Grade Repetition and Household Responses in a Low Income Setting

Tahir Andrabi[*], Ethan Matlin[†], Gabrielle Vasey[‡]

January 8, 2026

Abstract

Millions of children in low- and middle-income countries repeat a grade each year, yet little is known about how grade repetition shapes household behavior and student outcomes. In many such settings, promotion decisions are discretionary and made informally by teachers rather than governed by standardized rules. We study grade repetition in Pakistan using matched panel data that link students to their households, teachers, and schools. We interpret retention as an organic information signal that conveys new information about a child's academic ability to parents and students. Our empirical strategy exploits variation in retention decisions arising from differences in promotion thresholds and noise in teachers' assessments of ability, controlling flexibly for prior achievement and student characteristics. We find that retention lowers test scores by 0.27 to 0.44 standard deviations and increases dropout by 7 percentage points. Parents respond by revising downward their beliefs about their child's ability and reducing both short- and long-run educational investments, while retained students become less confident in the returns to academic effort. In contrast, we find little evidence that teachers treat repeaters differently after the retention decision. These results suggest that grade repetition operates as a powerful information signal that shapes household decisions and contributes to persistent educational disadvantage.

Keywords: Education Policy, Achievement, Grade Retention, Parental Investments.

[*] Pomona College

[†] Harvard University

[‡] Concordia University

1 Introduction

Parents play a central role in shaping their children's educational outcomes through decisions about enrollment, attendance, study time, tutoring, and long-term investments. These decisions are guided by beliefs about a child's ability and the returns to effort and schooling, a central insight in models of parental investment and intergenerational transmission (Becker and Tomes, 1976; Tomes, 1981; Behrman et al., 1994; Todd and Wolpin, 2003; Cunha and Heckman, 2007). A growing empirical literature shows that parents often hold biased or incomplete beliefs, and that providing new information can substantially alter expectations and educational choices (Nguyen, 2008; Hastings and Weinstein, 2008; Jensen, 2010; Kraft and Rogers, 2015; Rogers and Feller, 2018; Bergman, 2021). Much of this evidence comes from designed information interventions, such as report cards or returns-to-schooling experiments, that deliberately shift the information environment faced by households.

Schools, however, also generate information signals outside the context of explicit information policies. Grades, rankings, and promotion decisions convey meaningful information about student performance, even when they are not designed as explicit information policies. We know far less about how households interpret and respond to these organic school-generated signals, particularly in settings where performance standards are informal and decisions rely heavily on teacher judgment. Understanding how such signals affect beliefs and behavior is important, as they may shape educational trajectories in persistent ways.

Grade repetition is one of the most salient organic signals produced by schools. Being retained conveys information both to the student and to the parent that the child's performance falls below an implicit standard. Retention is also a policy decision with direct academic consequences. Despite its importance, grade repetition remains relatively understudied, especially in low- and middle-income countries where promotion decisions are often discretionary rather than governed by centralized rules or standardized assessments.

Each year, approximately 24 million children repeat at least one grade during primary school, and repetition rates in low- and middle-income countries are often five to ten times higher than in high-income countries (UNESCO 2024).[1] Existing evidence on the impacts of retention comes primarily from high-income settings, particularly the United States, where retention is typically governed by automatic promotion policies and standardized testing (e.g., Jacob and Lefgren, 2004, 2009; Eide and Showalter, 2001; Eren et al., 2022; Figlio and Özek, 2020; Borghesan et al., 2022). In contrast, grade repetition in many low-income settings reflects informal, subjective judgments by teachers rather than centralized rules. A small but growing literature from low- and middle-income countries studies the consequences of grade repetition and promotion policies using policy variation and quasi-experimental designs, including evidence from Brazil, Uruguay, Colombia, and Mexico (Gomes-Neto and Hanushek, 1994; Manacorda, 2012; Ferreira Sequeda et al., 2018; Cabrera-Hernandez, 2022). Work from Senegal finds that repetition increases dropout risk (Glick and Sahn, 2010) and studies from China and Pakistan document the discretionary nature of promotion decisions in the absence of formal thresholds (Hu and Hannum, 2020; King et al., 2016). Despite this progress, relatively little is known about how students, parents, and teachers interpret and respond to retention decisions, particularly in settings where promotion is highly discretionary.

This paper studies grade repetition in Pakistan, where promotion decisions are made at the discretion of individual teachers and are not governed by centralized rules or standardized thresholds (Chohan and Qadir, 2011). Using matched panel data from the Learning and Educational Achievement in Punjab Schools project, we follow a single cohort of students over four years. The data link students to their households, teachers, and schools, and include repeated test scores on a common item response theory scale, detailed parental beliefs and educational investments, and teacher assessments measured before and after the retention decision.

We find that grade repetition leads to large and persistent declines in academic performance. In the year following retention, repeaters score between 0.27 and 0.44 standard

[1] Data from 2022 http://data.uis.unesco.org.

deviations lower across three core subjects and are 7 percentage points more likely to drop out than their peers. These effects persist for at least two years and do not vary meaningfully by gender. Beyond academic outcomes, we document substantial behavioral responses to the retention decision. Parents revise downward their beliefs about their child's ability and reduce both short- and long-run educational investments. Retained students become less likely to believe that effort in school leads to improved life outcomes. In contrast, we find little evidence that teachers treat repeaters differently after the retention decision, conditional on achievement.

To guide our empirical strategy, we develop a simple conceptual framework in which teachers observe a noisy signal of student ability and retain students whose perceived ability falls below a subjective threshold. In the absence of formal promotion guidelines, these thresholds vary across teachers and schools, and measurement error in perceived ability introduces quasi-random variation in retention within classrooms. Our empirical strategy exploits this variation, using detailed controls for lagged achievement, teacher ratings, and student characteristics, and comparing students within the same school to account for systematic differences in promotion standards across schools.

This paper makes three contributions. First, we provide new evidence on the consequences of grade repetition in a low-income country where promotion decisions are informal and teacher-driven, a setting that reflects how most repetition decisions occur globally. Second, we contribute to a growing literature on parental belief formation and educational investment by showing that organic school-generated signals can powerfully shift expectations and behavior, even in the absence of explicit information interventions. Third, by leveraging rich linked panel data, we move beyond academic effects to document how different actors respond to retention, highlighting household belief updating as a key mechanism through which early academic setbacks may persist.

The paper proceeds as follows. Section 2 describes the institutional setting and the LEAPS data. Section 3 outlines our empirical strategy and presents descriptive evidence on retention decisions. Section 4 reports the main results, including impacts on academic outcomes, household responses, student beliefs, and teacher behavior. Section 5 discusses robustness checks and threats to identification. Section 6 concludes.

2 Data and Setting

This paper uses data from the Learning and Education Achievement in Punjab Schools (LEAPS) project, a longitudinal panel survey following primary school students, their households, teachers, and schools in rural Punjab, Pakistan. The LEAPS dataset has been used extensively to study schooling, learning, and household decision-making in Pakistan, and its sampling design and data collection procedures are described in detail in prior work (e.g., Andrabi et al. (2007); Andrabi et al. (2008); Andrabi et al. (2013); Andrabi et al. (2017)). We therefore provide only a brief overview here and focus on the institutional features and data elements most relevant for studying grade retention and household responses.

2.1 Setting

The setting for this paper differs in important ways from that of high-income countries. Education levels in our sample are in general low: 72% of mothers did not complete primary school and 64% received no formal education. Mothers with no formal education spend approximately zero minutes per day on children's educational activities at home. Learning levels for students are also low: by the end of 3rd grade, just over 50 percent of children have mastered the Mathematics curriculum for 1st grade (Andrabi et al. (2007)).

Schools in this setting also look very different from those in high income countries on average. Classes and schools are relatively small: the median class has 13 students and the median school has 119 students (across multiple grades). Most schools operate only a single Grade 4 classroom: 84 percent of schools in our estimation sample have just one Grade 4 teacher. Classrooms have large age-ranges of students: the average classroom has a three to four year gap between the youngest and oldest student. There is a large low-cost private school sector, making up 40% of the schools in our sample and costing only a dime a day.

Grade repetition after fourth grade is common in this setting (approximately 9% of students in our sample), and promotion decisions at this stage are largely discretionary. While Pakistan has automatic promotion policies in government schools through Grade 3, these rules no longer apply beginning in Grade 4.[2] As a result, promotion from Grade 4 to Grade 5 is the earliest grade transition at which a substantial share of students are formally retained, making it a natural margin at which to study repetition decisions and their consequences (Chohan and Qadir (2011)). Because these retention decisions occur relatively early in students' schooling careers, any effects of repetition have the potential to shape subsequent educational trajectories over many years.

Survey evidence from LEAPS indicates that classroom teachers play a central role in these decisions. In the school survey, 97% of schools report that the classroom teacher plays either an important or very important role in determining whether a student is promoted, and 98% cite academic performance as the primary reason for grade repetition. However, there are no centrally determined achievement thresholds governing promotion at this grade, and schools do not rely on standardized test scores when making retention decisions. Instead, teachers base promotion decisions on their own assessments of student performance, implying substantial scope for discretion and heterogeneity across classrooms.

This institutional environment contrasts sharply with high-income settings, where grade retention is relatively rare and typically governed by explicit rules or test-based thresholds (Eisemon (1997)).[3]

[2] In government schools in Pakistan, automatic promotion applies through Grade 3. In Grade 4, there is no centrally determined grade retention policy. Standardized public examinations, such as the Punjab Examination Commission (PEC) exam in Grade 8, begin to influence promotion decisions in government schools at that stage. In contrast, private schools are largely unregulated and continue to rely on teacher or school discretion even in later grades.

[3] Countries with clearer and properly enforced standards for repetition tend to have lower repetition rates than those with no or loosely enforced state-mandated criteria. Repetition rates are highest in Sub-Saharan Africa (approximately 22% of primary students repeat a grade at any given time) and also high in the Middle East and North Africa (12%), while in developed countries only 1–5% of students repeat a grade (Eisemon, 1997).

2.2 Description of Data

LEAPS is a five-year panel survey of schools, school-children, and their households in the Attock, Faisalabad, and Rahim Yar Khan districts of Punjab province,[4] conducted between 2004 and 2011. A key feature of the data for our analysis is that it follows the same cohort of students over time, allowing us to observe academic outcomes, household responses, and beliefs both before and after the grade retention decision.

In the first year of the survey, 13,735 students in 804 schools offering third grade instruction were sampled. Of these schools, 485 (60%) were government schools and 319 private (40%).[5] At sampled schools, every third-grade student was surveyed, and these initial students are followed over time without adding new students to the sample (see Figure 1). Our analysis focuses on grade retention that occurs after the second year of the survey, which corresponds to promotion from Grade 4 to Grade 5 for the majority of students. Our main outcomes are measured in the third year of the survey.

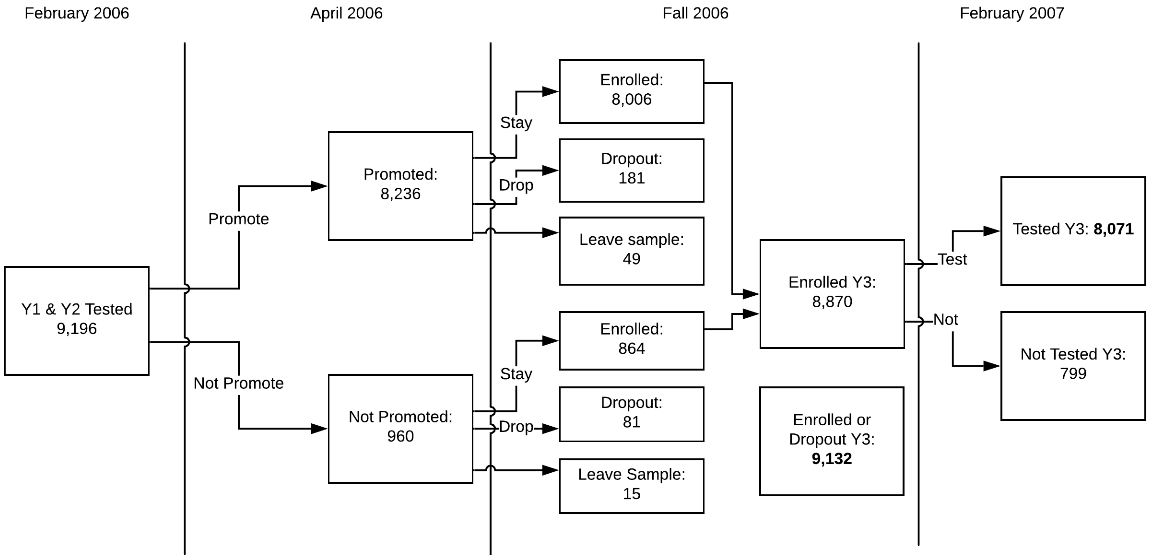
A second important feature of the data is the availability of standardized measures of student learning that are common across schools and classrooms. Surveyors proctored tests in three subjects (Urdu, English, and Math) to assess student knowledge and learning. These tests were designed to cover mostly basic topics typically learned in lower grades,[6] allowing us to measure learning levels rather than curriculum completion. The Urdu and English tests covered letters, word recognition, sentence construction, and reading comprehension, while the Math tests covered counting, addition, subtraction, multiplication, division, fractions, and word problems. Tests were scored using item response theory and normalized around zero using the first-year score distribution. Importantly, these assessments were designed for research purposes and were not part of the formal education system. They were not graded or incorporated into school records, and there is no evidence that they were used by teachers

[4] Punjab is the most populous of Pakistan's provinces, containing 56% of the population.

