WORKING PAPER 2026-3
FALL 2026

Identifying Retention Effects in the Presence of Remediation-Based Exemptions

Marcus A. Winters, Ph.D., Boston University
Abigail Francis, Boston University

Boston University Wheelock College of Education & Human Development
Wheelock Educational Policy Center

ABSTRACT

Test-based promotion policies that retain students who fail a gateway exam are often evaluated using a fuzzy regression-discontinuity (RD) design that exploits the discontinuity in retention probability at the passing threshold. This approach requires an exclusion restriction that fails when, as in many states, schools must provide remedial instruction to students who are promoted despite failing the gateway test. We show that the reduced-form effect of failing the test is a weighted average of the effects of retention and remediation, and we use this relationship to recover informative bounds on the causal effect of retention. An upper bound, equivalent to the conventional fuzzy RD estimand, follows from assuming remediation is not harmful. A lower bound follows from the empirically grounded assumption that schools negatively select failing students into retention rather than remediation. Applying this framework to administrative data from Indiana's third-grade promotion policy, we find that even under our most conservative assumptions, policy-induced retention raised fourth-grade test scores by approximately 0.3–0.4σ, with effects persisting through middle school. Our results illustrate how institutional rules that allocate individuals across multiple treatment margins complicate the interpretation of standard quasi-experimental estimators.

ACKNOWLEDGMENTS

We thank the Indiana Department of Education for providing data and valuable feedback. We received helpful comments from Cory Koedel, Kevin Lang, Joshua Goodman, Eric Taylor, Jesse Bruhn, and John Fallon, as well as seminar participants at Boston University's Applied Microeconomics Seminar and the Northeastern Economics of Education Workshop. The research-practice partnership under which this research was conducted benefits considerably from generous financial support from the William T. Grant Foundation and administrative support from Sidrah Baloch and Meagan Comb. All remaining errors are our own. The opinions expressed are those of the authors and do not represent views of the Indiana Department of Education.

INTRODUCTION

Many states use test-based promotion policies to prevent early-grade students from advancing before they master foundational literacy skills.[1] Under these policies, students who fail a gateway exam may be retained in grade, while those who pass typically are promoted on schedule. Prior studies find that policy-induced retention in early grades often raises short-run test scores, though effects frequently attenuate over time and vary across settings (Hwang and Koedel, 2025; Schwerdt et al., 2017; Jacob and Lefgren, 2004; Eren et al., 2017; Mariano and Martorell, 2013; Slungaard Mumma and Winters, 2023; Berne et al., 2026; Zhong, 2026).[2]

Prior studies identify the causal effect of retention by applying a fuzzy RD design that leverages plausibly exogenous variation in the probability a student is retained if they score on either side of the passing threshold on the gateway test. In a recent study using data from Michigan's retention policy, Berne et al. (2026) raise the concern that failing the test likely does not satisfy the necessary exclusion restriction in cases where the policy requires schools to provide remedial services to students who are promoted despite failing—a common policy feature. Treating students who benefit from receiving such remedial services as if they are non-compliers in the fuzzy RD design will produce upwardly biased estimates for the LATE of policy-induced retention. In this paper, we take the next step by showing how to recover informative bounds on the causal effect of retention in this context. Applying this approach to administrative data from Indiana, we provide evidence that even under conservative assumptions for the nature of selection bias, retention under the state's policy led to substantial and persistent improvements in student test scores.

Our strategy relies on the fact that the net effect of failing the gateway test on later outcomes is the sum of the effect of retention and the effect of remediation weighted by their respective share of the complier population. We use an RD strategy to identify the reduced-form effect of failing on later test scores and to distinguish the share of compliers who are retained and remediated. With these parameters in hand, we solve for the causal effect of retention that coincides with selected assumed values for the effect of remediation.

To generate an upper-bound for the effect of retention, we impose the assumption that remediation in the fourth grade is not harmful. This approach is equivalent to the fuzzy RD estimand. Consistent with prior studies evaluating the effect of retention in the context of programs with high exemption rates, the upper-bounds for the effect of retention are large but estimated so imprecisely that they are often insignificant (Berne et al., 2026; Slungaard Mumma and Winters, 2023).

Our estimates for the lower-bound for the effect of retention hinge on the assumption that students who fail the gateway test are negatively sorted to be retained rather than remediated. That is, when comparing observationally similar students, we anticipate that schools are more likely to retain those they expect would achieve worse outcomes without intervention. In addition to being highly intuitive, we show that failing students who were retained were more likely to be eligible for free or reduced price lunch and had substantially lower baseline test scores than failing students who were remediated instead, conditional on the forcing variable and school fixed-effects.

We apply this assumption and a deconstructed reduced-form estimate to produce an empirically grounded plausible lower-bound for the effect of retention under the policy. The resulting lower-bound estimates for the causal effect of retention on later student test scores are significant and large. Our most conservative estimate finds that policy-induced third grade retention increased fourth grade ELA scores by about 0.31σ. We find similar impacts in later grades as students progress through middle school. Our results are robust to a variety of alternative specifications.

Our findings have direct implications for the design of test-based promotion policies. Some states have expanded remediation and other exemptions, while others have strengthened retention requirements.[3] Our analyses re-frame the focus of this common exemption away from its potential to bias RD estimates for the effect of retention to the more pressing question facing schools and policymakers of whether the marginal student who has not demonstrated the required level of literacy skills is best served by being retained, remediated, or typically promoted.

More broadly, our paper contributes to a growing literature demonstrating that instrumental variables and related quasi-experimental estimators can be difficult to interpret when policy environments generate multiple treatment margins. Recent work shows that when individuals can respond to a policy through distinct treatment pathways, conventional IV estimands generally recover weighted averages of effects across different complier groups rather than a single causal parameter (Mountjoy, 2022; Kline and Walters, 2016; Mogstad et al., 2021; Bhuller and Sigstad, 2024). In our setting, students who fail the gateway test may be retained or remediated while promoted. Consequently, the discontinuity generated by the promotion policy does not solely alter exposure to retention, but rather it reallocates students across alternative instructional pathways. The parameter identified by a conventional fuzzy RD therefore depends not only on the effect of retention itself, but also on the prevalence and effectiveness of remediation-based exemptions. We show how institutional knowledge about these competing pathways can be used to recover informative bounds on the treatment margin of primary policy interest.

Viewed through this lens, our analysis extends beyond test-based promotion policies. Many public programs assign individuals to a menu of interventions rather than a simple treated-versus-untreated choice. In such settings, understanding how policy rules shift individuals across treatment pathways is essential both for interpreting quasi-experimental estimates and for designing effective policy. Our results illustrate how seemingly modest implementation decisions can fundamentally change both the parameter identified by an RD design and the policy conclusions drawn from it.

SETTING

Indiana's test-based promotion policy requires students to demonstrate a minimal level of literacy on the state's gateway reading test—the IREAD—before they are fully promoted to the fourth grade. The IREAD is an untimed, computer adaptive test. Both its basic design and the passing score have remained constant over time.

Students from the cohorts we analyze were first administered the IREAD in early- to mid-March of third grade.[4] Students who passed this spring administration were promoted and did not retake the test, while those who failed were required to attend summer school and then retook the IREAD. Students who passed the summer re-take were promoted to the fourth grade, and those who failed were designated for retention.[5]

Beginning with third graders in 2016–17, the state provided new guidance granting districts discretion to promote students who failed the summer IREAD.[6] Schools were required to provide students who were promoted despite failing IREAD with remedial third grade reading instruction while attending fourth grade. The state did not specify the form of these remedial services, and we do not observe the services that individual schools provided. The quality and implementation of these services likely varied across schools.[7] We refer to these students as "remediated," although we observe only that they were required to receive remedial services.

Policymakers intended schools to use their discretion to promote and remediate students who failed the IREAD as a targeted exception. In practice, following this change, the retention rate among students who failed the summer IREAD declined from 77% to 9.6%. The state legislature revised the policy to more forcefully require schools to retain students who fail the summer IREAD beginning with entering third graders in 2024–25.

DATA AND SAMPLE

We analyze longitudinal administrative data on the universe of public school students from 2016 through 2024 supplied by the Indiana Department of Education. For each year, we observe each student's grade level, school attended, and a variety of demographic characteristics. We classify students according to the year in which they first entered the third grade, and we identify as retained students who were again in the third grade the following year.

We observe student scores on the IREAD each time they were administered the test. Our analyses use scores from the summer administration, which is only given to students who failed the spring administration. We use as the forcing variable for the RD analyses the difference between the student's summer IREAD score and the passing score on the test.[8]

The outcome for each analysis is the student's score on the state's summative standardized test in ELA or math, which are administered each spring to all students in grades three through eight. These summative tests are distinct from the IREAD. The state changed summative standardized tests during the period we analyze. Students took the ISTEP+ until the state changed to the ILEARN in 2019. Our primary analyses use test scores standardized by grade and year to have mean 0 and standard deviation 1 and control for whether the student took the ILEARN or ISTEP+.

A unique feature of our data is that because students take summative tests before they attend summer school and re-take the IREAD, scores on the third grade summative ELA and math tests are a baseline characteristic. We use this variable as a control in our regressions.[9]

Our primary analyses include students who initially entered the third grade in 2016–17 through 2021–22. We exclude students who entered third grade in 2019–20 because they were not administered the IREAD. We exclude English learners and students with disabilities because, while they take the IREAD, they are not required to pass it to be promoted under the policy.

Measuring the effect of grade retention on later student test scores is complicated by the fact that in subsequent years retained students are typically a grade lower than comparison students who were initially promoted to the fourth grade. Our primary analyses compare retained students to promoted students after completing a given grade. This same-grade comparison incorporates the additional year of instruction received as part of the retention treatment. However, a concern with the same-grade approach is that treated students are a year older than comparison students when they take a grade-level test, so differences in their test scores could in part reflect maturity benefits.

As a robustness check, we also report results from analyses that compare student outcomes a given number of years after they first entered the third grade. For these comparisons, most students who were retained in the third grade are a grade below comparison students who were initially promoted.[10] This same-year comparison is appealing because it removes maturity effects. But to the extent that the intended treatment includes exposure to an additional year of instruction, the same-year comparison will understate the treatment effect. Further, these same-year comparisons rely heavily on accurate vertical-alignment of standardized test scores across grades, which is difficult to accomplish in practice (Ballou, 2009).[11] Validity studies of both ISTEP+ and ILEARN raised concerns about their vertical scaling (Roeber and Briggs, 2016; Indiana Department of Education, 2021).

Table 1 reports summary statistics for students in the relevant cohorts who were administered the summer IREAD, both for the full population of students who took the summer IREAD and for students who scored within the RD bandwidth. For each of these groups, we present descriptive statistics for students who passed or failed the summer IREAD and students who were retained in third grade. Among students who failed the summer test, 9.6% were retained in third grade, while retention was rare among passers (0.8%). These unconditional comparisons find that students who failed the summer IREAD were more likely to be Black and eligible for free or reduced-price lunch, and had lower spring third-grade standardized test scores in both ELA and math than students who passed the summer IREAD.

