – new findings." *US Department of Education*.

Frandsen, Brigham R. (2017), "Party bias in union representation elections: Testing for manipulation in the regression discontinuity design when the running variable is discrete." In *Regression discontinuity designs*, Emerald Publishing Limited.

Fruehwirth, Jane Cooley, Salvador Navarro, and Yuya Takahashi (2016), "How the timing of grade retention affects outcomes: Identification and estimation of time-varying treatment effects." *Journal of Labor Economics*, 34, 979–1021.

Gilraine, Michael, Jiaying Gu, and Robert McMillan (2020), "A new method for estimating teacher value-added." Working Paper 27094, National Bureau of Economic Research, URL `http://www.nber.org/papers/w27094`.

Goldhaber, Dan and Richard Startz (2017), "On the distribution of worker productivity: The case of teacher effectiveness and student achievement." *Statistics and Public Policy*, 4, 1–12.

Goodman, Joshua (2014), "Flaking out: Student absences and snow days as disruptions of instructional time." Working Paper 20221, National Bureau of Economic Research, URL `http://www.nber.org/papers/w20221`.

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

Jacob, Brian A. and Lars Lefgren (2004), "Remedial education and student achievement: A regression-discontinuity analysis." *Review of Economics and Statistics*, 86, 226–244.

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

Jensen, Arthur R. (1999), "The g factor: The science of mental ability." *Psicothema*, 11, 445–446.

Kolesár, Michal and Christoph Rothe (2018), "Inference in regression discontinuity designs with a discrete running variable." *American Economic Review*, 108, 2277–2304.

Krueger, Alan B. and Diane M. Whitmore (2001), "The effect of attending a small class in the early grades on college-test taking and middle school test results: Evidence from Project STAR." *Economic Journal*, 111, 1–28.

Lagerlöf, Johan NM and Andrew J. Seltzer (2009), "The effects of remedial mathematics on the learning of economics: Evidence from a natural experiment." *Journal of Economic Education*, 40, 115–137.

Lavy, Victor and Analia Schlosser (2005), "Targeted remedial education for underperforming teenagers: Costs and benefits." *Journal of Labor Economics*, 23, 839–874.

Lee, David S. and David Card (2008), “Regression discontinuity inference with specification error.” *Journal of Econometrics*, 142, 655–674.

Lee, David S. and Thomas Lemieux (2010), “Regression discontinuity designs in economics.” *Journal of Economic Literature*, 48, 281–355.

Macartney, Hugh, Robert McMillan, and Uros Petronijevic (2018), “Teacher value-added and economic agency.” Working Paper 24747, National Bureau of Economic Research, URL `http://www.nber.org/papers/w24747`.

Manacorda, Marco (2012), “The cost of grade retention.” *Review of Economics and Statistics*, 94, 596–606.

Martorell, Paco and Isaac McFarlin Jr (2011), “Help or hindrance? The effects of college remediation on academic and labor market outcomes.” *Review of Economics and Statistics*, 93, 436–454.

Matsudaira, Jordan D. (2008), “Mandatory summer school and student achievement.” *Journal of Econometrics*, 142, 829–850.

McCrary, Justin (2008), “Manipulation of the running variable in the regression discontinuity design: A density test.” *Journal of Econometrics*, 142, 698–714.

Mehta, Jal and Sarah Fine (2019), *In search of deeper learning: The quest to remake the American high school*. Harvard University Press.

Mueller, Steffen (2013), “Teacher experience and the class size effect—experimental evidence.” *Journal of Public Economics*, 98, 44–52.

North Carolina Education Research Data Center (NCERDC) (2008-2019), “Student, class and personnel files.” URL `http://childandfamilypolicy.duke.edu/research/hc-education-data-center/`.

Oreopoulos, Philip, Robert S. Brown, and Adam M. Lavecchia (2017), “Pathways to education: An integrated approach to helping at-risk high school students." *Journal of Political Economy*, 125, 947–984.

Penney, Jeffrey (2013), "Hypothesis testing for arbitrary bounds." *Economics Letters*, 121, 492–494.

Petek, Nathan and Nolan Pope (2021), "The multidimensional impact of teachers on students.", URL http://www.econweb.umd.edu/~pope/Nolan_Pope_JMP.pdf. Unpublished.

Pischke, Jörn-Steffen (2007), "The impact of length of the school year on student performance and earnings: Evidence from the German short school years." *Economic Journal*, 117, 1216–1242.

Roderick, Melissa and Jenny Nagaoka (2005), "Retention under Chicago's high-stakes testing program: Helpful, harmful, or harmless?" *Educational Evaluation and Policy Analysis*, 27, 309–340.

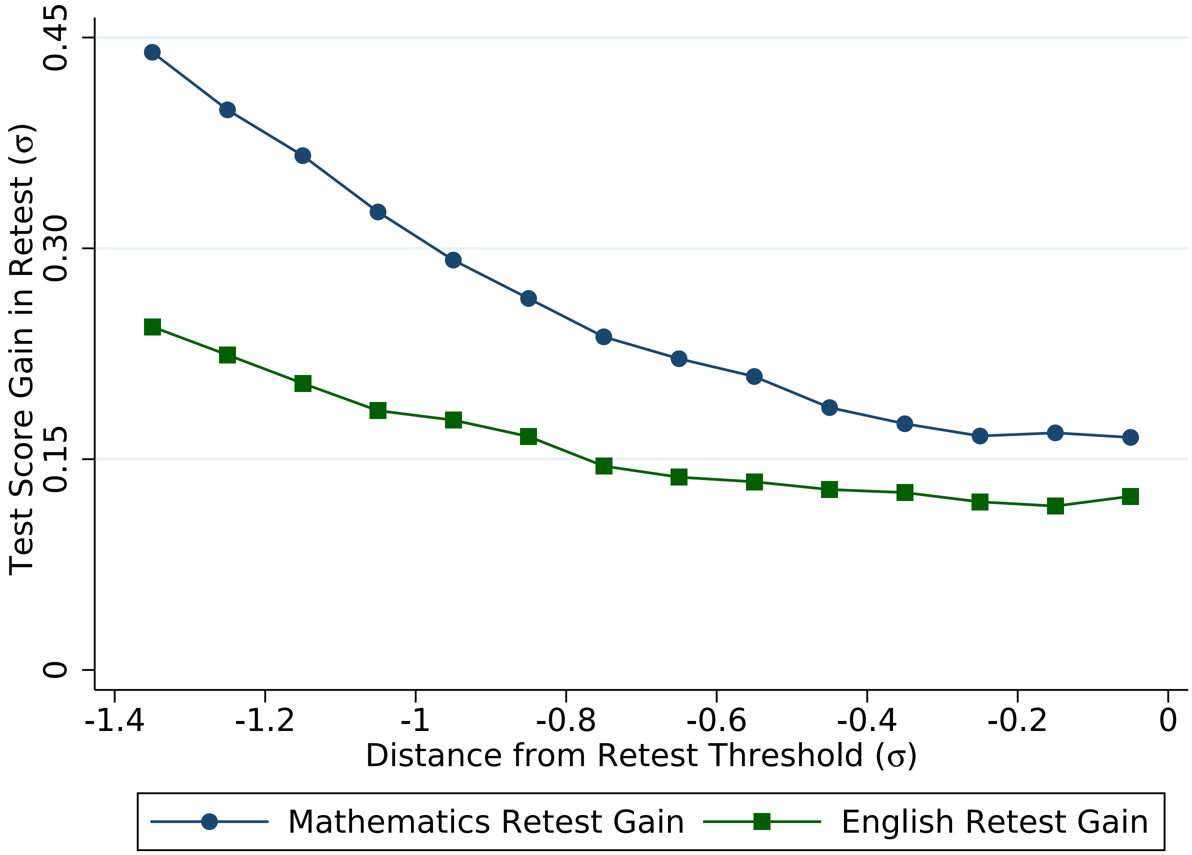
Rodriguez-Planas, Nuria (2012), "Longer-term impacts of mentoring, educational services, and learning incentives: Evidence from a randomized trial in the United States." *American Economic Journal: Applied Economics*, 4, 121–139.

Smith, Thomas M., Paul Cobb, Dale C. Farran, David S. Cordray, and Charles Munter (2013), "Evaluating math recovery: Assessing the causal impact of a diagnostic tutoring program on student achievement." *American Educational Research Journal*, 50, 397–428.

