Remediation and Student Learning: Quasi-Experimental Evidence from Pakistan

Zarak Sohail *, Ijaz Hussain Bajwa †

November 4, 2021

Please Click for the Latest Version

Abstract

While enrollments across the developing world have drastically improved in recent times, learning levels remain low. In this paper, we use a natural experiment to study the impact of a remediation program that provided customized instruction to the bottom 3 to 7 students in a class from grades 3 to 8 in Pakistan. Using a Fuzzy Regression Discontinuity Design, we show that students in these remediation classes gained 0.288σ and 0.217σ in English and Math, but not in Urdu. Importantly, we find that remediation classes were equally beneficial at the primary and the secondary grades levels and across genders and class sizes. We investigate the mechanisms driving these findings, including the null result for Urdu, using a survey of school principals in our sample. The survey revealed that some schools assigned non-subject specialist teachers to Urdu remediation classes. In cases where subject-specialist teachers led these classes, we find positive effects of remediation on Urdu test scores. We conclude that remediation classes can be effective in mitigating the within-classroom learning disparities prevalent across the developing world. However, their impact is sensitive to program design and implementation.

*Department of Economics, University of California, Irvine. Contact: zaraks@uci.edu.

†The Citizens Foundation. We would like to thank Damon Clark for his constant guidance and comments. We would like to thank Matthew Freedman, Arman Rezaee, Sidra Haye, Usman Ghaus, Mark Hup, Nishtha Sharma, the participants at the Applied Micro Workshop at UCI, the participants at the UCI Labor-Public Seminar, the seminar participants at the Lahore Univesity of Managment Sciences, the Institute of Business Administration Karachi for their useful comments. We would also like to thank Sarfaraz Naeem and Kanwal Ilyas for providing excellent research assistance. Funding for this project was generously provided by the American Institute of Pakistan Studies Short-term Research Grant.

1 Introduction

While primary and secondary enrollment rates in developing countries have seen significant improvements in recent times, there is now convincing evidence that attending school does not guarantee learning (WorldBank, 2018; Pritchett & Sandefur, 2020). For instance, almost half of fifth grade students in Pakistan cannot read simple sentences or solve simple arithmetic expected of students in the second grade (ASERPakistan, 2019)[1]. Low levels of learning are perturbing since many researchers argue that test scores are a better predictor of economic growth and development than years of schooling (Hanushek & Kimko, 2000; Barro, 2001; Woessmann, 2016; Hanushek, 2013)[2].

One binding constraint in converting schooling into learning in developing countries is the mismatch between the level of instruction in the classroom, determined by the curriculum, and student learning levels (Banerjee et al., 2007; WorldBank, 2018; Muralidharan et al., 2019). Policies that led to the expansion of education in developing countries did not lead to an accompanying adjustment of the curriculums or the pedagogical practices to cater to the influx of first-generation learners. Consequently, students in these countries exhibit significantly higher variations in within-class learning levels compared to developed countries (Glewwe & Muralidharan, 2016; Glewwe et al., 2009). These learning gaps subsequently widen as students progress towards higher grades and as curriculums become progressively demanding (Muralidharan & Zieleniak, 2014; Muralidharan et al., 2019)[3]. Given that learning is a key link between schooling and human capital formation, a central question for policymakers in the developing world is to identify cost-effective tools to improve student learning.

In this paper, we study the impact of a low-cost remediation program on the test scores of poor-performing students in Pakistan. The *Uraan* program was implemented by The Citizens Foundation (TCF), the largest education non-governmental organization (NGO) in Pakistan that operates over 1,500 schools spread across the country and provides low-cost

[1]Pakistan also ranks only above the Philippines in terms of ranking in the international TIMSS test for math and science achievement for grade 4 (Mullis et al., 2020)

[2]Hanushek (2013), for instance, shows that once measures of cognitive skills (proxied by test scores) are included in cross country regressions of growth, the coefficient on years of schooling is small and not significant. Thus only that portion of schooling that is directly related to skills has any impact on cross country differences in growth

[3]Muralidharan et al. (2019) show that not only are students in India several grades below their expected grade levels in terms of their learning, student achievement varies significantly within peers and this variation in learning exacerbates at higher grade levels.

education to approximately 220,000 children from poor households. The program provides customized remedial instruction to a small group of poor-performing students, ranging from 3 to 7 students, in grades 3 to 8 for English, Urdu, and Mathematics[4]. This restriction to the number of available seats in each class gives rise to a school-grade-subject specific rank cutoff that determines a student's eligibility into the program. We show that the probability of being assigned to remediation increases sharply at this cutoff, and we leverage this using a fuzzy regression discontinuity design. Intuitively, this identifies the causal effects of remediation assignment by comparing learning outcomes among students just below and just above this cutoff.

There are several reasons why remediation classes may be an effective tool to mitigate within-class learning disparities. First, a growing literature reveals that learning in developing countries is generally unresponsive to more inputs such as reduced class sizes (E. Duflo et al., 2015; Banerjee et al., 2007), textbooks (Glewwe et al., 2009), flip charts (Glewwe et al., 2004), and diagnostic feedback to teachers on student learning (Muralidharan & Sundararaman, 2010) unless provided in a targetted manner. Second, remediation programs such as the *Uraan* are very low-cost and thus scalable. For instance, while teachers did not get paid for the additional teaching time under the *Uraan* program, we estimate the shadow cost of the program to range between \$1.4 to \$1.7 per child per term[5]. Third, there is some evidence that targeted pedagogical interventions such as remediation classes improve learning. For instance, Banerjee et al. (2007) study the impact of a remedial education program for 3rd and 4th-grade students in India and find a 0.28σ improvement in Math and language test scores[6]. Similarly, Banerjee et al. (2010) find that after-school reading camps for students unable to read improved their ability to read letters by 7.9 percentage points. Moreover, A. Duflo et al. (2021) find a 0.14σ improvement in Math and English test scores in grades 2 and 3 as a result of remediation classes in Ghana.

However, important questions remain. First, it is unclear whether the results from the past literature extrapolate outside of the limited number of settings where these programs

[4]Data exploration and informal interviews of principals reveal that class sizes mostly remain fixed at the school-grade level from one term to the next. Furthermore, the size of these remediation classes varies across schools but remains similar within schools and across grades and subjects

[5]Section 6 details a speculative cost and benefit analysis of the program

[6]Their instrumental variables analysis shows that these effects are as high as 0.60σ for students assigned to remediation classes

were evaluated[7]. Second, while pedagogical interventions such as remediation have been successful at the primary level, it is unclear whether such interventions improve learning at higher grades (Muralidharan et al., 2019). Understanding whether remediation impacts learning at the post-primary levels is crucial. On the one hand, learning variations at post-primary levels may be too high for remediation classes to be effective. On the other hand, these wide variations in learning may imply potentially larger returns to such interventions at higher grades. Third, student-teacher ratios may be an important factor contributing to learning in pedagogical interventions such as remediation classes that involve focused teaching. Despite this, the relationship between student-teacher ratios and learning in remediation classes has received little attention from the previous literature. Lastly, while the previous studies indicate that volunteers and contract teachers can improve student performance through remediation classes in settings where the focus is on foundational competencies, it is unclear whether teacher type matters for learning in settings where both foundational and grade-level topics are the focus.

Our analyses reveal four main sets of findings. First, consistent with the previous literature, we find that remediation improved student test scores by 0.288σ and 0.217σ for English and Math. Contrary to the gains in English and Math scores, however, we find little evidence of improvements in Urdu test scores from Urdu remediation. The coefficients for Urdu test scores are small and statistically not significant. The test score gains for Math and English are equivalent to around 11 percent of the test score gap between the 10th and the 90th percentile student in the initial test score distribution in our setting. We show that our estimates are robust against a range of validity checks, including variations in the bandwidth around the cutoff point, alternative definitions of the outcome variable, and alternate functional forms of the running variable.

Second, we find that the effects of remediation are statistically similar across primary and secondary grade levels and across class sizes. These results suggest that remedial classes may be an effective pedagogical intervention to mitigate the learning variations within peers even at higher grades where curriculums become progressively demanding. Additionally, to the extent that these estimates can be extrapolated to larger class sizes, our results suggest

[7]For instance, there is wide variation in the impact of remediation *within* India with Banerjee et al. (2007) finding larger effects in Vadodara than Mumbai, primarily due to differences in baseline achievement. Furthermore, Pakistan has much lower enrollment rates than India or other comparable countries at the primary and secondary levels. As enrollments in Pakistan increase, greater variations in within-class learning are expected.

that the size of these classes may be increased to cater to a larger group of students[8]

Third, we do not find any evidence of spillover effects of remediation (either positive or negative) on the test scores for other subjects. For instance, remediation for Urdu does not affect tests scores for other subjects that are taught in Urdu (these include Science, Social Studies, and Islamic Studies). Similarly, we do not find any evidence for substitution in effort from one subject to another. For instance, remediation for Math does not affect test scores for Urdu or English. Furthermore, while remediation classes improved test scores for English and Math, our results suggest that these classes did not affect time-in-school outcomes such as attendance and drop-out rates. We also find some evidence of fade-out in the effects of remediation although these estimates are noisy due to high drop-out rates in our setting[9]

We investigate some of the mechanisms driving the effects we estimate, including the null findings for Urdu scores, through a survey of principals for the schools in our sample[10]. Through our survey, we gathered information on teacher appointments to these remediation classes, and the principals' perception regarding the success of the program and the motivation of teachers and students involved in the program. Our survey results reveal that contrary to English and Math, many of the teachers assigned to Urdu remediation classes were not subject specialist teachers. Instead, any available teacher may be appointed to teach remediation classes for Urdu. Using this variation in teacher assignment for Urdu, we find that the null results for Urdu are driven primarily by the appointment of non-specialist teachers to Urdu remediation classes. For the sample of schools where Urdu teachers taught Urdu remediation classes, we find that the impact of remediation was 0.222σ on Urdu test scores[11].

We draw two main conclusions from our results. First, we argue that remediation classes can be a cost-effective tool to mitigate learning disparities at both the primary and the

[8] Banerjee et al. (2007) show that remediation classes improved learning in remediation classes as large as 15-20 students.

[9] For instance, the positive gains from remediation faded out to around 0.12σ for Math and English one term later, both not significant. We show that a student's remediation status in the current cycle does not affect her future remediation status, suggesting that repeated exposure to remediation does not drive these results.

[10] COVID-19 restrictions dictated that these be conducted over the telephone. Out of the 75 schools in our sample, we were able to reach 61 principals. We were unable to survey the rest of the principals because they had left the organization.

[11] We are cautious not to interpret these estimates as causal since teacher type is not randomly assigned. Non-specialized teachers may generally be worse than specialized teachers and schools that assign non-specialized teachers may generally be lower in quality than schools that assign specialized teacher to these classes. Nevertheless, we carry out placebo checks in Table A.2 to mitigate these concerns.

secondary levels of education. In doing so, we mitigate concerns regarding the effectiveness of Teaching at the Right Level (TaRL) interventions such as remediation classes at post-primary grade levels (Muralidharan et al., 2019). A speculative cost and benefit analysis reveals a Net Present Value (NPV) of the future labor market gains from the program that ranges between \$44 and \$107 making the *Uraan* program highly cost-effective[12]. Given the low running costs and the logistical simplicity of implementing remediation classes, such programs can also prove useful in mitigating COVID-19 related learning losses across the developing world as schools begin to re-open[13].

Second, we directly contribute to the discussion on the design and implementation of effective educational interventions by showing that the positive impact of remediation classes does not depend on the size of these classes but is sensitive to the type of teacher teaching these classes, at least in our setting. While previous literature shows that volunteers and contract teachers may be effective in teaching basic-level competencies in remediation classes at primary levels of education, we argue that such teachers may not be as effective in situations where the content taught includes grade-level material and is flexible to the needs of the students, especially at higher grade levels.

The paper proceeds as follows. Section 2 outlines the details of the *Uraan* remediation program. Section 3 introduces the data; Section 4 outlines the empirical strategy used; Section 5 presents the results; Section 6 sheds light on the mechanisms and details a cost and benefit analysis and Section 7 concludes.

2 The *Uraan* Remediation Program

The *Uraan* remediation program was implemented by The Citizens' Foundation (TCF), the largest school-based Non-Governmental Organization in Pakistan. TCF operates around 1500 schools across Pakistan that provide low-cost education to more than 220,000 under-

[12]These estimates are sensitive to assumptions about wage levels, wage growth, the discount rate, the length of the stream of income, and the age at which these students are expected to start working. We provide the details for these calculations in section 6

[13]Many developing countries moved to remote learning methods as schools closed due to COVID-19. However, the access and use of remote learning technologies has been sub-optimal, particularly in low-income households (Akmal et al., 2020). For instance, studies reveal that the uptake of distant-learning tools has been limited in Pakistan during COVID-19 related school closures. A World Bank survey suggested as little as 30% of the households in Punjab knew about remote learning opportunities and only 10% of the households utilized such resources for learning (Geven & Hasan, 2020)

privileged children. The *Uraan* remediation program provides customized remedial education to the bottom 3 to 7 students in grades 3 to 8 for English, Urdu, and Math. These classes are held twice a week for each subject and are approximately 30 minutes long. Teachers identify the academic weakness of the students based on previous tests and focus their instruction around those topics.