Table 1: Observed Characteristics by Group

	ALL SUMMER STUDENTS				RD SAMPLE (BASELINE BANDWIDTH)			
	All	Passed	Failed	Retained	All	Passed	Failed	Retained
Male	0.506 (0.500)	0.509 (0.500)	0.503 (0.500)	0.504 (0.500)	0.503 (0.500)	0.506 (0.500)	0.498 (0.500)	0.500 (0.500)
Black	0.326 (0.469)	0.242 (0.428)	0.414 (0.493)	0.371 (0.483)	0.303 (0.459)	0.279 (0.448)	0.338 (0.473)	0.296 (0.457)
Hispanic	0.091 (0.287)	0.085 (0.279)	0.096 (0.295)	0.069 (0.254)	0.092 (0.289)	0.087 (0.282)	0.099 (0.299)	0.062 (0.242)
White	0.506 (0.500)	0.596 (0.491)	0.411 (0.492)	0.477 (0.500)	0.527 (0.499)	0.556 (0.497)	0.485 (0.500)	0.571 (0.495)
FRL	0.699 (0.459)	0.640 (0.480)	0.761 (0.426)	0.807 (0.395)	0.691 (0.462)	0.674 (0.469)	0.715 (0.451)	0.782 (0.413)
G3 ELA	-0.985 (0.644)	-0.742 (0.580)	-1.234 (0.610)	-1.281 (0.656)	-0.919 (0.556)	-0.847 (0.546)	-1.025 (0.553)	-1.081 (0.605)
G3 Math	-0.894 (0.715)	-0.618 (0.643)	-1.176 (0.674)	-1.235 (0.682)	-0.800 (0.628)	-0.725 (0.624)	-0.911 (0.617)	-0.986 (0.629)
Failed Summer Test	0.494 (0.500)	0.000 (0.000)	1.000 (0.000)	1.000 (0.000)	0.407 (0.491)	0.000 (0.000)	1.000 (0.000)	1.000 (0.000)
Retained in Grade 3	0.052 (0.221)	0.008 (0.090)	0.096 (0.295)	1.000 (0.000)	0.038 (0.192)	0.008 (0.091)	0.082 (0.275)	1.000 (0.000)
N	43,201	21,845	21,356	2,057	18,191	10,796	7,395	608

Notes: Table reports means and standard deviations (in parentheses) for students in their third-grade year. Sample is restricted to non-exempt students subject to Indiana's test-based promotion policy (treat g4= 1, exempt= 0). "Passed" and "Failed" refer to the summer IREAD re-test. "Retained" students failed the summer IREAD and repeated third grade. The RD sample reports baseline (grade-3) characteristics for the subset of students who fall within the MSE-optimal bandwidth, around the summer IREAD forcing variable, in their grade-4 record. Because reaching a grade-4 record occurs later, in calendar time, for retained students than for non-retained students, the RD sample reflects whichever students are actually observed at grade 4, not merely those in-band at the time of the summer test. G3 ELA and G3 Math are standardized spring third-grade test scores (mean zero, standard deviation one within grade and year).

MULTIPLE TREATMENT MARGINS FOR STUDENTS WHO FAIL THE GATEWAY TEST

The existence of a threshold on the gateway test that triggers assignment to treatment naturally points the researcher towards using an RD design to measure the effect of retention. The exclusion restriction necessary to apply the fuzzy RD design in this setting requires that students who were promoted despite failing the gateway test experienced no treatment effect under the policy. That failing students who are not retained receive remedial services that they would not have received had they passed the gateway test poses a problem. The observed net treatment effect for students who failed the summer IREAD is a weighted average for the effect experienced by students who were retained and students who were remediated. Formally, we can write the net effect of failing the summer IREAD test as:

$$\underbrace{\tau_{net}}_{\text{Net Effect of Failing Summer IREAD}} = \alpha \underbrace{\tau_{ret}}_{\text{Effect of 3rd Grade Retention}} + (1-\alpha) \underbrace{\tau_{rem}}_{\text{Effect of Remediation in 4th Grade}} \quad (1)$$

where α is the share of compliers assigned to the full treatment under the policy. The exclusion restriction necessary for the fuzzy RD design to identify the LATE of policy-induced retention requires the strong assumption that $\tau_{rem} = 0$.[12]

We can describe the observed differences in student outcomes by their treatment pathway by estimating reduced-form regressions that replace the indicator for failing the summer IREAD with indicators that distinguish failing students who were retained from failing students who were remediated. The outcomes for students in these subgroups are measured relative to the omitted comparison group comprised of students who passed the summer IREAD. Specifically, we estimate:

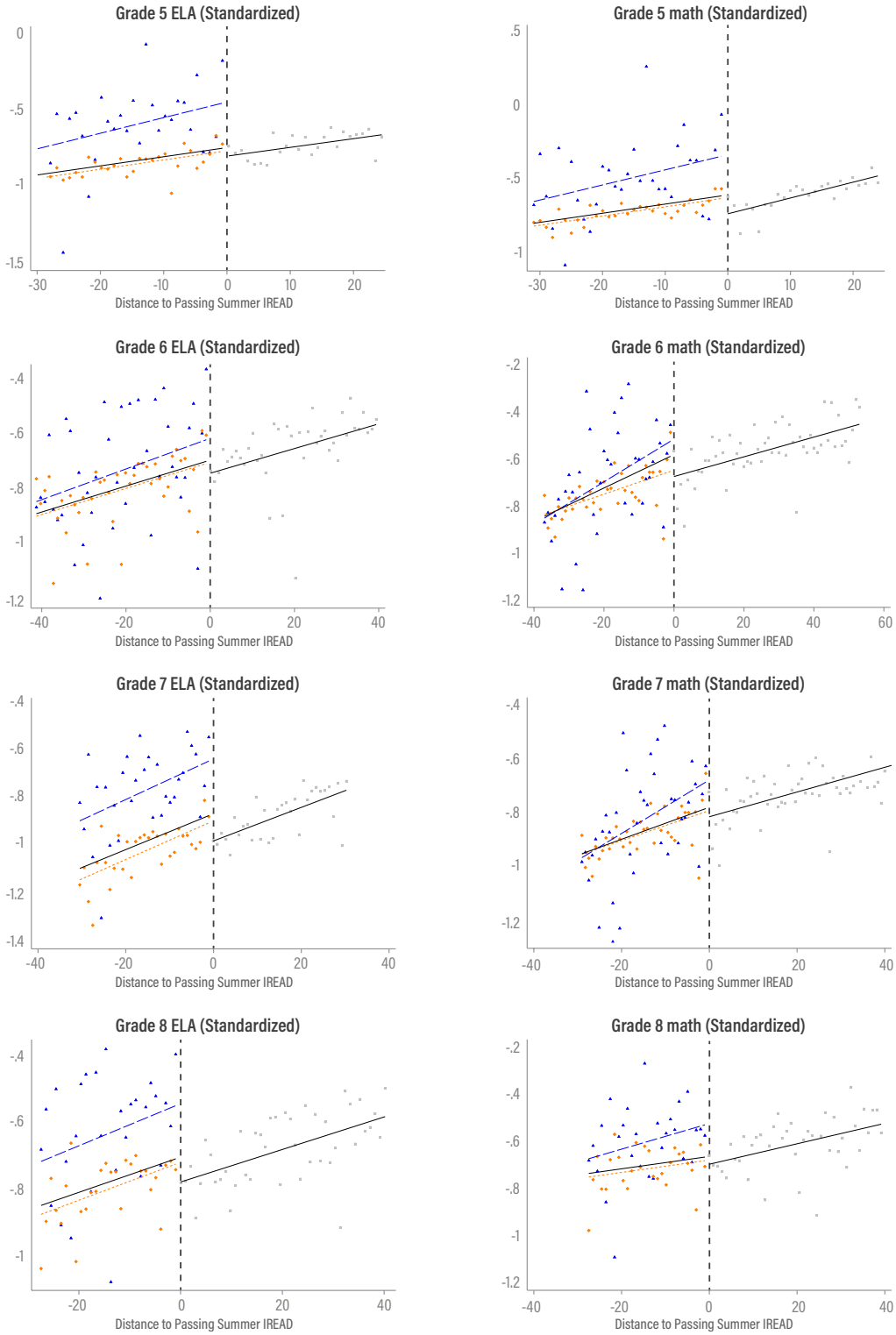
$$y_i = \zeta + f(a_i) + \lambda_{ret} Ret_i + \lambda_{rem} Rem_i + \epsilon_i \quad (2)$$

Figure 1 illustrates the relationship between the distance between the student's score on the summer IREAD from the passing threshold and later test scores overall and by treatment pathway. The grey dots to the right of the figure illustrate the average outcome for students with a given passing summer IREAD score. To the left of the threshold, we separately illustrate the average outcome at each failing Summer IREAD score for students who received the full treatment (blue triangles) and students who received an exemption (orange diamonds). We run regression lines through each group.

The difference between the solid black lines at the threshold illustrates the reduced-form relationship between failing the test and the outcome. If the classical RD assumptions hold, then we can interpret this difference as a causal estimate for the intention to treat. This is itself a policy relevant parameter.

But our primary goal is to isolate the causal effect of retention under the policy. The substantial difference in average math and ELA scores between retained and remediated students is striking. On average, retained students scored substantially higher than remediated students who had the same score on the summer IREAD test. The pattern is consistent across subjects and subsequent grades. In each case, the average outcome for both treated subgroups is above the line for students who passed the summer IREAD. However, while the jump at the threshold is consistently large for retained students, in several cases the discontinuity in the average test scores between passing students and failing students who were remediated appears to be small and likely insignificant.

Figure 1: Deconstructing the Effect of Failing Summer IREAD by Treatment Pathway



Notes: Figure illustrates results from estimating Equation (4). X-axis is the distance between student's score on the summer IREAD and the passing score. Y-axis is the student's score on the respective test, standardized by grade and year. Points on the figure illustrate average score by group for students with a particular summer IREAD score. Grey squares represent students with passing scores; Blue triangles are students who failed, repeated third grade, and entered the respective grade a year behind their original cohort; orange diamonds are students who failed and received the exemption for their respective cohort type. Lines represent local regressions fit for the respective observations. Black lines illustrate separately estimated reduced form regressions for the effect of failing summer IREAD that do not distinguish between student types.