[5] More specifically, there were 303 private schools and 16 NGO/Trust schools, which we group together for our analysis.

[6] This decision in test construction was made to reflect students' lagged learning rather than to assess grade-level mastery.

Figure 1: Evolution of the sample



Note: There are 9,196 students between 7.5 and 15.5 years old in year 2 (4th graders are usually 9-10) with non-missing promotion status and test scores for both years 1 and 2. After year 2 (baseline), the teacher/school decided on each student's promotion status: should they be demoted to a lower grade, retained in their current grade, or promoted to the next grade? Promoted means the student advanced to the next grade while not promoted means the student was either retained in the current grade or demoted. 8,236 are promoted of which 8,006 remain enrolled and 181 dropout. 960 are not promoted of which 864 remain enrolled and 81 dropout. Thus, in year 3, there are 8,870 total enrolled students. Of these, 8,071 students are tested in year 3.

in promotion or retention decisions.

In addition to test scores and enrollment outcomes, the LEAPS data provide rich information on students, households, and teachers. For all students, we observe sex, age, and mother's education. For a subsample, we also observe height and weight, as well as students' reported feelings and attitudes toward school. A matched household survey, administered to a subsample of students, provides information on parental perceptions of their child, an asset-based wealth index, detailed educational expenditures, assessments of teacher quality, and expectations of child ability. Finally, teachers were surveyed and asked to evaluate a subsample of students in their class, responding to the question: "On a scale of 1 to 10, how good would you say that this student is in his/her studies?"

Figure 1 illustrates the evolution of the sample over time and the timing of the retention decision. After the second survey year, teachers and schools determined each student's promotion status: promotion to the next grade, retention in the current grade, or demotion to a lower grade. Promoted students advanced to the next grade, while students who were not promoted were either retained or demoted. In total, 8,236 students were promoted and 960 were not promoted. Of these students, 8,870 remained enrolled in year 3, and 8,071 were tested in that year.

We primarily use four subsamples of the data. Sample 1 includes students tested in all three survey years and is used for our main academic outcome analyses (N = 8,071).[7] Sample 2 includes students tested in all three years whose teachers also rated their performance (N = 5,529). Sample 3 includes students tested in all three years and matched to households that completed the detailed household survey (N = 698). Sample 4, used for the analysis of mechanisms and household responses, includes students tested in the first two years and matched to the detailed household survey (N = 741). Table 1 presents baseline characteristics for each subsample. Across samples, 7–9% of students repeat a grade in year 3, just under half are female, and most students are approximately 10 years old. Teacher and parent ratings are slightly inflated on average but broadly comparable across subsamples, and we observe no substantial differences in baseline characteristics.

3 Identification and Empirical Strategy

3.1 How Are Retention Decisions Made? Descriptive Evidence

We begin by documenting how grade retention decisions are made in practice, and how students who are retained differ from those who are promoted prior to the retention decision. This descriptive evidence motivates our identification strategy and the conceptual framework

[7] This number is lower because some students were either new to the sample in the second year or were absent on the day of the test in year 1.

Table 1: Baseline characteristics by sample

Variable	Child Sample	Teacher Sample	Parent Sample	Mech Sample
English score (Pre-)	0.306 (0.937)	0.363 (0.923)	0.252 (0.975)	0.238 (0.989)
Math score (Pre-)	0.254 (1.131)	0.304 (1.101)	0.304 (1.044)	0.287 (1.070)
Urdu score (Pre-)	0.380 (0.960)	0.420 (0.948)	0.373 (0.946)	0.352 (0.975)
Not Promoted	0.088 (0.284)	0.077 (0.266)	0.070 (0.256)	0.076 (0.264)
Female	0.467 (0.499)	0.470 (0.499)	0.446 (0.497)	0.441 (0.497)
Age	10.484 (1.394)	10.576 (1.404)	10.517 (1.409)	10.553 (1.488)
Mom Educated	0.277 (0.448)	0.355 (0.478)	0.287 (0.452)	0.265 (0.441)
Rating by teacher	- (-)	6.098 (2.323)	6.255 (2.354)	6.267 (2.377)
Parental Perception of Whether Child is Hard Working	- (-)	- (-)	3.304 (0.622)	3.294 (0.632)
Parental Perception of Child Intelligence	- (-)	- (-)	3.340 (0.621)	3.333 (0.625)
Inteligence Rating	-	-	-	3.333 (0.625)
Hardworking Rating	-	-	-	3.294 (0.632)
School Performance Rating	-	-	-	3.296 (0.656)
Educational Expenditures (short-run)	-	-	-	133.142 (100.163)
Lagged Educational Expenditures (long-run)	-	-	-	916.161 (413.651)
Family Help (any)	-	-	-	0.340 (0.474)
Hours of help with studies from family (median)	-	-	-	2.452 (3.887)
Hours reading or telling stories (median)	-	-	-	0.000 (0.532)
Any tutoring	-	-	-	0.182 (0.386)
Hrs. Tutoring	-	-	-	0.000 (4.749)
Know name of child's teacher	-	-	-	0.565 (0.496)
Met child's teacher	-	-	-	0.591 (0.492)
Look HW	-	-	-	0.560 (0.497)
Observations	8071	5529	698	741

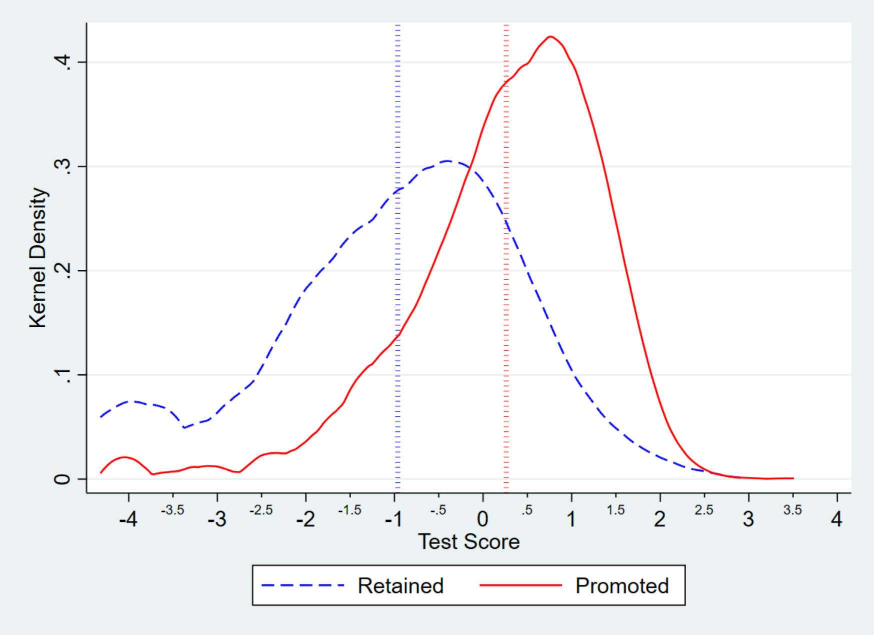
Note: Entries are means with standard deviations in parentheses. Dashes indicate not available for that sample. "Observations" reports the number of units in each column. Column (1) includes all students with three years of test scores. Column (2) includes all students with three years of test scores whose teacher also rated their performance. Column (3) includes all students with three years of test scores matched to households that completed a more detailed household survey. Column (4) includes all students with two years of test scores matched to households that completed a more detailed household survey.

that follows.

Table 2 shows that, prior to the retention decision, students who are later retained score significantly lower on average across all three tested subjects. Teachers also rate these students substantially lower than their promoted peers. These patterns are consistent with a setting in which retention decisions are based, at least in part, on perceptions of academic ability.

However, these averages mask substantial heterogeneity. Figure 2 plots the distribution of pre-retention test scores for retained and promoted students and shows considerable overlap between the two groups. Many retained students outperform some promoted students prior to the decision. If teachers rely on ability when making retention decisions, this overlap implies either that implicit promotion thresholds vary across teachers or that the signal

Figure 2: Densities of test scores before promotion decision



Note: Kernel density estimation of math test scores for retained (blue) and promoted (red) students. Vertical lines indicate group means. Scores are from year 2.

teachers observe is noisy.

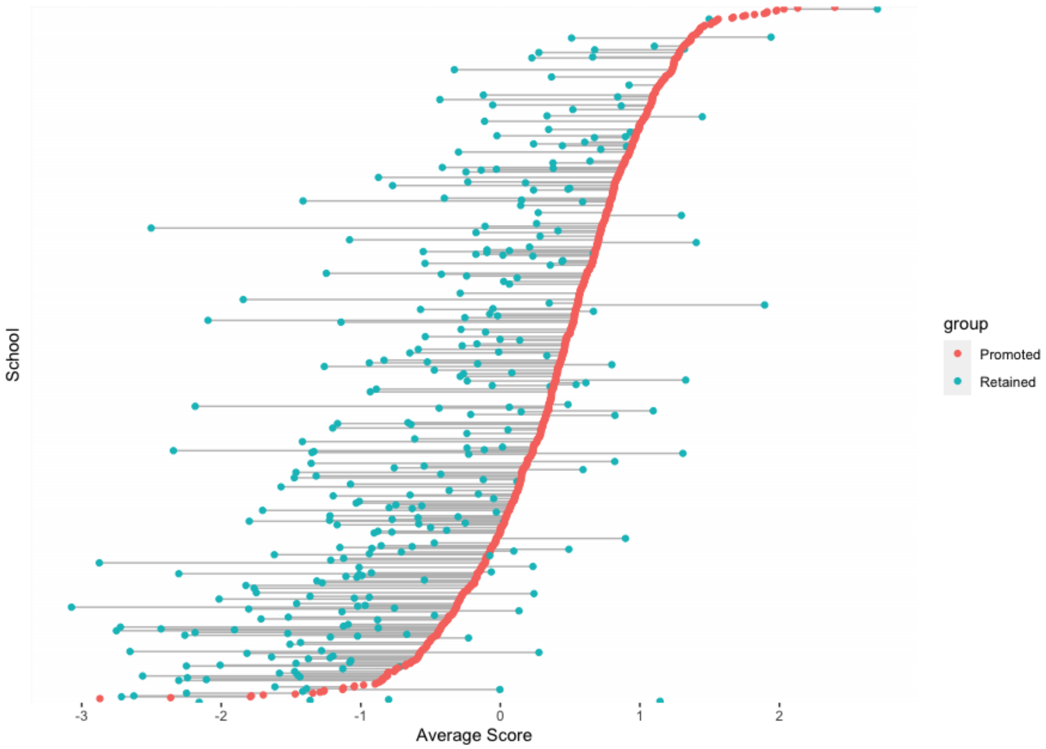
Table 2: Differences in test scores and teacher ratings before the retention decision

	Promoted	Not Promoted	Difference	Total N
English score	0.367 (0.893)	-0.532 (1.111)	-0.900***	9196
Math score	0.338 (1.051)	-0.910 (1.410)	-1.248***	9196
Urdu score	0.440 (0.914)	-0.581 (1.173)	-1.021***	9196
Rating by teacher	6.190 (2.264)	4.373 (2.422)	-1.817***	6269

Note: Entries in first two columns are means with standard deviations in parentheses. All variables are from year 2 (the retention decision occurs between year 2 and year 3). The sample for the first three rows includes all students with test scores in years 1 and 2. The sample for the fourth row includes students with test scores in years 1 and 2 and a teacher rating in year 2.

To examine variation in promotion standards across schools, Figure 3 plots, for each school, the average test score of promoted students and the average test score of retained students. There is substantial variation in school standards: in some schools, retained students have higher average test scores than promoted students in other schools, and in a small number of schools, retained students outperform promoted students within the same

Figure 3: Within-school test scores before promotion decision



Note: For each school, two averages are shown: the average test score of promoted students (red) and retained students (blue), connected by a gray line.

school. We interpret this as evidence of both heterogeneous promotion thresholds and noise in teachers' assessments, as well as the effect of small samples from schools with few grade 4 students.

To quantify how well retention decisions can be predicted using observable characteristics, we estimate a series of linear probability models with retention as the outcome. Table 3 reports R^2 and adjusted R^2 values for specifications that include flexible controls for two lags of test scores and progressively add demographic characteristics, teacher ratings, and parent ratings. Without school fixed effects, observables explain less than 20% of the variation in retention decisions. Including school fixed effects substantially increases R^2, but adjusted R^2 remains low, reflecting the limited within-school sample sizes.

Together, these results indicate that while retention decisions are related to academic performance, they are only weakly predicted by observables and vary substantially across

Table 3: Predicting retention with observables and school fixed effects

Specification	R^2	Adj. R^2	R^2 (FE)	Adj. R^2 (FE)	N
Child Characteristics	.14	.14	.36	.28	8071
Teacher Rating	.14	.14	.38	.27	5529
Parent Ratings	.2	.16	.67	.27	698
Child + Teacher + Parent	.18	.09	.77	.21	501

Note: Each row reports results from linear probability models with retention as the outcome. Columns 1–2 exclude school fixed effects; Columns 3–4 include them. All specifications include two lags of test scores.

schools, suggesting an important role for teacher discretion and idiosyncratic judgment.