The program was operationalized in two cycles each academic year, with each cycle coinciding with an academic term. A clear score threshold does not dictate the eligibility for enrollment in these classes. Instead, principals decide the class sizes for each cycle and identify low-performing students based on the scores in previous standardized exams to fill these seats. While the class sizes can vary between 3 and 7 students, we show below that the modal class has 5 students and that these class sizes usually remain fixed over time and within schools and across subjects[14]. This system of class size selection implies that the size of these classes is not endogenous to the unobservable characteristics of the students such as the unobserved ability.

Figure 1 illustrates the structures of the remediation program and the selection criteria for remediation. The figure shows that the students eligible for remediation in cycle two are identified based on their scores on the midterm exam at the end of term one. Similarly, students in cycle one of the program are selected based on their scores at the end of term two of the previous academic year. Note that both the midterm and the final exam are similar across schools as they are centrally set and distributed across schools. Furthermore, teachers are provided with detailed rubrics to grade these tests. The eligibility criteria, combined with the fact that each remediation class has a limited capacity give rise to a rank threshold such that those below the rank cutoff are eligible for remediation while those above it are not. We use this feature of the program to estimate the causal effects of remediation on student learning.

Several features differentiate the *Uraan* remediation program from previously tested remedial programs. First, instead of focusing on a pre-determined standardized curriculum, the *Uraan* remediation program identifies the academic weaknesses of each incoming batch of students and organizes instruction around those weak points. While this feature ensures customized targeted instruction for each child, we show below that the success of such a

[14] That is, while the size of these classes may vary across schools, they mostly are similar within a given school suggesting that the principals stick with the same number of students in their schools

system relies heavily on implementation. Second, the *Uraan* program's eligibility criteria and the variation in the size of the regular classes lead to variations in the position of the marginal student in the baseline score distribution. This feature allows us to speak to the external validity of the program to other settings. Lastly, remediation is provided across primary and post-primary grade levels which allows us to estimate the impact of remediation at the various grade levels.

3 Data

Data for this study come from the administrative records of The Citizens' Foundation. These pertain to 75 schools in Karachi, the largest city in Pakistan. We link standardized test scores to remediation assignment data across two remediation cycles for the 2018-2019 and 2019-2020 academic years. We make two sampling restrictions. First, we exclude dropouts from our analysis because we do not have post-remediation scores for these students. To alleviate the concern that remediation could impact dropout and thereby bias our estimates, we show that remediation has little impact on dropout. Second, we exclude newly admitted students from our analysis. This is because these students are assigned to remediation classes based on a test taken upon admission to the school and because we do not observe these admission test scores.

Table 1 provides an overview of the data and the variables used in the analysis that follows. Columns 1 and 2 provide the summary statistics for the full sample, pooled across the two remediation cycles and 75 schools. Columns 3 and 4 limit the sample to those students who fall in the -7 to +15 relative ranks for Math. Similarly, columns 5 and 6 limit the sample to those students who fall within -7 to +15 relative ranks for English, and columns 7 and 8 for Urdu. Panel A presents the pre-remediation raw test scores and the attendance rates of students. It shows that students generally perform better in Urdu than either English or Math, scoring roughly 47 percent in Urdu. Furthermore, students perform much better in languages than Math, scoring 35 percent on average in Math.

As expected, students in the restricted sample perform worse on average for each of the subjects. For instance, an average student in the full sample scores 46 percent in Urdu whereas an average student in the restricted sample scores around 39 percent. Students in the restricted bandwidths also score lower in other non-remediation subjects such as Islamic

Students, Science, and Social Studies. However, they are very similar in terms of their age (11 years) and attendance (87 percent).

Panel B of the table shows that a higher share of students (54 percent) in the sample are in primary school (grades 3 to 5). This ratio remains consistent within the restricted samples for the three subjects. One exception to this is Urdu where 62.1 percent of the students at the low end of the performance distribution are in primary school. Panel B also shows that the ratio of girls in the restricted samples for Urdu and English is lower than the ratio of girls in the full sample. This trend is consistent with the fact that school-going girls outperform boys, especially in languages. This difference in performance may be due to the selection of girls into schools given the gender disparities in enrollments in Pakistan.

Panel C presents statistics on the remediation classes. The mean number of students in these remediation classes is 5 and attendance rates in these classes are comparable to attendance rates in regular classes. Panel D presents the pre-remediation test scores for girls and boys separately. Consistent with the above observation, Panel D shows that girls greatly out-perform boys in languages. On average, girls score almost 44.5 percent and 49.5 percent in English and Urdu standardized tests while boys score 37.5 percent and 41 percent. The difference between the test scores of girls and boys in Math scores is less pronounced.

Lastly, Panel E divides the sample by the level of education. As expected, students generally perform better in primary school than in secondary school (grades 6-8). Worryingly, students grossly under-perform in Math when they reach higher grades, scoring only 25 percent on average. Additionally, the table shows that the variation in scores is higher at post-primary levels suggesting that the learning gaps widen at higher grades. Such a drop in test scores underscores the need for targeted pedagogical interventions such as remedial classes.

We combine the administrative data on test scores with a survey of principals for the schools in our sample. Out of the 75 schools, we were able to reach the principals of 61 schools. We were not able to survey the rest of the principals primarily because they had left the school. The primary aim of the survey was to assess the quality of program implementation. First, we asked the principals regarding the student selection criteria for each grade and subject. In line with the official policy, nearly all the principals answered previous performance as being the main criteria for eligibility for remediation. Next, we asked the principals regarding the assignment of teachers to these remediation classes. While subject specialist teachers were exclusively assigned to teach English and Math remediation classes, around half of the schools appointed non-subject specialist teachers to Urdu remediation classes. This variation in teacher assignments may be due to there not being an official policy for teacher assignments. We use this variation in teacher assignments to these remediation classes to explain our results later on.

4 Empirical Strategy

In the absence of an actual experiment, a key challenge in estimating the causal effects of remediation on test scores is that the assignment to remediation may not be random. The students' remediation status may correlate with some unobservable characteristics that may also affect learning outcomes and hence the probability of being assigned to remediation. This correlation with unobservable characteristics will bias the estimates of the program effect. We circumvent this issue by exploiting the existence of a rank threshold that emerges due to the limited number of students in a given remediation class. Given this rank threshold, we generate the rank for each student relative to the last student eligible for remediation for English, Urdu, and Math separately. These relative ranks serve as the forcing variable such that students with a relative rank of 0 and below are eligible for remediation while students with a positive relative rank are ineligible for remediation.

We estimate the causal impact of remediation on student learning outcomes using a Regression Discontinuity Design (RDD). This research methodology compares students' learning outcomes very close to a rank cutoff point (a relative rank of zero in our setting) that presents a scenario where the assignment into remediation classes is as-if random.

Figure 2 plots the relationship between the probability of being assigned to remediation and relative ranks for the three subjects. The graphs show large discontinuities at the cutoff point. Students falling below the threshold are approximately 48 to 51 percent more likely to be enrolled in a remediation class.

Additionally, the graphs depict a 'fuzziness' in the assignment to remediation. That is, being above or below the cutoff point does not fully determine a student's remediation status. Instead, it leads to a discontinuity in the probability of being assigned to remediation. Data exploration suggests that this is driven primarily by new admissions. Newly enrolled students are enrolled in these classes based on their school entrance exam scores. In cases where the new students get assigned to these remediation classes, the marginal regular student eligible for remediation does not get enrolled in these classes. This also explains the downward trend on the left-hand side of the cutoff. Since we do not have information on admission test scores of the new admissions, we drop these students from our analysis sample. Nevertheless, they do affect the size of the remediation classes and hence influence the first stage estimates[15].

Given this setting, we implement a fuzzy regression discontinuity (FRD) design to estimate the causal impact of remediation. Our empirical strategy follows that of Card & Giuliano (2016) closely. Specifically, we use the student's eligibility (an indicator being above or below the rank threshold) as an instrument for actual assignment to remediation. The causal impact of remediation is then the effect of remediation classes on the scores of students who got remediation *because* they were eligible for them (that is, the compliers). The first stage equation takes the following form:

$$d_{ijk} = \gamma_1 z_{ijk} + \gamma_2 r_{ijk} + \gamma_3 r_{ijk} z_{ijk} + \Gamma X_{ijk} + \epsilon_{ijk} \tag{1}$$

Equation 1 relates the probability of getting remediation(d_{ijk}) to the eligibility status (z_{ijk}) of student i in grade j in school k. Furthermore, r_{ijk} represents the rank of the student relative to the cutoff rank that determines eligibility for remediation. Hence, z_{ijk} takes the value of 0 if r_{ijk} is less than zero for a given student in a given subject, and one otherwise. $r_{ijk} z_{ijk}$ is the interaction term between the eligibility dummy and the relative rank. The interaction term allows for the slopes of the relationship to be different on either side of the cutoff point. Finally, X_{ijk} is a vector of student-level controls that include age and gender, and ϵ_{ijk} is the error term. The coefficient of interest in the first stage is γ_1 which represents the jump in the probability of being assigned to remediation given the eligibility status of the student. The second stage equation takes the following form:

$$y_{ijk} = \beta_1 \hat{d_{ijk}} + \beta_2 r_{ijk} + \beta_3 r_{ijk} z_{ijk} + \Lambda X_{ijk} + v_{ijk} \tag{2}$$

Equation 2 relates the the predicted values $\hat{d_{ijk}}$ from the first stage to the test scores y_{ijk}. The coefficient of interest is β_1, which provides us with the causal impact of remediation on student test scores. For the empirical model to be valid, we assume the eligibility status to

[15]Given the curvature of the points on the left-hand side of the cutoff, especially for Urdu, in table A.3 we test for the robustness of the estimates to a second-order polynomial of the relative ranks and find that our results are consistent.

be uncorrelated with the error term $[cov(z_{ijk}, \epsilon_{ijk}) = 0]$.

Validity of the RD Design

The validity of the RD design hinges on the inability of the students to self-select into either side of the cutoff point. If students move to either side of the cutoff point, then the two groups are incomparable. Figure A.1 plots the relationship between a student's relative rank in each of the three subjects and several pre-remediation outcomes. Each column represents RD plots for the subsample within the -7 to +15 ranks for Urdu, English, and Math and each row of graphs represents a pre-remediation outcome. The graphs show no discontinuities in gender, age, pre-remediation test scores, or attendance rates at the cutoff. Table A.1 quantifies these estimates and reiterates the smoothness of observed variables at the cutoff point[16].

While most remediation classes have 5 students, there is some variation in the size of these remediation classes. One concern that may stem from this is that these variations in class sizes lead to a discontinuity in the unobservable characteristics of students. For instance, principals might set remediation class sizes by identifying students that remediation the most (based on unobserved ability), leading to a bias in our estimates. One way to address this concern is to check if the class size varies from one term to another at the grade-school level. If, in practice, principals have a fixed class size that they prefer from one cycle to the next that will mitigate any concerns regarding selection.

We show the raw data comparing the remediation class sizes from one term to another in Table A.9. The table shows that around half of the remediation classes remain similar in size from one remediation cycle to another. Roughly 82 percent of the classes maintain a plus or minus one class size from the previous term. In cases where there is variation in class size from one term to another, we show that the change is the same across subjects in around 88 percent of such cases. This system implies that class sizes are not endogenous to the unobserved characteristics of students such as need and ability. Nevertheless, to check for the robustness of our estimates to this variation in the number of students, Appendix A presents results from the donut-hole RD approach that leaves out observations that are ± 1

[16] An alternate test is to plot the histograms of the running variables and check for discontinuity in the density of the running variable at the cutoff point. Such a test will not be reliable in our setting since the nature of the relative ranks dictate that there will be no discontinuity in the density at the threshold point

of the cutoff (Bajari et al., 2011)[17]. We show that our results remain robust to this check.

There remain, however, roughly 18% of the grades-schools that have a difference of more than one student in remediation classes across the two cycles for which we have data. To address the concerns of selection arising from the variation in class sizes over the two terms across a small sample of grades, we carry out an additional robustness check by using a simulated threshold for each grade school. More specifically, we use term 1's number of students in a remediation class to form the threshold for term 2's remediation session. Appendix B presents these results and shows our results remain consistent with these restrictions in place.

An additional cause for concern in this setting might be that students drop out *because* of being assigned to remediation. Differential attrition at the cutoff point will bias our estimates upwards because students who remain in school may stand to gain more from remediation. Figure A.2 addresses these concerns and shows that students do not drop out deferentially at the cutoff point.