Taken at face value, the patterns illustrated in Figure 1 suggest that retention under the policy substantially increased a student's later test scores, while remediation under the policy had either a small or no significant impact relative to passing and being typically promoted to the fourth grade. But these analyses do not clarify the extent to which these observed outcome differences among the retained and remediated groups are driven by differing treatment effects or the type of failing student who is assigned to one or the other pathway. While it is plausible that students scoring at the threshold are as-good-as-randomly assigned to receive some treatment, we suspect that students who fail the summer IREAD are endogenously sorted to either be retained or remediated. One strategy to address this problem is to identify an instrument that impacts the likelihood that a failing student is assigned to be retained but is otherwise unrelated to later outcomes. But identifying a plausible instrument can be difficult. Instead, our approach is to use a combination of well-identified parameters and empirically-grounded assumptions for the nature of selection bias to generate plausible bounds for τ_{ret}.[13]

BOUNDING THE EFFECT OF POLICY-INDUCED RETENTION

Our goal is to estimate the causal effect of retention under Indiana's policy on later test scores. Our first strategy for achieving this goal is to combine causal estimates for τ_{net} and α with extreme assumed values for τ_{rem} to generate plausible bounds τ_{ret}. Given this purpose, it is helpful to re-write Equation 1 as:

$$\tau_{ret} = \frac{\tau_{net} - [(1-\alpha)\tau_{rem}]}{\alpha} \quad (3)$$

We begin by generating causal estimates for τ_{net} and α. We do this by applying an RD analysis that exploits the discontinuity in treatment under the policy at the passing threshold on the summer IREAD. We first measure the reduced-form relationship between failing the summer IREAD and later outcomes by estimating:

$$y_{is} = \zeta + \theta_s + \gamma X_i + f(a_i) + \beta_{net}\mathbb{1}(a_i < 0) + \epsilon_i \quad (4)$$

where y_{is} is the outcome for student i who attended school s during their initial third grade year; θ_s is a fixed effect for the school the student attended during their initial third grade year; X is a vector of observed baseline covariates that includes race/ethnicity, gender, eligibility for free or reduced price lunch, and the student's spring of third grade scores on the state's summative ELA and math tests; $f(a)$, the forcing variable, is a function of the distance in scale scores from the passing threshold; $1\ (a < 0)$ indicates that the student failed the test; and ϵ is a stochastic term. Our preferred specification treats $f(a)$ as a linear function with a uniform kernel allowed to vary on either side of the threshold. Each sample is comprised of eligible students with a summer IREAD score that falls within the MSE-optimal bandwidth for the respective outcome (Calonico et al., 2017). We rebuild the bandwidth for each comparison.

We next adjust this RD approach to measure the effect of failing the summer IREAD on the probability that a student is retained in the third grade:

$$Ret = \zeta + \theta_s + \gamma X_i + f(a_i) + \beta_{fs}\mathbb{1}(a_i < 0) + \epsilon_i \quad (5)$$

We are primarily interested in $\hat{\beta}_{net}$ and $\hat{\beta}_{fs}$, which under the classical RD assumptions respectively identify the causal effect of failing the summer IREAD on later student outcomes and on the probability of repeating the third grade. These coefficients serve as our estimates for τ_{net} and α, respectively. Plugging $\hat{\beta}_{net}$ and $\hat{\beta}_{fs}$ into Equation 3 gives:

$$\tau_{ret} = \frac{\hat{\beta}_{net} - [(1 - \hat{\beta}_{fs})\tau_{rem}]}{\hat{\beta}_{fs}} \quad (6)$$

We can use Equation 6 to calculate τ_{ret} for any assumed value of τ_{rem}. To generate an upper-bound estimate for τ_{ret}, we apply the assumption that $\tau_{rem} \geq 0$. This assumption requires only that providing fourth grade students with supplemental reading instruction does not reduce their later test scores. Since $\frac{\partial \tau_{ret}}{\partial \tau_{rem}} \leq 0$, we can generate an upper-bound estimate for τ_{ret} by setting $\tau_{rem} = 0$, which leaves:

$$\tau_{ret}^{upper} = \frac{\hat{\beta}_{net}}{\hat{\beta}_{fs}} \quad (7)$$

Generating a lower-bound estimate for τ_{ret} requires grounding a plausible maximum value for τ_{rem}. We do this by estimating a regression that deconstructs the reduced-form relationship across the retention and remediation pathways. This regression is similar to Equation 2, but we now incorporate controls for observed baseline characteristics, including a fixed-effect for the school the student attended in their initial third grade year. Formally:

$$y_i = \zeta + \theta_s + \gamma X_i + f(a_i) + \lambda_{ret} Ret_i + \lambda_{rem} Rem_i + \epsilon_i \quad (8)$$

As before, we do not interpret λ_{ret} or λ_{rem} as causal estimates for τ_{ret} and τ_{rem} because we suspect that schools non-randomly sort failing students across the retention and remediation pathways. But while we cannot credibly estimate the magnitude of such selection bias without a compelling instrument for how failing students are sorted to be retained or remediated, we can plausibly sign its direction. We argue that students who fail the summer IREAD are negatively selected to be retained rather than remediated, and thus to the extent that selection bias exists, Equation 8 will overstate the effect of remediation. This leads to the key assumption that $\hat{\lambda}_{rem} \geq \tau_{rem}$.

The intuition underlying our assumption that failing students are negatively sorted to be retained is straightforward. Schools have historically viewed retention as an extreme intervention reserved for students facing exceptional challenges. Consistent with this view, once given discretion, schools remediated rather than retained most students who failed the summer IREAD. Parents also historically prefer for their child to be promoted from third to fourth grade, and it is reasonable to suspect that lower resourced parents are in a weaker position to advocate for their child to be remediated.[14]

Thus, we suspect that when faced with observationally equivalent students who failed the summer IREAD, schools are more likely to use their discretion under the policy to retain students whose unobserved-to-the-researcher attributes suggests that they face disproportionately difficult challenges relative to the average student who failed the summer IREAD. Said another way, this assumption is violated only if schools are systematically more inclined to retain among the pool of students who failed the summer IREAD those who would make larger academic progress if they received no further intervention. We provide some empirical support for the plausibility of this assumption in Probing Key Assumptions section.

If $\hat{\lambda}_{rem} \geq \tau_{rem}$, then we can generate a lower-bound estimate for τ_{rem} by plugging $\hat{\lambda}_{rem}$ into Equation 6.[15]

$$\tau_{ret}^{lower} = \frac{\hat{\beta}_{net} - [(1 - \hat{\beta}_{fs})\hat{\lambda}_{rem}]}{\hat{\beta}_{fs}} \quad (9)$$

We bootstrap the entire system of equations to generate confidence intervals for τ_{ret}^{upper} and τ_{ret}^{lower}. We draw 300 random samples from the observations used to estimate Equation 4 for each respective outcome. To ensure sufficient observations to estimate the fixed-effects, we cluster the sample by the student's third grade school. Within each of these samples we estimate Equations 4, 5, and 8 and use the relevant coefficients to calculate τ_{ret}^{upper} and τ_{ret}^{lower}. The standard deviation of the resulting distribution of the lower- and upper-bound estimates is the bootstrapped standard error.

Probing Key Assumptions

Identifying Assumptions for Reduced-Form and First-Stage Relationships

The key identifying assumption for both $\hat{\beta}_{net}$ and $\hat{\beta}_{fs}$ is that absent the policy the relationship between summer IREAD scores and the outcome would have been smooth at the passing threshold. One way for this assumption to be violated is if students or schools manipulated summer IREAD scores to avoid failing the test. The nature of the summer IREAD as a computer-administered standardized test makes it unlikely that students or schools could systematically manipulate the scores. Density tests illustrated on Figure B1 find no evidence of a break in the distribution of summer IREAD scores at the passing threshold.

We can also test the validity of the identifying assumption by investigating whether the relationship between observed student characteristics and the forcing variable is smooth at the threshold. To do this, we estimate Equation 4 using student characteristics as the outcome and removing all covariates other than the forcing variable. The results, reported in the top panel of Table 2, find no significant differences at the threshold in student gender, race/ethnicity, eligibility for free or reduced price lunch, or scores on summative ELA and math tests administered in the spring of their initial third grade year. The coefficients are all near-zero and precisely estimated.

Table 2: Test for Covariate Balance

	Male	Black	Hispanic	White	FRL	G3 ELA	G3 Math
Pass vs Fail							
Failed	-0.010 (0.015)	-0.020 (0.014)	0.011 (0.009)	0.007 (0.015)	0.000 (0.014)	-0.005 (0.016)	0.008 (0.018)
N	18231	18231	18231	18231	18231	18199	18241
Retained vs Remediated							
Retained	-0.005 (0.023)	-0.059*** (0.021)	-0.022* (0.012)	0.079*** (0.022)	0.058*** (0.019)	-0.079*** (0.026)	-0.091*** (0.029)
Remediated	-0.010 (0.015)	-0.016 (0.014)	0.016* (0.009)	-0.004 (0.015)	-0.005 (0.014)	-0.008 (0.016)	0.003 (0.018)
N	18231	18231	18231	18231	18231	18199	18241
Retained vs Remediated, Include 3rd Grade FE							
Retained	0.007 (0.024)	-0.010 (0.016)	-0.008 (0.013)	0.009 (0.018)	0.046** (0.020)	-0.096*** (0.027)	-0.112*** (0.031)
Remediated	-0.002 (0.015)	-0.019* (0.011)	0.014 (0.009)	0.006 (0.012)	-0.000 (0.013)	-0.014 (0.016)	0.010 (0.018)
N	18231	18231	18231	18231	18231	18199	18241

Notes: Table reports estimates from regressions that use student baseline characteristics as the outcome. The top panel reports RD estimates for the effect of failing the summer IREAD at the passing threshold. The middle and bottom panels decompose the net effect of failing across failing students who were retained and failing students who were remediated, relative to students who passed the summer IREAD. All regressions control for the difference between the student's summer IREAD score and the passing score. The bottom panel also controls for a fixed effect for the school the student attended in their initial third-grade year. Table A1 reports results from similar regressions that include all covariates used in the main regression specification and that include all covariates except for summative third grade ELA and math scores. Robust standard errors are reported in parentheses. *$p < 0.10$, **$p < 0.05$, ***$p < 0.01$

Assumption that Failing Students are Negatively Selected to Retention Pathway

The central assumption underlying our bounding strategy is that retained students are negatively selected to be retained rather than remediated. Table 2 demonstrates precisely such negative selection on observed dimensions.

The remaining panels of Table 2 report the results from regressions that estimate Equation 8 but use student baseline observed characteristics as the outcome. The middle panel reports the results from regressions that only control for the forcing variable, and the bottom panel reports results from regressions that also include a fixed effect for the school the student attended in their initial third grade year.[16]

When we split students across treatment pathways we find that the characteristics of failing students who were remediated are similar to students who passed the summer IREAD. In contrast, failing students who were retained are significantly less likely to be Black or Hispanic, more likely to be white and eligible for free or reduced price lunch, and had substantially lower spring of third grade summative math and ELA scores than students who passed the summer IREAD. These differences in observed baseline characteristics across these groups suggest that failing students are non-randomly sorted across the retention and remediation treatment pathways.

At first blush, it seems odd that retained students appear more advantaged according to their race/ethnicity but more disadvantaged on free lunch eligibility and baseline test scores than remediated students with the same score on the gateway test. The results reported in the bottom panel show that controlling for the student's initial third grade school eliminates the difference in student race/ethnicity at the threshold. This pattern suggests that the finding that on average failing white students are more likely to be retained is driven by across-school differences in the propensity to remediate failing students that correlates with a higher proportion of white students in the school.[17] When making within-school comparisons, the significant and substantial differences in free or reduced price lunch status and baseline summative test scores remains.

The key finding from Table 2 is that, relative to failing students with the same summer IREAD score and attending the same school, students who come from a lower-income household and those who performed worse on the spring summative ELA and math tests are more likely to be retained rather than remediated. That conditional on summer IREAD score retained students have substantially lower baseline ELA and math scores than both remediated and passing students is especially telling. This pattern indicating that failing students are negatively sorted to retention based on their observed characteristics lends credibility to our assumption that they are also negatively sorted to be retained based on unobserved characteristics.[18]

Results

Table 3 reports upper- and lower-bounds for the effect of retention along with the results for $\hat{\beta}_{net}$, $\hat{\beta}_{fs}$, $\hat{\lambda}_{rem}$, and $\hat{\lambda}_{ret}$ used to generate the point estimates.