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

Wagner, Lindsay (2013), "Budget cuts impact franklin county middle school." *NC Policy Watch*, URL http://www.ncpolicywatch.com/2013/03/18/budget-cuts-impact-franklin-county-middle-school/.

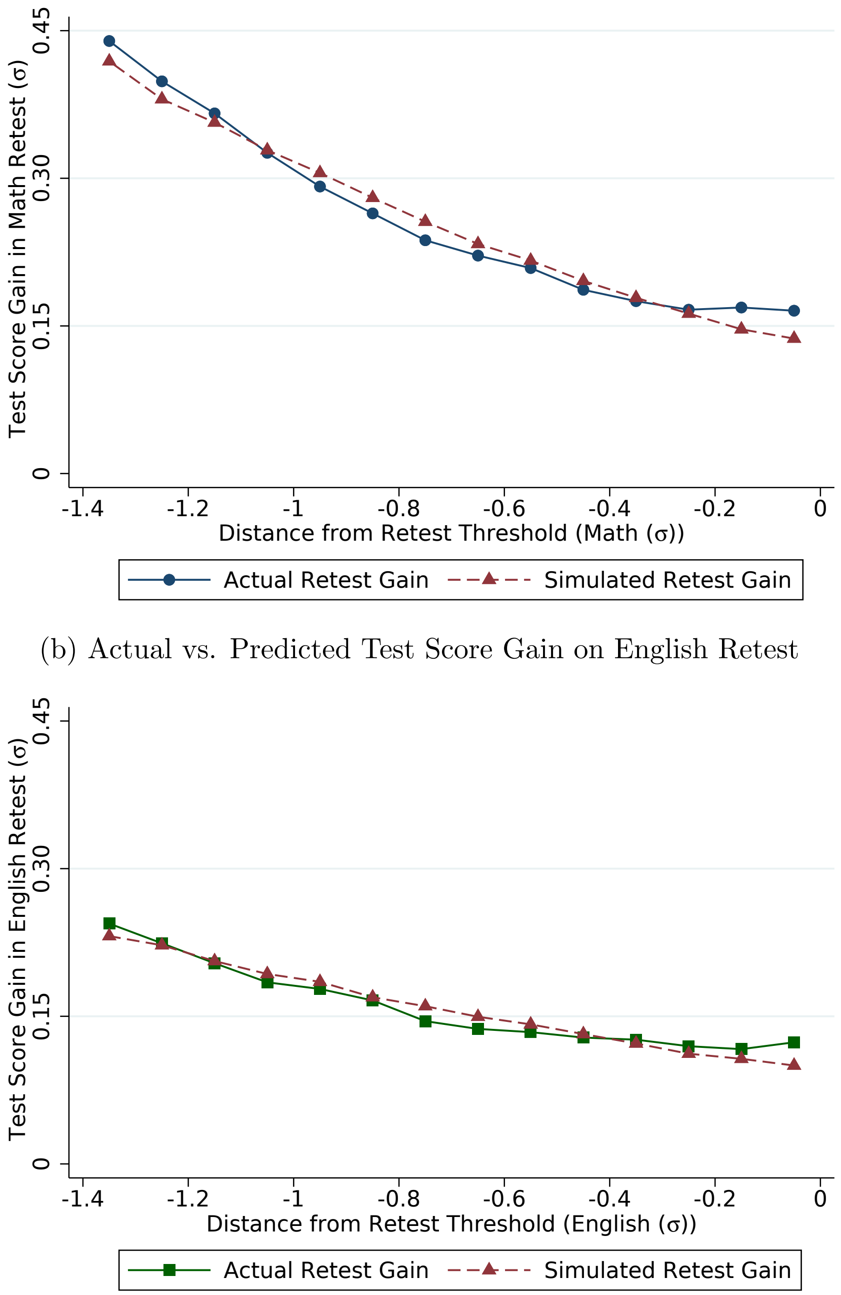
Figure 1: Test Score Gain on Retest



Notes: Figure shows the average test score increase retested students achieve on the retest in comparison to the initial test by distance to the retest threshold. Test scores for both the initial test and the retest are standardized by grade-year based on the mean and standard deviation of the initial test and are discretized into 0.1σ bins. Test score gains on the y-axis thereby indicate the difference between a student's initial test score and their retest score in terms of standard deviations of their grade-year distribution on the initial test. The units on the x-axis represent the distance of a student's initial test score is from the retest threshold in terms of standard deviation units of their grade-year test score distribution. The figure relies on 424,618 and 560,188 observations for mathematics and English, respectively.

Figure 2: Retest Score Gain (Actual vs. Predicted)

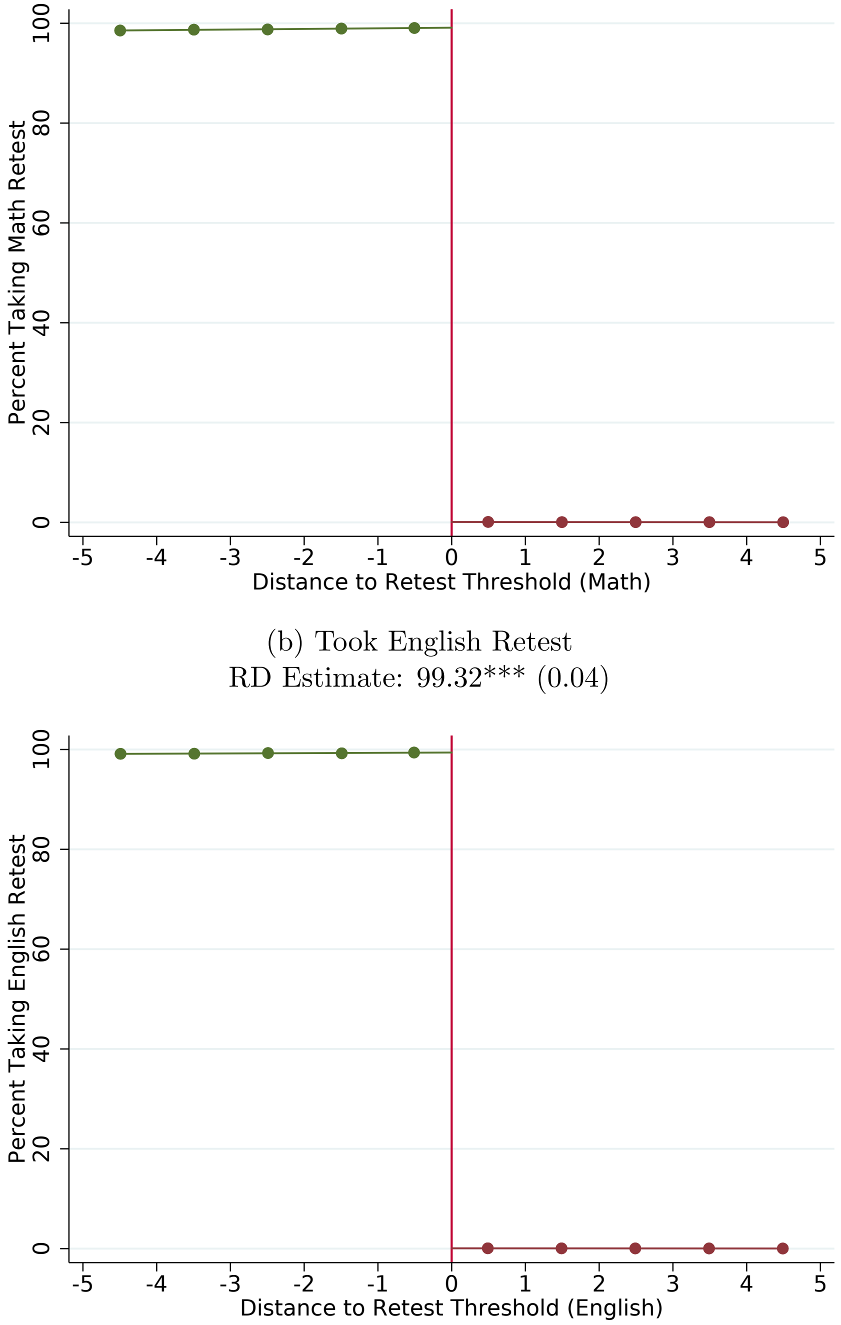
(a) Actual vs. Predicted Test Score Gain on Mathematics Retest



(b) Actual vs. Predicted Test Score Gain on English Retest

Notes: Figure shows the average test score increase retested students achieve on the retest in comparison to the initial test by distance to the retest threshold in both the raw and simulated data. The raw test gains indicated by the solid line are identical to those in Figure 1. We then superimpose our simulated test score gains using a dashed line. The figures therefore indicate the ability of our structural model to fit the underlying data. Simulated test score gains are based on a simulation of 2.5 million students. Test scores for both the initial test and the retest are standardized by grade-year based on the mean and standard deviation of the initial test. Both the actual and simulated test scores are discretized into 0.1σ bins to map our continuous simulated initial test scores to the observed discrete test scores.