5 Results

5.1 The Impact of Remediation on Test Scores

This section tests whether the *Uraan* remediation program improved the learning outcomes of students in our setting. Our main outcome of interest are test scores defined as the residualized post-remediation test scores. Essentially, we regress post-remediation test scores on a flexible function of pre-remediation test scores and school fixed effects. We then take the difference between the predict scores from the regressions and the actual scores. These residuals then form our outcome and represent the amount students scored above or below what was expected given their past peformance (Kane, 2017; Castellano & Ho, 2013).

Figure 3 plots the relationship between the students' relative ranks and their test scores for English, Math, and Urdu. The RD plots reveal clear discontinuities for English and Math test scores. The magnitude of the discontinuities ranges between 0.13σ and 0.15σ for English and Math, both significant at least the 5 percent level. Contrary to the positive effect of

[17] This approach achieves two goals. The first is to assess the influence of the few observation closest to the cutoff point. Second, if there is endogenous sorting across the rank threshold, such sorting might only happen among students closest to the cutoff point. These observations are then excluded from the analysis (M. Cattaneo & Rocio, in press)

remediation on English and Math test scores, we find that remediation had little impact on Urdu test scores. The discontinuity in Urdu test scores is very small and not statistically different from zero.

Table 2 complements figure 3 by reproducing the estimates from the RD plots (panel A, column 2) and checking their robustness to additional controls. Column 3 adds student-level controls such as gender and age, column 4 adds students' prior scores to the list of controls, and column 5 adds grade fixed effects[18]. The table shows that the reduced form estimates are robust to the addition of various controls. In summary, our reduced-form estimates suggest that remediation improved the test scores for English and Math but did not seem to impact Urdu test scores. These estimates, however, do not take into account the fuzzy first stage, a task we take up next.

Panel B of Table 2 takes into account the fuzzy first stage and presents the main results of the paper. Results show that the *Uraan* remediation program has been successful in improving student test scores for English and Math. Estimates show a 0.288σ to 0.217σ gain in English and Math, both highly significant. The program, however, did not have any effect on Urdu scores with the estimates remaining small and statistically insignificant.

One concern regarding our results may be that these are a compound effect of both remediation and additional class time. Although isolating each effect is not possible in our setting, we present two arguments why extra class time is not driving these results. First, if additional class time is positively associated with test scores, we should observe gains in test scores for all remediated subjects. Instead, we find no effect for Urdu. Second, studies in other settings have shown that private tuition or extra class times and smaller classes did not have an impact on student learning (Banerjee et al., 2007; Berry & Mukherjee, 2016).

There are various ways to think about these test score gains. Firstly, the magnitude of these improvements in test scores for Math and English is equivalent to around 11 percent of the initial test score gap between the 10th and the 90th percentile student in the initial test score distribution[19]. Secondly, Bau et al. (2021) estimate the learning trajectories in Pakistan

[18]Note that schools fixed effects are excluded from these regressions since differences across schools were accounted for in the creation of the outcome

[19]We calculate these numbers as follows

$$\frac{coefficient}{90^{th} Percentilescore - 10^{th} Percentilescore} * 100$$

and find a 1.19 S.D gain in test scores as a result of 4 years of schooling. Consequently, the test score gains observed in our setting translate roughly into 0.8 years of additional schooling for students that received remediation[20]. Alternatively, one can think of these improvements in tests scores in terms of future labor market returns, assuming that these test score gains translate into future human capital (Muralidharan & Sundararaman, 2011). Using estimates on the wage returns to test score gains in Pakistan for girls and boys from Aslam et al. (2011), our results approximate to 3.45%and 8.72% gains in wages for boys and girls[21].

In Table 2, we restrict the sample to a bandwidth of 7 ranks to the left of the cutoff and 15 to the right[22]. A key concern is whether these estimates remain robust to variation in bandwidths around the cutoff rank. Moving further away from the bandwidth would reduce the variation in the estimates because of a larger sample size but reduce their reliability. Contrarily, narrowing the bandwidth will improve the reliability of the estimates but increase the variation due to the smaller sample size. Figure 4 plots the 2SLS coefficients (y-axis) along with a 90 percent confidence interval for various bandwidths (x-axis) across the cutoff point. Due to sample size limitations, we fix the left-hand bandwidth to 7 relative ranks and vary the right-hand bandwidth from 5 to 20 relative ranks. The figure also plots the optimal bandwidths found using the cross-validation method (Imbens & Lemieux, 2008)[23]. Reassuringly, the figures show that our 2SLS estimates are robust to the choice of the bandwidth for both outcomes. Consistent with our earlier findings, the estimates for the effect of remediation on Urdu scores remain insignificant across the range of the bandwidth considered. Naturally, the estimates are noisier at narrower bandwidths due to restricted sample sizes but remain significant for English and Math at the 10 percent level for a large

[20]score gains are converted to years of schooling as follows:

$$\frac{coefficient}{1.19} * 4$$

[21]Aslam et al. (2011) estimate the returns to a 1 SD improvement in test score in Math and English to be 14 and 13 percent for boys and 36 and 32 percent for girls. Using these estimates, we covert the test score gains in our setting to wage returns for girls as follows:

$$\frac{gain_{math} * 0.32 + gain_{english} * 0.36}{2} * 100$$

A similar calculation is done for boys

[22]This bandwidth was chosen because on the left hand side of the cutoff we do not have much margin to vary the bandwidth while on the right hand side a bandwidth of 15 splits the data in half.

[23]Other methods of optimal bandwidth selection cannot be used in our setting because the running variable is discrete.

range of bandwidths across the cutoff point.[24]

5.2 Heterogeneity in the Impact of Remediation

The above analysis underscores the success of the program in improving student test scores for Math and English. A key question is whether these gains are similar for various groups of students, a task we take up in this section.

5.2.1 Effect of Remediation by Gender

In this subsection, we test whether remediation affected test scores for girls and boys differently. More specifically, we regress test scores on a fully interacted model of covariates and the gender dummy. The coefficient of interest is the one on the interaction between the indicator variable for remediation and the indicator for female. We instrument this term with the interaction between the indicator for eligibility and the indicator for gender.

Consistent with previous literature, we find that remediation was equally effective in improving test scores for English and Math for girls and boys. Figure 5 plots test scores for girls and boys against their relative rank for each of the three subjects. The graphs provide evidence of similar discontinuities in test scores for both girls and boys. Consistent with our findings above, the RD plots show no discontinuity in test scores at the cutoff point in Urdu for either girls or boys. However, these RD plots mask the fuzziness in remediation assignments and do not formally test the differences in the effects of remediation between girls and boys. To formally test the differences in the impact of remediation between girls and boys, Panel A of Table 3 presents the reduced form and the 2SLS estimates corresponding to the RD plots in Figure 5. We test for the significance of the interaction term between the indicator for remediation and that for gender. The coefficients on the interaction terms are small and statistically not significant providing evidence against heterogeneity in the effects of remediation across genders[25].

[24] Appendix A reports further robustness checks. These include using a second-order polynomial of the running variables, varying the flexibility of the pre-remediation test score function when forming the residualized scores, and using the Donut-hole RD approach that leaves out observations extremely close to or at the cutoff. We find that our results remain consistent to these checks.

[25] figure A.9 tests for the robustness of the coefficients on the interaction term against variations in the bandwidth. The estimates are robust to this check.

5.2.2 Effect of Remediation by Grade Level

Next, we ask whether remediation is equally effective at the primary and secondary levels of education. Previous research on remedial education in developing countries limits itself to lower grades. For instance, Banerjee et al. (2007) study the effects of the *Balsakhi* program on students in grades 3 and 4 in India. While learning at the primary level is crucial, some researchers argue that as developing countries move towards universal primary education, the returns to post-primary education may outweigh the returns to primary education. Despite this, there exists very little evidence on effective pedagogical interventions at the post-primary level. Furthermore, several studies highlight that learning gaps widen as students progress through higher grades (Muralidharan & Zieleniak, 2014). On the one hand, large variations in learning accompanied by difficult curriculums imply that it may be too late to mitigate within-classroom learning variations at higher grades. Alternatively, larger variations may also imply that targeted interventions such as remediation may have greater returns at the post-primary levels. It is thus crucial to empirically understand whether remediation programs are equally beneficial at the secondary level, as they are at the primary level. Since the *Uraan* remediation program is implemented in all grades from 3 through 8, we formally test this difference in the impact of remediation across grade levels using a methodology similar to the one described in the previous subsection. The results are presented in panel B of Table 3 with the accompanying RD plots in figure 6.

The table presents the reduced-form estimates and the IV estimates of the effect of remediation by grade level. The results suggest that remediation has been equally successful both at the primary and the secondary grade levels. The coefficients on the interaction between the indicator for remediation and the indicator for secondary level are small and statistically not significant. Taken together, these results suggest that the wide learning gaps that emerge as students progress through higher grades may potentially be mitigated by identifying weak students and teaching them at their level.

5.2.3 Effect of Remediation by Remediation Class Size

Since one of the primary functions of remediation classes is to focus more on the needs of the poor-performing students, an important question from a program implementation angle is whether the size of these classes matters for learning. Despite its potential importance, student-teacher ratios have received little attention in the past remediation literature. The variation in class sizes across schools and grades in our setting allows us to empirically test the impact of the remediation class sizes on students' test scores. Figure 7 presents the effects of remediation by class size. As expected, the coefficients in classes with 3 or 7 students are noisier given that very few schools have classes with 3 or 7 students. The graphs suggest that the effects of remediation do not vary with class sizes. Complementing the graphs, panel C of table 3 presents the regression estimates from a regression of the test scores on a fully interacted model with class size. Consistent with the graphical depiction in figure 7, the coefficients on the interaction term are very small and statistically not significant, implying that the size of the remediation classes does not impact learning. To the extent that our estimates can be extrapolated to slightly larger class sizes, our results suggest that more students can be accommodated in these classes without negatively affecting the learning of the existing students.

5.2.4 Heterogeneity by Variation in the Marginal Student

The unique nature of the eligibility criteria of the *Uraan* remediation program, where eligibility depends on the size of a remediation class and students' past performance, leads to variation in the position of the marginal student in the baseline performance distribution. For instance, the location of the marginal Math student (the student with a relative rank of 0 in Math) varies between the first and the fourth decile of the initial score distribution. For all three subjects, 89 percent of the marginal students lie in either the first or the second decile of the initial score distribution. An additional 10 percent lie in the third decile. This feature of the program allows us to speak to the generalizability of our results. Figure A.3 presents these results for each subject. Each point on the graph represents the effect of remediation on test scores given that the marginal student was positioned in that decile. For instance, the first point on the graph on the top-right shows that the effect of remediation was approximately 0.4σ in classes where the marginal student was in the first decile of the initial score distribution. Similarly, the second point represents the effect of remediation in classes where the marginal student was in the second decile of the initial score distribution and so on.

The graphs for Math and Urdu show that the effects of remediation are not sensitive to the position of the cutoff point in the initial score distribution with the effects for Math being positive and, for the most part, statistically significant and the effect for Urdu being statistically not distinguishable from 0. The effect for English is positive if the marginal student lies in the first two deciles of the initial score distribution. While the coefficient for the third decile is negative, it is not statistically different from those in the second and the first deciles. These results are reassuring since they imply that remediation works regardless of the position of the marginal student in the initial score distribution and provides credence to the external validity of the findings.

5.3 Spillover Effects to Other Subjects

In this subsection, we test whether the effects of remediation spillover on other non-targeted subjects. Testing for spillover effects is important because the existence of spillover effects would imply that the overall effect of remediation may be under-or over-estimated in studies that do not incorporate spillover effects into their calculations. *A priori*, the magnitude and the direction of these spillover effects are ambiguous in our setting. On the one hand, the language of instruction serves as an important factor that either promotes or impedes learning. For instance, Yip et al. (2003) and Brock-Utne (2013) argue that students would learn much more if they are taught in their mother language. Urdu is the national language, serves as the language of instruction for all subjects across TCF schools. We should thus expect a complementarity between Urdu remediation and test scores in subjects that have Urdu as the medium of instruction. On the other hand, it may be the case that due to remediation, students focus more on one subject and neglect the others, thus performing poorly in other subjects. For instance, students getting remediation for Math may underperform in Urdu or English.

We thus test for these cross-subject effects (or spillover effects) of remediation in Table 4. Each row in Table 4 represents the effect of remediation for a targetted subject on the test scores for other subjects. For instance, the first row presents the effect of getting Urdu remediation on Urdu, English, Math, Science, Islamic Studies, and Social Studies test scores. The diagonal values in columns 1-3 are the same as the estimates in Panel B of Table 2. The results provide no evidence of spillover effects, either positive or negative. Remediation for Urdu, for instance, does not affect test scores for other subjects. This result makes sense since the Urdu remediation did not improve the Urdu language skills of students enrolled in these classes. We also do not find evidence of any substitution in effort from non-targeted subjects towards targetted subjects. For instance, remediation for Math does not affect test scores for English or Urdu. While there might be students who get remediation for multiple subjects, we show in figure A.5 that the eligibility for anyone subject does not affect the probability of remediation for another subject.