The results from the reduced-form regressions suggests that, on average, failing the summer IREAD leads to modest improvements in later ELA and math scores. The estimates in the second panel on the table show that the modest net intention-to-treat impact for the average student masks substantial heterogeneity in the outcomes for failing students who were retained or remediated. The conditional average score for students who failed the summer IREAD and were remediated is at best slightly higher than the comparison group of students who passed the summer IREAD. In contrast, failing students who were retained on average score between 0.12–0.39σ higher than students who passed the gateway test. Failing students who were retained comprise only between 5–12% of the complier population.

The bottom panel of Table 3 reports the lower- and upper-bounds for the effect of retention. As we previously described, we derive the upper-bound estimate by dividing the net effect of failing by the share of compliers who were retained. Adjusting the reduced-form by the modest first-stage produces very large but very imprecisely estimated upper-bounds. Though most of the upper-bound estimates are above 0.5σ, the standard errors are so large that only a few are statistically significant.[19]

Table 3: Estimated Bounds for the Effect of Retention

	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
	English Language Arts					Mathematics				
Reduced Form: Net Effect of Failing										
$\hat{\beta}_{net}$	0.038* (0.021)	0.043 (0.026)	0.047 (0.039)	0.112*** (0.032)	0.039 (0.036)	0.042** (0.020)	0.106** (0.044)	0.053 (0.040)	0.067* (0.029)	0.040 (0.046)
Decomposition										
$\hat{\lambda}_{ret}$	0.362*** (0.032)	0.310*** (0.037)	0.122*** (0.047)	0.357*** (0.041)	0.230*** (0.044)	0.395*** (0.032)	0.307*** (0.045)	0.150*** (0.041)	0.329*** (0.037)	0.181*** (0.052)
$\hat{\lambda}_{rem}$	0.023 (0.020)	0.022 (0.027)	0.036 (0.039)	0.071** (0.031)	0.012 (0.035)	0.023 (0.019)	0.090** (0.043)	0.031 (0.041)	0.016 (0.028)	0.038 (0.048)
Effect of Failing on Probability Retained										
$\hat{\beta}_{fs}$	0.052*** (0.006)	0.068*** (0.008)	0.092*** (0.010)	0.121*** (0.013)	0.114*** (0.013)	0.051*** (0.006)	0.064*** (0.008)	0.092*** (0.009)	0.127*** (0.013)	0.114*** (0.013)
Estimated Lower and Upper Bounds for Effect of Retention										
τ_{ret}^{lower}	0.312*** (0.094)	0.336*** (0.091)	0.153* (0.090)	0.406*** (0.086)	0.251*** (0.085)	0.387*** (0.095)	0.336*** (0.118)	0.279*** (0.086)	0.410*** (0.083)	0.057 (0.089)
τ_{ret}^{upper}	0.733* (0.398)	0.630* (0.374)	0.511 (0.445)	0.927*** (0.290)	0.343 (0.289)	0.822*** (0.345)	1.652*** (0.672)	0.581 (0.412)	0.523* (0.271)	0.354 (0.419)
N	15017	10388	9738	7352	7908	15401	10320	10619	7679	6760

Notes: Table reports estimates used to bound the effect of third-grade retention under Indiana's test-based promotion policy. The first panel reports RD estimates of the reduced-form effect of failing the summer IREAD on later test scores ($\hat{\beta}_{net}$). The second panel reports coefficients from regressions that decompose the reduced-form relationship into outcomes for failing students who were retained ($\hat{\lambda}_{ret}$) and failing students who were remediated while promoted ($\hat{\lambda}_{rem}$), measured relative to students who passed the summer IREAD. The third panel reports RD estimates of the effect of failing the summer IREAD on the probability of retention ($\hat{\beta}_{fs}$). The bottom panel reports estimated lower and upper bounds for the effect of retention calculated using Equations 9 and 7. Upper bounds assume that remediation is not harmful ($\tau_{rem} \geq 0$). Lower bounds assume that failing students are negatively selected into the retention pathway, implying that $\hat{\lambda}_{rem}$ weakly overstates the causal effect of remediation. Confidence intervals for the lower- and upper-bound estimates are generated by bootstrapping the full system of equations used to calculate the bounds. Regressions include school fixed effects, student demographic controls, baseline third-grade ELA and math achievement, and the forcing variable. Robust standard errors are reported in parentheses. *$p < 0.10$,**$p < 0.05$, ***$p < 0.01$

We are primarily interested in the lower-bound estimates for the effect of retention, which most directly address whether retained students' larger observed gains can be explained by non-random sorting across treatment pathways. Accounting for the small improvements relative to those who passed the test made by the large share of failing students who were remediated has a substantial influence on our estimate for the causal effect of retention. In most cases, the lower-bound is less than half the size of the respective upper-bound estimate. Nonetheless, even under our most conservative assumption for the nature of selection bias we find that retention led to large improvements in later student test scores. At the lower-bound, in the fourth grade retained students scored about 0.312σ higher in ELA and 0.387σ higher in math. Students maintained these benefits as they progressed through middle school. All but one of the lower-bound estimates is statistically significant.

Robustness

Tables A2 through A7 in the Online Appendix investigate the robustness of our result to a variety of alternative specifications. We estimate regressions that apply alternative bandwidths; use a quadratic form for the forcing variable; exclude the control for third grade summative ELA and math scores; and exclude all covariates. None of these specification changes materially alters the pattern of our results.

Table A8 in the Online Appendix reports results from analyses that compare students a given number of years after entering the third grade.[20] These regressions use the vertically aligned scale score on the ILEARN as the outcome. The pattern of results is similar to what we observe when making across-grade comparisons. The lower-bound estimates are consistently positive and substantial. Though a clean conversion is not possible because students in the sample are taking the test in different grades, a back-of-the-envelope calculation of the magnitude of the same-year estimates is similar to the results from the respective same-grade estimates. The lower-bound estimates are significant in all but one case per subject. Though we prefer the same-grade estimates for the reasons described in Data and Sample section, that we continue to find benefits when making same-year comparisons addresses the concern that our primary results are driven by maturity benefits for retained students.

CONCLUSION

A growing body of research uses an RD design embedded within test-based promotion policies to estimate the causal effect of grade retention on later student outcomes. Recent work has highlighted that these designs may not identify the effect of retention when students who fail the gateway test but are nonetheless promoted receive meaningful remedial services. In such settings, failing the gateway test does not simply alter exposure to retention. Instead, it reallocates students across multiple instructional pathways, raising important questions about the interpretation of conventional fuzzy RD estimates.

This paper develops and applies a simple framework for recovering informative bounds on the effect of retention in the presence of remediation-based exemptions. We show that the reduced-form effect of failing the gateway test can be expressed as a weighted average of the effects of retention and remediation. Combining RD estimates of the reduced-form and first-stage relationships with empirically grounded assumptions about the effect of remediation allows us to recover plausible bounds for the causal effect of retention. Applying this approach to Indiana's third-grade reading policy, we find that retention generated substantial and persistent improvements in student achievement. Even under conservative assumptions that maximize the potential benefits of remediation, our lower-bound estimates suggest that retained students experienced gains of approximately 0.3–0.4σ in both ELA and math that persisted through middle school.

These findings have direct implications for policy design. Policymakers often debate whether struggling readers should be retained, promoted with remediation, or allowed to progress without additional intervention. Our results suggest that these choices should not be viewed solely through the lens of how they affect the number of students retained. The structure of exemption pathways fundamentally shapes the impacts generated by the policy.

More broadly, our analysis illustrates a challenge that extends well beyond grade retention policies. Many public programs do not create a simple distinction between treated and untreated individuals. Instead, they shift individuals across multiple treatment pathways that differ in both intensity and effectiveness. In such settings, conventional estimators often recover weighted averages across distinct treatment margins rather than the effect of a single intervention. Understanding how institutional rules allocate individuals across competing pathways is therefore essential both for interpreting causal estimates and for designing effective policy.

REFERENCES

Ballou, D. (2009). Test scaling and value-added measurement. *Education Finance and Policy,* 4(4):351–383.

Berne, J. S., Jacob, B. A., Weiland, C., and Strunk, K. O. (2026). Assessing the impact of grade retention: A cautionary tale of exclusion restriction violations. *Journal of Public Economics,* 253:105534.

Bhuller, M. and Sigstad, H. (2024). 2sls with multiple treatments. *Journal of Econometrics,* 242(1):105785.

Calonico, S., Cattaneo, M. D., Farrell, M. H., and Titiunik, R. (2017). rdrobust: Software for Regression-Discontinuity designs. *Stata Journal,* 17(2):372–404.

Colorado Department of Education (2023). Read Act annual report.

Eren, O., Depew, B., and Barnes, S. (2017). Test-based promotion policies, dropping out, and juvenile crime. *Journal of Public Economics,* 153:9–31.

Greene, J. P. and Winters, M. A. (2009). The effects of exemptions to Florida's test-based promotion policy: Who is retained?: Who benefits academically? *Economics of Education Review,* 28(1):135–142.

Hwang, N. and Koedel, C. (2025). Helping or hurting: The effects of retention in the third grade on student outcomes. *Educational Evaluation and Policy Analysis,* 47(1):65–88.

Indiana Department of Education (2021). ILEARN 2020–2021 technical report. Technical report, Indiana Department of Education.

Jacob, B. A. and Lefgren, L. (2004). Remedial education and student achievement: A regression-discontinuity analysis. *Review of Economics and Statistics,* 86(1):226–244.

Kline, P. and Walters, C. R. (2016). Evaluating public programs with close substitutes: The case of Head Start. *Quarterly Journal of Economics,* 131(4):1795–1848.

Mariano, L. T. and Martorell, P. (2013). The academic effects of summer instruction and retention in New York City. *Educational Evaluation and Policy Analysis,* 35(1):96–117.

Michigan Department of Education (2024). Read by Grade Three guide.

Mogstad, M., Torgovitsky, A., and Walters, C. R. (2021). The causal interpretation of two-stage least squares with multiple instrumental variables. *American Economic Review,* 111(11):3663–3698.

Mountjoy, J. (2022). Community colleges and upward mobility. *American Economic Review,* 112(8):2580–2611.

North Carolina Department of Public Instruction (2023). Read to Achieve annual report.

Roeber, E. and Briggs, D. C. (2016). An investigation into the validity of the Indiana 2015 ISTEP+ assessment program. Technical report, Indiana State Board of Education.

Schwerdt, G., West, M. R., and Winters, M. A. (2017). The effects of test-based retention on student outcomes over time: Regression-discontinuity evidence from Florida. *Journal of Public Economics,* 152:154–169.

Slungaard Mumma, K. and Winters, M. (2023). The effect of retention under Mississippi's test-based promotion policy. Technical report, Wheelock Educational Policy Center.

Winters, M. A. (2026). Student responses to the threat of grade retention. Working paper, Boston University.

Zhong, J. (2026). Early grade retention harms adult earnings. *American Economic Journal: Applied Economics.* Forthcoming.