3.2 Conceptual Framework

The descriptive evidence above motivates a simple conceptual framework to formalize how discretionary retention decisions generate identifying variation. In the institutional setting we study, teachers decide whether to retain students based on their own assessments of readiness, without reference to centralized promotion criteria or standardized test thresholds.

Let each student i in period t have a latent ability θ_{it} that is not directly observed by the teacher. Instead, teacher j observes a noisy signal

$$s_{itj} = \theta_{it} + \varepsilon_{itj},$$

where ε_{itj} is mean-zero measurement error. The teacher compares this signal to a teacher-specific threshold τ_j and retains the student if

$$R_{itj} = \mathbf{1}(s_{itj} < \tau_j).$$

Retention probabilities therefore depend on both student ability and teacher thresholds. Thresholds may vary across teachers and schools, and noise in perceived ability generates within-classroom variation. Together, heterogeneity in thresholds and idiosyncratic noise create the variation we exploit to identify the causal effects of retention.

Rather than modeling outcome production explicitly, we take a reduced-form approach and estimate whether being retained shifts academic outcomes, household investments, and beliefs measured in the subsequent year. This allows us to remain agnostic about mechanisms while capturing responses to the negative signal of retention.

3.3 Estimating Equation and Identification Assumptions

Guided by this framework, our empirical strategy compares outcomes for students with similar observed characteristics who experience different retention decisions.

Our main estimating equation is

$$y_{ijt+1} = \beta R_{itj} + f(y_{ijt}, y_{ijt-1}) + X_i'\gamma + \mu_j + \varepsilon_{ijt}, \tag{1}$$

where y_{ijt+1} is the outcome measured in the year after the retention decision, R_{itj} indicates retention, $f(\cdot)$ is a fourth-order polynomial in two lags of test scores, X_i includes age, gender, and teacher ratings,[8] and μ_j are school fixed effects.[9] Standard errors are clustered at the school level.

Identification relies on the assumption that, conditional on observed characteristics, retention decisions are independent of students' potential outcomes. In particular, our key

[8] While the conceptual framework abstracts from these characteristics for simplicity, we include them in the empirical specification to flexibly account for observable dimensions that may influence teachers' assessments and retention decisions. We control for age because younger students may be less intellectually mature and more likely to be held back (Mahjoub, 2017; Eide and Showalter, 2001), while older students may prioritize school less or be retained for behavioral reasons. We control for gender because boys may be more disruptive in class, leading to retention for behavioral reasons, though such behavior may also distract them from their studies, potentially biasing coefficients upward. We also control for the teacher's rating of the student, which may capture behavioral or socio-emotional skills, classroom engagement, or other dimensions of performance that are not fully reflected in test scores but are salient to teachers when forming assessments of student readiness. Note that, as shown in Tables A.1 and 6, we find no differences by gender in either the probability of retention or in differential outcomes.

[9] Promotion decisions are modeled as teacher-specific in the conceptual framework. In the data, however, approximately 84 percent of schools have only a single Grade 4 classroom, so school fixed effects coincide with teacher fixed effects for the majority of observations. In multi-teacher schools, school fixed effects absorb common institutional standards while identification comes from within-school variation in retention decisions. Appendix Section A.1.3 shows that results are robust to restricting the sample to single-teacher schools or using teacher fixed effects directly.

identifying assumption is that, conditional on flexible controls for lagged test scores, teacher ratings, and school fixed effects, whether a student is retained is as good as random. Under this assumption, retained and promoted students with similar observed characteristics would have followed similar outcome trajectories in the absence of retention, allowing us to interpret differences in subsequent outcomes as causal effects of grade repetition.

This identifying assumption is closely related to the conditional independence assumption commonly invoked in matching estimators and related selection-on-observables approaches. While our main estimates are obtained from a regression framework with rich controls and fixed effects, the identifying content of the design is the same: once we condition on observed proxies for ability and school fixed effects, remaining variation in retention status reflects idiosyncratic noise or discretion rather than systematic differences in underlying potential outcomes. To assess the robustness of our main results to this identifying assumption, we complement our regression-based estimates with propensity score matching and sensitivity analyses based on the method of Oster (2019), both of which are discussed in Section 5.

3.4 Are Retained and Promoted Students Comparable?

Our identification strategy assumes that, conditional on observed characteristics, retained and promoted students are comparable. In this subsection, we assess the plausibility of this assumption using conditional balance tests.

Specifically, we re-estimate Equation 1 using baseline characteristics as outcomes. If the coefficient on the retention indicator is statistically indistinguishable from zero in these regressions, this indicates that, once we condition on prior achievement and teacher fixed effects, retained and promoted students are similar along observed dimensions. Such patterns are consistent with the identifying assumption that retention decisions are as good as random conditional on observables.

Table 4 presents conditional balance tests for a broad set of baseline covariates. Conditional on prior achievement and school fixed effects, retained and promoted students are similar along multiple observed dimensions, including student characteristics (age and gender), family socioeconomic status, parental perceptions of the child, parental investments, and anthropometric measures. Notably, once we control flexibly for lagged test scores and teacher-specific promotion standards, we find no systematic differences between the two groups across any of these domains.

Table 4: Conditional balance tests

Outcome Variable	**Coef. on Retained**	**SE**	**p-value**
Female (1 = girl)	0.0021	0.0117	0.8579
Age (in years)	0.0249	0.0726	0.7315
Mother educated (1 = yes)	-0.0164	0.0237	0.4886
Parent rating: School performance	0.0607	0.1910	0.7511
Parent rating: Intelligence	0.1251	0.2063	0.5446
Asset-based wealth index	0.2555	0.4174	0.5410
Hours of family help	0.0435	1.6931	0.9795
Met child's teacher (1 = yes)	-0.0030	0.0223	0.8917
Height-for-age z-score	-0.0119	0.0706	0.8657
BMI z-score	-0.0263	0.0586	0.6537

Note: Each row reports estimates from a regression of the indicated baseline characteristic on an indicator for whether the student was retained, controlling for two lags of test scores, teacher ratings, student age and gender (where not the outcome), and school fixed effects. Standard errors are clustered at the school level. Sample sizes vary across rows due to differences in survey coverage and data availability, particularly for variables drawn from the household survey and anthropometric measurements. None of the coefficients are statistically distinguishable from zero, consistent with retained and promoted students being comparable along observed dimensions conditional on controls.

While these balance tests cannot rule out selection on unobserved characteristics, they provide strong evidence that retention decisions are not driven by observable differences in student background, household resources, or parental beliefs once prior achievement is accounted for. Together, this evidence supports the internal validity of our empirical strategy: conditional on observables, differences in subsequent outcomes between retained and promoted students can plausibly be interpreted as reflecting the causal effect of grade repetition. We further assess the sensitivity of our results to potential unobserved selection in Section 5.

4 Results

We begin by documenting the impact of grade repetition on academic outcomes. We then turn to the mechanisms underlying these effects, asking how teachers, parents, and students themselves respond to the negative signal generated by retention. For every outcomes, we estimate Equation 1, which includes school fixed effects, teacher ratings of the student, and basic student-level controls. Lagged (and double-lagged) test scores enter in 4th order polynomial form to allow flexibility, and standard errors are clustered at the school level (Abadie et al. (2023) and Cameron and Miller (2015)). We then present robustness checks on these estimates, to validate our specification.

4.1 How are academic outcomes impacted by repeating?

Understanding how grade repetition affects academic outcomes is a necessary first step. These effects provide both a benchmark for comparison with existing work and a reference point for interpreting the behavioral responses we analyze next. We therefore begin by examining the impact of grade repetition on test scores (Math, English, and Urdu) and on dropout, measured one and two years after the retention decision.[10]

Results are presented in Table 5 and show that grade repetition persistently lowers academic achievement and increases the probability of dropping out in the year immediately following the retention decision. Across all three subjects, retained students score approximately 0.3 standard deviations lower than their promoted peers, and these achievement gaps remain fairly stable even two years after retention. Retention also increases the probability of dropping out the following year by 7 percentage points, relative to a baseline dropout rate of 3%. However, this increase in dropout is not persistent: two years after the retention

[10] By "dropout," we mean whether the student drops out before year $t+1$. The timeline is as follows. At the end of year t, teachers assign a promotion status to each student (either promoted to next grade or repeat current grade). Between years t and $t+1$, each student makes a decision to stay in school or dropout (they can also leave the mauza or leave the sample for unspecified reasons, however we ignore those options here). Our dropout variable is equal to one if the child dropped out of school and equal to zero if they are still enrolled in school.

decision, the estimated effect on dropout is a precisely estimated zero.

Table 5: The effect of repetition on test scores and dropout

	Math		English		Urdu		Dropout	
	(1) 1 yr	(2) 2 yrs	(3) 1 yr	(4) 2 yrs	(5) 1 yr	(6) 2 yrs	(7) 1 yr	(8) 2 yrs
Not Promoted	**-0.440*****	**-0.322*****	**-0.278*****	**-0.338*****	**-0.371*****	**-0.364*****	**0.070*****	**0.009**
	(0.055)	(0.076)	(0.036)	(0.050)	(0.043)	(0.056)	(0.014)	(0.019)
Lagged Score	**0.467*****	**0.555*****	**0.404*****	**0.364*****	**0.467*****	**0.432*****	**-0.005**	**-0.004**
	(0.027)	(0.037)	(0.023)	(0.033)	(0.024)	(0.028)	(0.007)	(0.011)
2yr Lagged Score	**0.247*****	**0.240*****	**0.195*****	**0.142*****	**0.226*****	**0.174*****	**-0.018****	**-0.010**
	(0.027)	(0.035)	(0.020)	(0.032)	(0.027)	(0.030)	(0.008)	(0.010)
Age/Gender Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Teacher Rating	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
School Fixed Effect	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Lag/Double Lag Polynomial	4th Order	4th Order	4th Order	4th Order	4th Order	4th Order	4th Order	4th Order
N	5529	3992	5529	3992	5529	3992	6222	4633
R^2	0.72	0.70	0.78	0.75	0.75	0.73	0.19	0.25

Note: Coefficient estimates bolded, standard errors (clustered at school-level) in parentheses below. We run separate regressions for each school subject (math, English, and Urdu). Lagged score refers to a student's previous year score on the appropriate school subject (the lagged outcome variable).
*** Significant at 1% level, ** Significant at 5% level, * Significant at 10% level.

The results in Table 5 control for gender, but it is possible that girls and boys react differently when retained. In this setting, girls and boys are not retained at different rates (see Table A.1). Table 6 then considers how the impacts of retention differ between girls and boys by allowing an interaction term between the retention dummy and a gender dummy. The results suggest that girls fare better than boys after being retained, in that their test scores do not decrease by such a large amount. However, the results are noisier than the main effects presented in Table 5 and not as large. For the remainder of the analysis, we will control for gender but we will not show the results separately.

4.2 Mechanism - Teachers

If grade repetition leads to worse outcomes because teachers subsequently stigmatize repeaters or reduce effort, we would expect retained students to be viewed more negatively by teachers even after accounting for achievement. To proxy for a teacher's opinion of a given student, we will use the teacher rating.[11] In Table 7 we present results where the outcome

[11] In the main specification, we include teacher rating as a control. In these specifications we have removed the control, and instead use it as a dependent variable.

Table 6: The effect of repetition on test scores and dropout - By gender

	Math		English		Urdu		Dropout	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Not Promoted × Female		**0.1644**		**0.1695****		**0.1806****		**0.0036**
		(0.1062)		(0.0701)		(0.0845)		(0.0275)
Not Promoted	**-0.4403*****	**-0.5225*****	**-0.2783*****	**-0.3634*****	**-0.3706*****	**-0.4614*****	**0.0701*****	**0.0684*****
	(0.0553)	(0.0850)	(0.0364)	(0.0575)	(0.0434)	(0.0708)	(0.0145)	(0.0196)
Female	**-0.0341**	**-0.0391**	**0.0404***	**0.0353**	**0.0604****	**0.0550****	**-0.0036**	**-0.0037**
	(0.0279)	(0.0280)	(0.0218)	(0.0218)	(0.0256)	(0.0257)	(0.0080)	(0.0079)
Lagged Score	**0.4673*****	**0.4671*****	**0.4042*****	**0.4046*****	**0.4668*****	**0.4659*****	**-0.0046**	**-0.0046**
	(0.0270)	(0.0269)	(0.0231)	(0.0231)	(0.0238)	(0.0237)	(0.0074)	(0.0074)
2yr Lagged Score	**0.2473*****	**0.2457*****	**0.1951*****	**0.1946*****	**0.2263*****	**0.2256*****	**-0.0177****	**-0.0178****
	(0.0269)	(0.0270)	(0.0199)	(0.0198)	(0.0271)	(0.0271)	(0.0085)	(0.0084)
Constant	**0.7768*****	**0.7805*****	**0.6076*****	**0.6091*****	**0.6420*****	**0.6440*****	**0.0157****	**0.0158****
	(0.0314)	(0.0316)	(0.0261)	(0.0262)	(0.0255)	(0.0255)	(0.0080)	(0.0080)
AgeControls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Teacher Rating	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
School Fixed Effect	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Lag Polynomial	4th Order	4th Order	4th Order	4th Order	4th Order	4th Order	4th Order	4th Order
N	5529	5529	5529	5529	5529	5529	6222	6222
R^2	0.72	0.72	0.78	0.78	0.75	0.75	0.19	0.19

Note: Coefficient estimates bolded, standard errors (clustered at school-level) in parentheses below. We run separate regressions for each school subject (math, English, and Urdu). Lagged score refers to a student's previous year score on the appropriate school subject (the lagged outcome variable).
*** Significant at 1% level, ** Significant at 5% level, * Significant at 10% level.

variable is the teacher's rating of the student, with specifications using the rating of the contemporaneous teacher (the teacher who holds the student back) as well as the next year's teacher (the teacher after the retention decision).