5.4 Effect of Remediation on Time-in-School Outcomes

We next estimate the effect of remediation on two time-in-school variables: the attendance rates, and the drop-out rate one period later. On the one hand, remediation classes improve the test scores of students for English and Math, leading to an increase in the marginal benefit of going to school. On the other hand, the additional time spent in school attending the remediation classes increases the opportunity cost of schooling leading to an increase in the marginal cost of going to school for a student enrolled in remediation. Hence, it is unclear whether the improvements in test scores correspond to improvements in time-in-school outcomes.

Figure A.4 presents the relationship between attendance and drop-out rates one period later on relative ranks of the students. The figures do not show any discontinuities in the outcomes suggesting the remediation did not affect the attendance rates or the drop-out rates of students enrolled in remediation classes. Table 5 presents the IV estimates corresponding to the RD plots. The estimates confirm that remediation did not impact the students' time in school. These results are consistent with the finding of Banerjee et al. (2007) who find that remediation did not affect student attendance in India.

5.5 Persistence in Effects

A key question from a policy perspective is whether the positive effects of programs such as remediation persist after the duration of the program. In this subsection, we test the persistence of the effects of remediation in Table 6 which relates the test scores of students in time t+1 (next term) to their remediation status in time t. More specifically, we regress the scores of students one academic term later on their present remediation status.

Column 1 of the table reproduces our main results while column 2 tests for persistence. The estimates suggest that some fade out in the effects of remediation with remediated students performing around 0.12 standard deviations better than their peers one term later in

English and Math. Although the coefficients are modest and positive, they are statistically not different from zero. These estimates are noisy due to large dropout rates in our setting. The persistence in the effects of remediation, however, is not attributable to repeated enrolment in remediation classes. In column 3 of table 6, we regress the probability of being enrolled in term t+1 on the student's remediation status in term t and find that remediation in the present cycle is not indicative of future remediation enrollment.

The "fade-out" in the effects of remediation is consistent with previous educational interventions aimed at improving test scores. For instance, Krueger & Whitmore (2001) and Chetty et al. (2011) estimate the long-run effects of Project STAR in student test scores and find significant fade-out in the effects of smaller and high-quality classes, with larger reductions in effects happening immediately after the end of the program. Similarly, Banerjee et al. (2007) find a significant reduction in the effects of remediation a year after the *Balsakhi* program had ended. This fade-out may occur because teachers may be teaching to the test which may result in students performing better in the exam and being removed from the remediation classes but this may not lead to substantial gains in learning. It could also be the case that students require repeated exposure to these remediation classes for them to make learning gains that persist, although longer-term exposure to remediation has also been unable to translate into gains that persist (Banerjee et al., 2007).

While the fade-out in the effects of remediation on test scores is prevalent in educational interventions, many studies show differences in later life outcomes as a result of such interventions. For instance, Chetty et al. (2011) show long-term benefits of class reductions as a result of the Tennessee Student/Teacher Achievement Ratio program despite a fade-out in the test score gains. Similar patterns are observed in other early childhood interventions such as the Head Start program (Deming, 2009) and the Fast Track program (Dodge et al., 2015) in the United States. Hence in section 6, we attempt a cost and benefit analysis of the program by estimating labor market returns of the test score gains.

6 Discussion

6.1 Principals Survey and Mechanisms

Our full-sample results suggest that while remediation was successful in English and Math test scores, it had no effect on test scores for Urdu. Banerjee et al. (2007) find a similar result for a sample with an already high level of basic competency in the language. While students in our setting perform much better in Urdu, the differences in program structures between their study and ours imply that such ceiling effects are unlikely to explain the null effects of remediation on Urdu in our setting. Specifically, the *Uraan* program involves teaching students a combination of basic-level competencies as well as grade-level topics. This implies the existence of some students at the bottom end of the score distribution that may benefit from remediation that involves teaching grade-level topics even if these students are competent in foundational topics.

To better understand the program structure and the mechanisms driving our results, including the null effects for Urdu, we survey the principals in our sample. Due to COVID-19 restrictions, we conducted these surveys over the phone. Out of the 76 schools in our sample, we were able to reach principals of 61 schools. We were unable to survey the rest of the principals because they had left the organization. Through our survey, we gathered information on student selection and teacher appointments in these remediation classes, teacher turnover for different subjects, the principals' perception regarding the effectiveness of the program, and their perception regarding teacher and student motivation for these classes.

The survey revealed that in almost half of the schools, teachers assigned to Urdu remediation classes were not subject specialists. These non-subject specialist teachers were any teachers that were available during the Urdu remediation class time for that term. This is contrary to English and Math remediation classes where subject-specialist teachers taught these classes in all schools in our survey sample. Furthermore, the survey results regarding the principal's perception of the teachers' motivation show that a majority of the principals (63 percent) perceive the teachers teaching Urdu remediation classes to have the lowest motivation out of the three subjects. Further data exploration reveals that these low motivation Urdu remediation teachers are concentrated in schools where non-specialized teachers teach

Urdu remediation classes[26].

The variation in teacher assignment is critical in our setting because of three main reasons. First, teachers assigned to remediation classes are not provided with additional training to conduct these classes. Second, remediation involves identifying the academic weaknesses of the students, a task non-subject specialist teachers might find hard to perform. However, qualitative interviews with the principals suggest that Urdu-specialist teachers do assist in identifying the weak topics even when they are not teaching these classes. Thirdly, the *Uraan* program provides remediation at both the primary and secondary grade levels. While the content taught in lower grades may be easy enough for non-specialized teachers, the content at higher grades might be difficult for such teachers to teach. This is key because the teachers are not provided with lesson guides to follow in these classes. Previous literature shows that para teachers and volunteers may also lead to improvements in test scores through remediation classes in settings where remediation involves teaching foundational competencies to lower grade students (Banerjee et al., 2007). However, it is unclear whether such teachers may be effective in settings where the content is flexible to the needs of the students and remediation classes are implemented at higher grades as well.

To study whether such teacher assignments explain the differences in the impact of remediation on Urdu test scores, in Table 7 we test for the significance of the interaction term between the Urdu remediation dummy and the teacher type dummy. Column 1 presents the results for the full sample. It shows that the interaction term is positive and significant, suggesting that the remediated students in schools where specialized teachers teach Urdu remediation classes perform 0.398σ better than those where non-specialized teachers teach these classes [27]

One concern regarding our statistical approach may be that teacher assignment is not random. For instance, schools that assign Urdu specialist teachers may generally be better than schools that assign non-specialist teachers. While we do not claim causality for our estimates, we placate these concerns by conducting a placebo test where we estimate whether

[26] There are no differences in principal's perception about Math and English teachers' motivation across the two type of schools. Furthermore, no clear patterns emerge regarding perceptions of student motivation across the two types of schools. The survey also confirmed that students were selected based on their prior performance on the midterm or the final exam. Qualitative interviews also revealed that principals usually choose the same class size across different subjects at the school level.

[27] Splitting the sample by teacher type reveals that the effect of remediation on Urdu in schools where a specialized teacher taught these classes is also individually statistically significant with an effect size of approximately 0.222σ, comparable to those for English and Math.

the type of Urdu teacher affects the impact of remediation for English and Urdu. These results are presented in Table A.2. Specifically, we test for the significance of the interaction between the indicator for remediation (English or Math) and the indicator for Urdu teacher type. The results provide evidence against differences in school quality driving the impact of Urdu remediation by teacher type.

We explore these heterogeneities in the impact of remediation by teacher type further in two main ways. First, we study the impact of teacher type separately for primary and secondary grade levels. These results are presented in columns 2 and 3 of Table 7. Similar to the full sample results, specialized teachers result in greater score gains than non-specialized teachers for both primary and secondary grade levels. However, the impact of having a specialized teacher is greater at the secondary level. Qualitative interviews with the principals suggest that at the primary grade levels, teachers believe that the content is easy enough to be covered within the regular class time. However, the content becomes difficult at the post-primary level. These differences in curriculum difficulty may explain the varying impact of remediation at the primary and secondary levels.

Second, we study whether the positive effect of remediation on Urdu at the post-primary level also translates into non-targeted subjects for which Urdu is the medium of instruction. Hence in columns 4-6 of Table 7, we estimate the impact of Urdu remediation on Science, Social Studies, and Islamic Studies scores. While the estimates are noisy due to sample restrictions, we find suggestive evidence that students that were taught by subject specialist teachers also had improved test scores in other subjects. This provides suggestive evidence in favor of positive spill-overs of remediation. Taken together, these results highlight the sensitivity of the impact of remediation to program design and implementation (Banerjee et al., 2016).

6.2 Cost and Benefit Analysis

While teachers do not get paid for the additional teaching involved in the remediation classes, we can still estimate the shadow cost of the program by assuming that teachers get compensated for the total time spent in school. Using data on monthly salaries of primary and post-primary level teachers, with the latter being slightly higher, we can estimate the per student-term cost of running the program. We estimate the per student-term shadow cost of remediation to be around \$1.48 for primary grades and \$1.73 for secondary grades. These numbers are similar to the costs of other remediation programs. For instance, Banerjee et al. (2007) estimate the cost of their remediation program to be \$2.25 per child per annum, which implies a cost of roughly \$1.2. The Uraan remediation program is also much cheaper than other interventions such as hiring additional teachers (Muralidharan & Sundararaman, 2013), performance incentives for teachers (Muralidharan & Sundararaman, 2011), and Computer Assisted Learning (CAL) programs (Banerjee et al., 2007). However, the program's running costs are higher than interventions such as report cards (Andrabi et al., 2017).

We estimate the benefits of the program through the potential labor market returns to test score gains. Aslam et al. (2011) estimate the wage returns to a 1σ increase in Math test scores to be 14% for boys and 36% for girls. Similarly, they estimate the wage returns to a 1σ increase in English test scores to be 13% for boys and 32% for girls in Pakistan. Given these numbers, we estimate that the test score gains in our setting imply an average wage returns of around 3.45% for boys and 8.7% for girls. The average return of 6.08% is slightly lower than the 7.7% found by Muralidharan and Sunandararaman (2011) for a performance pay program. These returns are also lower than 9.4% returns estimated using score gains in Banerjee et al. (2007). These differences in returns arise because of the differences in remediation eligibility criteria between our setting and the Balsakhi program studied by Banerjee et al. (2007). Under that program, all eligible students are provided remediation for language and Math. In the program we study, not all students get remediation for multiple subjects. These differences in program design lead to differences in the methods used to estimate the returns. Specifically, while Banerjee et al. (2007) and Muralidharan et al. (2019) sum the returns from Math and language score gains, we take the average of these.

Using various estimates of wages in Pakistan, the returns in our setting yield \$11.05 to \$21.74 increase in annual wages for males and a \$27.9 increase for females. These predictions are conservative since the used wages in agriculture, which are lower than those in other sectors in Pakistan. Assuming a steady stream of fixed earnings over 40 years for boys and 30 years for girls (due to breaks in employment due to child-rearing etc.) and a discount rate of 10%, we estimate a net present value (NPV) between \$44 and \$107, and an internal rate of return (IRR) ranging between 38% to 48%. These estimates are speculative and sensitive to the assumptions about wage levels, the age at which students start working (20 years), and the number of years for which they are expected to work. Nevertheless, these IRRs are large, even though we use conservative estimates of wages, primarily because the costs of the program are extremely low.

7 Conclusion

In this paper, we present quasi-experimental evidence on the impact of a remediation program that provided customized remediation classes to the bottom 3 to 7 poor performing students in grades 3 to 8 in Pakistan. This limit to the number of students in a remediation class leads to a class specific rank threshold. We show that students that fall below the rank threshold have around 50 percent higher probability of being assigned to a remediation class. Using administrative data on student test scores in a fuzzy regression discontinuity design, we find that remediation classes improved test performance for English (0.288σ) and Math (0.217σ) but not for Urdu. Importantly, we find that these positive effects of remediation are similar across grade levels, gender, and class sizes. However, we find some evidence of a fade-out in the effects remediation, although these estimates are very noisy due to smaller sample sizes.

To understand some of the mechanisms driving our results, we conduct a phone survey of principals in our sample. Our survey results reveal that the null effects of Urdu remediation are driven primarily by the assignment of non-specialized teachers to teach Urdu classes. For the subset of schools where Urdu teachers led Urdu classes, we find the remediation improved test scores for Urdu (by around 0.22σ). For these schools, we also find suggestive evidence of positive spillovers of remediation to non-targeted subjects for which Urdu was the medium of instruction.

However, there are certain limitations to remediation classes when implementing them more broadly. Firstly, remediation classes by nature target only those students that are at the bottom of the performance distribution. While they may help improve the learning outcomes of those at the bottom, they may not lead to distributional shifts that other interventions such as the Mindspark program tested in Muralidharan et al. (2019) potentially do. Secondly, our results suggest that the success of the remediation program may be sensitive to program design and implementation. Other studies have also documented these limitations (Banerjee et al., 2016).