ONLINE APPENDIX

A. Tables

Table A1: Robustness Test for Covariate Balance

	Male	Black	Hispanic	White	FRL	G3 ELA	G3 Math
Retained vs Remediated, Full Covariates							
Retained	0.014 (0.024)	-0.017 (0.016)	-0.012 (0.013)	0.021 (0.018)	0.045** (0.020)	-0.084*** (0.026)	-0.086*** (0.029)
Remediated	-0.002 (0.015)	-0.019* (0.011)	0.014 (0.009)	0.006 (0.012)	0.002 (0.013)	-0.021 (0.016)	0.015 (0.017)
N	18092	18092	18092	18092	18092	18092	18092
Retained vs Remediated, No G3 Score Controls							
Retained	0.007 (0.024)	-0.009 (0.016)	-0.010 (0.013)	0.012 (0.018)	0.053*** (0.020)	-0.121*** (0.027)	-0.133*** (0.030)
Remediated	-0.001 (0.015)	-0.018* (0.011)	0.013 (0.009)	0.006 (0.012)	0.002 (0.013)	-0.018 (0.016)	0.006 (0.017)
N	18231	18231	18231	18231	18231	18119	18160

Notes: Table reports estimates from regressions that use student baseline characteristics as the outcome. Both panels decompose the effect of failing the summer IREAD into failing students who were retained and failing students who were remediated, relative to students who passed the summer IREAD. All regressions control for difference between summer IREAD score and the passing score and a fixed effect for school attended during initial third grade year. The first panel also controls for all covariates included in the main regression specification—except for indicators for different values of the same characteristic (e.g. when Black is the outcome the model does not control for if the student is white.) The second panel includes the full set of demographic covariates but omits third-grade ELA and math score controls. Third-grade ELA and math scores are standardized within grade and year. Robust standard errors are reported in parentheses. *$p < 0.10$, **$p < 0.05$, ***$p < 0.01$

Table A2: Bounding the Effect of Retention (Half Bandwidth)

	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
	English Language Arts					Mathematics				
Reduced Form: Net Effect of Failing										
$\hat{\beta}_{net}$	0.043 (0.031)	0.095*** (0.034)	-0.021 (0.064)	0.071 (0.049)	0.037 (0.049)	0.040 (0.029)	0.156** (0.068)	0.041 (0.067)	0.099** (0.047)	0.049 (0.074)
Decomposition										
$\hat{\lambda}_{ret}$	0.369*** (0.049)	0.361*** (0.054)	0.065 (0.068)	0.321*** (0.064)	0.191*** (0.063)	0.398*** (0.047)	0.387*** (0.071)	0.172** (0.067)	0.365*** (0.058)	0.176** (0.080)
$\hat{\lambda}_{rem}$	0.035 (0.030)	0.080** (0.033)	-0.031 (0.062)	0.068 (0.048)	0.024 (0.049)	0.028 (0.028)	0.153** (0.067)	0.007 (0.067)	0.055 (0.045)	0.029 (0.073)
Effect of Failing on Probability Retained										
$\hat{\beta}_{fs}$	0.042*** (0.009)	0.063*** (0.012)	0.085*** (0.014)	0.105*** (0.020)	0.100*** (0.019)	0.045*** (0.010)	0.059*** (0.012)	0.083*** (0.014)	0.108*** (0.020)	0.096*** (0.021)
Estimated Lower and Upper Bounds for Effect of Retention										
τ_{ret}^{lower}	0.240 (0.158)	0.321** (0.152)	0.089 (0.173)	0.095 (0.149)	0.157 (0.153)	0.289* (0.154)	0.202 (0.157)	0.419** (0.205)	0.460*** (0.137)	0.239 (0.179)
τ_{ret}^{upper}	1.029* (0.612)	1.513*** (0.550)	-0.246 (0.757)	0.674 (0.540)	0.376 (0.521)	0.889 (0.555)	2.637** (1.065)	0.495 (0.858)	0.913* (0.489)	0.511 (0.859)
N	7,919	5,454	5,091	3,797	4,162	7,937	5,494	5,609	3,992	3,500

Notes: Robustness check: Half Bandwidth. The first panel reports RD estimates of the reduced-form effect of failing the summer IREAD on later test scores ($\hat{\beta}_{net}$). The second panel decomposes the reduced-form into outcomes for retained ($\hat{\lambda}_{ret}$) and remediated-while-promoted ($\hat{\lambda}_{rem}$) students, relative to passers. The third panel reports the effect of failing on retention probability ($\hat{\beta}_{fs}$). The bottom panel reports lower and upper bounds for the causal effect of retention. Upper bounds assume remediation is not harmful ($\tau_{rem} \geq 0$). Lower bounds assume failing students are negatively selected into retention. Bootstrap standard errors in parentheses for bound rows; robust standard errors for all other rows. Bandwidth is set to half the MSE-optimal bandwidth from the baseline specification, computed separately by grade and subject. Regressions include the full set of demographic covariates and third-grade ELA/math scores, matching the preferred specification. *p < 0.10, **p < 0.05, ***p < 0.01

Table A3: Bounding the Effect of Retention (Double Bandwidth)

	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
	English Language Arts					Mathematics				
Reduced Form: Net Effect of Failing										
$\hat{\beta}_{net}$	0.032** (0.016)	0.034* (0.020)	0.020 (0.030)	0.081*** (0.023)	0.054* (0.030)	0.034** (0.016)	0.061** (0.026)	0.025 (0.025)	0.045** (0.021)	0.045 (0.034)
Decomposition										
$\hat{\lambda}_{ret}$	0.334*** (0.024)	0.281*** (0.028)	0.069** (0.035)	0.376*** (0.031)	0.247*** (0.042)	0.402*** (0.022)	0.285*** (0.031)	0.114*** (0.032)	0.334*** (0.027)	0.154*** (0.049)
$\hat{\lambda}_{rem}$	0.016 (0.016)	0.016 (0.020)	0.019 (0.030)	0.032 (0.022)	0.041 (0.031)	0.012 (0.016)	0.044* (0.026)	0.013 (0.026)	-0.007 (0.020)	0.045 (0.036)
Effect of Failing on Probability Retained										
$\hat{\beta}_{fs}$	0.053*** (0.005)	0.073*** (0.006)	0.101*** (0.007)	0.131*** (0.009)	0.110*** (0.009)	0.052*** (0.005)	0.072*** (0.007)	0.138*** (0.009)	0.108*** (0.020)	0.113*** (0.010)
Estimated Lower and Upper Bounds for Effect of Retention										
τ_{ret}^{lower}	0.327*** (0.055)	0.267*** (0.062)	0.031 (0.068)	0.404*** (0.052)	0.165 (0.102)	0.426*** (0.050)	0.292*** (0.059)	0.137** (0.066)	0.370*** (0.047)	0.050 (0.083)
τ_{ret}^{upper}	0.605** (0.270)	0.470* (0.257)	0.196 (0.296)	0.619*** (0.167)	0.494** (0.223)	0.652** (0.258)	0.851*** (0.273)	0.248 (0.268)	0.326** (0.143)	0.402 (0.330)
N	24,928	17,669	15,445	11,953	12,032	25,152	17,597	16,146	12,306	11,152

Notes: Robustness check: Double Bandwidth. The first panel reports RD estimates of the reduced-form effect of failing the summer IREAD on later test scores ($\hat{\beta}_{net}$). The second panel decomposes the reduced-form into outcomes for retained ($\hat{\lambda}_{ret}$) and remediated-while-promoted ($\hat{\lambda}_{rem}$) students, relative to passers. The third panel reports the effect of failing on retention probability ($\hat{\beta}_{fs}$). The bottom panel reports lower and upper bounds for the causal effect of retention. Upper bounds assume remediation is not harmful ($\tau_{rem} \geq 0$). Lower bounds assume failing students are negatively selected into retention. Bootstrap standard errors in parentheses for bound rows; robust standard errors for all other rows. Bandwidth is set to double the MSE-optimal bandwidth from the baseline specification, computed separately by grade and subject. Regressions include the full set of demographic covariates and third-grade ELA/math scores, matching the preferred specification. *p < 0.10, **p < 0.05, ***p < 0.01

Table A4: Bounding the Effect of Retention (CER-Optimal Bandwidth)

	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
	English Language Arts					Mathematics				
Reduced Form: Net Effect of Failing										
$\hat{\beta}_{net}$	0.041 (0.027)	0.075** (0.031)	0.011 (0.065)	0.093** (0.043)	0.024 (0.051)	0.046* (0.026)	0.142** (0.059)	0.070 (0.051)	0.092** (0.038)	-0.001 (0.065)
Decomposition										
$\hat{\lambda}_{ret}$	0.372*** (0.041)	0.355*** (0.051)	0.109 (0.071)	0.359*** (0.053)	0.174*** (0.065)	0.387*** (0.041)	0.358*** (0.062)	0.180*** (0.053)	0.352*** (0.048)	0.121 (0.077)
$\hat{\lambda}_{rem}$	0.027 (0.027)	0.057* (0.031)	0.001 (0.063)	0.062 (0.042)	0.001 (0.051)	0.032 (0.025)	0.136** (0.058)	0.038 (0.051)	0.037 (0.038)	-0.014 (0.063)
Effect of Failing on Probability Retained										
$\hat{\beta}_{fs}$	0.049*** (0.008)	0.058*** (0.011)	0.089*** (0.014)	0.115*** (0.017)	0.099*** (0.018)	0.040*** (0.009)	0.056*** (0.011)	0.093*** (0.011)	0.122*** (0.016)	0.100*** (0.020)
Estimated Lower and Upper Bounds for Effect of Retention										
τ_{ret}^{lower}	0.297*** (0.111)	0.371** (0.153)	0.114 (0.154)	0.330*** (0.116)	0.231* (0.136)	0.378*** (0.132)	0.257* (0.148)	0.385*** (0.121)	0.491*** (0.115)	-0.006 (0.713)
τ_{ret}^{upper}	0.832 (0.508)	1.301*** (0.453)	0.122 (0.885)	0.803** (0.387)	0.242 (0.562)	1.143** (0.525)	2.525** (0.983)	0.755 (0.614)	0.754** (0.361)	-0.006 (0.713)
N	9,672	6,210	5,582	4,691	3,875	8,395	6,525	7,811	5,100	3,286

Notes: Robustness check: CER-Optimal Bandwidth. The first panel reports RD estimates of the reduced-form effect of failing the summer IREAD on later test scores ($\hat{\beta}_{net}$). The second panel decomposes the reduced-form into outcomes for retained ($\hat{\lambda}_{ret}$) and remediated-while-promoted ($\hat{\lambda}_{rem}$) students, relative to passers. The third panel reports the effect of failing on retention probability ($\hat{\beta}_{fs}$). The bottom panel reports lower and upper bounds for the causal effect of retention. Upper bounds assume remediation is not harmful ($\tau_{rem} \geq 0$). Lower bounds assume failing students are negatively selected into retention. Bootstrap standard errors in parentheses for bound rows; robust standard errors for all other rows. Bandwidth is re-estimated using the CER-optimal selector, computed separately by grade and subject. Regressions include the full set of demographic covariates and third-grade ELA/math scores, matching the preferred specification. *$p < 0.10$, **$p < 0.05$, ***$p < 0.01$