For the contemporaneous teacher, we find that they have a significantly lower rating for the child than we would predict given the student's lagged test scores and other information. This can be seen in column (1) of Table 7. If we consider the student's teacher for the next year, we see that they also have a negative view of the student (column (2)), however, when controlling for the students test scores in that year, the fact that they were retained has no significant impact on the teacher rating. In fact, comparing the R^2 of columns (3) and (4), we learn that promotion status has no additional predictive value on the subsequent-year teacher's rating of the student above test scores.

These results tell us two things. First, teachers do not seem to label students as "repeaters" and have a negative bias towards them. If anything, it seems as if each year the teacher views the students as i.i.d draws conditional on their test scores. Second, if we were worried that the contemporaneous teacher retained the student because of something that

Table 7: The effect of repetition on teacher ratings

	(1) Contemporaneous	(2) Next Year	(3) Next Year	(4) Next Year
Not Promoted	**-0.8575***** (0.160)	**-0.3729***** (0.137)	**-0.0398** (0.140)	
Lagged Average Score	**1.1570***** (0.095)	**1.1220***** (0.088)	**0.5880***** (0.093)	**0.5891***** (0.093)
2yr lagged Average Score	**0.4842***** (0.106)	**0.5342***** (0.099)	**0.2781***** (0.093)	**0.2786***** (0.093)
Age	**-0.0791***** (0.026)			
Average Score: Y3			**1.0268***** (0.086)	**1.0318***** (0.084)
Constant	**6.0961***** (0.289)	**5.4368***** (0.085)	**4.7356***** (0.106)	**4.7292***** (0.104)
Age/Gender Controls	Yes	Yes	Yes	Yes
Fixed Effect	School	School	School	School
Lag Polynomial	4th Order	4th Order	4th Order	4th Order
N	5529	5446	5446	5446
R^2	0.45	0.49	0.51	0.51

Note: Coefficient estimates bolded, standard errors (clustered at school-level) in parentheses below. The first column refers to the teacher who retained the student, and the remaining three columns refer to the teacher who taught the student the following year. Lagged score refers to a student's previous year average test score. Column (4) illustrates that promotion does not increase the R-squared of the regression.
*** Significant at 1% level, ** Significant at 5% level, * Significant at 10% level.

was observable to them but unobservable to us, it seems as if these traits were also unobservable to the student's teacher the following year. This supports our theory that teachers are trained to teach, not to recognize which students should be retained, and that some of the retention that we see in the data is caused by random noise.

4.3 Mechanism - Household

In settings where parents play a central role in shaping their child's educational trajectory, grade repetition may trigger important household responses. How parents interpret and respond to retention matters because parental beliefs and expectations directly influence subsequent educational investments and support. If parents view repetition as a negative signal of a child's ability or future returns to schooling, they may revise beliefs and reallocate resources away from the retained child. We use our linked household–student–school panel data to examine how parental perceptions, expectations, and investments change following grade repetition.

We find that parents treat grade repetition as a signal of lower intellectual ability or potential. Table 8 has three parental perceptions as the outcome variables. Parents decrease their ratings of their child's performance in school, work ethic, and intelligence by 0.15 to 0.32 standard deviations after their child is retained. This effect holds even after controlling for lagged parent ratings, demonstrating that there is a clear change in parent evaluations of a child after she has been retained.

Table 8: The effect of repetition on parent perceptions

	School Performance		Work Ethic		Intelligence	
	(1)	(2)	(3)	(4)	(5)	(6)
Not Promoted	-0.1516**	-0.1321**	-0.3281**	-0.3120*	-0.2552**	-0.2405*
	(0.062)	(0.063)	(0.158)	(0.160)	(0.129)	(0.124)
Lagged Average Score	0.1811***	0.1668***	0.2553***	0.2476***	0.3758***	0.3485***
	(0.039)	(0.038)	(0.090)	(0.089)	(0.079)	(0.078)
2yr lagged Average Score	-0.0366	-0.0586	0.0309	0.0225	-0.1467	-0.1673
	(0.050)	(0.049)	(0.112)	(0.112)	(0.112)	(0.110)
Lagged School Performance		0.1532***				
		(0.027)				
Lagged Hardworking				0.0830		
				(0.061)		
Lagged Inteligence						0.2366***
						(0.061)
Age/Gender Controls	Yes	Yes	Yes	Yes	Yes	Yes
Lag Polynomial	4th Order	4th Order	4th Order	4th Order	4th Order	4th Order
N	741	741	741	741	741	741
R^2	0.13	0.17	0.08	0.08	0.10	0.12

Note: Coefficient estimates bolded, standard errors (clustered at school-level) in parentheses below. Lagged score refers to a student's previous year average test score. Outcome questions are listed in full in the text.
*** Significant at 1% level, ** Significant at 5% level, * Significant at 10% level.

Table 9 examines perceptions two years after the retention, and finds mixed results depending on the question. For school performance, there is no significant effect, perhaps signalling that parents have revised downward their expectations and their child is no longer "disappointing" them. The impacts on work ethic are persistent and similar in size two years after the retention. The impacts on perceived intelligence are no longer significant, but in terms of magnitude are still half the size as the initial estimates. Together, this shows that having their child repeat a grade causes parents to significantly downgrade their perceptions of their child, and some of these shifts are long-lasting.

Table 9: The effect of repetition on parent perceptions - Two years after repeating

	School Performance		Work Ethic		Inteligence	
	(1)	(2)	(3)	(4)	(5)	(6)
Not Promoted	**-0.0467**	**-0.0330**	**-0.3141***	**-0.2730***	**-0.1355**	**-0.1180**
	(0.136)	(0.134)	(0.165)	(0.161)	(0.165)	(0.161)
Lagged Average Score	**0.2392*****	**0.2323*****	**-0.0381**	**-0.0560**	**0.0262**	**0.0056**
	(0.067)	(0.067)	(0.092)	(0.092)	(0.093)	(0.091)
2yr lagged Average Score	**-0.0453**	**-0.0562**	**0.1236**	**0.1057**	**0.0713**	**0.0568**
	(0.080)	(0.080)	(0.108)	(0.106)	(0.123)	(0.122)
Lagged School Performance		**0.0850***				
		(0.044)				
Lagged Hardworking				**0.2042*****		
				(0.062)		
Lagged Inteligence						**0.2039*****
						(0.062)
Age/Gender Controls	Yes	Yes	Yes	Yes	Yes	Yes
Lag Polynomial	4th Order	4th Order	4th Order	4th Order	4th Order	4th Order
N	710	710	710	710	710	710
R^2	0.09	0.09	0.03	0.05	0.05	0.06

Note: Coefficient estimates bolded, standard errors (clustered at school-level) in parentheses below. Lagged score refers to a student's previous year average test score. Outcome questions are listed in full in the text.
*** Significant at 1% level, ** Significant at 5% level, * Significant at 10% level.

While the parental perception questions are insightful, they are also somewhat vague. In the fifth round of surveys, approximately 5 years after the retention decision, a more concrete survey questions specifically asked parents what score they believe that their child could achieve on a test (they were asked for a minimum, maximum, and average). Table 10 presents suggestive evidence that parents substantially decreased all three of these measures if their child was retained.[12] This, along with the previous two tables, offers evidence that parents significantly lower their expectations of their child once they are retained, and that this effect persists for many years.

In addition to revising their beliefs and expectations downwards, parents also reallocate resources away from repeating students. Table 11 shows annual expenditure on a repeater's

[12] The sample size shrinks in this table because the question was asked in a follow-up survey for which not all households could be found and matched.

Table 10: The effect of repetition on parent beliefs

	(1) Average	(2) Maximum	(3) Minimum
Not Promoted	**-4.9701***	**-5.9614***	**-3.9929**
	(2.761)	(3.084)	(2.550)
Lagged Average Score	**1.1569**	**1.6451**	**0.3857**
	(2.246)	(2.510)	(2.120)
2yr lagged Average Score	**3.0027**	**2.4747**	**2.8306**
	(2.658)	(2.820)	(2.484)
Constant	**50.2885*****	**61.3707*****	**39.6159*****
	(1.944)	(2.097)	(1.803)
Age/Gender Controls	Yes	Yes	Yes
Lag Polynomial	4th	4th	4th
N	513	513	514
R^2	0.43	0.43	0.40

Note: Coefficient estimates bolded, standard errors (clustered at school-level) in parentheses below. Lagged score refers to a student's previous year average test score. Outcome refers to how many questions parents expect a student to answer correctly on a test.
*** Significant at 1% level, ** Significant at 5% level, * Significant at 10% level.

education (which includes longer-term educational expenses including annual school fees, school uniforms, textbooks, and school supplies) decreases by approximately 176 PKR, or 20% of mean annual educational expenditure (column 3).[13] While there appears to be less of a decrease in monthly educational expenditures (which includes shorter-term expenses such as transportation costs, private tutoring, and pocket money for school), the point estimate is negative.[14] The magnitude of the decrease in expenditure gets even larger two years after the retention.

To confirm that these effects represent a resource reallocation away from the repeating student specifically (rather than from some confounding household-level shock), we estimate the same regression but change the outcome variable to be the average expenditure on all other siblings in the household. Table 12 shows that household expenditure on the siblings of repeaters may even rise, though the effect is not statistically distinguishable from zero

[13] This drop in educational expenditure is not due to dropouts. All children in this regression's sample remained in school before and after grade repetition.

[14] We are lacking statistical power for this regression as it is identified off of only 56 repeaters in the smaller subsample of students whose household was included in the household survey.

Table 11: The effect of repetition on parent investments

	Monthly $ $(t+1)$		Annual $ $(t+1)$		Monthly $ $(t+2)$		Annual $ $((t+2))$	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Not Promoted	**-17.7139**	**-15.0460**	**-176.1899*****	**-111.4625***	**-47.2153****	**-46.2378****	**-209.9473***	**-199.9287***
	(14.908)	(18.070)	(59.553)	(57.743)	(22.513)	(21.416)	(115.382)	(114.009)
Lagged Average Score	**12.3195**	**-0.0491**	**-15.5160**	**-20.4853**	**13.0877**	**8.5559**	**122.1426**	**121.3734**
	(15.489)	(13.301)	(47.249)	(45.470)	(17.020)	(16.707)	(77.506)	(77.480)
2yr lagged Average Score	**33.6261**	**10.5395**	**102.3851***	**65.1619**	**20.5382**	**12.0793**	**32.8975**	**27.1361**
	(22.248)	(15.056)	(59.911)	(57.953)	(21.311)	(21.183)	(102.077)	(101.723)
Lagged Monthly $		**0.8893*****				**0.3258*****		
		(0.149)				(0.091)		
Lagged Anual $				**0.4742*****				**0.0734**
				(0.055)				(0.085)
Age/Gender Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Lag Polynomial	4th Order	4th Order	4th Order	4th Order	4th Order	4th Order	4th Order	4th Order
N	741	741	741	741	741	741	741	741
R^2	0.05	0.33	0.05	0.16	0.04	0.06	0.05	0.05

Note: Coefficient estimates bolded, standard errors (clustered at school-level) in parentheses below. Lagged score refers to a student's previous year average test score. Full descriptions of short-term and long-term investments are in the text.
*** Significant at 1% level, ** Significant at 5% level, * Significant at 10% level.

(columns 1-4).[15] Correspondingly, parent perceptions of the repeater's siblings also rise or stay constant (Table 12 columns 5-10).

Table 12: The effect of repetition on investments in and perceptions of siblings

	Monthly $ $(t+1)$		Annual $ $(t+1)$		Performance (t+1)		Work Ethic (t+1)		Inteligence (t+1)	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Not Promoted	**0.5588**	**5.3413**	**29.2969**	**74.7752**	**0.1453**	**0.0234**	**0.2836***	**0.2352**	**0.0218**	**0.0073**
	(29.768)	(27.152)	(78.805)	(69.771)	(0.136)	(0.116)	(0.168)	(0.170)	(0.158)	(0.156)
Lagged Average Score	**26.8716**	**13.8832**	**92.1976**	**61.4781**	**0.1592***	**0.1126***	**0.1884****	**0.1850****	**0.0803**	**0.0649**
	(16.515)	(14.155)	(60.140)	(56.643)	(0.086)	(0.066)	(0.083)	(0.080)	(0.079)	(0.073)
2yr lagged Average Score	**19.2244**	**14.8686**	**18.2714**	**32.7411**	**-0.0202**	**-0.0223**	**-0.2928*****	**-0.2922*****	**-0.0812**	**-0.0163**
	(21.413)	(17.907)	(74.761)	(69.083)	(0.106)	(0.082)	(0.107)	(0.104)	(0.105)	(0.100)
Lagged Sibling Monthly $		**1.9400*****								
		(0.267)								
Lagged Sibling Annual $				**1.1220*****						
				(0.223)						
Lagged Sibling Overall						**0.4903*****				
						(0.036)				
Lagged Sibling Hardworking								**0.2441*****		
								(0.089)		
Lagged Sibling Inteligence										**0.3269*****
										(0.079)
Age/Gender Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Lag Polynomial	4th Order	4th Order	4th Order	4th Order	4th Order	4th Order	4th Order	4th Order	4th Order	4th Order
N	719	719	719	719	719	702	719	702	719	702
R^2	0.05	0.20	0.07	0.22	0.07	0.34	0.04	0.05	0.02	0.05

Note: Coefficient estimates bolded, standard errors (clustered at school-level) in parentheses below. The table shows results from a student-level regression of average household expenditure on all siblings (leaving out the student) on whether the student repeated or not (columns 1-4) and of average parent evaluation of siblings (leaving out the student) on whether the student repeated or not. Lagged score refers to a student's previous year average test score. Full descriptions of short-term and long-term investments are in the text.
*** Significant at 1% level, ** Significant at 5% level, * Significant at 10% level.