Figure 3: First Stage (Student Took Retest)

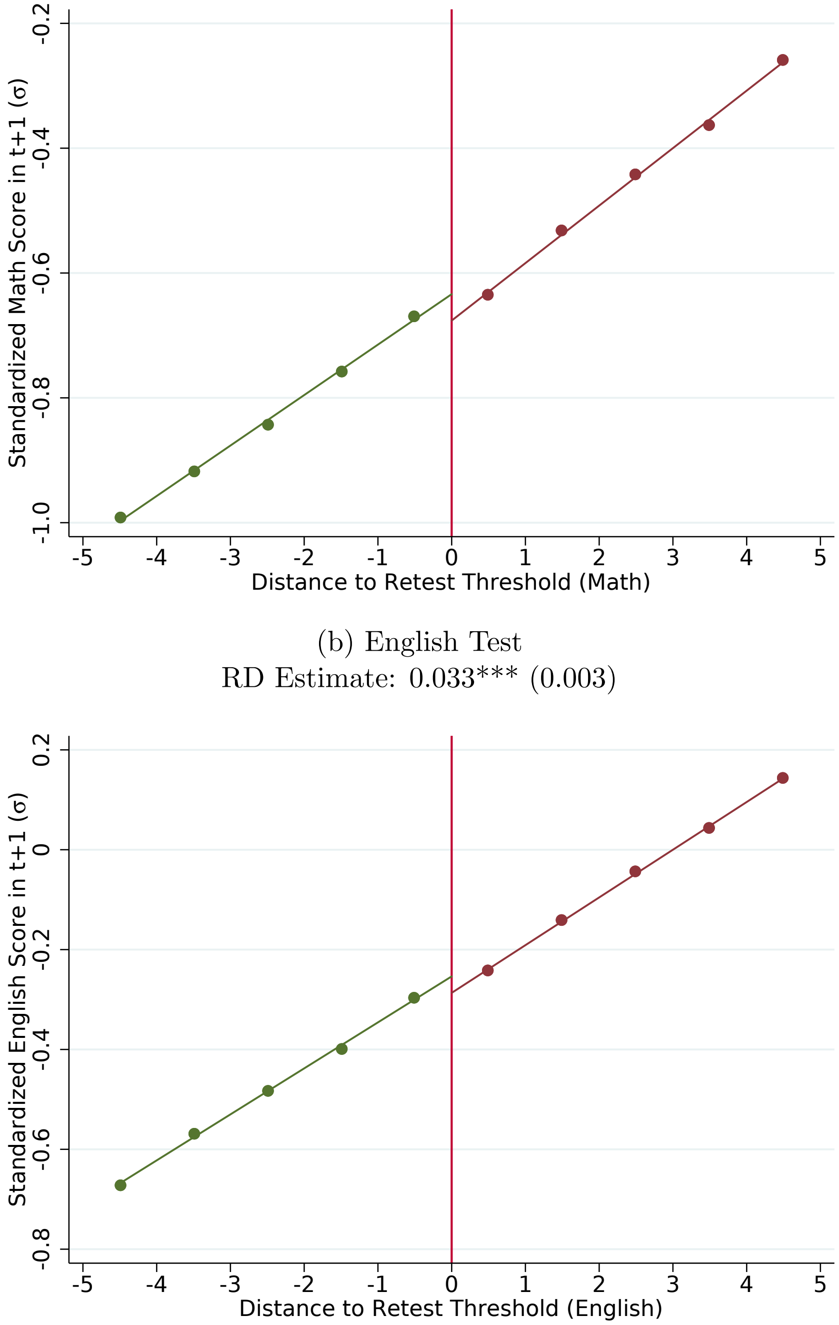


(a) Took Math Retest
RD Estimate: 99.02*** (0.05)

(b) Took English Retest
RD Estimate: 99.32*** (0.04)

Notes: Figures display the percent of students taking a retest by their test score distance to the retest threshold. Figures 3(a) and 3(b) are based on 628,302 and 751,411 observations, respectively. The vertical line denotes the retest threshold. RD estimates from a local linear regression allowing for different functions on either side of the threshold are reported above the figures. Standard errors are two-way clustered by student and school. ***,** and * denote significance at the 1%, 5% and 10% levels, respectively.

Figure 4: Reduced Form (Standardized Test Scores in $t+1$)



(a) Math Test
RD Estimate: 0.042*** (0.003)

(b) English Test
RD Estimate: 0.033*** (0.003)

Notes: Figures display students' standardized test scores in their first sitting of the end-of-grade test in the *following* year by their test score distance to the retest threshold. Test scores are standardized at the grade-year level to be mean zero and have a variance of one. Figures 4(a) and 4(b) are based on 628,302 and 751,411 observations, respectively. The vertical line denotes the retest threshold. RD estimates from a local linear regression allowing for different functions on either side of the threshold are reported above the figures and are the same as those in columns (1) and (3) of Panel A in Table 3. Standard errors are two-way clustered by student and school. ***,** and * denote significance at the 1%, 5% and 10% levels, respectively.

Table 1: Summary Statistics

	Full Sample[1]	Math RD Sample[2]	Math Retest Sample[3]	English RD Sample[2]	English Retest Sample[3]
	(1)	(2)	(3)	(4)	(5)
Mean of Student Characteristics					
Math Score (σ)	0.00	-0.63	-1.23	-0.22	-0.66
Reading Score (σ)	0.00	-0.50	-0.92	-0.25	-0.87
% White	54.0	42.2	33.6	47.9	36.5
% Black	25.9	35.6	43.8	30.6	39.0
% Hispanic	12.2	14.8	15.5	13.9	17.0
% Asian	2.5	1.3	1.2	1.9	1.7
% Disadvantaged	52.0	66.7	74.9	60.1	71.5
% English Learners	7.1	10.0	12.2	7.5	12.6
% Gifted	14.4	1.2	0.3	4.0	1.0
% with Disability	12.0	14.6	21.0	10.9	16.3
% Repeating Grade	1.1	1.8	2.7	1.4	2.6
# of Students	884,488	379,164	261,515	443,617	339,012
Observations[6] (student-year)	2,076,143	628,302	424,618	751,441	560,188

Notes:

[1] Data coverage: grades 3-7 from 2008-09 through 2011-12. Sample is restricted to students with valid current and subsequent test scores in at least one subject.

[2] Sample is restricted to students within five scale points of that subject's retest threshold.

[3] Sample is restricted to students who scored below the retest threshold, within 1.45 standard deviations of the retest threshold, and retook the retest.

Table 2: Model Estimates

Outcome Variable:	Mathematics Test Score (1)	English Test Score (2)
Parameters of Interest (Knowledge Gain)		
Average Retest Knowledge Gain	0.140***	0.083***
($\bar{z}$)	(0.002)	(0.003)
Variance-Covariance Parameters		
θ_0	0.154**	0.257***
	(0.009)	(0.008)
θ_1	0.160***	0.041***
	(0.009)	(0.009)
Observations	261,515	339,012

Notes: This table shows estimates of the model parameters that are estimated using the minimum-distance procedure described in Section 3.2. The parameter of interest is $\bar{z}$ which gives the average retest knowledge gain among students taking the retest. The two parameters θ_0 and θ_1 then govern the variance of error distribution and its correlation with student ability to match the 'U-shaped' psychometric estimated error distribution. In particular, the error term, ϵ_{igt}, is assumed to be distributed $\epsilon_{igt} \sim \mathcal{N}(0, \sigma_\epsilon^2)$, where $\sigma_\epsilon \equiv \theta_0 + \theta_1 |\alpha_{igt}|$. Standard errors are calculated via bootstrap. ***,** and * denote significance at the 1%, 5% and 10% levels, respectively.