Table A5: Bounding the Effect of Retention (Quadratic Forcing Variable)

	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
	English Language Arts					Mathematics				
Reduced Form: Net Effect of Failing										
$\hat{\beta}_{net}$	0.061**	0.079**	0.013	0.105**	0.040	0.053*	0.149**	0.071	0.104**	-0.002
	(0.031)	(0.037)	(0.065)	(0.048)	(0.051)	(0.031)	(0.059)	(0.063)	(0.044)	(0.076)
Decomposition										
$\hat{\lambda}_{ret}$	0.385***	0.343***	0.093	0.347***	0.232***	0.406***	0.341***	0.153***	0.350***	0.165**
	(0.038)	(0.043)	(0.062)	(0.050)	(0.050)	(0.035)	(0.055)	(0.053)	(0.046)	(0.070)
$\hat{\lambda}_{rem}$	0.050*	0.056	0.003	0.060	0.014	0.036	0.127**	0.034	0.039	0.020
	(0.029)	(0.036)	(0.059)	(0.044)	(0.047)	(0.028)	(0.055)	(0.059)	(0.041)	(0.069)
Effect of Failing on Probability Retained										
$\hat{\beta}_{fs}$	0.043***	0.053***	0.089***	0.113***	0.093***	0.045***	0.058***	0.080***	0.117***	0.089***
	(0.010)	(0.012)	(0.014)	(0.020)	(0.019)	(0.010)	(0.012)	(0.014)	(0.019)	(0.020)
Estimated Lower and Upper Bounds for Effect of Retention										
τ_{ret}^{lower}	0.328	0.487***	0.119	0.456**	0.294	0.420**	0.498**	0.494**	0.595***	-0.232
	(0.243)	(0.164)	(0.279)	(0.191)	(0.230)	(0.170)	(0.233)	(0.200)	(0.167)	(0.415)
τ_{ret}^{upper}	1.441**	1.488***	0.150	0.929**	0.431	1.171**	2.582**	0.887	0.894**	-0.026
	(0.722)	(0.562)	(0.995)	(0.461)	(0.654)	(0.568)	(1.022)	(0.844)	(0.387)	(0.955)
N	15,017	10,388	9,738	7,352	7,908	15,401	10,320	10,619	7,679	6,760

Notes: Robustness check: Quadratic Forcing Variable. The first panel reports RD estimates of the reduced-form effect of failing the summer IREAD on later test scores ($\hat{\beta}_{net}$). The second panel decomposes the reduced-form into outcomes for retained ($\hat{\lambda}_{ret}$) and remediated-while-promoted ($\hat{\lambda}_{rem}$) students, relative to passers. The third panel reports the effect of failing on retention probability ($\hat{\beta}_{fs}$). The bottom panel reports lower and upper bounds for the causal effect of retention. Upper bounds assume remediation is not harmful ($\tau_{rem} \geq 0$). Lower bounds assume failing students are negatively selected into retention. Bootstrap standard errors in parentheses for bound rows; robust standard errors for all other rows. Bandwidth is fixed at the baseline MSE-optimal bandwidth. Regressions include the full covariate set as in the preferred specification, but allow a quadratic relationship between the outcome and the forcing variable, interacted with failing status. *$p < 0.10$, **$p < 0.05$, ***$p < 0.01$

Table A6: Bounding the Effect of Retention (No Grade-3 Score Controls)

	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
	English Language Arts					Mathematics				
Reduced Form: Net Effect of Failing										
$\hat{\beta}_{net}$	0.048** (0.021)	0.049* (0.027)	0.042 (0.039)	0.104*** (0.033)	0.046 (0.035)	0.048** (0.023)	0.124*** (0.044)	0.050 (0.041)	0.055* (0.032)	0.030 (0.047)
Decomposition										
$\hat{\lambda}_{ret}$	0.313*** (0.032)	0.304*** (0.037)	0.073 (0.047)	0.321*** (0.042)	0.210*** (0.044)	0.280*** (0.037)	0.294*** (0.048)	0.068 (0.044)	0.263*** (0.040)	0.124** (0.053)
$\hat{\lambda}_{rem}$	0.034 (0.021)	0.026 (0.027)	0.033 (0.039)	0.063** (0.032)	0.016 (0.034)	0.034 (0.022)	0.104** (0.044)	0.032 (0.042)	0.006 (0.031)	0.027 (0.049)
Effect of Failing on Probability Retained										
$\hat{\beta}_{fs}$	0.055*** (0.006)	0.068*** (0.008)	0.095*** (0.010)	0.124*** (0.013)	0.117*** (0.013)	0.054*** (0.007)	0.064*** (0.008)	0.095*** (0.010)	0.130*** (0.013)	0.117*** (0.013)
Estimated Lower and Upper Bounds for Effect of Retention										
τ_{ret}^{lower}	0.288*** (0.080)	0.367*** (0.079)	0.118 (0.102)	0.394*** (0.083)	0.265*** (0.096)	0.300*** (0.088)	0.417*** (0.088)	0.223** (0.087)	0.391*** (0.082)	0.051 (0.101)
τ_{ret}^{upper}	0.878** (0.365)	0.727* (0.378)	0.436 (0.455)	0.841*** (0.318)	0.389 (0.373)	0.896** (0.380)	1.948*** (0.638)	0.524 (0.462)	0.428* (0.246)	0.256 (0.451)
N	15,148	10,507	9,789	7,435	7,994	15,531	10,433	10,671	7,764	6,841

Notes: Robustness check: No Grade-3 Score Controls. The first panel reports RD estimates of the reduced-form effect of failing the summer IREAD on later test scores ($\hat{\beta}_{net}$). The second panel decomposes the reduced-form into outcomes for retained ($\hat{\lambda}_{ret}$) and remediated-while-promoted ($\hat{\lambda}_{rem}$) students, relative to passers. The third panel reports the effect of failing on retention probability ($\hat{\beta}_{fs}$). The bottom panel reports lower and upper bounds for the causal effect of retention. Upper bounds assume remediation is not harmful ($\tau_{rem} \geq 0$). Lower bounds assume failing students are negatively selected into retention. Bootstrap standard errors in parentheses for bound rows; robust standard errors for all other rows. Bandwidth is fixed at the baseline MSE-optimal bandwidth. Regressions include demographic covariates but omit third-grade ELA and math score controls. *p < 0.10, **p < 0.05, ***p < 0.01

Table A7: Bounding the Effect of Retention (No Covariates)

	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
	English Language Arts					Mathematics				
Reduced Form: Net Effect of Failing										
$\hat{\beta}_{net}$	0.049** (0.021)	0.047* (0.027)	0.056 (0.041)	0.116*** (0.033)	0.052 (0.035)	0.051** (0.023)	0.119*** (0.043)	0.046 (0.041)	0.058* (0.032)	0.035 (0.046)
Decomposition										
$\hat{\lambda}_{ret}$	0.311*** (0.033)	0.292*** (0.038)	0.079* (0.047)	0.333*** (0.042)	0.222*** (0.044)	0.289*** (0.038)	0.292*** (0.050)	0.066 (0.043)	0.271*** (0.040)	0.133** (0.053)
$\hat{\lambda}_{rem}$	0.035* (0.021)	0.026 (0.027)	0.051 (0.041)	0.077** (0.032)	0.024 (0.034)	0.038* (0.023)	0.099** (0.042)	0.030 (0.042)	0.010 (0.032)	0.032 (0.048)
Effect of Failing on Probability Retained										
$\hat{\beta}_{fs}$	0.055*** (0.006)	0.067*** (0.008)	0.096*** (0.010)	0.124*** (0.013)	0.117*** (0.012)	0.054*** (0.007)	0.064*** (0.008)	0.096*** (0.009)	0.129*** (0.013)	0.117*** (0.013)
Estimated Lower and Upper Bounds for Effect of Retention										
τ_{ret}^{lower}	0.292*** (0.079)	0.339*** (0.083)	0.110 (0.104)	0.390*** (0.084)	0.255*** (0.098)	0.283*** (0.089)	0.397*** (0.088)	0.202** (0.082)	0.386*** (0.082)	0.055 (0.100)
τ_{ret}^{upper}	0.892** (0.374)	0.699* (0.388)	0.589 (0.492)	0.938*** (0.322)	0.440 (0.391)	0.953** (0.401)	1.849*** (0.602)	0.482 (0.451)	0.454* (0.244)	0.298 (0.433)
N	15,229	10,587	9,892	7,509	8,097	15,612	10,513	10,788	7,840	6,929

Notes: Robustness check: No Covariates. The first panel reports RD estimates of the reduced-form effect of failing the summer IREAD on later test scores ($\hat{\beta}_{net}$). The second panel decomposes the reduced-form into outcomes for retained ($\hat{\lambda}_{ret}$) and remediated-while-promoted ($\hat{\lambda}_{rem}$)students, relative to passers. The third panel reports the effect of failing on retention probability ($\hat{\beta}_{fs}$). The bottom panel reports lower and upper bounds for the causal effect of retention. Upper bounds assume remediation is not harmful ($\tau_{rem} \geq 0$). Lower bounds assume failing students are negatively selected into retention. Bootstrap standard errors in parentheses for bound rows; robust standard errors for all other rows. Bandwidth is fixed at the baseline MSE-optimal bandwidth. Regressions include no demographic or baseline-achievement covariates; school fixed effects and the forcing variable are retained in all specifications, as noted above. *$p < 0.10$, **$p < 0.05$, ***$p < 0.01$

Table A8: Bounds on the Effect of Retention: Same-Year Comparisons

	YEAR 2		YEAR 3		YEAR 4		YEAR 5	
	ELA	Math	ELA	Math	ELA	Math	ELA	Math
τ_{ret}^{lower}	15.858** (7.163)	35.236*** (9.001)	2.145 (10.516)	32.163** (13.807)	25.442*** (7.654)	39.548*** (6.178)	44.393*** (14.293)	11.464 (13.913)
τ_{ret}^{upper}	22.131 (34.924)	125.142** (57.581)	14.097 (75.544)	75.371 (82.070)	75.520*** (24.878)	58.614*** (21.277)	89.016** (34.524)	45.827 (73.664)

Notes: Bootstrapped standard errors (300 replications, clustered by grade-3 school) in parentheses. *$p < 0.10$, **$p < 0.05$, ***$p < 0.01$. Outcomes are scores on the ILEARN test measured on the test's vertical scale. For context, though the specific amount differs by grade and year, for the population of test-takers in 2019 a standard deviation within a given grade is about 75 in both ELA and math. Sample restricted to test-score observations from the 2018–19 school year or later, after the state changed from ISTEP+ to ILEARN. Students from all entering third-grade cohorts are included, but only their post-2018–19 (ILEARN) outcomes are used, since ISTEP+ and ILEARN scores are not on a comparable vertical scale. Years refer to years since grade 3 entry.