The previous evidence indicates that parents place the blame of retention on their child,

[15] The sample sizes in this table are smaller than in tables 10 and 11 due to students who have no siblings.

and the following tables suggest that they do not put any on their child's teachers. In the parent surveys, parents were asked questions about how they view their child's teachers. Three of these questions were "How regular is your child's class-teacher overall?", "How good would you say that your child's class-teacher is in his/her teaching skills?", and "How good would you say that your child's class-teacher is overall?". Parents were allowed to respond that they "do not know", which is why the sample size decreases from the previous tables. Table 13 shows that parents of repeating students do not hold worse views of their child's teacher than parents of promoted students (conditional on expressing a view of their child's teacher). All coefficients are fairly precisely estimated zeros. Table 14 shows the same analysis, but for teachers the following year (the year during which the student is repeating the grade). The sample size is small, but it does not seem as if parents have any different views about the teachers if their child is held back or repeating the grade.

Table 13: The effect of repetition on parental ratings of teachers (current year)

	How Regular		Good Teacher		Overall	
	(1)	(2)	(3)	(4)	(5)	(6)
Not Promoted	**-0.0538**	**0.0158**	**0.0130**	**0.0381**	**-0.0408**	**-0.0548**
	(0.054)	(0.062)	(0.091)	(0.108)	(0.092)	(0.109)
Lagged Average Score		**0.0062**		**-0.0268**		**-0.0785**
		(0.033)		(0.059)		(0.060)
2yr lagged Average Score		**0.0773***		**0.1743****		**0.0616**
		(0.045)		(0.075)		(0.074)
Age/Gender Controls	No	Yes	No	Yes	No	Yes
Lag Polynomial	-	4th	-	4th	-	4th
N	700	631	699	630	686	620
R^2	0.00	0.07	0.00	0.04	0.00	0.04

Note: Coefficient estimates bolded, standard errors (clustered at school-level) in parentheses below.
Lagged score refers to a student's previous year average test score.
*** Significant at 1% level, ** Significant at 5% level, * Significant at 10% level.

Table 14: The effect of repetition on parental ratings of teachers (next year)

	How Regular		Good Teacher		Overall	
	(1)	(2)	(3)	(4)	(5)	(6)
Not Promoted	**-0.0078**	**0.0184**	**-0.0102**	**0.0655**	**0.0711**	**0.0136**
	(0.074)	(0.086)	(0.104)	(0.116)	(0.107)	(0.121)
Lagged Average Score		**0.0876***		**0.0781**		**0.0765**
		(0.047)		(0.065)		(0.068)
2yr lagged Average Score		**0.0300**		**0.1659****		**0.0891**
		(0.061)		(0.084)		(0.088)
Age/Gender Controls	No	Yes	No	Yes	No	Yes
Lag Polynomial	-	4th	-	4th	-	4th
N	820	739	826	744	820	739
R^2	0.00	0.05	0.00	0.06	0.00	0.04

Note: Coefficient estimates bolded, standard errors (clustered at school-level) in parentheses below. Lagged score refers to a student's previous year average test score.
*** Significant at 1% level, ** Significant at 5% level, * Significant at 10% level.

4.4 Mechanisms - Student

Academic outcomes capture only part of the impact of grade repetition. Retention may also affect students' beliefs, motivation, and sense of agency, which are harder to observe but potentially central to longer-run outcomes. In particular, if retention contains a discretionary or noisy component conditional on achievement, students may interpret being held back as an arbitrary or discouraging signal. Such a signal could weaken beliefs about the returns to effort in school, even if it does not affect broader attitudes toward effort or control more generally.

To examine these channels, we estimate our main specification using student belief measures as outcomes. These questions come from the fifth round follow-up survey, conducted approximately five years after the retention decision, resulting in a smaller sample. Results are presented in Table 15. Column (1) shows that retention significantly reduces the likelihood that a student agrees with the statement "If I study hard at school I will be rewarded with a better job in the future."[16] We interpret this as evidence that grade repetition

[16] To obtain the outcome variable for this regression, we recode a Likert scale response into a binary indicator for agree versus disagree.

discourages students by weakening beliefs about the returns to academic effort.

Columns (2) through (5) show that retention does not seem to discourage students in other areas of their life. There are no significant impacts of retention on any of the four other questions. The questions for each column are the following: (2) “If I try hard, I can improve my situation in life"; (3) “Other people in my family make all the decisions about how I spend my time"; (4) “I like to make plans for my future studies and work"; (5) “I have no choice about the work I do - I must work”. This specific decrease in belief that effort in school will benefit their life goals could be one reason why students who are retained see a decrease in their test scores.

Table 15: The effect of repetition on student beliefs

	(1) Study Hard	(2) Try Hard	(3) Others Decisions	(4) Make Plans	(5) Must Work
Not Promoted	**-0.2602****	**-0.0375**	**0.0110**	**-0.0584**	**0.0772**
	(0.102)	(0.032)	(0.084)	(0.087)	(0.091)
Lagged Average Score	**0.0210**	**-0.0226**	**-0.0378**	**0.0323**	**-0.0243**
	(0.046)	(0.015)	(0.061)	(0.043)	(0.057)
2yr lagged Average Score	**-0.0017**	**0.0277**	**-0.0312**	**0.0200**	**-0.0800**
	(0.062)	(0.021)	(0.078)	(0.056)	(0.068)
Constant	**0.8329*****	**0.9918*****	**0.5317*****	**0.8900*****	**0.1568*****
	(0.052)	(0.006)	(0.054)	(0.042)	(0.051)
Age/Gender Controls	Yes	Yes	Yes	Yes	Yes
Lag Polynomial	4th Order	4th Order	4th Order	4th Order	4th Order
N	578	578	578	578	578
R^2	0.35	0.27	0.30	0.32	0.20

Note: Coefficient estimates bolded, standard errors (clustered at school-level) in parentheses below. Lagged score refers to a student’s previous year average test score. Outcome questions are listed in full in the text.
*** Significant at 1% level, ** Significant at 5% level, * Significant at 10% level.

5 Robustness Checks for the Main Results

The interpretation of our baseline estimates relies on the assumption that, after conditioning on school fixed effects, prior achievement, teacher ratings, and other observed characteristics, students who are retained are comparable to those who are promoted. This section examines the sensitivity of the main results to potential violations of this assumption. We focus on two main concerns. First, teachers’ promotion decisions may reflect observable factors that are not fully captured by test scores, including systematic differences by age or gender. Second, retention decisions may be influenced by unobserved student traits that also affect later outcomes. We assess the importance of these concerns using a range of complementary approaches, including balance tests, placebo outcomes, formal sensitivity analysis, and alternative estimators.

5.1 Selection on Observables and Measurement Error

Our baseline specifications control flexibly for observable student characteristics that may influence both retention decisions and subsequent outcomes, including age, gender, two lags of test scores, and the teacher's pre-decision rating of the student. These controls address concerns that certain groups of students may be systematically over- or underestimated, or that retention decisions reflect information observed by teachers but not directly recorded in test scores. Conditional balance tests reported in Table 4 show that, after conditioning on these controls, retained and promoted students are similar along observable dimensions.

Because promotion decisions are made by individual teachers, one might be concerned that using school fixed effects rather than teacher fixed effects could place undue weight on the subset of schools with multiple Grade 4 teachers. Our baseline estimates include school fixed effects, which account for systematic differences in promotion standards across schools. In practice, 86% of schools in our sample have only a single Grade 4 class, so school and teacher fixed effects coincide in the majority of cases. As robustness checks, we re-estimate our preferred specifications restricting the sample to schools with only one Grade 4 teacher, and separately replacing school fixed effects with teacher fixed effects in the decision year. Results are shown in Appendix A.1.3, and in both cases, the estimated impacts of grade repetition on test scores are similar or slightly larger in magnitude, while dropout effects remain stable. These findings suggest that our results are not driven by heterogeneity in promotion standards across teachers within schools.

5.2 Selection on Unobservables

To assess robustness to selection on unobserved characteristics, we first examine placebo outcomes among siblings. Table 12 shows that a child's retention has no effect on parental beliefs or investments toward siblings. If retention and subsequent outcomes were jointly driven by unobserved household-level shocks, we would expect spillovers to other children in the household.

A related concern is that teachers may retain students based on persistent traits that are not observed in the data, such as behavior or motivation. Table 7 provides evidence against this channel. In the year following the retention decision, teachers do not rate repeaters differently than their peers, conditional on recent achievement. This pattern suggests that retention decisions are unlikely to be driven by time invariant unobserved traits that continue to influence teacher assessments.

We also implement the sensitivity analysis proposed by Oster (2019). The results, reported in Table 16, indicate that the implied values of δ are generally above one, suggesting that unobserved factors would need to be at least as important as observed controls to fully explain away the estimated effects. Given the richness of the baseline specification, including multiple lags of achievement and teacher ratings, it is difficult to identify plausible omitted variables of comparable importance.

Table 16: Oster test for robustness of main results

	Math	English	Urdu	Dropout
Coef. (Not promoted)	-0.440	-0.278	-0.371	0.070
Oster bound (β)	-0.020	0.021	-0.034	0.064
Delta	1.039	0.937	1.085	3.583
Observations	5529	5529	5529	6222

Note: This table reports results from the Oster (2019) test for robustness to omitted variable bias, corresponding to the estimates in Table 5. We restrict attention to the impacts measured one year after grade retention. Reported coefficients are the baseline estimates of not being promoted. The bound β^* indicates the estimated coefficient after adjusting for potential selection on unobservables, with δ measuring the relative degree of selection on unobservables versus observables. Following standard practice, we set R^* equal to 1.3 times the observed R^2.

5.3 Alternative Estimators and Identification Strategies

5.3.1 Propensity Score Matching

As an additional robustness check, we re-estimate the short-run impacts of grade repetition using propensity score matching (PSM). The goal of this exercise is not to address selection on unobserved characteristics, but rather to verify that our baseline results are not driven by differences in observable characteristics between retained and promoted students or by functional form assumptions imposed by the regression model.

Let $D_i = 1$ indicate that student i is retained between years t and $t+1$, and let $Y_i(1)$ and $Y_i(0)$ denote potential outcomes under retention and promotion, respectively. Propensity score matching relies on the conditional independence assumption

$$\Big(Y_i(1), Y_i(0)\Big) \;\perp\; D_i \;\mid\; X_i, \tag{2}$$

where X_i includes pre-retention characteristics such as lagged and double-lagged test scores, teacher ratings measured prior to the retention decision, basic student characteristics, and school fixed effects. This assumption is identical in content to the identification assumption underlying our baseline regression specifications.

We estimate the average treatment effect on the treated (ATT) by matching retained students to promoted students with similar predicted probabilities of retention, imposing common support to ensure that comparisons are made only among observationally comparable students. As shown in Table 17, the matching estimates for test scores and dropout one year after the retention decision closely mirror our baseline regression results. These findings reinforce the conclusion that the negative short-run effects of grade repetition are not driven by imbalance in observable characteristics or by modeling assumptions. Additional details on balance and propensity score overlap are reported in Appendix A.1.4.

Table 17: Propensity score matching estimates of the effect of grade repetition (1-year outcomes)

	Math	English	Urdu	Dropout
	(1) 1 yr	(2) 1 yr	(3) 1 yr	(4) 1 yr
Not Promoted	**-0.551*** **	**-0.450*** **	**-0.433*** **	**0.078*** **
	(0.087)	(0.064)	(0.066)	(0.023)
School Fixed Effect	Yes	Yes	Yes	Yes
Lag/Double Lag Polynomial	4th Order	4th Order	4th Order	4th Order
Common Support	Yes	Yes	Yes	Yes
Exact Match (School)	Yes	Yes	Yes	Yes
N treated	684	684	684	684
N control	6670	6670	6670	6670

Note: This table reports average treatment effects on the treated (ATT) from propensity score matching. Retained students are matched to promoted students with similar predicted probabilities of retention using nearest-neighbor matching within schools and restricting attention to regions of common support. Propensity scores are estimated using flexible functions of lagged and double-lagged test scores, teacher ratings measured prior to the retention decision, and basic student characteristics. Standard errors are obtained via bootstrap and are reported in parentheses below coefficient estimates.
*** Significant at 1% level, ** Significant at 5% level, * Significant at 10% level.