Table 3: Regression-Discontinuity Estimates on Future Same-Subject Scores

	Mathematics Retest		*English Retest*	
	No Covariates	Covariates	No Covariates	Covariates
	(1)	(2)	(3)	(4)
Panel A. Retest Effect on Same Subject Score in $t+1$				
RD Estimate	0.042***	0.041***	0.033***	0.032***
	(0.003)	(0.003)	(0.003)	(0.003)
Observations	628,302	628,302	751,441	751,441
Panel B. Retest Effect on Same Subject Score in $t+2$				
RD Estimate	0.031***	0.025***	0.014***	0.013***
	(0.004)	(0.003)	(0.003)	(0.003)
Observations	476,871	476,871	558,872	558,872
Panel C. Retest Effect on Same Subject Score in $t+3$				
RD Estimate	0.013***	0.015***	0.005	0.013***
	(0.004)	(0.004)	(0.004)	(0.004)
Observations	341,963	341,963	403,212	403,212

Notes: Table reports regression results from the RD regression defined by equation (3.6). The bandwidth used is five. Covariates include math and English scores on the initial test interacted with grade dummies, gender, ethnicity, English learner status, economically disadvantaged status, disability status, English and mathematics gifted status, whether the student is repeating the current grade, and grade-by-year fixed effects. When initial test scores in the other subject are missing, we set the other subject score to zero and include an indicator for missing data in the other subject interacted with initial own-subject test scores. Standard errors are two-way clustered by student and school. ***,** and * denote significance at the 1%, 5% and 10% levels, respectively.

Table 4: Regression-Discontinuity Estimates on Future Other-Subject Scores

	Mathematics Retest		*English Retest*	
	No Covariates	Covariates	No Covariates	Covariates
	(1)	(2)	(3)	(4)
Panel A. Retest Effect on Other-Subject Score in $t+1$				
RD Estimate	0.017***	0.013***	0.014***	0.012***
	(0.004)	(0.003)	(0.003)	(0.002)
Observations	621,530	621,530	749,842	749,842
Panel B. Retest Effect on Other-Subject Score in $t+2$				
RD Estimate	0.017**	0.011***	0.005	0.005*
	(0.005)	(0.004)	(0.004)	(0.003)
Observations	472,495	472,495	557,988	557,988
Panel C. Retest Effect on Other-Subject Score in $t+3$				
RD Estimate	0.002	0.005	-0.001	0.004
	(0.005)	(0.004)	(0.005)	(0.004)
Observations	339,801	339,801	402,814	402,814

Notes: Table reports regression results from the RD regression defined by equation (3.6) where the score for the other subject is used as the outcome (i.e., look at the impact of the mathematics retest on English scores). The bandwidth used is five. Covariates include math and English scores on the initial test interacted with grade dummies, gender, ethnicity, English learner status, economically disadvantaged status, disability status, English and mathematics gifted status, whether the student is repeating the current grade, and grade-by-year fixed effects. When initial test scores in the other subject are missing, we set the other subject score to zero and include an indicator for missing data in the other subject interacted with initial own-subject test scores. Standard errors are two-way clustered by student and school. ***,** and * denote significance at the 1%, 5% and 10% levels, respectively.

Table 5: Regression-Discontinuity Estimates on Long-Run Outcomes

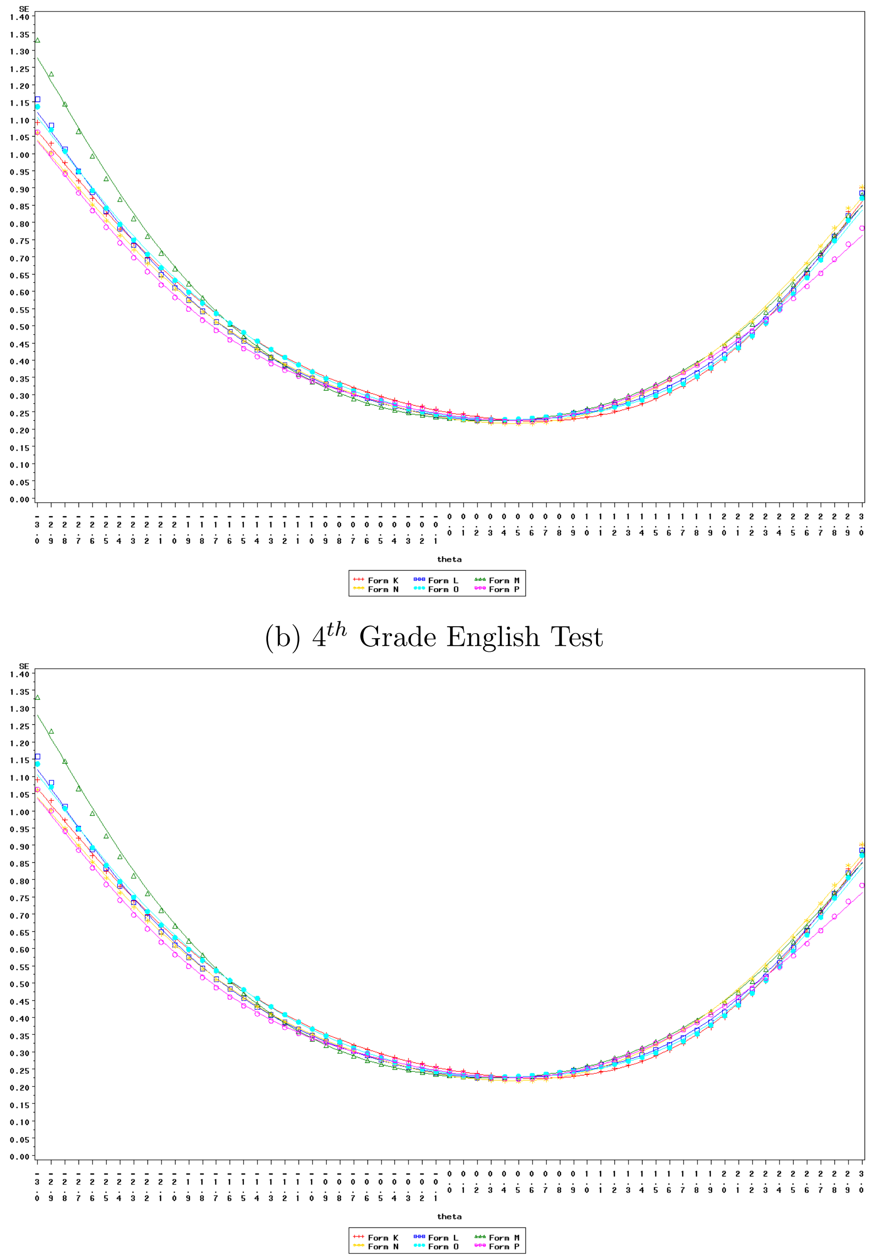
	% Taking PSAT	PSAT Score	% Taking SAT	SAT Score	% Graduating High School
	(1)	(2)	(3)	(4)	(5)
Panel A. Math Retest					
RD Estimate	0.08 (0.24)	3.78** (1.38)	-0.18 (0.22)	0.61 (1.17)	-0.07 (0.27)
Observations	566,601	187,724	541,303	123,729	448,782
Panel B. ELA Retest					
RD Estimate	0.42* (0.22)	1.99* (1.18)	0.04 (0.22)	1.37 (0.92)	-0.14 (0.23)
Observations	680,327	265,859	651,206	191,753	539,650
Panel C. Combined Math and ELA Retests (i.e., 'Stacked')					
RD Estimate	0.22 (0.16)	2.67*** (0.87)	-0.07 (0.15)	1.11 (0.70)	-0.14 (0.18)
Observations	1,246,928	453,583	1,192,509	315,482	988,432

Notes: Table reports regression results from the RD regression defined by equation (3.6) where the test score outcome is replaced by various long-run outcomes. Panel A reports results when the mathematics retest RD is used, Panel B when the English retest is used, and Panel C stacks the data and so reports the impact of taking one additional retest on the long-run outcome. The number of observations in Panel C equals the sum of the observations in Panels A and B since the data is stacked. Since our PSAT data cover 2013-19, we restrict the sample to cohorts who appear in grade 10-11 during that time frame (drops the 2008-09 seventh grade cohort and 2012-13 third grade cohort). Similarly, our SAT and high school education data cover 2009-19. For the SAT, we restrict the sample to cohorts who have reached grade 12 by 2019 (drops the 2011-12 and 2012-13 third grade cohorts), while we restrict the sample to cohorts who have reached grade 12 by 2018 for high school graduation to allow for one year of grade repetition (drops the 2010-11 through 2012-13 third grade cohorts). Our data only records a student graduation if they receive a graduation certificate or diploma from the state of North Carolina. The bandwidth used is five. All regressions include covariates, which includes math and English scores on the initial test interacted with grade dummies, gender, ethnicity, English learner status, economically disadvantaged status, disability status, English and mathematics gifted status, whether the student is repeating the current grade, and grade-by-year fixed effects. When initial test scores in the other subject are missing, we set the other subject score to zero and include an indicator for missing data in the other subject interacted with initial own-subject test scores. Standard errors are two-way clustered by student and school. ***,** and * denote significance at the 1%, 5% and 10% levels, respectively.