The positive gains due to a "Teach at the Right Level" (TaRL) pedagogical intervention in our setting suggest that the current curriculums and pedagogical practices are binding constraints in converting schooling into learning. Remediation then provides a cost-effective option to alleviate such a constraint by teaching students at their level of competence. The finding that remediation was successful at both the primary and secondary levels of education suggests that such programs can mitigate the wide within-class learning disparities at higher grades. However, we argue that the success of such programs hinges on program design and implementation. While the previous literature shows that non-specialized teachers such as volunteers and contract teachers can lead to improvements in test scores in settings where remediation involves teaching basic competencies, we argue that such teachers may not be very effective in higher grade levels and in settings where the content is flexible to needs of the students.

Despite their limitations, remediation classes are a cost-effective tool with the potential to mitigate the within-class learning gaps in low-resource settings where other interventions may prove expensive to implement. Overall, our study demonstrates that pedagogical interventions that teach at the right level may improve learning outcomes of poor-performing students at both the primary and post-primary levels of education in a very cost-effective manner. They may also prove effective in mitigating COVID-19 related learning losses in settings where the uptake of remote learning technologies was low.

References

Akmal, M., Crawfurd, L., Hares, S., & Minardi, A. L. (2020). Covid-19 in pakistan: A phone survey to assess education, economic, and health-related outcomes. *Center for Global Development Policy Paper*, *188*, 1–36.

Andrabi, T., Das, J., & Khwaja, A. I. (2017). Report cards: The impact of providing school and child test scores on educational markets. *American Economic Review*, *107*(6), 1535–63.

ASERPakistan. (2019). Annual status of education report: Aser pakistan 2019. *Lahore: SAFED*.

Aslam, M., De, A., Kingdon, G., & Kumar, R. (2011). Economic returns to schooling and skills–an analysis of india and pakistan. *Education Outcomes and Poverty in the South*.

Bajari, P., Hong, H., Park, M., & Town, R. (2011). *Regression discontinuity designs with an endogenous forcing variable and an application to contracting in health care* (Tech. Rep.). National Bureau of Economic Research.

Banerjee, A., Banerji, R., Berry, J., Duflo, E., Kannan, H., Mukherji, S., ... Walton, M. (2016). *Mainstreaming an effective intervention: Evidence from randomized evaluations of "teaching at the right level" in india* (Tech. Rep.). National Bureau of Economic Research.

Banerjee, A., Banerji, R., Duflo, E., Glennerster, R., & Khemani, S. (2010). Pitfalls of participatory programs: Evidence from a randomized evaluation in education in india. *American Economic Journal: Economic Policy*, *2*(1), 1–30.

Banerjee, A., Cole, S., Duflo, E., & Linden, L. (2007). Remedying education: Evidence from two randomized experiments in india. *The Quarterly Journal of Economics*, *122*(3), 1235–1264.

Barro, R. J. (2001). Human capital and growth. *American economic review*, *91*(2), 12–17.

Bau, N., Das, J., & Chang, A. Y. (2021). New evidence on learning trajectories in a low-income setting. *International Journal of Educational Development*, *84*, 102430.

Berry, J., & Mukherjee, P. (2016). Pricing of private education in urban india: Demand, use and impact. *Unpublished manuscript. Ithaca, NY: Cornell University*.

Brock-Utne, B. (2013). Language of instruction and learning in mathematics and science in some african countries. *African and Asian Studies*, *12*(1-2), 83–99.

Card, D., & Giuliano, L. (2016). Can tracking raise the test scores of high-ability minority students? *American Economic Review*, *106*(10), 2783–2816.

Castellano, K. E., & Ho, A. D. (2013). A practitioner's guide to growth models. *Council of Chief State School Officers*.

Cattaneo, M., & Rocio, T. (in press). *Regression discontinuit designs* (Tech. Rep.). Annual Review of Economics.

Cattaneo, M. D., Idrobo, N., & Titiunik, R. (2019). *A practical introduction to regression discontinuity designs: Foundations*. Cambridge University Press.

Chetty, R., Friedman, J. N., Hilger, N., Saez, E., Schanzenbach, D. W., & Yagan, D. (2011). How does your kindergarten classroom affect your earnings? evidence from project star. *The Quarterly journal of economics*, *126*(4), 1593–1660.

Deming, D. (2009). Early childhood intervention and life-cycle skill development: Evidence from head start. *American Economic Journal: Applied Economics*, *1*(3), 111–34.

Dodge, K. A., Bierman, K. L., Coie, J. D., Greenberg, M. T., Lochman, J. E., McMahon, R. J., ... Group, C. P. P. R. (2015). Impact of early intervention on psychopathology, crime, and well-being at age 25. *American journal of psychiatry*, *172*(1), 59–70.

Duflo, A., Kiessel, J., & Lucas, A. (2021). *Experimental evidence on alternative policies to increase learning at scale* (Tech. Rep.). National Bureau of Economic Research.

Duflo, E., Dupas, P., & Kremer, M. (2015). School governance, teacher incentives, and pupil–teacher ratios: Experimental evidence from kenyan primary schools. *Journal of public Economics*, *123*, 92–110.

Geven, K., & Hasan, A. (2020). Learning losses in pakistan due to covid-19 school closures.

Glewwe, P., Kremer, M., & Moulin, S. (2009). Many children left behind? textbooks and test scores in kenya. *American Economic Journal: Applied Economics*, *1*(1), 112–35.

Glewwe, P., Kremer, M., Moulin, S., & Zitzewitz, E. (2004). Retrospective vs. prospective analyses of school inputs: the case of flip charts in kenya. *Journal of development Economics*, *74*(1), 251–268.

Glewwe, P., & Muralidharan, K. (2016). Improving education outcomes in developing countries: Evidence, knowledge gaps, and policy implications. In *Handbook of the economics of education* (Vol. 5, pp. 653–743). Elsevier.

Hanushek, E. A. (2013). Economic growth in developing countries: The role of human capital. *Economics of Education Review*, *37*, 204–212.

Hanushek, E. A., & Kimko, D. D. (2000). Schooling, labor-force quality, and the growth of nations. *American economic review*, *90*(5), 1184–1208.

Imbens, G. W., & Lemieux, T. (2008). Regression discontinuity designs: A guide to practice. *Journal of econometrics*, *142*(2), 615–635.

Kane, M. T. (2017). Measurement error and bias in value-added models. *ETS Research Report Series*, *2017*(1), 1–12.

Krueger, A. B., & Whitmore, D. M. (2001). The effect of attending a small class in the early grades on college-test taking and middle school test results: Evidence from project star. *The Economic Journal*, *111*(468), 1–28.

Mullis, I., Martin, M., Foy, P., Kelly, D., & Fishbein, B. (2020). Timss 2019 international results in mathematics and science. *Boston College, TIMSS & PIRLS International Study Center*.

Muralidharan, K., Singh, A., & Ganimian, A. J. (2019). Disrupting education? experimental evidence on technology-aided instruction in india. *American Economic Review*, *109*(4), 1426–60.

Muralidharan, K., & Sundararaman, V. (2010). The impact of diagnostic feedback to teachers on student learning: Experimental evidence from india. *The Economic Journal*, *120*(546), F187–F203.

Muralidharan, K., & Sundararaman, V. (2011). Teacher performance pay: Experimental evidence from india. *Journal of political Economy*, *119*(1), 39–77.

Muralidharan, K., & Sundararaman, V. (2013). *Contract teachers: Experimental evidence from india* (Tech. Rep.). National Bureau of Economic Research.

Muralidharan, K., & Zieleniak, Y. (2014). Chasing the syllabus: Meauring learning trajectories in developing countries with longitudinal data and item response theory: Uc san diego.

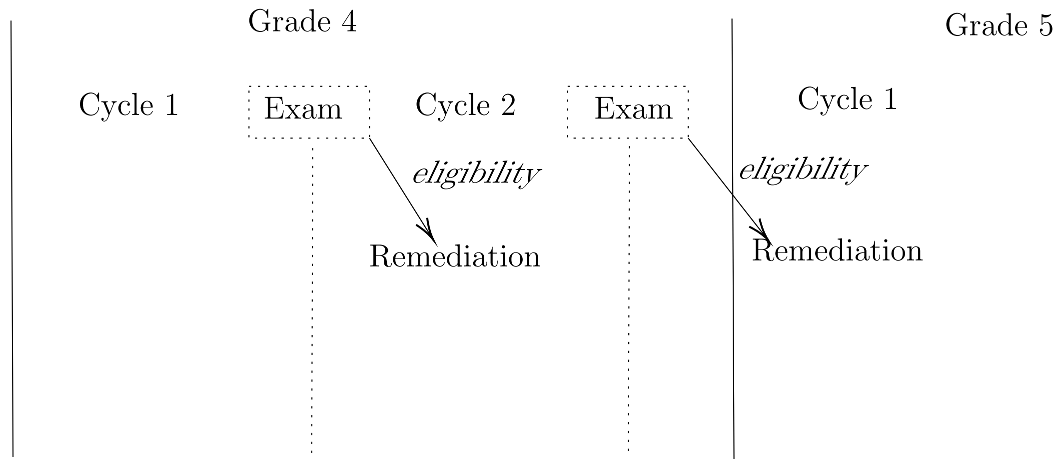
Pritchett, L., & Sandefur, J. (2020). Girls' schooling and women's literacy: schooling targets alone won't reach learning goals. *International Journal of Educational Development*, *78*, 102242.

Woessmann, L. (2016). The economic case for education. *Education Economics*, *24*(1), 3–32.

WorldBank. (2018). Learning to realize education's promise. *World Development Report. The World Bank*.

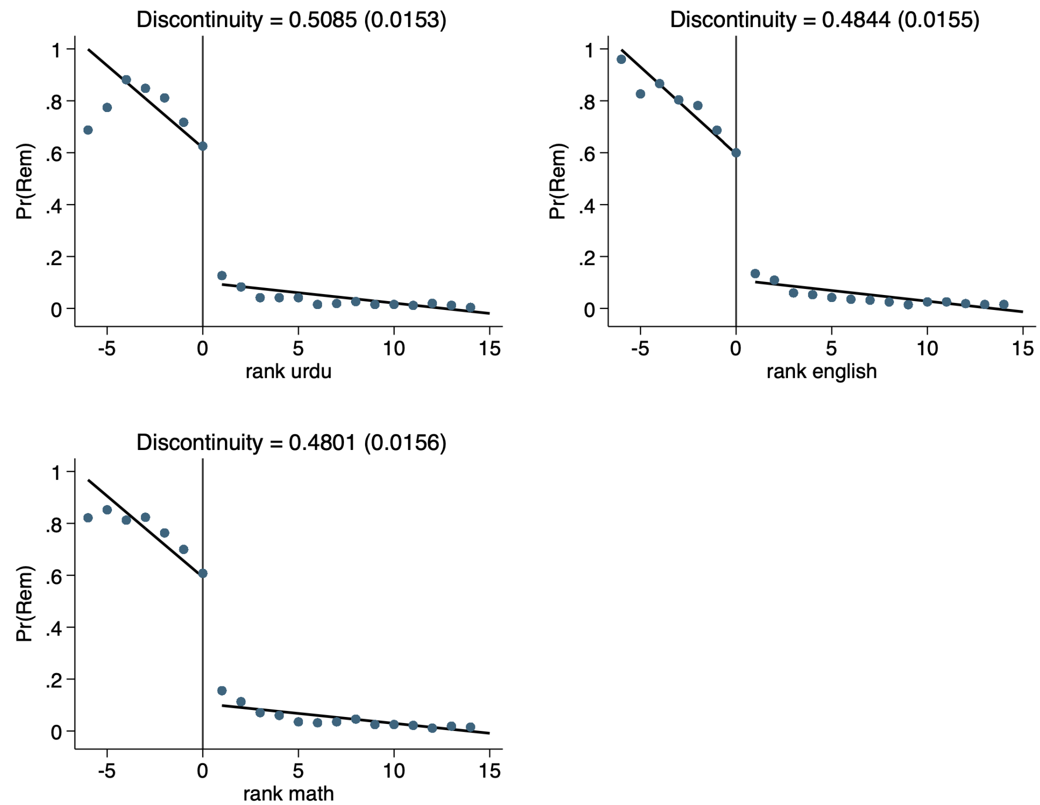
Yip, D. Y., Tsang, W. K., & Cheung, S. P. (2003). Evaluation of the effects of medium of instruction on the science learning of hong kong secondary students: Performance on the science achievement test. *Bilingual Research Journal*, *27*(2), 295–331.