5.3.2 Teacher Strictness Instrument

Given the discretionary nature of promotion decisions, one might view this setting as amenable to a judge-style instrumental variables design based on teacher strictness. However, several features of the institutional setting limit the applicability of such an approach, including non-random assignment to teachers, the absence of a clear monotonicity condition, and the fact that many schools have only a single Grade 4 class. We therefore do not rely on a judge IV as a primary specification. As a sensitivity check, we nonetheless explore an instrumental variables strategy based on cross-school variation in retention strictness; the details and results are discussed in Appendix A.1.5.

6 Conclusion

This paper provides new evidence on the consequences of grade repetition in a low-income country, where promotion decisions are made by teachers based on informal criteria. Using detailed administrative and survey data from public and private schools in Pakistan, we include a rich set of controls for student ability and school fixed effects to estimate the impacts of retention on student outcomes and household responses.

We find that repeating a grade has a negative impact on test scores the next year, and these effects persist into the second year after the retention decision. In contrast, the effects on dropout are large but only immediately after being retained.

To understand the mechanisms behind these academic responses, we examine how teachers, households, and students react to retention using survey data. The results suggest that parents revise downward their beliefs about their child's academic ability and the value of investing in their education. Importantly, these responses are child-specific: parents do not reduce investments in siblings, nor do they shift blame toward the school or teacher. We also find that students themselves revise their beliefs about the relationship between effort and academic success, although beliefs about their own ability remain largely unchanged.

Taken together, these findings highlight that the effects of grade retention are not limited to academic outcomes, but also operate through behavioral channels in the household and for the student. This is a setting where the consequences of falling behind are steep and retention is viewed as a signal of failure. These results underscore the need to consider how teachers and schools are using grade retention as a policy, and how the optimal retention strategy may differ substantially depending on the setting.

References

Abadie, A., S. Athey, G. W. Imbens, and J. M. Wooldridge (2023). When should you adjust standard errors for clustering? *The Quarterly Journal of Economics 138*(1), 1–35.

Andrabi, T., J. Das, and A. I. Khwaja (2008). A dime a day: The possibilities and limits of private schooling in pakistan. *Comparative Education Review 52*(3), 329–355.

Andrabi, T., J. Das, and A. I. Khwaja (2013). Students today, teachers tomorrow: Identifying constraints on the provision of education. *Journal of Public Economics 100*, 1–14.

Andrabi, T., J. Das, and A. I. Khwaja (2017). Report cards: The impact of providing school and child test scores on educational markets. *American Economic Review 107*(6), 1535–1563.

Andrabi, T., J. Das, A. I. Khwaja, T. Vishwanath, and T. Zajonc (2007). Learning and educational achievements in punjab schools (leaps): Insights to inform the education policy debate. *World Bank, Washington, DC*.

Becker, G. S. and N. Tomes (1976). Child endowments and the quantity and quality of children. *Journal of Political Economy 84*(4, Part 2), S143–S162.

Behrman, J. R., M. R. Rosenzweig, and P. Taubman (1994). Endowments and the allocation of schooling in the family and in the marriage market: the twins experiment. *Journal of Political Economy 102*(6), 1131–1174.

Bergman, P. (2021). Parent-child information frictions and human capital investment: Evidence from a field experiment. *Journal of Political Economy 129*(1), 286–322.

Borghesan, E., H. Reis, and P. E. Todd (2022). Learning through repetition? a dynamic evaluation of grade retention in portugal. *PIER Working Paper* (22-030), 1–80.

Cabrera-Hernandez, F. (2022). Leave them kids alone! the effects of abolishing grade repetition: evidence from a nationwide reform. *Education Economics 30*(4), 339–355.

Cameron, A. C. and D. L. Miller (2015). A practitioner's guide to cluster-robust inference. *Journal of Human Resources 50*(2), 317–372.

Chohan, B. I. and S. A. Qadir (2011). Automatic promotion policy at primary level and mdg-2. *Journal of Research & Reflections in Education (JRRE) 5*(1).

Cunha, F. and J. Heckman (2007). The technology of skill formation. *American economic review 97*(2), 31–47.

Eide, E. R. and M. H. Showalter (2001). The effect of grade retention on educational and labor market outcomes. *Economics of Education Review 20*(6), 563–576.

Eisemon, T. O. (1997). *Reducing repetition: Issues and strategies.* Unesco, International Institute for Educational Planning.

Eren, O., M. F. Lovenheim, and H. N. Mocan (2022). The effect of grade retention on adult crime: Evidence from a test-based promotion policy. *Journal of Labor Economics 40*(2), 361–395.

Ferreira Sequeda, M., B. H. Golsteyn, and S. Parra-Cely (2018). The effect of grade retention on secondary school performance: Evidence from a natural experiment. Technical report, IZA Discussion Papers.

Figlio, D. and U. Özek (2020). An extra year to learn english? early grade retention and the human capital development of english learners. *Journal of Public Economics 186*, 104184.

Glick, P. and D. E. Sahn (2010). Early academic performance, grade repetition, and school attainment in senegal: A panel data analysis. *The World Bank Economic Review 24*(1), 93–120.

Gomes-Neto, J. B. and E. A. Hanushek (1994). Causes and consequences of grade repetition: Evidence from brazil. *Economic Development and Cultural Change 43*(1), 117–148.

Hastings, J. S. and J. M. Weinstein (2008). Information, school choice, and academic achievement: Evidence from two experiments. *The Quarterly Journal of Economics 123*(4), 1373–1414.

Hu, L.-C. and E. Hannum (2020). Red flag: Grade retention and student academic and behavioral outcomes in china. *Children and Youth Services Review 113*, 104896.

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

Jacob, B. A. and L. Lefgren (2009). The effect of grade retention on high school completion. *American Economic Journal: Applied Economics 1*(3), 33–58.

Jann, B. (2017). kmatch: Stata module for multivariate distance and propensity-score matching. *Statistical Software Components.*

Jensen, R. (2010). The (perceived) returns to education and the demand for schooling. *The Quarterly Journal of Economics 125*(2), 515–548.

King, E. M., P. F. Orazem, and E. M. Paterno (2016). Promotion with and without learning: Effects on student enrollment and dropout behavior. *The World Bank Economic Review 30*(3), 580–602.

Kraft, M. A. and T. Rogers (2015). The underutilized potential of teacher-to-parent communication: Evidence from a field experiment. *Economics of Education Review 47*, 49–63.

Mahjoub, M.-B. (2017). The treatment effect of grade repetitions. *Education Economics 25*(4), 418–432.

Manacorda, M. (2012). The cost of grade retention. *Review of Economics and Statistics 94*(2), 596–606.

Nguyen, T. (2008). Information, role models and perceived returns to education: Experimental evidence from madagascar. *Unpublished manuscript 6*(5).

Oster, E. (2019). Unobservable selection and coefficient stability: Theory and evidence. *Journal of Business & Economic Statistics 37*(2), 187–204.

Rogers, T. and A. Feller (2018). Reducing student absences at scale by targeting parents' misbeliefs. *Nature Human Behaviour 2*(5), 335–342.

Todd, P. E. and K. I. Wolpin (2003). On the specification and estimation of the production function for cognitive achievement. *The Economic Journal 113*(485).

Tomes, N. (1981). The family, inheritance, and the intergenerational transmission of inequality. *Journal of Political Economy 89*(5), 928–958.

A.1 Appendix

A.1.1 Additional descriptive tables and figures

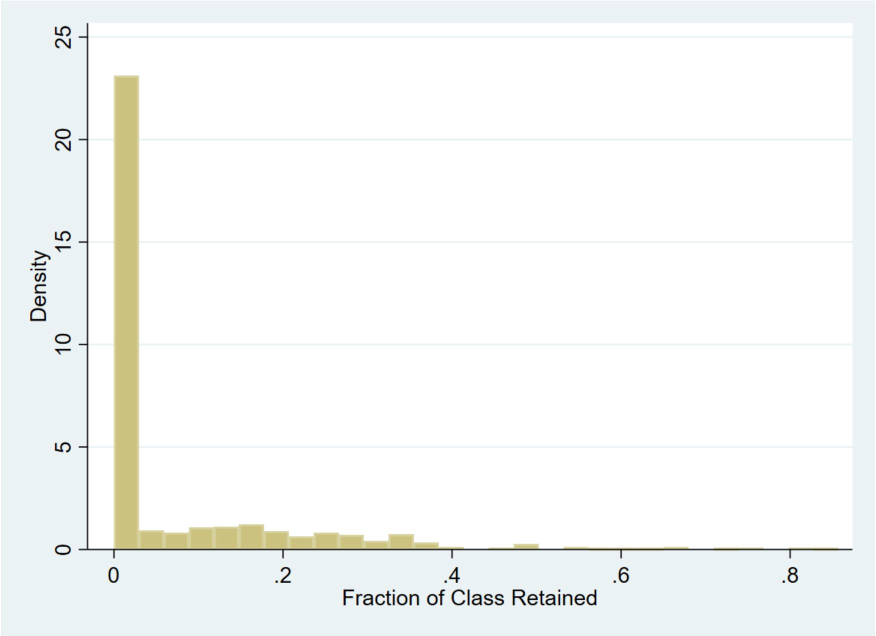
The first section of the Appendix provides additional tables and figures to support Section 3. First, Table A.1 shows that there is no difference in retention patterns across genders. Figure A.1 shows the distribution of the fraction of each class that was retained. Figure A.2 shows that teacher's ratings of students do not always align with their retention decisions. It shows one example classroom and the histogram plots the ratings of the students and whether they were retained or not. The full regression results associated with Table 3 can be found in Table A.2. Finally Table **??** provides additional conditional balance tests to supplement those shown in Table 4.

Table A.1: The probability of repeating by gender

	(1)	(2)	(3)
Female	**0.0064**	**-0.0095**	**0.0003**
	(0.0057)	(0.0107)	(0.0117)
Lagged Score			**-0.0762*** **
			(0.0091)
2yr Lagged Score			**-0.0493*** **
			(0.0106)
Constant	**0.1035*** **	**0.0979*** **	**0.0652*** **
	(0.0038)	(0.0072)	(0.0117)
AgeControls	No	Yes	Yes
Teacher Rating	No	No	Yes
Fixed Effect	-	School	School
Lag Polynomial	-	-	4th
N	11869	11361	6113
R^2	0.00	0.18	0.34

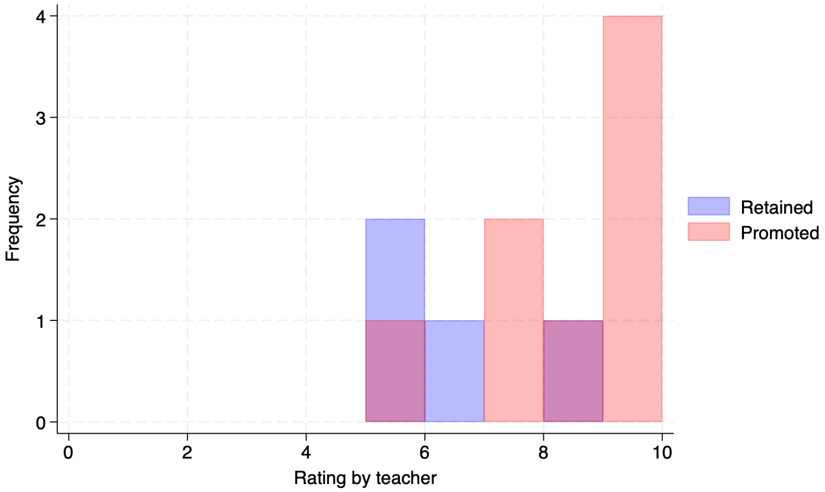
Note: Coefficient estimates bolded, standard errors (clustered at school-level) in parentheses below. The outcome variable is a dummy variable equal to one if the student repeats the grade. Lagged score refers to a student's previous year average score.
*** Significant at 1% level, ** Significant at 5% level, * Significant at 10% level

Figure A.1: Across-teacher distribution of the fraction of class retained



Note: Histogram of the fraction of each teacher's classroom who is retained (including classrooms without any retained students).