A Appendix Figures and Tables

Figure A.1: Psychometric Estimated Error Distribution: North Carolina 4th Grade Tests

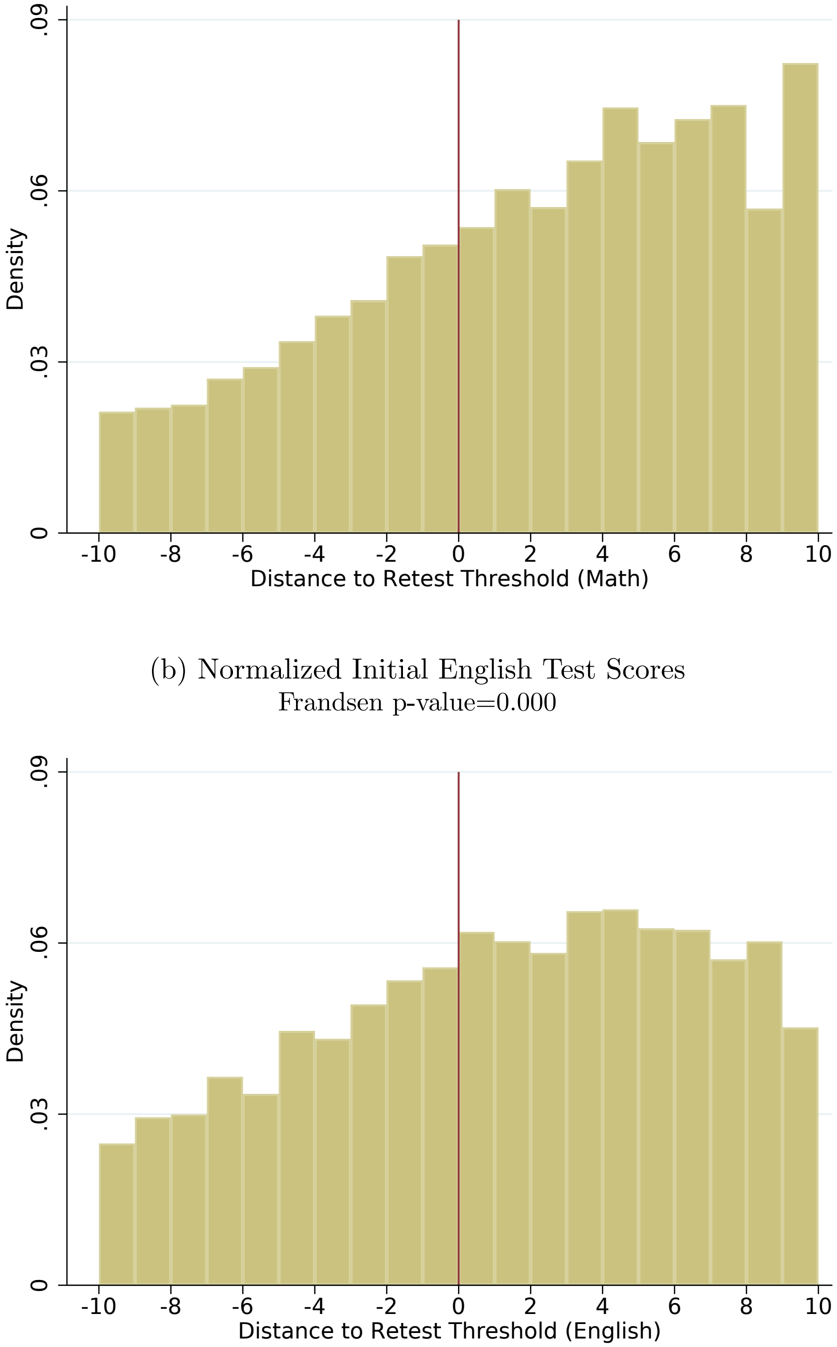
(a) 4th Grade Mathematics Test



(b) 4th Grade English Test

Notes: Figures display the estimated standard error of measurement of North Carolina's 4th Grade tests on the y-axis against the (estimated) true test-taking ability of the student on the x-axis. Figure A.1(a) does so for the mathematics test, with Figure A.1(a) doing so for the English test. Estimates are reported separately for each of the "forms" components of the test. Estimates are created by the psychometricians who designed the test and rely on test-retest and parallel forms reliability. Figure A.1(a) is available at `https://web.archive.org/web/20110407021418/ncpublicschools.org/docs/accountability/reports/mathtechmanualdrafted2.pdf` and Figure A.1(b) is available at `https://web.archive.org/web/20150403103551/http://www.ncpublicschools.org/docs/accountability/testing/reports/eogreadingtechman3.pdf`.