Figure 1: Remediation Program Structure and Eligibility



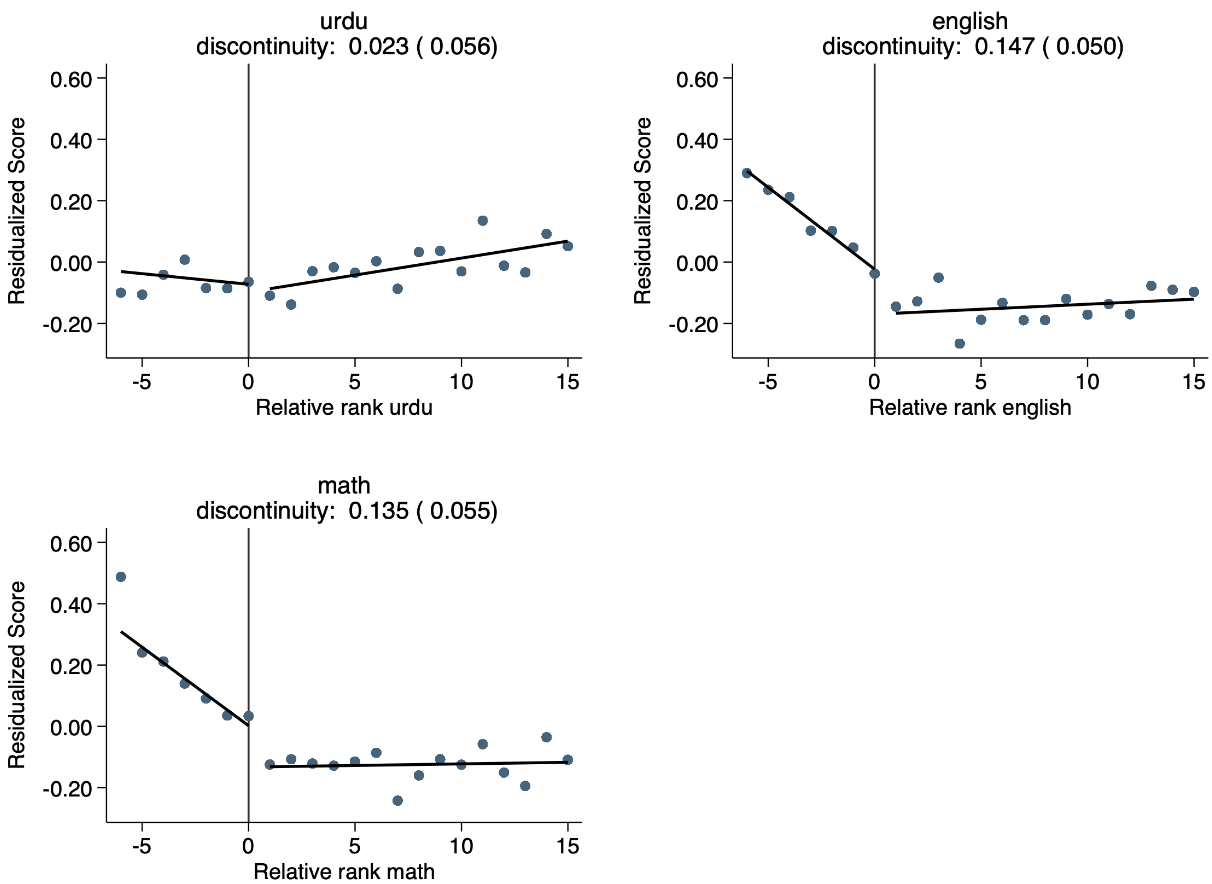
Notes- The figure illustrates the eligibility criteria for remediation classes. Remediation classes are operationalized in two cycles per academic year. Eligibility for cycle 1 depends on the score on final exams of the previous grade while eligibility for cycle 2 depends on the score on the midterm of the previous term. Both the midterm and the final exams are set centrally and distributed across schools. Teacher are provided with detailed rubrics to grade these exams.

Figure 2: First Stage Relationship between Relative Ranks and Enrollment in Remediation



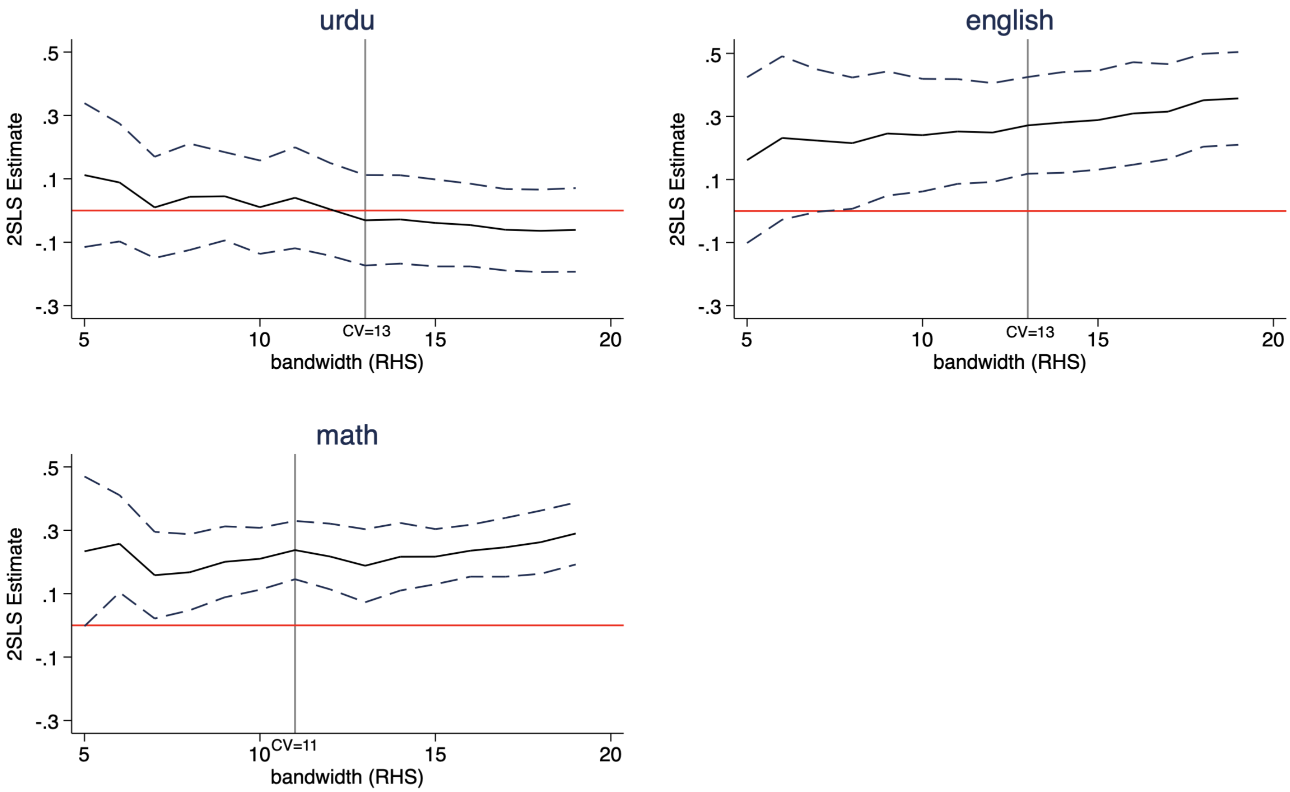
Notes- The figure presents the first stage relationship between relative ranks for English, Urdu and Math, and the probability of being assigned to remediation for that subject. Each point represents a rank mean and the lines represent a linear fit on either side of the cutoff point. The RD plots show a clear discontinuity at the threshold with students below the rank cutoff being around 48-50 percent more likely to be assigned to remediation.

Figure 3: Reduced Form Relationship between Relative Ranks and Test Scores



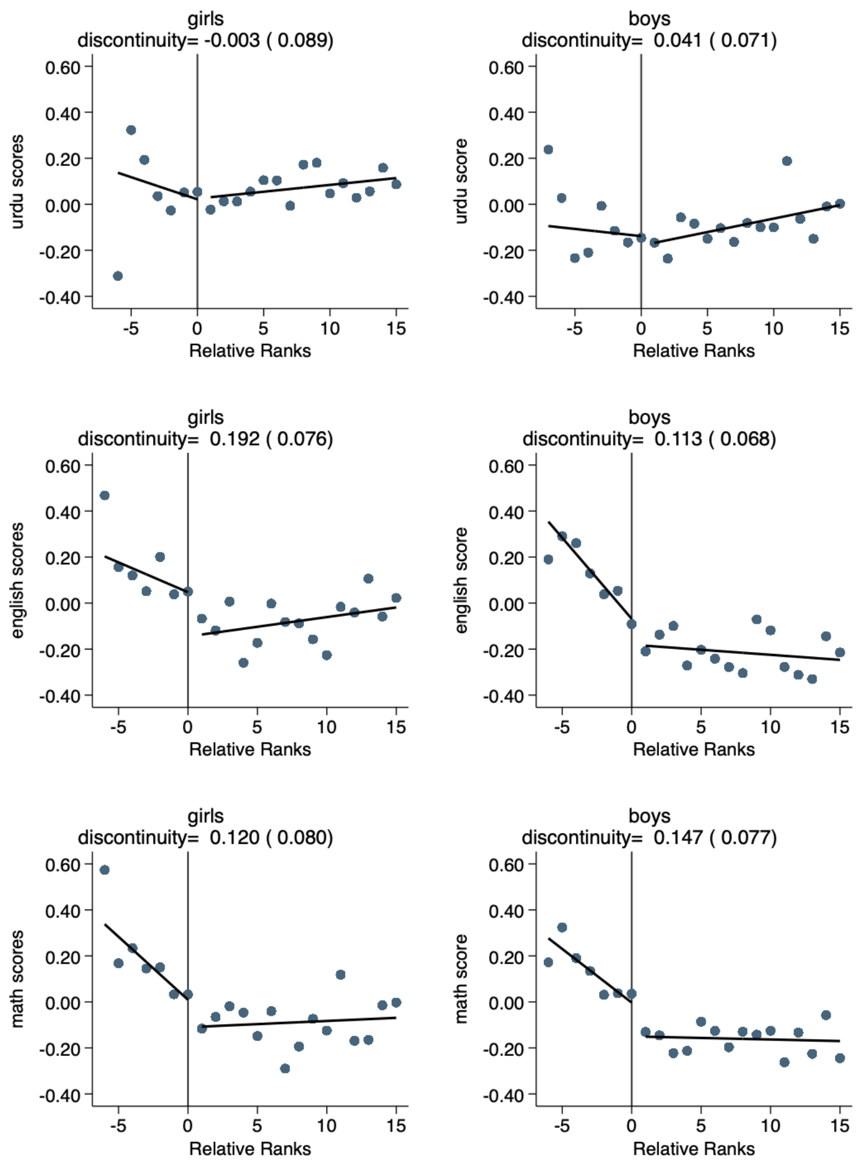
Notes- The figure presents the RD plot for the relationship between relative rank of students and their performance as measured residualized test scores. Each point represents a rank mean and the lines represent a linear fit on either side of the cutoff point. The corresponding estimates are presented in table 2.

Figure 4: Robustness of RD estimates to Variations in Bandwidth



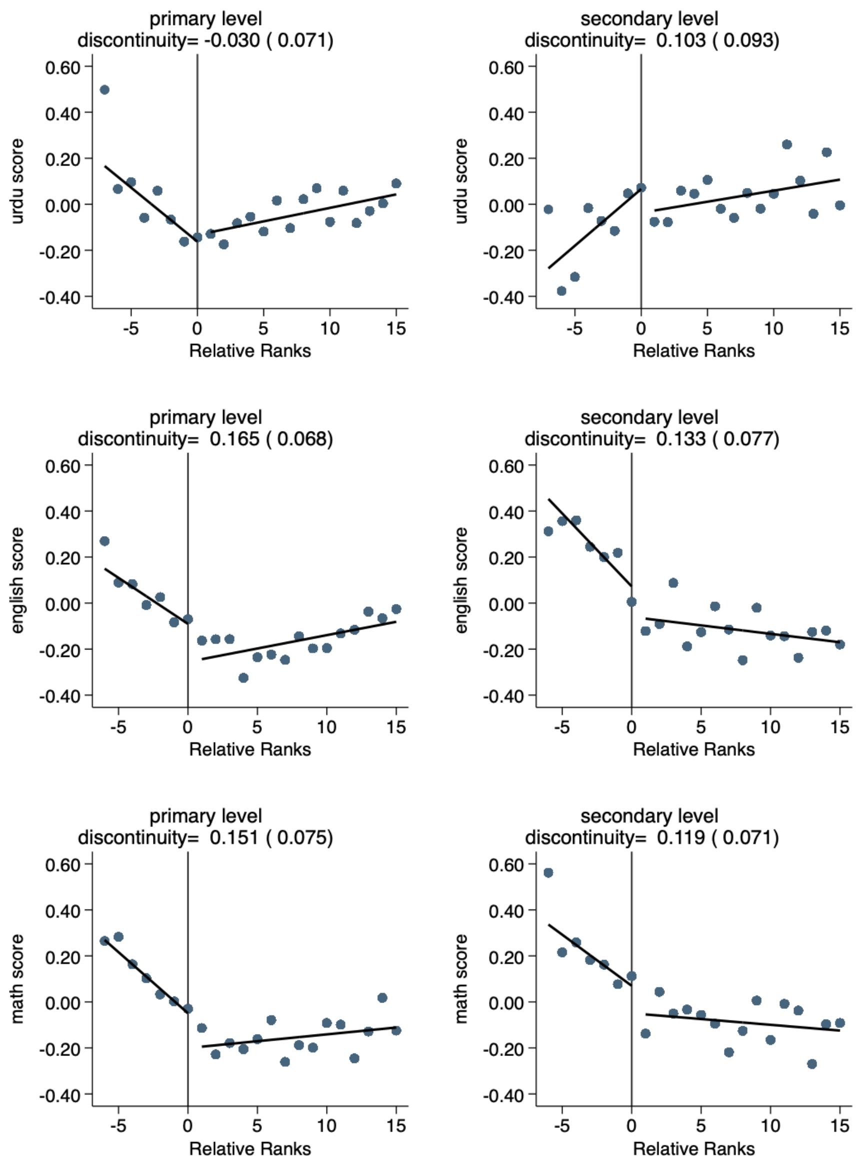
Notes- This figure tests for the robustness of our estimates to variations in the bandwidth. More specifically, we vary the bandwidth on the right hand side of the cutoff from 5 to 20 relative ranks. The figure show that our estimates remain consistent across the range of bandwidths considered.