Figure A.2: Example teacher: frequency histogram of number of promoted and retained students with each rating



Note: Plot shows the number of retained and promoted students (y-axis) with different teacher ratings (shown on x-axis)

Table A.2: Regression Predicting Repetition

	(1)	(2)	(3)	(4)	(5)	(6)	(7) int	(8) int
Lagged Average Score	**-0.1099*** (0.0079)**	**-0.1037*** (0.0094)**	**-2.3451*** (0.4743)**	**-0.0872*** (0.0095)**	**-0.1226*** (0.0324)**	**-0.1304*** (0.0337)**	**-0.0389 (0.1111)**	**-0.0085 (0.1931)**
2yr lagged Average Score	**-0.0463*** (0.0095)**	**-0.0444*** (0.0111)**	**0.4857 (0.5080)**	**-0.0372*** (0.0110)**	**-0.0328 (0.0303)**	**-0.0199 (0.0450)**	**-0.1719 (0.1433)**	**-0.2791 (0.2704)**
Mom Educated		**0.0010 (0.0078)**	**0.0117 (0.1393)**		**0.0011 (0.0076)**			**-0.0771 (0.0594)**
Child Wealth		**-0.0038 (0.0022)**	**-0.0458 (0.0389)**					**-0.0036 (0.0214)**
Height		**-0.0020 (0.0046)**	**-0.0786 (0.0778)**					**-0.0001 (0.0396)**
Weight		**-0.0002 (0.0052)**	**-0.0136 (0.0875)**					**0.0071 (0.0444)**
Child Talent								
= 1				**0.1831*** (0.0206)**	**0.1756*** (0.0208)**			**0.1585 (0.1536)**
= 2				**0.1180*** (0.0186)**	**0.1114*** (0.0190)**			**-0.1605 (0.1339)**
= 3				**0.0713*** (0.0161)**	**0.0687*** (0.0162)**			**-0.1470 (0.1230)**
= 4				**0.0231 (0.0131)**	**0.0224 (0.0133)**			**0.1044 (0.1299)**
= 6				**-0.0208 (0.0117)**	**-0.0253** (0.0125)**			**-0.1889 (0.1063)**
= 7				**-0.0095 (0.0123)**	**-0.0216 (0.0135)**			**-0.1121 (0.1191)**
= 8				**-0.0070 (0.0126)**	**0.0052 (0.0155)**			**-0.0778 (0.1150)**
= 9				**0.0048 (0.0142)**	**-0.0022 (0.0184)**			**-0.0678 (0.1257)**
= 10				**-0.0019 (0.0162)**	**0.0314 (0.0227)**			**0.1208 (0.2219)**
Hardworking								
Below Average						**0.1650*** (0.0629)**	**0.1945*** (0.0656)**	**0.3568*** (0.1211)**
Above Average						**-0.0330 (0.0470)**	**-0.0600 (0.0521)**	**-0.0636 (0.0949)**
Intelligent								
Below Average						**-0.2381*** (0.0746)**	**-0.2619*** (0.0790)**	**-0.6469** (0.2744)**
Above Average						**0.0080 (0.0457)**	**0.0331 (0.0508)**	**-0.0598 (0.1267)**
Age/Gender Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Lag Polynomial 4th Order	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Test Score Components	No	No	Yes	No	Yes	No	Yes	Yes
Interact Scores, Child Chars	No	No	Yes	No	No	No	No	Yes
Interact Scores, TeachRatings	No	No	No	No	Yes	No	No	Yes
Interact Scores, ParRatings	No	No	No	No	No	No	Yes	Yes
Interact Par Ratings, TeachRatings	No	No	No	No	No	No	No	Yes
N	8071	5442	5434	5529	5529	698	698	487
Pseudo R^2	0.35	0.37		0.38	0.39	0.66	0.69	0.85

Note: This table shows coefficients from the regressions used to predict repetition used for computing the R^2's shown in table 3.

A.1.2 Robustness of main results

Our rich data contains many variables, however not all students in the main sample have responses recorded for all variables. This means that we face a tradeoff: including more variables or having a larger sample. For the main analysis, we have included the most important variables. In the following subsections we replicate the main results while including other variables that readers may find relevant. Note that the sample size usually decreases, at times substantially. The following tables all replicate our main results for academic outcomes, shown in Table 5. Our results are robust to all of the inclusions.

A.1.2.1 Controlling for parent ratings

We add baseline parent ratings separately to the specification from teacher ratings since only very small subsample received ratings from *both* their teachers and their parents. We have two separate ratings from parents: one of a child's intelligence and one of a child's work ethic. Since the two are highly correlated, we introduce them to the model one at a time.

While inclusion of parent rating in the regression substantially reduces power,[17] the effect of repetition remains large and significant. Table A.3 (columns 2, 3, 5, 6, 8, and 9) shows that, accounting for lagged parental ratings, grade repetition reduces test scores by between .29 and .44 standard deviations for the three subjects.[18] Comparing the second and third columns for each subject to the first, we see that the repetition coefficient decreases by at most only 0.024 standard deviations when including the parent ratings, suggesting that our baseline model isn't missing any important dimension of latent student ability.[19]

[17] These variables come from a household survey which was administered only to ten randomly selected households per village. The sample is all tested students matched with a surveyed household. Due to the reduced sample and lack of intra-school variation, we switch from school- to mauza-level fixed effects.

[18] Note that due to the smaller sample size, we switch from school- to mauza-level fixed effects here–otherwise, identification comes from only 36 repeaters and 58 non-repeaters in classrooms for which we observe both repeaters and non-repeaters.

[19] Columns 10-12 show the specification with dropping out as the outcome variable. While we lack sufficient power to claim an effect, it is reassuring that the coefficient barely changes between column 10 (the baseline specification on the smaller subsample) and columns 11 and 12 (the specification including parent ratings). Hence, the lack of significance seems to be due to small sample size rather than because the effect disappears when controlling for parent ratings.

Additionally, none of the lagged parent ratings significantly explain test score variation, controlling for lagged scores.

Table A.3: Effect of repetition on test scores - Considering characteristics observable to parents

	Math			English			Urdu			Dropout		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Not Promoted	**-0.433****	**-0.409****	**-0.436****	**-0.305****	**-0.291****	**-0.311****	**-0.393*****	**-0.369*****	**-0.389*****	**0.045**	**0.045**	**0.044**
	(0.179)	(0.181)	(0.179)	(0.143)	(0.143)	(0.142)	(0.143)	(0.137)	(0.141)	(0.036)	(0.036)	(0.036)
Lagged Score	**0.554*****	**0.549*****	**0.538*****	**0.481*****	**0.472*****	**0.484*****	**0.491*****	**0.478*****	**0.483*****	**0.001**	**0.001**	**0.002**
	(0.058)	(0.059)	(0.058)	(0.047)	(0.049)	(0.048)	(0.057)	(0.056)	(0.057)	(0.009)	(0.009)	(0.009)
2yr Lagged Score	**0.321*****	**0.317*****	**0.324*****	**0.317*****	**0.322*****	**0.317*****	**0.199*****	**0.196*****	**0.197*****	**-0.007**	**-0.007**	**-0.007**
	(0.067)	(0.067)	(0.065)	(0.053)	(0.053)	(0.052)	(0.064)	(0.062)	(0.063)	(0.007)	(0.006)	(0.007)
Hardworking (Lag)												
Below Average		**-0.196**			**-0.130**			**-0.146**			**-0.007**	
		(0.140)			(0.104)			(0.109)			(0.006)	
Above Average		**0.058**			**0.027**			**0.102****			**-0.008**	
		(0.061)			(0.055)			(0.050)			(0.006)	
Intelligent (Lag)												
Below Average			**-0.303**			**-0.112**			**-0.053**			**-0.001**
			(0.173)			(0.134)			(0.127)			(0.005)
Above Average			**0.054**			**-0.049**			**0.069**			**-0.009**
			(0.058)			(0.049)			(0.047)			(0.006)
Age/Gender Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Mauza Fixed Effect	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Lag Polynomial	4th Order	4th Order	4th Order	4th Order	4th Order	4th Order	4th Order	4th Order	4th Order	4th Order	4th Order	4th Order
N	698	698	698	698	698	698	698	698	698	760	760	760
R^2	0.68	0.69	0.69	0.73	0.73	0.73	0.74	0.74	0.74	0.25	0.26	0.26

Note: This table replicates the results shown in Table 5. The first column for each outcome variable is the exact same regression as in the main text, however the sample size is restricted to students who have are associated with a household survey, and therefore have parental responses. The results are robust.

A.1.2.2 Specification checks on controlling for two lags of test scores

The purpose of the value-added model is to “difference out” time-invariant factors. One way to check whether it succeeds is by testing if results are sensitive to inclusion of covariates that (1) should bias results in the absence of the value-added model and (2) we expect to remain constant year to year. If coefficients remain largely unchanged whether we include the variable or not and the time-invariant covariate doesn’t significantly explain student performance, it should be taken as evidence that the value-added model is effectively accounting for time-invariant factors. We consider three different time-invariant confounding factors at the household-, child-, and parent-level. Two of these (household wealth and child health) may be subject to small shocks year to year. In each case, absent the value-added specification, lack of inclusion would bias results. However, using the value-added model, results remain largely unchanged, indicating that the model is successful.

The first test variable is household wealth. It affects both achievement and repetition since wealthier households may spend more resources to help their children, including pressuring schools to promote their children to the next grade. Household wealth should also be relatively time-invariant (affecting past test scores as much as present test scores) so it should have little predictive power using the value-added model. We compute a principal components child wealth index based on 20 survey questions asking whether the child's household owned a variety of products.[20] Table A.4, column 2 shows results from our baseline specification plus the wealth index. The repetition coefficients remains similar in magnitude (differing by at most .002 standard deviations). Furthermore, as expected, the wealth index is a poor predictor of test scores, conditional on lagged test scores.

[20] These include beds, radio, television, refrigerator, bicycle, plough, small agricultural tools, tables, fans, tractor, cattle, goats, chicken, watches, motor rickshaw, motorcycle/scooter, car/taxi/van/pickup, telephone, and tubewell

Table A.4: Effect of repetition on test scores - Considering household wealth

	Math		English		Urdu		Dropout	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Not Promoted	-0.468***	-0.468***	-0.313***	-0.313***	-0.400***	-0.402***	0.074***	0.074***
	(0.054)	(0.054)	(0.037)	(0.037)	(0.042)	(0.041)	(0.015)	(0.015)
Lagged Score	0.501***	0.501***	0.440***	0.440***	0.493***	0.492***	-0.008	-0.008
	(0.026)	(0.027)	(0.023)	(0.023)	(0.024)	(0.024)	(0.007)	(0.007)
2yr Lagged Score	0.262***	0.262***	0.213***	0.213***	0.249***	0.251***	-0.017**	-0.017**
	(0.027)	(0.027)	(0.020)	(0.020)	(0.027)	(0.027)	(0.008)	(0.008)
Child Wealth		-0.002		-0.001		-0.009**		-0.000
		(0.005)		(0.004)		(0.004)		(0.002)
Constant	0.775***	0.774***	0.641***	0.641***	0.661***	0.659***	0.019***	0.019***
	(0.026)	(0.026)	(0.021)	(0.021)	(0.021)	(0.021)	(0.007)	(0.007)
Age/Gender Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
School Fixed Effect	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Lag Polynomial	4th Order	4th Order	4th Order	4th Order	4th Order	4th Order	4th Order	4th Order
N	5543	5543	5543	5543	5543	5543	6239	6239
R^2	0.72	0.72	0.78	0.78	0.75	0.75	0.19	0.19

Note: This table replicates the results shown in Table 5. The first column for each outcome variable is the exact same regression as in the main text, however the sample size is restricted to students who have a measure of household wealth. The results are robust.

Second, child health affects both achievement and promotion decisions (malnourished children may be held back due to perceived immaturity and also perform worse due to lower energy/cognitive capacity), however it also should be relatively time-invariant (aside from short-term shocks). Again, table A.5 shows that the value-added model is successful: after inclusion of the health variables, coefficients change by at most 0.015 standard deviations and become *more negative* for all three subjects. Additionally, height/weight explain little of the variation in test scores.

Table A.5: Effect of repetition on test scores - Considering child health

	Math		English		Urdu		Dropout	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Not Promoted	-0.463***	-0.478***	-0.309***	-0.313***	-0.389***	-0.396***	0.075***	0.078***
	(0.055)	(0.056)	(0.037)	(0.037)	(0.042)	(0.042)	(0.015)	(0.015)
Lagged Score	0.503***	0.500***	0.440***	0.441***	0.496***	0.497***	-0.007	-0.005
	(0.026)	(0.027)	(0.023)	(0.023)	(0.024)	(0.024)	(0.007)	(0.007)
2yr Lagged Score	0.261***	0.266***	0.213***	0.212***	0.249***	0.250***	-0.017**	-0.018**
	(0.027)	(0.027)	(0.020)	(0.020)	(0.027)	(0.027)	(0.008)	(0.008)
Height		-0.015		-0.012		-0.003		0.005
		(0.012)		(0.009)		(0.009)		(0.003)
Weight		-0.001		0.005		-0.015		0.008
		(0.013)		(0.010)		(0.010)		(0.004)
Constant	0.773***	0.770***	0.641***	0.642***	0.658***	0.651***	0.019***	0.025***
	(0.026)	(0.027)	(0.020)	(0.021)	(0.021)	(0.022)	(0.007)	(0.007)
Age/Gender Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
School Fixed Effect	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Lag Polynomial	4th Order	4th Order	4th Order	4th Order	4th Order	4th Order	4th Order	4th Order
N	5493	5442	5493	5442	5493	5442	6178	6123
R^2	0.72	0.72	0.78	0.78	0.75	0.75	0.19	0.20

Note: This table replicates the results shown in Table 5. The first column for each outcome variable is the exact same regression as in the main text, however the sample size is restricted to students who have a measure of child health. The results are robust.