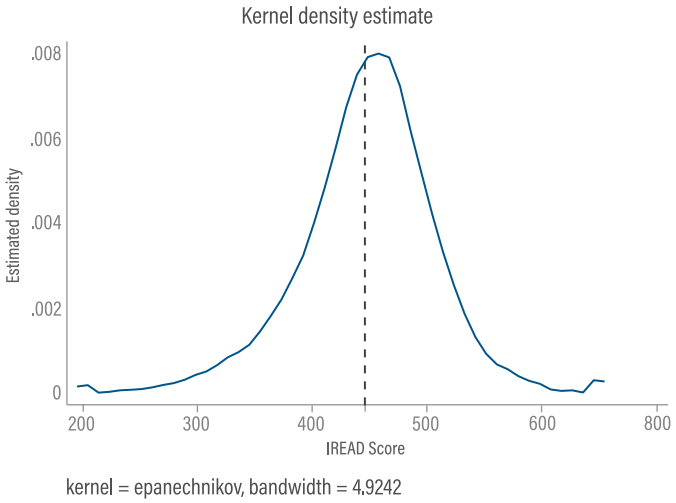
Table A9: Third-Grade Reading Retention Policy Characteristics by State

	Timeline			Policy Type	Promotion Pathways and Exemptions							
State	First	Expand	Limit	Hard Gate	IEP	ELL	Prev. Ret.	Alt. Assess.	Port.	Sum. Sch.	Discr. Override	Remed.
Alabama	2023	0	0	1	1	1	1	1	1	1	0	1
Arizona	2013	0	1	0	1	1	0	1	0	1	0	1
Delaware	2012	0	0	0	1	0	1	1	1	1	1	1
Florida	2002	1	0	1	1	1	1	1	1	1	0	1
Indiana	2011	1	1	1	1	1	1	0	0	0	0	1
Iowa	2012	0	0	0	1	1	1	1	1	1	0	1
Michigan	2021	0	1	0	1	1	1	1	1	1	1	1
Mississippi	2014	1	1	1	1	1	1	1	1	1	0	1
North Carolina	2013	1	0	1	1	1	1	1	1	1	0	1
Ohio	2013	0	1	0	1	1	1	1	0	1	1	1
Oklahoma	2013	1	1	0	1	1	1	1	1	1	1	1
South Carolina	2014	1	0	0	1	1	1	1	1	1	1	1
Tennessee	2022	1	1	0	1	1	1	1	0	1	1	1
DC	2014	0	0	1	0	0	0	0	0	0	0	1

Notes: Sample includes states with an identifiable statewide third-grade test-based reading retention policy that was operational at any point during 2016–2024. Coding reflects the most expansive version of each state's policy implemented during the sample period; states that subsequently weakened or repealed provisions are coded according to the pre-limitation policy. States are excluded when available evidence does not establish an operational statewide third-grade test-based retention gate during the sample period, including cases involving repeal prior to implementation, absence of a comparable statewide retention statute, or policy designs that do not meet the study's coding criteria. Entries are binary indicators except First, which reports the first fall cohort actually subject to the state's retention policy; this may postdate statutory enactment where implementation was delayed or suspended. Hard Gate indicates that failure to meet the statutory reading requirement creates a default retention consequence absent a qualifying statutory exemption. Discr. Override indicates additional promotion authority through mechanisms such as placement committees, superintendent or district-level appeals, or parental written waivers. Remed. indicates statutory requirements for intensive reading intervention or remediation following failure to meet the reading benchmark. Authors' coding is based on state statutes, state department of education guidance, and Education Commission of the States summaries.

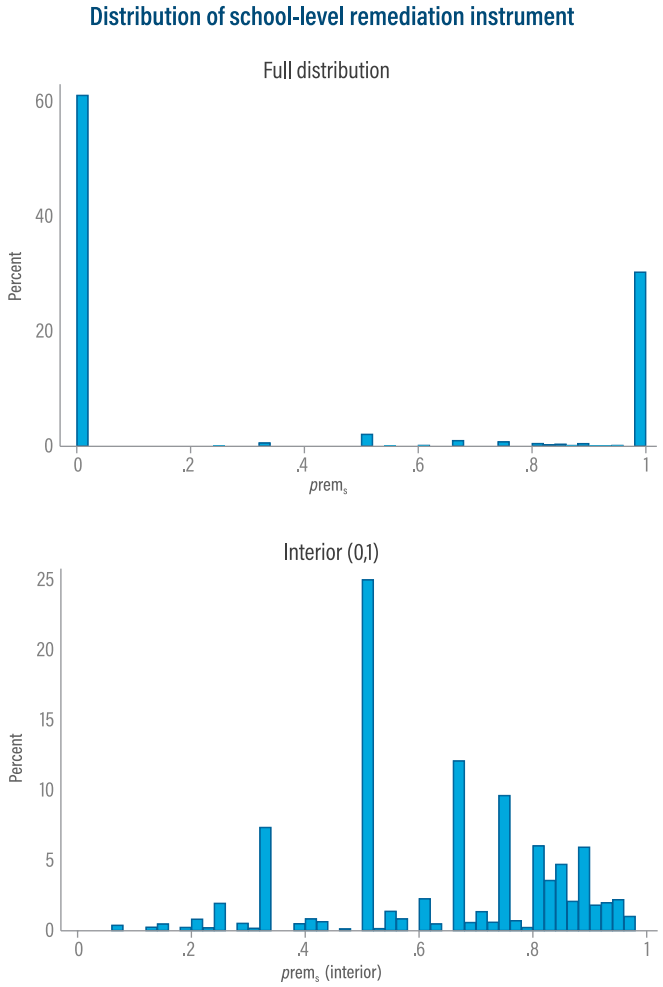
B. Figures

Figure B1: Distribution of Summer IREAD Scores



Notes: Kernel density of summer IREAD scores. The vertical line denotes the passing score (446). P-value for test for discontinuity at the passing score is 0.991.

Figure B2: Histogram for Lagged % Remediated



Notes: Each panel displays the distribution of remediation assignment rates across schools (left) and the interior distribution (right), where each bar represents the share of institutions with a given percentage of students assigned to remediation.

C. Instrumental Variables Approach

In this Appendix, we apply an IV approach as an alternative identification strategy to measure the effect of retention. We use the proportion of failing students within a school the prior year who were remediated in the fourth grade, Z_{prem}, to instrument for whether a failing student is retained. Formally, within a sample restricted to students who failed the summer IREAD test, we use 2SLS to estimate:

$$Ret_i = \zeta + \gamma X_i + f(a_i) + \delta_{fs} Z_{prem} + \mu_i \quad (10)$$

$$y_i = \zeta + \gamma X_i + f(a_i) + \delta_{iv} Ret_i + \mu_i \quad (11)$$

We are primarily interested in δ_{iv}, which we interpret as the causal effect of retention relative to remediation. The result is a LATE that applies to compliers who failed the summer IREAD test and were retained because they attended a school that the previous year had a lower propensity for remediating students who failed the test.

Data limitations make it so that we can only apply this strategy to a much more restricted sample of students than we are able to include in our primary analyses. The sample is limited to students who entered the third grade in either 2017–18 or 2018–19. We are also only able to apply this strategy to measure the effect of retention on sixth and seventh grade test scores.

The issues that lead to greater data restrictions arise because the instrument represents the school's propensity to remediate failing students in the prior year, and they are amplified by the pandemic. We lose students from the first entering third grade class (2016–17) because we do not observe a lagged remediation rate that year. As in our primary analysis, we exclude students who entered third grade in 2019–20 because they did not take IREAD. For this analyses we also lose students who entered third grade in 2020–21 because schools did not remediate students in 2019–20, and thus we do not observe a lagged exemption rate for them. To generate a direct comparison, to be included in the sample measuring the effect of retention in a given later grade we require that we observe test scores in that grade for both retained and remediated students who are then cumulatively promoted. Because students were not administered summative tests in 2019–20, and they were also not retained (and thus not given an exemption to be remediated) that year, we do not observe both retained and remediated groups of students from either the 2017–18 or 2018–19 cohorts until the 6th grade.

Probing the Identifying Assumptions

The exclusion restriction necessary for $\hat{\delta}_{iv}$ to identify the LATE for retention under the policy requires that, conditional on the covariates, Z_{prem} affects outcomes only by influencing the proba-bility that a failing student is remediated rather than retained. The primary threat to identification is the potential that a school's propensity to exempt students is related to the impact it has on later student outcomes in ways not accounted for by the covariates.

Table C10 investigates balance of observed characteristics on Z_{prem} for students included in the sample. It is encouraging that Z_{prem} is balanced along several key observed characteristics including percent of students who are Black, white, and eligible for free or reduced price lunch. However, students in schools with higher propensity to remediate failing students are significantly more likely to be Hispanic and had significantly lower scores on the IREAD

as well as spring of third grade ELA and math tests. Our analyses control for these factors directly, which we expect should account for major differences across schools. However, their imbalance is concerning because it opens the possibility that the schools might also differ in unobserved ways that could influence later student outcomes, even after controlling for demographics and third grade scores.

Table C10: Measuring Covariate Balance on Zprem

	Male	Black	Hispanic	White	FRL	G3 ELA	G3 Math	IREAD
Z_{prem}	0.017 (0.030)	-0.021 (0.049)	0.046** (0.018)	-0.047 (0.047)	-0.039 (0.025)	-0.177*** (0.043)	-0.102** (0.043)	-6.563*** (2.165)
N	3893	3893	3893	3893	3893	3892	3892	3893

Notes: Table reports estimates from regressions that use baseline student characteristics and summer IREAD scores as outcomes. The independent variable, Zprem, is the lagged proportion of students in the school who failed the summer IREAD and were remediated rather than retained in the previous year. Regressions are estimated on the sample of students who failed the summer IREAD. Robust standard errors are reported in parentheses. *$p < 0.10$, **$p < 0.05$, ***$p < 0.01$

We provide a direct test of the exclusion restriction by estimating regressions that modify the RD approach to examine whether the relationship between Z_{prem} and later achievement differs for students just above and just below the promotion threshold. The premise of this exercise is that if Z_{prem} captures broader differences in school quality or unobserved instructional inputs then it should predict outcomes for both failing and passing students, but if Z_{prem} operates solely by reallocating failing students across post-failure pathways it should only affect the outcomes for students who failed the test. Students who passed summer IREAD are not sorted to be retained or remediated, and therefore provide a natural placebo group.

Specifically, within the RD sample we used to estimate Equation 4 to measure the net effect of failing the summer IREAD and later outcomes in the Bounding the Effect of Policy-Induced Retention section, we estimate:

$$y_i = \zeta + \gamma X_i + f(a_i) + \lambda Z_{prem} + \beta_{net}\mathbb{1}(a_i < 0) + \delta(\mathbb{1}(a_i < 0) \times Z_{prem}) + \epsilon_i \quad (12)$$

where all variables are as previously defined. Under the classical RD assumptions $\hat{\beta}_{net}$ identifies the average causal effect of failing the summer IREAD for students who attend schools where all students who failed the test were retained. We are primarily interested in λ, which measures the relationship between Z_{prem} and the outcome for

students who passed the summer IREAD. If the exclusion restriction necessary to identify $\hat{\delta}_{iv}$ holds, we would expect to observe $\hat{\beta}_{net} > 0$ and $\lambda = 0$. We would also anticipate finding $\delta < 0$, which would imply that the effect of failing the summer IREAD increases as a higher proportion of failing students are retained rather than remediated.