Figure A.2: Distribution of Normalized Test Scores

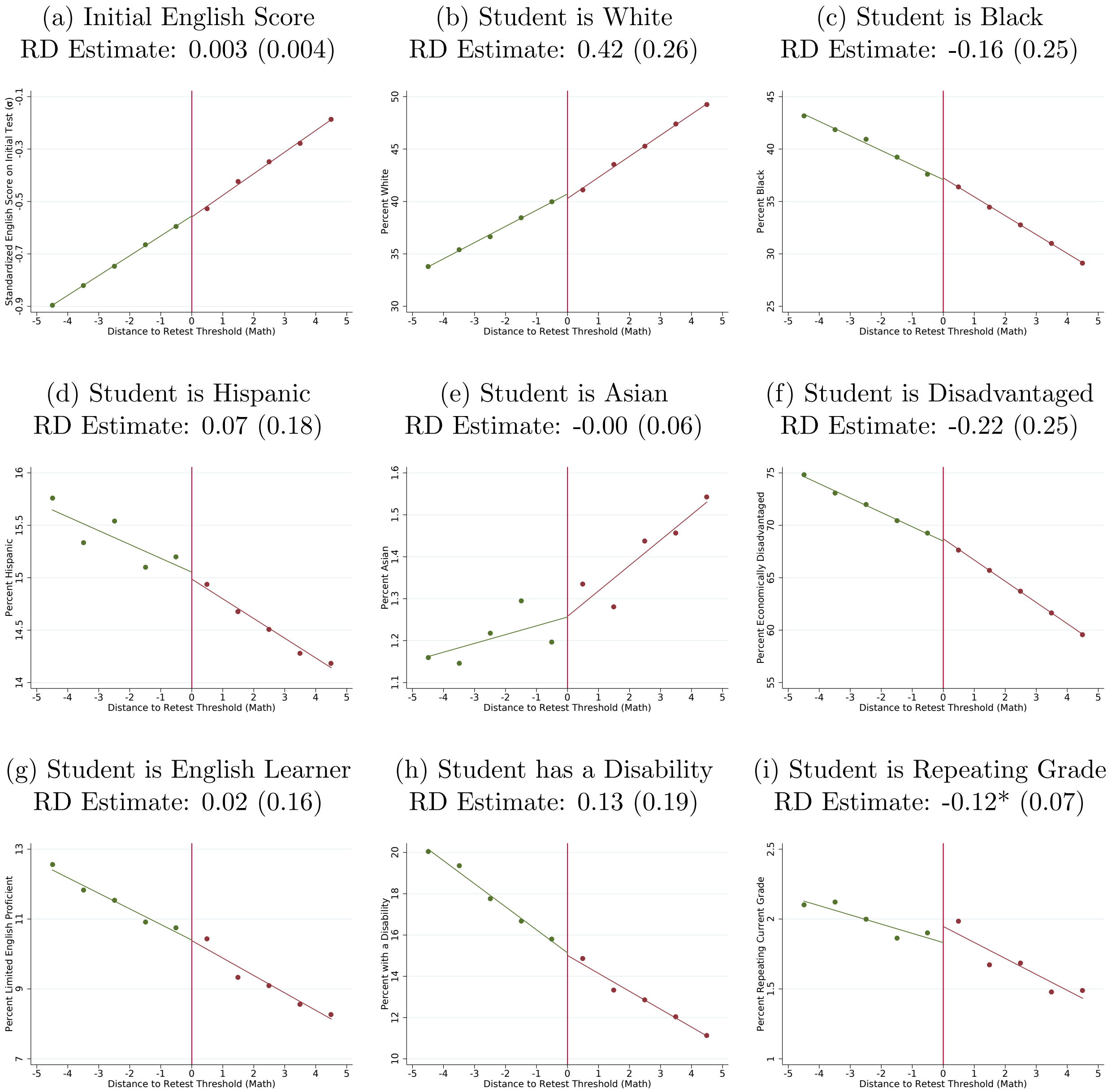


(a) Normalized Initial Mathematics Test Scores
Frandsen p-value=0.526

(b) Normalized Initial English Test Scores
Frandsen p-value=0.000

Notes: Figures show the distribution of normalized test scores by distance to the retest threshold for mathematics and English. The vertical line indicates the retest threshold. Figure A.2(a) and A.2(b) are based on 1,202,172 and 1,346,039 observations, respectively. The 'Frandsen p-value' then reports the p-value of Frandsen (2017)'s manipulation test where we choose $k = 0.032$, corresponding to eight support points within one standard deviation of the threshold.

Figure A.3: Covariate Balance (Math Retest Threshold)



(a) Initial English Score
RD Estimate: 0.003 (0.004)

(b) Student is White
RD Estimate: 0.42 (0.26)

(c) Student is Black
RD Estimate: -0.16 (0.25)

(d) Student is Hispanic
RD Estimate: 0.07 (0.18)

(e) Student is Asian
RD Estimate: -0.00 (0.06)

(f) Student is Disadvantaged
RD Estimate: -0.22 (0.25)

(g) Student is English Learner
RD Estimate: 0.02 (0.16)

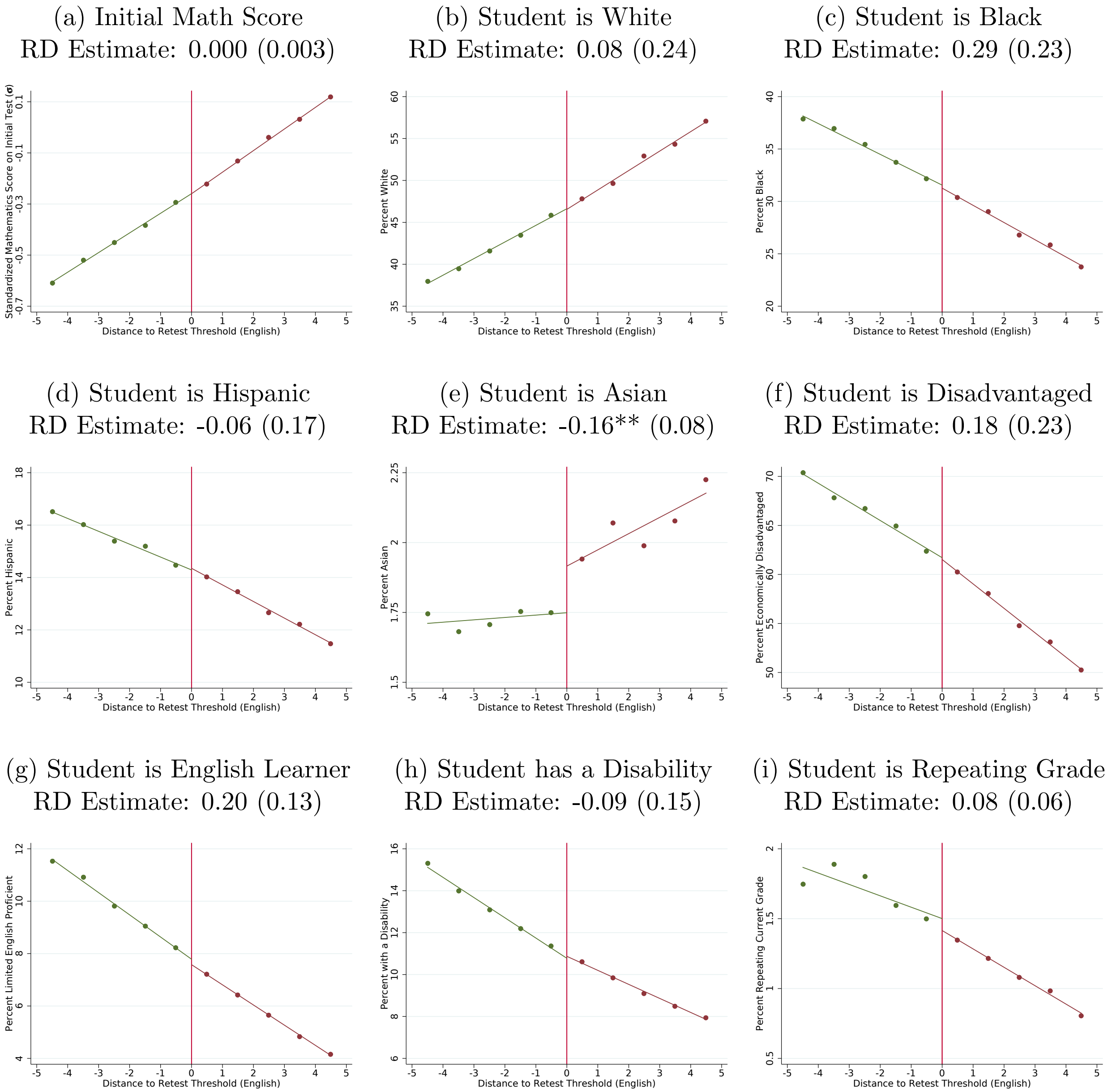
(h) Student has a Disability
RD Estimate: 0.13 (0.19)

(i) Student is Repeating Grade
RD Estimate: -0.12* (0.07)

Standardized English Score on Initial Test (σ)
-0.9
-0.7
-0.5
-0.3
-0.1
Percent White
30
35
40
45
50
Percent Black
25
30
35
40
45
Percent Hispanic
14
14.5
15
15.5
16
Percent Asian
1.1
1.2
1.3
1.4
1.5
1.6
Percent Economically Disadvantaged
55
60
65
70
75
Percent Limited English Proficient
7
9
11
13
Percent with a Disability
10
12
14
16
18
20
Percent Repeating Current Grade
1
1.5
2
2.5
-5
-4
-3
-2
-1
0
1
2
3
4
5
Distance to Retest Threshold (Math)

Notes: All figures are based on 628,302 observations with the exception of Figure A.3(a) which is based on 622,365 observations. Each RD estimate comes from a separate local linear regression allowing for different functions on either side of the threshold. The bandwidth used is five. Standard errors are two-way clustered by student and school. ***,** and * denote significance at the 1%, 5% and 10% levels, respectively.

Figure A.4: Covariates (Read Retest Threshold)


(a) Initial Math Score
RD Estimate: 0.000 (0.003)

(b) Student is White
RD Estimate: 0.08 (0.24)

(c) Student is Black
RD Estimate: 0.29 (0.23)

(d) Student is Hispanic
RD Estimate: -0.06 (0.17)

(e) Student is Asian
RD Estimate: -0.16** (0.08)

(f) Student is Disadvantaged
RD Estimate: 0.18 (0.23)

(g) Student is English Learner
RD Estimate: 0.20 (0.13)