A.1.2.3 Controlling for nonlinearities in teacher rating

We also show results are identical when including teacher rating as a categorical rather than continuous variable (only showed it as continuous in main table for convenience of exposition)

Table A.6: Main specification with factorized teacher rating

	Math		English		Urdu		Dropout	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Not Promoted	-0.472***	-0.440***	-0.313***	-0.278***	-0.402***	-0.371***	0.073***	0.070***
	(0.055)	(0.055)	(0.037)	(0.036)	(0.043)	(0.043)	(0.014)	(0.014)
Lagged Score	0.502***	0.467***	0.439***	0.404***	0.496***	0.467***	-0.007	-0.005
	(0.027)	(0.027)	(0.023)	(0.023)	(0.024)	(0.024)	(0.007)	(0.007)
2yr Lagged Score	0.258***	0.247***	0.213***	0.195***	0.245***	0.226***	-0.018**	-0.018**
	(0.027)	(0.027)	(0.020)	(0.020)	(0.027)	(0.027)	(0.009)	(0.008)
Child Talent								
= 1		-0.122		-0.060		-0.114**		0.027
		(0.066)		(0.050)		(0.056)		(0.019)
= 2		-0.096		-0.098**		-0.066		-0.001
		(0.054)		(0.040)		(0.043)		(0.013)
= 3		-0.120***		-0.052		-0.094**		0.021
		(0.045)		(0.036)		(0.037)		(0.014)
= 4		-0.071		-0.019		-0.036		0.003
		(0.039)		(0.028)		(0.031)		(0.009)
= 6		0.026		0.052**		0.052**		0.004
		(0.030)		(0.024)		(0.024)		(0.007)
= 7		0.033		0.081***		0.049**		0.001
		(0.030)		(0.025)		(0.024)		(0.007)
= 8		0.093***		0.133***		0.102***		0.000
		(0.033)		(0.025)		(0.026)		(0.008)
= 9		0.123***		0.166***		0.150***		-0.007
		(0.038)		(0.030)		(0.029)		(0.009)
= 10		0.176***		0.234***		0.166***		-0.009
		(0.040)		(0.032)		(0.031)		(0.010)
Age/Gender Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
School Fixed Effect	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Lag Polynomial	4th Order	4th Order	4th Order	4th Order	4th Order	4th Order	4th Order	4th Order
N	5529	5529	5529	5529	5529	5529	6222	6222
R^2	0.72	0.72	0.78	0.78	0.75	0.75	0.19	0.19

Note: This table replicates the results shown in Table 5. The results are robust.

A.1.3 Robustness to School Fixed Effects

Table A.7: Robustness checks: math outcomes

	Baseline		Single-teacher schools		Teacher FE (Y2)	
	(1) 1 yr	(2) 2 yrs	(3) 1 yr	(4) 2 yrs	(5) 1 yr	(6) 2 yrs
Not Promoted	**-0.440*****	**-0.322*****	**-0.539*****	**-0.420*****	**-0.508*****	**-0.459*****
	(0.055)	(0.076)	(0.075)	(0.106)	(0.059)	(0.085)
Fixed effects			School FE	School FE	Teacher FE (Y2)	Teacher FE (Y2)
Sample			Single-teacher schools	Single-teacher schools	All schools	All schools
N	5529	3992	4235	2997	5529	3992

Table A.8: Robustness checks: english outcomes

	Baseline		Single-teacher schools		Teacher FE (Y2)	
	(1) 1 yr	(2) 2 yrs	(3) 1 yr	(4) 2 yrs	(5) 1 yr	(6) 2 yrs
Not Promoted	**-0.278*****	**-0.338*****	**-0.306*****	**-0.371*****	**-0.304*****	**-0.388*****
	(0.036)	(0.050)	(0.045)	(0.065)	(0.039)	(0.051)
Fixed effects			School FE	School FE	Teacher FE (Y2)	Teacher FE (Y2)
Sample			Single-teacher schools	Single-teacher schools	All schools	All schools
N	5529	3992	4235	2997	5529	3992

Table A.9: Robustness checks: urdu outcomes

	Baseline		Single-teacher schools		Teacher FE (Y2)	
	(1) 1 yr	(2) 2 yrs	(3) 1 yr	(4) 2 yrs	(5) 1 yr	(6) 2 yrs
Not Promoted	**-0.371*****	**-0.364*****	**-0.414*****	**-0.406*****	**-0.412*****	**-0.419*****
	(0.043)	(0.056)	(0.056)	(0.073)	(0.044)	(0.055)
Fixed effects			School FE	School FE	Teacher FE (Y2)	Teacher FE (Y2)
Sample			Single-teacher schools	Single-teacher schools	All schools	All schools
N	5529	3992	4235	2997	5529	3992

Table A.10: Robustness checks: dropout outcomes

	Baseline		Single-teacher schools		Teacher FE (Y2)	
	(1) 1 yr	(2) 2 yrs	(3) 1 yr	(4) 2 yrs	(5) 1 yr	(6) 2 yrs
Not Promoted	**0.070*****	**0.009**	**0.090*****	**0.015**	**0.072*****	**0.009**
	(0.014)	(0.019)	(0.019)	(0.028)	(0.015)	(0.020)
Fixed effects			School FE	School FE	Teacher FE (Y2)	Teacher FE (Y2)
Sample			Single-teacher schools	Single-teacher schools	All schools	All schools
N	6222	4633	4789	3512	6222	4633

A.1.4 Propensity Score Matching

This appendix provides additional details on the implementation of the propensity score matching (PSM) robustness check discussed in Section 5. The objective of this exercise is to assess whether the estimated short-run effects of grade repetition are sensitive to differences in observable characteristics between retained and promoted students or to functional form assumptions imposed by the regression framework.

We implement propensity score matching using the `kmatch` package in Stata (Jann, 2017). For each outcome, we estimate each student's propensity score, $p(X_i) = \Pr(D_i = 1 \mid X_i)$, where $D_i = 1$ indicates that the student is retained between years t and $t+1$. The covariate vector X_i includes flexible functions of lagged and double-lagged test scores in the relevant subject, teacher ratings measured prior to the retention decision, basic student characteristics such as age and gender, and school fixed effects.

We match retained students to promoted students using nearest-neighbor matching within schools, imposing exact matching on school to ensure that all comparisons are made among students facing the same institutional environment. To further limit extrapolation, we restrict the analysis to regions of common support by discarding observations whose estimated propensity scores lie outside the region of overlap between retained and promoted students. We focus on the average treatment effect on the treated (ATT), which captures the effect of grade repetition for students who were actually retained. Standard errors are obtained via bootstrap.

Figure A.3 illustrates the distribution of estimated propensity scores for retained and promoted students for the English outcome one year after the retention decision. Consistent with the relatively low incidence of grade repetition, the propensity score distributions are right-skewed. Nevertheless, there is substantial overlap between retained and promoted students over the region of common support, indicating that retained students are matched to observationally similar promoted students with comparable predicted probabilities of retention.

Table A.11 reports covariate balance statistics before and after matching for the English outcome one year after the retention decision. Prior to matching, retained and promoted students differ substantially in terms of lagged achievement and teacher ratings, with standardized differences exceeding conventional thresholds. After matching, these differences

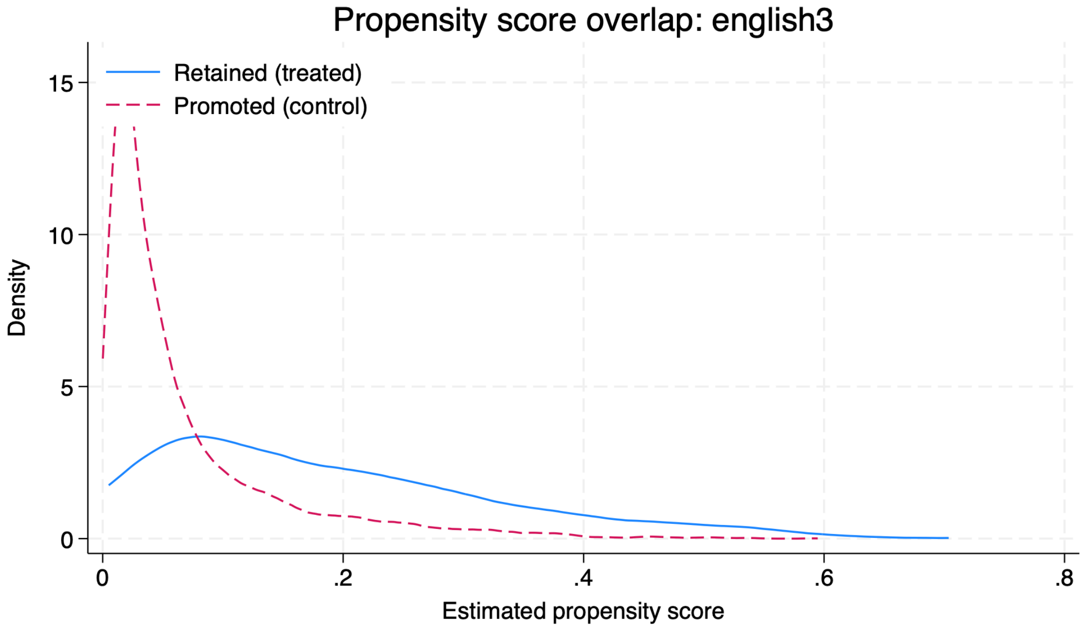


Figure A.3: Propensity score overlap for English (1-year outcome)

are markedly reduced. In particular, retained and promoted students are well balanced on teacher ratings and basic demographic characteristics, and balance improves substantially for lagged and double-lagged test scores. Balance patterns for math, Urdu, and dropout outcomes are qualitatively similar and are therefore omitted for brevity.

Table A.11: Covariate balance for propensity score matching (English, 1-year outcome)

	Raw	Matched (ATT)
	Std. diff.	Std. diff.
Lagged English score (t)	-0.904	-0.214
Double-lagged English score ($t-1$)	-0.743	-0.174
Teacher rating	-0.794	-0.008
Female	0.047	-0.016
Mean |Std. diff.| (all covariates)	0.252	0.078
Max |Std. diff.| (all covariates)	0.904	0.214

Note: This table reports standardized differences in covariates between retained (treated) and promoted (control) students before matching (Raw) and after matching (Matched ATT). Summary rows report the mean and maximum absolute standardized difference across all covariates included in the propensity score.

Together, these diagnostics confirm that the matching procedure compares retained students to a plausibly comparable group of promoted students and that the resulting ATT estimates are not driven by lack of common support or imbalance in observable characteristics.

A.1.5 Instrumenting Retention with School-Level Strictness: Additional Details

This appendix provides additional details on an instrumental variables (IV) robustness exercise based on variation in retention strictness across teachers and schools. The goal of this analysis is not to replace the baseline identification strategy, but rather to explore whether the main findings are sensitive to an alternative approach that instruments the retention decision using differences in promotion standards.

The institutional setting may appear, at first glance, amenable to a judge-style IV design, in which treatment assignment is instrumented by the strictness of the decision-maker. However, several features of the data and institutional environment limit the applicability of such an approach. Students are not randomly assigned to teachers, many schools have only a single Grade 4 class, and retention decisions are inherently relative rather than governed by a common cutoff, making the monotonicity assumption unlikely to hold. We therefore do not treat this setting as a canonical judge IV. Instead, we implement a more limited robustness exercise that exploits cross-school variation in retention strictness and interpret the resulting estimates with caution.

Construction of the strictness measure. We construct a measure of retention strictness based on the extent to which students are retained at rates higher or lower than would be predicted given observable characteristics. Specifically, we first estimate a linear probability model for the probability that student i is retained between years t and $t+1$:

$$D_i = \alpha_s + X_i'\beta + \varepsilon_i, \tag{3}$$

where $D_i = 1$ indicates retention, α_s denotes school fixed effects, and X_i includes flexible functions of lagged and double-lagged test scores, teacher ratings measured prior to the retention decision, and basic student characteristics such as age and gender. This specification captures systematic differences in retention risk arising from observed achievement, demographics, and school-level factors.

Using the fitted values from this model, we compute student-level residuals,

$$\widehat{u}_i = D_i - \widehat{D}_i, \tag{4}$$

which measure whether a student was retained more or less often than predicted based on observables.

We then define a leave-one-out measure of retention strictness for each student as the average residual retention among other students taught by the same Grade 4 teacher:

$$Z_i = \frac{1}{N_j - 1} \sum_{k \in j,\, k \neq i} \widehat{u}_k, \tag{5}$$

where j indexes teachers and N_j is the number of students taught by teacher j in the estimation sample. This leave-one-out construction ensures that the instrument is not mechanically correlated with student i's own retention outcome.

Intuitively, Z_i captures whether a student is exposed to a teacher who retains students more or less frequently than would be expected given observable student characteristics. Teachers with positive values of Z_i are relatively stricter, while those with negative values are relatively more lenient.

Identification and interpretation. Because many schools in the sample have only a single Grade 4 class, variation in the strictness measure is largely driven by differences across schools rather than within schools. Consequently, the IV specifications are estimated without school fixed effects and are identified from cross-school variation in retention strictness. The first stage is strong, indicating that the strictness measure is a powerful predictor of retention.

At the same time, this approach relies on substantially weaker identifying assumptions than the baseline regressions. Differences in retention strictness across schools may reflect unobserved institutional characteristics or educational practices that directly affect student outcomes, violating the exclusion restriction. Moreover, retention decisions do not satisfy the monotonicity condition required for a judge IV, as teachers may apply different criteria across students and contexts.

Consistent with these limitations, the resulting IV estimates for test scores and dropout one year after the retention decision are imprecise and centered near zero. While these results do not overturn the baseline findings, they are not sufficiently informative to draw strong causal conclusions. We therefore view this exercise as a supplementary robustness check that underscores the importance of within-school identification and motivates our focus on the main regression and matching-based analyses.