Figure 5: Impact of Remediation by Gender.



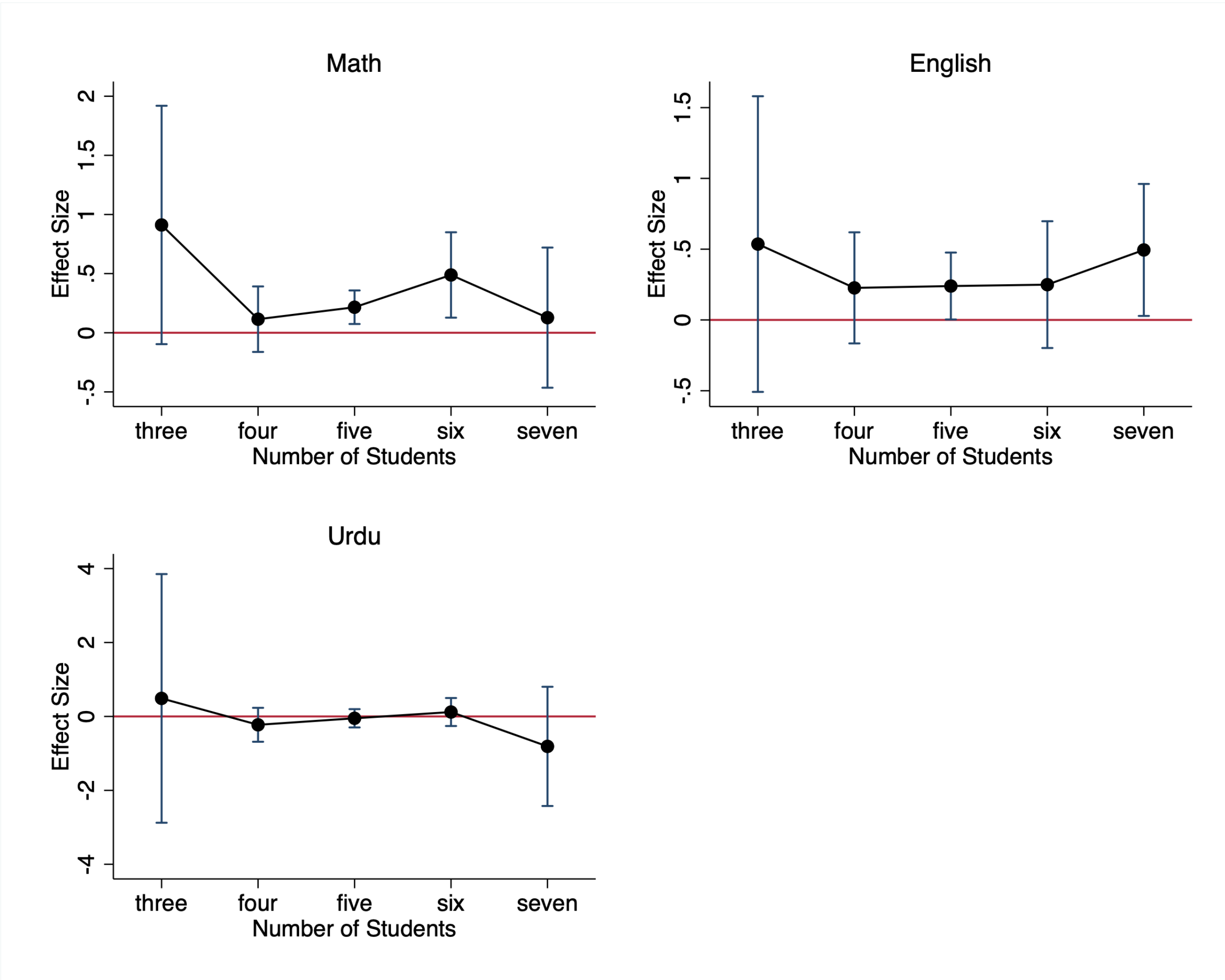
Notes- This figure presents RD plots for the effect of remediation on test scores for girls and boys separately. Column 1 presents RD plots for girls while columns 2 presents the plots for boys. Each row represents a separate subject. Each point represents a rank mean and the lines represent a linear fit on either side of the cutoff point.

Figure 6: Impact of Remediation by Grade Level.



Notes- This figure presents RD plots for the effect of remediation on test scores for primary and secondary grade levels separately. Column 1 presents RD plots for primary grades (grades 3-5) while column 2 presents the plots for secondary grades (grades 6-8). Each row represents a separate subject. Each point represents a rank mean and the lines represent a linear fit on either side of the cutoff point.

Figure 7: Impact of Remediation by Number of Students in a Remediation Class



Notes- The figure plots the 2SLS estimates (along with a 90 percent confidence interval) against the number of students in a remediation class for English, Urdu, and Math. The figure shows that class size does not matter for learning in these remediation classes.

Table 1: Sample Characteristics

		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
		Full Sample		*Bandwidth: -7 to +15*					
				Math		*English*		*Urdu*	
		mean	sd	mean	sd	mean	sd	mean	sd
Panel A: Student Scores and Attendance (%)									
Urdu		46.58	19.82	42.43	18.86	41.70	18.42	38.86	17.21
English		42.22	19.26	38.62	18.00	35.99	16.44	37.82	17.81
Math		35.79	21.12	29.25	17.96	31.97	19.42	32.98	19.78
Islamic Studies		51.84	19.43	47.56	18.52	47.15	18.36	45.41	17.92
S.Std		37.62	21.48	33.20	19.85	32.68	19.72	31.47	19.43
Science		42.59	21.60	37.69	20.07	37.59	20.14	36.88	20.33
Regional		45.30	22.59	40.84	21.68	40.23	21.40	39.19	21.64
Attendance		87.51	11.06	86.44	11.62	86.73	11.58	86.55	11.44
Panel B: Student and School Characteristics									
age		11.67	2.073	11.74	2.133	11.70	2.133	11.50	2.077
female (%)		0.515	0.500	0.495	0.500	0.472	0.499	0.460	0.498
primary (%)		0.546	0.498	0.541	0.498	0.559	0.497	0.621	0.485
Observations		8,744		5,594		5,572		5,180	
Panel C: Remediation Classes									
Class Size	Urdu	4.951	0.989						
	English	5.032	0.931						
	Math	5.145	0.990						
Attendance (%)	Urdu	83.87	10.19						
	English	83.02	10.56						
	Math	85.96	10.43						
Panel D: Scores by Gender									
Urdu	Girls (4502)	49.50	19.52						
	Boys (4242)	41.02	19.04						
English	Girls	44.57	19.19						
	Boys	37.48	18.62						
Math	Girls	37.31	21.23						
	Boys	34.14	20.89						
Attendance	Girls	88.52	10.90						
	Boys	86.57	11.17						
Panel E: Score by Grade Level									
Urdu	Primary (4777)	47.65	18.22						
	Secondary (3967)	42.81	20.56						
English	Primary	44.87	17.62						
	Secondary	36.80	19.58						
Math	Primary	45.27	17.29						
	Secondary	24.62	19.10						
Attendance	Primary	88.07	11.20						
	Secondary	87.02	10.88						

Notes- The table presents that summary statistics of the sample. Columns 1 and 2 present the summary of the full sample, containing information on primary and secondary level students from 75 schools across three academic terms. Columns 3 to 8 limit the sample to within the -7 to +15 relative rank bandwidth for English, Urdu, and Math. Test scores are presented in raw percentages. Regional subject is Sindhi, the language spoken by most people in the province.

Table 2: The Effect of Remediation on Test Scores

	First-stage	Reduced Form	2SLS	2SLS	2SLS	N
Urdu	0.509***	0.0228	0.0435	-0.0555	-0.0390	5169
	(0.0153)	(0.0556)	(0.0778)	(0.0782)	(0.0832)	
English	0.484***	0.147***	0.303***	0.289***	0.288***	5562
	(0.0155)	(0.0503)	(0.100)	(0.0962)	(0.0952)	
Math	0.480***	0.135**	0.276***	0.226***	0.217***	5584
	(0.0156)	(0.0554)	(0.0545)	(0.0494)	(0.0527)	
Controls	No	No	Yes	Yes	Yes	
Prior Score	No	No	No	Yes	Yes	
Grade Fixed Effects	No	No	No	No	Yes	

Notes- This table estimates the impact of remediation on student test scores. Column 1 reports the first stage estimates, column 2 reports the reduced form estimates, and columns 3-5 present the 2SLS estimates that scale the reduced form estimates using the first-stage estimates. The outcome of interest are residualized test scores. Student controls include gender, age, and age^2.

Table 3: Heterogeneity in the Impact of Remediation

	Urdu	English	Math	Urdu	English	Math
	Reduced-Form Estimates			2SLS Estimates		
Panel A: Gender						
Outcome: Residualized Scores						
Remediation	0.016	0.123	0.130**	0.041	0.265**	0.273***
	(0.056)	(0.079)	(0.059)	(0.097)	(0.125)	(0.088)
Remediation x Female	0.021	0.051	-0.001	0.011	0.083	-0.017
	(0.082)	(0.109)	(0.076)	(0.137)	(0.143)	(0.091)
Panel B: Grade Level						
Outcome: Residualized Scores						
Remediation	-0.103	0.156***	0.140**	-0.059	0.219***	0.253***
	(0.053)	(0.034)	(0.043)	(0.104)	(0.069)	(0.077)
Remediation x Secondary	0.127	-0.021	-0.023	0.047	0.007	0.034
	(0.095)	(0.107)	(0.054)	(0.170)	(0.243)	(0.105)
Panel C: Class Size						
Outcome: Residualized Scores						
Remediation	-0.031	0.141	0.148	-0.0758	0.167	0.270
	(0.235)	(0.248)	(0.156)	(0.547)	(0.590)	(0.368)
Remediation x #students	0.055	-0.002	-0.006	-0.013	0.0240	-0.004
	(0.048)	(0.049)	(0.029)	(0.108)	(0.119)	(0.071)
N	5196	5562	5584	5196	5562	5584

Notes- This table tests for the heterogeneity in the impact of remediation by various sub-groups. Panel A tests whether the effects of remediation varied by gender, panel B does the same for grade-level. Panel C checks whether the size of these remediation classes affect learning in these classes. These are fully interacted models with the full set of controls.

Table 4: Spillover Effects of Remediation

	Urdu	English	Math	Science	Islamic Std.	Social Std.
Urdu Remediation	-0.0194	-0.0816	0.0963	0.118	-0.0350	0.0996
	(0.104)	(0.0835)	(0.107)	(0.0844)	(0.0889)	(0.0916)
English Remediation	0.00878	0.288**	0.0922			
	(0.0956)	(0.095)	(0.0758)			
Math Remediation	-0.106	0.0286	0.217***			
	(0.103)	(0.111)	(0.052)			

Notes- The table tests for spillover effect of remediation for one subject on test scores for other subjects. All regressions control for student charachteristics and prior scores, as well as school and grade fixed effects.

Table 5: Effect of Remediation on Time in School Variables

	(1)	(2)	(3)	(4)	(5)	(6)
		Attendance			Dropout (t+1)	
Urdu Remediation	0.276	0.585	0.646	0.0125	0.0489	0.0355
	(1.105)	(1.095)	(1.049)	(0.0360)	(0.0511)	(0.0362)
English Remediation	-0.661	-0.371	-0.284	0.0147	0.0356	0.0220
	(1.148)	(1.138)	(1.089)	(0.0366)	(0.0366)	(0.0373)
Math Remediation	0.204	0.174	0.454	-0.0244	0.00314	-0.0171
	(1.156)	(1.145)	(1.091)	(0.0361)	(0.0554)	(0.0391)
Controls	No	Yes	Yes	No	Yes	Yes
School Fixed Effects	No	No	Yes	No	No	Yes
Grade Fixed Effects	No	No	Yes	No	No	Yes

Notes- The table estimates the impact of remediation classes on time in school variables of attendance and dropout rate in the next time period. Column 1 does not include any controls, column two includes student-level controls, and column 3 additionally adds schools and grade fixed effects.