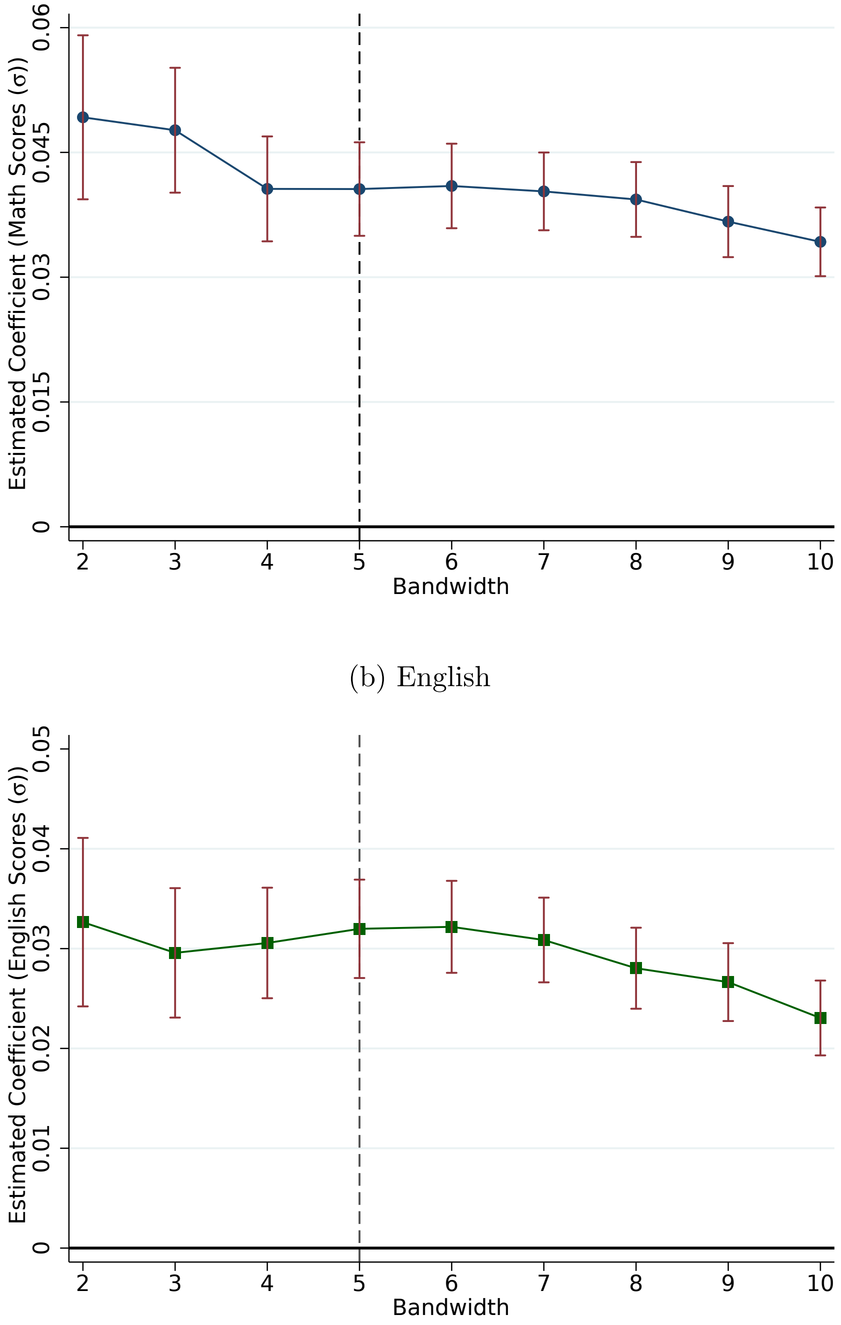
(h) Student has a Disability
RD Estimate: -0.09 (0.15)

(i) Student is Repeating Grade
RD Estimate: 0.08 (0.06)

Standardized Mathematics Score on Initial Test (σ)
-0.7
-0.5
-0.3
-0.1
0.1
Percent White
35
40
45
50
55
60
Percent Black
20
25
30
35
40
Percent Hispanic
10
12
14
16
18
Percent Asian
1.5
1.75
2
2.25
Percent Economically Disadvantaged
50
55
60
65
70
Percent Limited English Proficient
4
6
8
10
12
Percent with a Disability
6
8
10
12
14
16
Percent Repeating Current Grade
0.5
1
1.5
2
-5
-4
-3
-2
-1
0
1
2
3
4
5
Distance to Retest Threshold (English)

Notes: All figures are based on 751,411 observations with the exception of Figure A.3(a) which is based on 750,090 observations. Each RD estimate comes from a separate local linear regression allowing for different functions on either side of the threshold. The bandwidth used is five. Standard errors are two-way clustered by student and school. ***,** and * denote significance at the 1%, 5% and 10% levels, respectively.

Figure A.5: Bandwidth Robustness


(a) Mathematics

(b) English

Notes: Figures show robustness of our main RD estimates with respect to bandwidth. The horizontal line represents a effect size of zero while the vertical line indicates our chosen bandwidth of five. Covariates are included, and so point estimates at a bandwidth of five correspond to the estimates reported in columns (2) and (4) of Table 3 for mathematics and English, respectively. Whiskers indicate the 95 percent confidence of our point estimate under each bandwidth, with standard errors are clustered at the student and school level.

Table A.1: Regression-Discontinuity Robustness: Alternative Functional Forms

	Mathematics Retest			*English Retest*		
	Linear	Quadratic	Triangular	Linear	Quadratic	Triangular
			Kernel		Kernel	
	(1)	(2)	(3)	(4)	(5)	(6)
Retest Effect on Same Subject Score						
Panel A. Retest Effect on Same Subject Score in $t+1$						
RD Estimate	0.041*** (0.003)	0.049*** (0.005)	0.044*** (0.003)	0.032*** (0.003)	0.030*** (0.004)	0.031*** (0.003)
Panel B. Retest Effect on Same Subject Score in $t+2$						
RD Estimate	0.025*** (0.003)	0.030*** (0.006)	0.027*** (0.004)	0.013*** (0.003)	0.010** (0.005)	0.012*** (0.003)
Panel C. Retest Effect on Same Subject Score in $t+3$						
RD Estimate	0.015*** (0.004)	0.017*** (0.007)	0.017*** (0.004)	0.013*** (0.004)	0.010* (0.006)	0.012*** (0.004)
Retest Effect on Other-Subject Score						
Panel D. Retest Effect on Other-Subject Score in $t+1$						
RD Estimate	0.013*** (0.003)	0.011** (0.005)	0.013*** (0.003)	0.012*** (0.002)	0.006 (0.004)	0.010*** (0.003)
Panel E. Retest Effect on Other-Subject Score in $t+2$						
RD Estimate	0.011*** (0.004)	0.012** (0.006)	0.013*** (0.004)	0.005* (0.003)	0.005 (0.005)	0.006* (0.003)
Panel F. Retest Effect on Other-Subject Score in $t+3$						
RD Estimate	0.005 (0.004)	-0.005 (0.007)	0.003 (0.005)	0.004 (0.004)	0.006 (0.006)	0.008* (0.004)

Notes: Table reports regression results from the RD regression defined by equation (3.6) using different functional forms to control for the running variables. The bandwidth used is five. Covariates are included, and so point estimates for the linear functional form correspond to the estimates reported in columns (2) and (4) of Table 3 for same subject scores. For other-subject scores, the point estimates for the linear functional form correspond to those reported in columns (2) and (4) of Table 4. Covariates include math and English scores on the initial test interacted with grade dummies, gender, ethnicity, English learner status, economically disadvantaged status, disability status, English and mathematics gifted status, whether the student is repeating the current grade, and grade-by-year fixed effects. When initial test scores in the other subject are missing, we set the other subject score to zero and include an indicator for missing data in the other subject interacted with initial own-subject test scores. Standard errors are two-way clustered by student and school. ***,** and * denote significance at the 1%, 5% and 10% levels, respectively.

B Literature Exploring Focused Interventions

Paper	Data and Type of Remediation	Research Design	Findings
Jacob and Lefgren (2004)	Use data from Chicago Public Schools for the 1993-94 through 1998-99 school years to evaluate the effect of grade 3 and 6 students attending a six-week summer school program – with 10-20 percent of those students being retained a grade – on academic achievement.	Uses a regression discontinuity design using the fact that students must attain a predefined level in both math and reading to be promoted to the next grade.	Finds that summer school and grade retention increases student achievement by roughly 20% of a year's worth of learning.
Dynarski, James-Burdumy, Moore, Rosenberg, Deke, and Mansfield (2004)	Use data from 26 21st Century Elementary school centers which provide three hours of after-school activities (snack time, homework time, and academic and enrichment activities) five days a week. Average student attendance was 2 days a week.	Uses random assignment with a sample of 2,308 students and compares outcomes of students randomly assigned to the program to those who were not.	Find that 21st Century centers did not affect reading test scores or grades for elementary students
Lavy and Schlosser (2005)	Evaluate the effects of an after-school remedial education program that provided individualized instruction in small study groups of up to five students in grades 10-12 in Israel.	Use a difference-in-differences design that compares schools implementing the program earlier to those implementing later.	Find that the program raise the matriculation rate by 3.3 percentage points.
Banerjee, Cole, Duflo, and Linden (2007)	Report the results from a remedial education program catering to disadvantaged grade 3 or 4 students in Mumbai and Vadodara, India. The program takes children out of the regular classroom to work on basic skills with a young woman for two hours. They also investigate a computer-assisted learning program.	Implement a randomized experiment in the 2001-02 and 2002-03 school year and compare outcomes of students randomly assigned to the program to those who were not.	Find that the program increased average test scores by 0.14σ and 0.28σ in the first and second year of the program, respectively. Treatment effects were larger for weaker students.
Matsudaira (2008)	Uses data from a large Northeastern school district for the 2000-01 and 2001-02 school years to evaluate the effect of attending 20-30 days of summer school on student achievement.	Uses a regression discontinuity design using the fact that students who fail to attain math or reading proficiency on a standardized test must attend summer school.	Finds that summer school increases both math and reading achievement by about 0.12 standard deviations.