The results of this exercise, reported in Table C11, lend support to the exclusion restriction in our context. The top panel reports results from models that control only for summer IREAD score (the forcing variable), and the bottom panel reports results from models that also control for student demographics and third grade test scores. For each analysis the coefficient on Z_{prem} is near-zero and insignificant, implying that a school's propensity to remediate students who the prior year failed the summer IREAD is unrelated to the outcomes of students who passed the test. Further, the coefficient on the interaction term is negative and significant except for seventh grade ELA, implying that the effect of failing the summer IREAD on later outcomes becomes more negative as the proportion of failing students within the school who were remediated increases. Interestingly, in most cases the magnitude of the negative interaction coefficient mirrors the positive coefficient for failing the test, suggesting that the effect of failing the test entirely dissipates within schools that the previous year remediated all students who failed the summer IREAD.

Table C11: Relationship Between Z_{prem} and Outcomes on Either Side of Passing Threshold

	ELA		Math	
	Grade 6	Grade 7	Grade 6	Grade 7
Only Control for Forcing Variable				
Fail	0.215*** (0.067)	0.218*** (0.070)	0.210*** (0.064)	0.228*** (0.068)
Z_{prem}	0.001 (0.045)	0.002 (0.042)	0.008 (0.041)	0.017 (0.042)
Fail × Z_{prem}	-0.197*** (0.064)	-0.094 (0.069)	-0.204*** (0.062)	-0.125* (0.067)
N	4470	3949	4893	4118
Control for Forcing Variable and Other Covariates				
Fail	0.220*** (0.065)	0.232*** (0.067)	0.213*** (0.063)	0.242*** (0.065)
Z_{prem}	0.002 (0.042)	-0.012 (0.041)	0.008 (0.038)	0.001 (0.041)
Fail × Z_{prem}	-0.198*** (0.062)	-0.108 (0.068)	-0.202*** (0.061)	-0.135** (0.066)
N	4423	3907	4837	4076

Notes: Table reports estimates from regressions examining whether the relationship between the instrument (Z_{prem}) and student outcomes differs for students who failed versus passed the summer IREAD. Z_{prem} is the lagged proportion of students in the school who failed the summer IREAD and were remediated rather than retained in the previous year. The sample includes students with summer IREAD scores within the estimation bandwidth. The indicator (Fail) denotes students who scored below the passing threshold on the summer IREAD. Regressions control for the forcing variable, and the bottom panel additionally includes student demographic characteristics and baseline third-grade achievement. Robust standard errors clustered by entering third grade school are reported in parentheses. *$p < 0.10$, **$p < 0.05$, ***$p < 0.01$

Results

Table C12 reports the first-stage and IV estimates for the LATE of retention relative to remediation on sixth and seventh grade test scores for students who failed the summer IREAD test. Failing students who were retained because they attended a school that remediated a smaller proportion of its failing students the prior year on average scored between 0.44σ and 0.64σ in the sixth and seventh grade than if they had been remediated. The estimates are significant, though measured imprecisely.

Table C12: First Stage and IV Estimates: Effect of Retention on Later Test Scores

	ELA		Math	
	Grade 6	Grade 7	Grade 6	Grade 7
Panel A: First Stage (Dep. var.: Retained in Grade 3)				
School Propensity to Remediate	-0.294*** (0.036)	-0.286*** (0.037)	-0.296*** (0.037)	-0.290*** (0.037)
Panel B: IV Second Stage (Dep. var.: Test Score)				
Retention LATE	0.637*** (0.131)	0.443*** (0.130)	0.514*** (0.142)	0.457*** (0.129)
N	3,892	3,798	3,893	3,792

Notes: Panel A reports first-stage estimates relating school remediation propensity to third-grade retention. Panel B reports two-stage least squares estimates of the effect of retention on later achievement. The instrument is the school's lagged propensity to remediate eligible students. The sample includes non-exempt students failing the summer IREAD from the 2018 and 2019 third-grade cohorts. All specifications include student covariates and cohort fixed effects. Standard errors, clustered at the initial third-grade school, are reported in parentheses.

ENDNOTES

1 Table A9 in the Online Appendix characterizes state policies.

2 We focus on test-based promotion policies for students in early elementary grades—typically third grade. Some states have similar promotion requirements for students in middle school grades.

3 States differ substantially in how strictly they apply promotion requirements. Florida's third-grade promotion policy, for example, defines good-cause exemptions narrowly and historically retained a large share of students who failed the promotion exam (Greene and Winters, 2009; Schwerdt et al., 2017). By contrast, in Texas students who fail the state's reading assessment are frequently promoted following review by grade placement committees and provided mandated remediation (Jacob and Lefgren, 2004). In Colorado and North Carolina, portfolio pathways and discretionary exemptions similarly shift the policy toward conditional promotion with supplemental services rather than mandatory retention (Colorado Department of Education, 2023; North Carolina Department of Public Instruction, 2023). Policymakers often debate whether to soften retention mandates by expanding exemption categories or substituting remediation in higher-grade classrooms. For example, Michigan recently repealed its retention requirement but kept literacy supports for failing students in place (Michigan Department of Education, 2024).

4 In recent years the state adjusted the policy first to allow and now requiring schools to first administer IREAD to students in the second grade.

5 There are small groups of non-compliers who passed the IREAD but were nonetheless retained. We focus this discussion on the typical student pathways to enhance clarity.

6 Indiana first adopted the test-based promotion policy for students entering the third grade in 2011–12. However, prior to 2016–17 the state did not give schools discretion to promote students who failed the summer IREAD. Though there were non-compliers who failed the test and were nonetheless promoted, the state did not require such students receive remedial instruction while attending the fourth grade. Further, for students who entered the third grade under this initial iteration of the policy, about 38% of students who repeated the third grade then skipped the fourth grade, thus returning to their cohort. The different structure of the first iteration of this policy has interesting implications that we plan to investigate in a separate paper. In this paper, we focus on distinguishing the effect of retention relative to remediation under the policy, which is most directly associated to similar policies found across the U.S.

7 In the context of Michigan's similar policy, Berne et al. (2026) report survey evidence that schools did provide additional supports for struggling students, and that the type of supports varied.

8 A passing score was 446.

9 Using these data, Winters (2026) finds that the salient threat of retention under the policy from failing the spring IREAD led to measurable improvements on the third grade summative tests administered only four to six weeks later. However, these patterns do not threaten the analysis in this paper because all students who take the summer IREAD by definition also failed the spring IREAD and thus were subjected to this incentive effect. Similarly, students take the summer IREAD after they have completed summer school.

10 There is no consensus in the literature on whether to prefer the same-grade or same-year approach to measure the effect of retention.

11 Both ILEARN and ISTEP+ use a vertically aligned scale such that a score is designed to represent the same level of ability regardless of the grade level of the test. But scaled scores are not comparable across the tests. When making same-year comparisons reported in the Online Appendix, we thus restrict analyses to only students taking the ILEARN. The resulting reduction in sample size reduces available power, which contributes to imprecision in the estimates.

12 The proportion remediated would differ from $(1 - \alpha)$ if there are non-compliers. There are no non-compliers who failed the test because the policy required schools to provide remedial services to all students who failed the summer IREAD but were nonetheless promoted. Table 1 shows that only a trivial proportion of students who passed the summer IREAD were retained in the third grade.

13 As a robustness test, Online Appendix C describes and reports the results from applying an alternative IV strategy to estimate the effect of retention relative to remediation on later student test scores. These analyses use across-school differences in the propensity to retain students who failed the IREAD the previous year to instrument for whether a failing student is retained. As we describe in the Appendix, we can only use these analyses to estimate the impact of retention relative to remediation when students are in the sixth or seventh grade, or three or four years after their initial third grade year. We present evidence that the instrument satisfies the exclusion restriction—specifically, we demonstrate that the lagged percentage of failing students in the school who were remediated predicts this year's scores for students who failed the summer IREAD but not for students who passed and thus were not sorted across the retention or remediation treatments. The IV estimates measure the effect of retention for a different complier population and thus are not directly comparable to our primary results. We focus on the bounding approach in the main paper because the RD approach relies on weaker identifying assumptions than does IV that leverages arguably exogenous variation resulting from a natural experiment. However, that we find similar estimates from analyses that leverage a very different source of identifying variation lends credibility to the patterns we observe in our primary analysis.

14 There are cases where parents might advocate for their child to receive an intervention that is linked to additional resources. For example, a parent might advocate for their child to be classified with a disability so that they can receive additional services. Though grade retention does allocate some resources to a child, the cost of an additional year of schooling and disruption of peer groups is likely high enough that parents do not prefer for their child to repeat a grade. In addition, the "red shirting" phenomenon in which parents advocate to hold their child back for an additional year in Kindergarten is not typically thought to apply in later grades.

15 Our argument that failing students are negatively sorted to be retained rather than remediated similarly implies that $\hat{\lambda}_{ret} \leq \tau_{ret}$. We do not use λ_{ret} directly as the lower bound because doing so would treat pathway assignment as the object of estimation rather than using the RD-induced variation in treatment exposure. Our preferred bound combines the causal reduced-form effect of failing the test, the causal first-stage effect on retention, and an empirically grounded upper bound on the remediation effect, with uncertainty computed by bootstrapping the entire system. Our preferred lower-bound estimates are in all cases similar in magnitude but have substantially larger standard errors than the respective $\hat{\lambda}_{ret}$.

16 Table A1 reports results from similar regressions that control for all covariates included in the primary regressions and regressions that control for all covariates except for third grade summative ELA and math test scores. These regressions produce results similar to the analyses that control for the forcing variable and school fixed-effects.

17 Figure B2 illustrates that propensity to remediate varies by school. It is possible that school quality is correlated with its propensity for retaining students who fail the summer IREAD. We address this problem in our main analyses by including fixed effects for the school the student attended in their initial third grade year, thus holding school quality constant.

18 One potential concern is that spring of third grade ELA and math scores could be "too good" of a control. Perhaps schools are more likely to retain students with lower spring of third grade math and ELA scores than those with higher scores, but when comparing observationally similar students with the same third grade ELA and math scores are more likely to retain the student with more advantaged characteristics that are not observed in the data. The results in Tables A6 and A7 address this potential problem by showing that analyses that remove the controls for third grade test scores and analyses that include no covariates produce results similar to our primary estimates.

19 This pattern is similar to those found by prior studies that apply the fuzzy RD design to estimate the effect of retention on later outcomes within the context of policies where failing the test only marginally increases the likelihood a student is actually retained (Berne et al., 2026; Slungaard Mumma and Winters, 2023).

20 Because they do not use comparable vertical scales, we restrict these same-year to only observations in 2018-19 or later, after the state changed from the ISTEP+ to the ILEARN test.

SUGGESTED CITATION

Francis, A., Winters, M. A. (2026). *Identifying Retention Effects in the Presence of Remediation-Based Exemptions* (Working Paper 2026–3). Wheelock Educational Policy Center. Available at wheelockpolicycenter.org.

OUR MISSION

The Wheelock Educational Policy Center (WEPC) conducts and disseminates rigorous, policy-relevant education research in partnership with local, state, and federal policymakers and stakeholders to improve educational opportunities and holistic outcomes for underserved students.

www.wheelockpolicycenter.org
wheelockpolicy@bu.edu

BU

Boston University Wheelock College of Education & Human Development
Wheelock Educational Policy Center

WEPC
WHEELOCK EDUCATIONAL POLICY CENTER