Black, Somers, Doolittle, Unterman, and Grossman (2009)	Use data from 27 enhanced after-school centers which provided, on average, approx. 45 minutes of formal academic instruction for grade 2-5 students in either math or reading during after-school programs (about 30 percent increase in instruction above what is received during the regular school day)	Uses random assignment with a sample of 2,049 students and compares outcomes of students randomly assigned to the enhanced program to those who were not.	Finds that one year of enhanced instruction increases math scores by approximately one month's worth of extra math learning. No effects were found on reading scores.
Taylor (2014)	Uses data from Miami-Dade County Public Schools to investigate the effect of taking two math classes – one remedial, one regular – instead of just one class for students in grades 6-8.	Use a regression discontinuity design taking advantage of an administrative rule that students are identified as candidates for the two math class schedule if their score on their prior state math test fell below a pre-set score.	Finds that the remedial math class increases math test scores by 0.16-0.18σ, though decay about one-half of that effect the following year and one-third two years later.
Cortes, Goodman, and Nomi (2015)	Study Chicago Public Schools double-dose algebra policy which doubled instructional time for affected grade 9 students to 90 minutes every day.	Use a regression discontinuity design using that students were required to take the double dose algebra course if they scored below the national median on the math portion of the Iowa Tests of Basic Skills.	Find that the double-dose algebra course reduced algebra failure rates by 25% and also raised credits earned, test scores, high school graduation, and college enrollment rates.
Dougherty (2015)	Study a Hampton County Public School district program whereby struggling students were assigned to a supplementary reading course designed to complement their regular English course.	Use a regression discontinuity design capitalizing on the fact that students are assigned to an additional literacy course in middle school if their grade 5 Iowa Test score falls below some threshold.	Finds no overall effect, but persistently negative effects of the additional reading class for black students. Effects for white, Latino, and Asian students were positive, but not statistically significant.
Battistin and Meroni (2016)	Use data from Southern Italy for the 2009-10 and 2010-11 school years to investigate the effects of the European Social Fund which provides addtional instruction time (averaging about 45 houre per year) to lower secondary schools outside of normal school hours .	Use a difference-in-differences strategy to compare participating to non-participating groups both within and across schools.	Find that the intervention increased math test scores by 0.296 standard deviations for schools in the bottom tertile. No statistically significant effect found for reading.

C Fade-out of Various Educational Interventions

Paper	Intervention	Test Score Measure	Contemporaneous Gain	Gain Persisting to t+1 (% faded out from t)	Gain Persisting to t+2 (% faded out from $t+1$)
Krueger and Whitmore (2001)	Smaller class sizes	Composite math and reading score, SAT and CTBS (percentiles)	4.5	1.2 (73%)	1.3 (∞)
Jacob and Lefgren (2004)	Summer school and grade rentention	Math score, ITBS	untested	0.155	0.066 (57%)
		Reading score, ITBS	untested	0.082	0.032 (61%)
Deming (2009)	Head Start enrollment	Cognitive test scores using PPVT, PIATMT, PIATRR	0.145	0.133 (8%)	0.055 (59%)
Jacob, Lefgren, and Sims (2010)	Teacher quality	North Carolina math scores	1 (baseline)	0.27 (73%)	0.16 (41%)
		North Carolina reading scores	1 (baseline)	0.20 (80%)	0.18 (10%)
Chetty, Friedman, Hilger, Saez, Schanzenbach, and Yagan (2011)	Kindergarten class quality	Stanford Achievement Test math and reading scores	6.27 percentile points	1.50 percentile points (76%)	1.40† percentile points (7%)
Clements, Sarama, Wolfe, and Spitler (2013)	Building Blocks early mathematics curriculum	Research-based Early Math Assessment (REMA)	0.414 Rasch scale points	0.184 Rasch scale points (56%)	0.099 Rasch scale points (46%)
Smith, Cobb, Farran, Cordray, and Munter (2013)	Grade 1 mathematics tutoring program	MR proximal math assessment	0.29	0.07 (76%)	n/a
Chetty, Friedman, and Rockoff (2014)	Teacher quality	End-of-year math and English scores	1 (baseline)	0.53 (47%)	0.36 (32%)
Taylor (2014)	Extra remedial math class	Florida math test score	0.166	0.139 (16%)	0.066 (53%)
Bailey, Fuchs, Gilbert, Geary, and Fuchs (2020b)	First-grade intervention supporting arithmetic	Mathematics tests	1.42	0.62 (56%)	-0.03 (∞)
Petek and Pope (2021)	Teacher quality	Los Angeles math scores	0.19†	0.06† (68%)	0.03† (50%)

	Los Angeles English scores	0.13†	0.05† (62%)	0.03† (40%)

Notes: The † symbol indicates that the authors do not report a point estimate, but rather report the result in a figure and so the estimate is approximated visually from the figure. Effects are in test score standard deviations unless otherwise noted. The percentage of the effects that persist after the treatment are in parentheses. The periods t+1 and t+2 refer to one year and two years after treatment unless otherwise noted. For Krueger and Whitmore (2001), the program gain is in the SAT score (Stanford Achievement Test) in third grade, while the gain in t+1 and t+2 are measured in terms of the CTBS score for fourth and fifth grade respectively. For Deming (2009), the 'gain' column refers to the score at ages 5 to 6, the t+1 score to ages 7 to 10, and t+2 score to ages 11 to 14. The teacher quality interventions are the impact of teacher assignment in year t, which is set so that a teacher with one unit higher quality raises year t test scores by one. The intervention effects in Taylor (2014) refer to the treatment students receive for the 1/2 cut score. For Bailey et al. (2020b) we report results for the 'facts correctly retrieved' outcome, although results are similar if we use the 'number sets' outcome.

D Proofs

Proof of Proposition 1

Proof. The distribution of α is irrelevant since it is fixed and independent of ϵ. What is left is to demonstrate is that those who take the retest will, on average, have a lower value of ϵ_{igt} than the overall average: $\frac{1}{F(r)}\int_{-\infty}^{r}\epsilon dF(\epsilon) < \int_{-\infty}^{\infty}\epsilon dF(\epsilon)$. First, observe that taking the limit of the left-hand side expression to infinity yields the right-hand side: $\lim_{r\to\infty}\frac{1}{F(r)}\int_{-\infty}^{r}\epsilon dF(\epsilon) = \frac{1}{F(\infty)}\int_{-\infty}^{\infty}\epsilon dF(\epsilon) = \int_{-\infty}^{\infty}\epsilon dF(\epsilon)$ since $\frac{1}{F(\infty)} = 1/1 = 1$. What remains to be shown is that $\frac{1}{F(r)}\int_{-\infty}^{r}\epsilon dF(\epsilon)$ is increasing in r. Consider $\frac{1}{F(r')}\int_{-\infty}^{r'}\epsilon dF(\epsilon)$ where $r' > r$. Because $f(\epsilon)$ is a probability distribution function, it contains only positive mass. Therefore, the function taken to the limit r' will contain the same mass and probabilities of the function taken to r but will have additional mass on higher values of ϵ, and so $\frac{1}{F(r')}\int_{-\infty}^{r'}\epsilon dF(\epsilon) > \frac{1}{F(r)}\int_{-\infty}^{r}\epsilon dF(\epsilon)$, completing the proof. ■

Proof of Corollary 1

Proof. Follows directly from the fact that $\frac{1}{F(r)}\int_{-\infty}^{r}\epsilon dF(\epsilon)$ is increasing in r (as shown in the proof to Proposition 1). ■