Table 6: Persistence in Effects of Remediation

	$Score_t$	$Score_{t+1}$	$Remediation_{t+1}$	N
Urdu	-0.039	0.086	0.004	2551
	(0.083)	(0.252)	(0.045)	
English	0.288**	0.118	0.043	2618
	(0.095)	(0.113)	(0.073)	
Math	0.217***	0.121	0.013	2075
	(0.053)	(0.078)	(0.055	

Notes- The table tests for persistence in remediation effects two academic terms/semesters into the future. It shows that there is persistence for Math remediation but gains from English remediation do not last very long. Overall, we find evidence of significant fade-out in the effects of remediation in the medium term. The persistence in the effect of Math remediation, however, are not driven by repeated exposure to remediation classes.

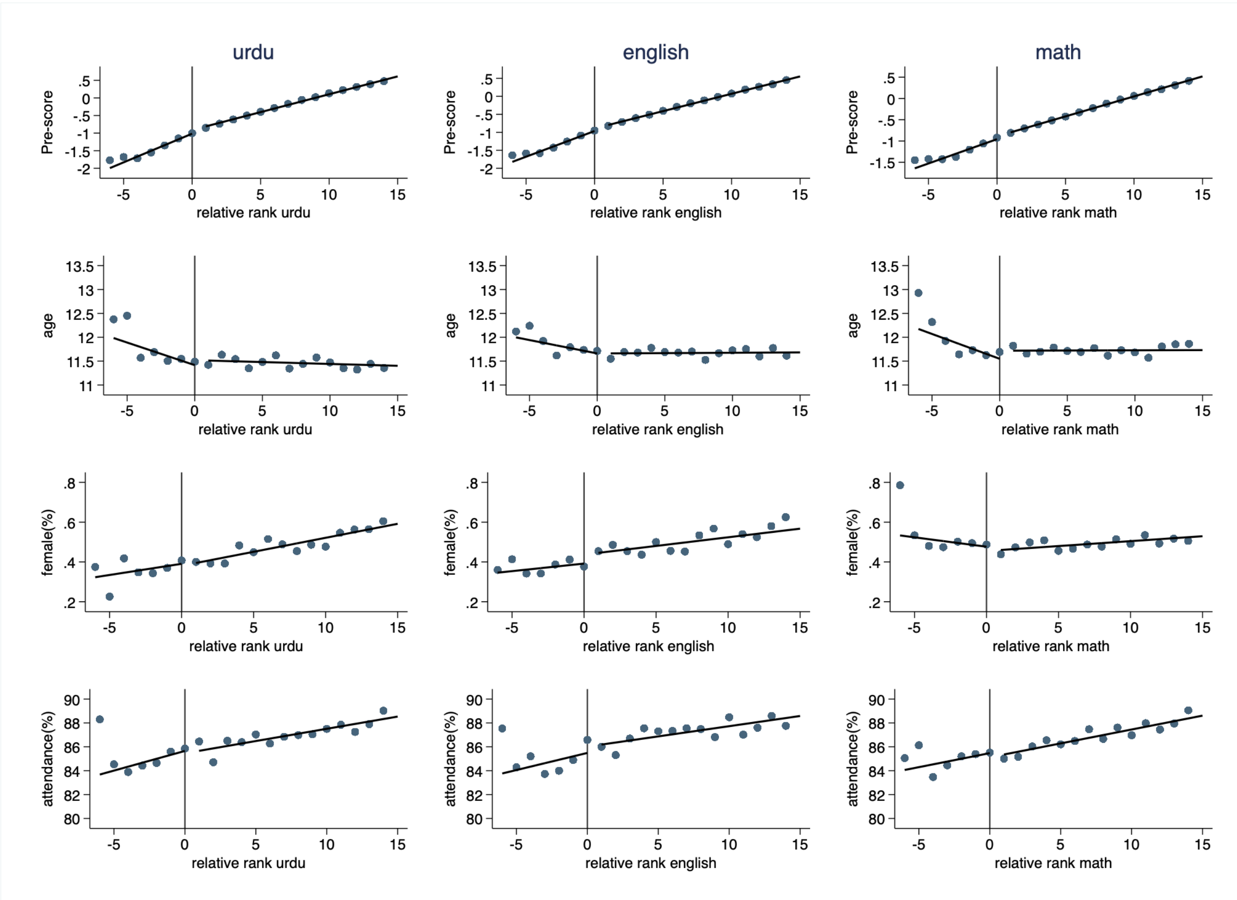
Table 7: Teacher Type and the Impact of Remediation

	(1) Full Sample	(2) Primary	(3) Secondary	(4) Total	(5) Science	(6) SSt.	(7) Ist.
Urdu Rem. (D)	-0.131	-0.185	-0.102	-0.083	0.097	0.079	-0.019
	(0.132)	(0.183)	(0.191)	(0.151)	(0.190)	(0.203)	(0.178)
D * Urdu Teacher	0.398**	0.282	0.657**	0.372*	0.421	0.150	0.428
	(0.203)	(0.263)	(0.329)	(0.209)	(0.311)	(0.322)	(0.293)
Observations	5,070	3,110	1,960	1,960	1,960	1,960	1960

Notes- The table presents the impact of remediation on Urdu test scores by the type of teacher assigned to these remediation classes. It shows that students in schools where Urdu teachers taught these classes gained more on Urdu test scores that in schools where non-specialized teachers taught these classes. There is also suggestive evidence that remediation for Urdu where Urdu teachers taught these classes also had a positive impact on test scores for subjects that had Urdu as the medium of instruction.

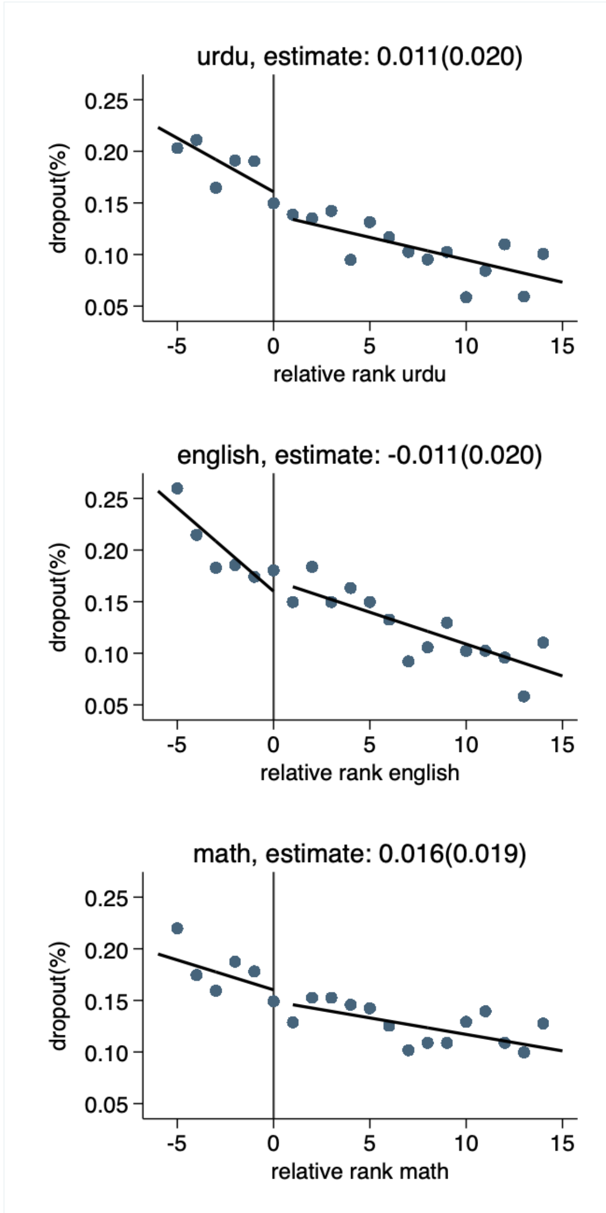
A Appendix

Figure A.1: RD plot for Pre-Remediation Variables



Notes- The figure plots the relationship between the relative ranks for each subject against several pre-remediation student-level variables. Each columns represents a different subject from Urdu, English and Math, and each row of the graphs considers a separate pre-remediation observable. We find no significant discontinuity in any of the pre-remediation outcomes. The accompanying estimates are presented in table A.1

Figure A.2: RD plot for dropout rates



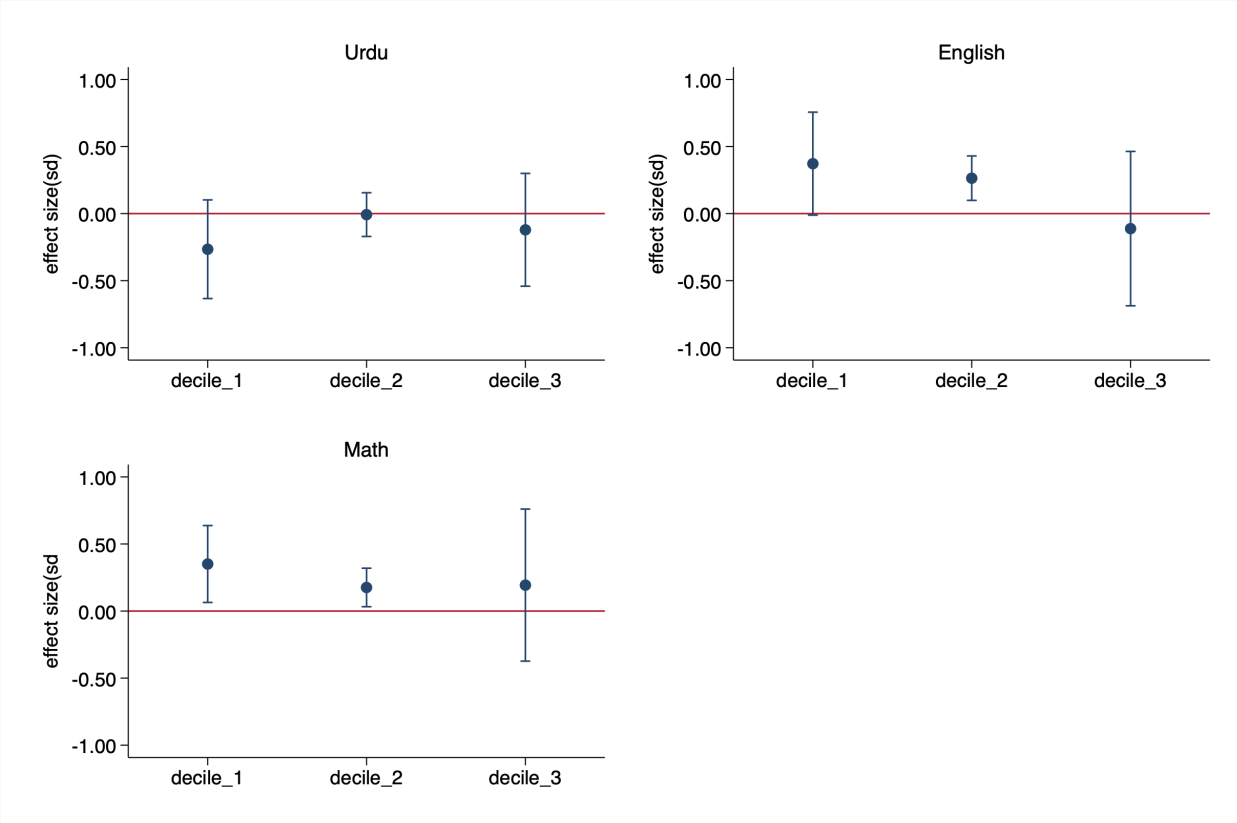
Notes- The figure plots the relationship between the relative ranks for each subject against dropout rates for Urdu, English, and Math. Each point represents a rank mean and the lines represent a linear fit on either side of the cutoff point. We find no significant discontinuity in dropouts. The accompanying estimates are presented in column 5 of table A.1

Table A.1: Testing discontinuity in pre-remediation variables

	(1) Pre-score	(2) Age	(3) Female(%)	(4) Attendance (%)	(5) Dropout	N
Urdu	-0.011	-0.0434	0.0202	-0.317	0.0115	5,426
	(0.053)	(0.117)	(0.0269)	(0.704)	(0.0201)	
English	-0.011	0.0406	-0.0120	-0.586	-0.0110	5,856
	(0.051)	(0.117)	(0.0262)	(0.704)	(0.0203)	
Math	-0.049	-0.190	0.0385	0.758	0.0163	5,886
	(0.047)	(0.118)	(0.0267)	(0.687)	(0.0194)	

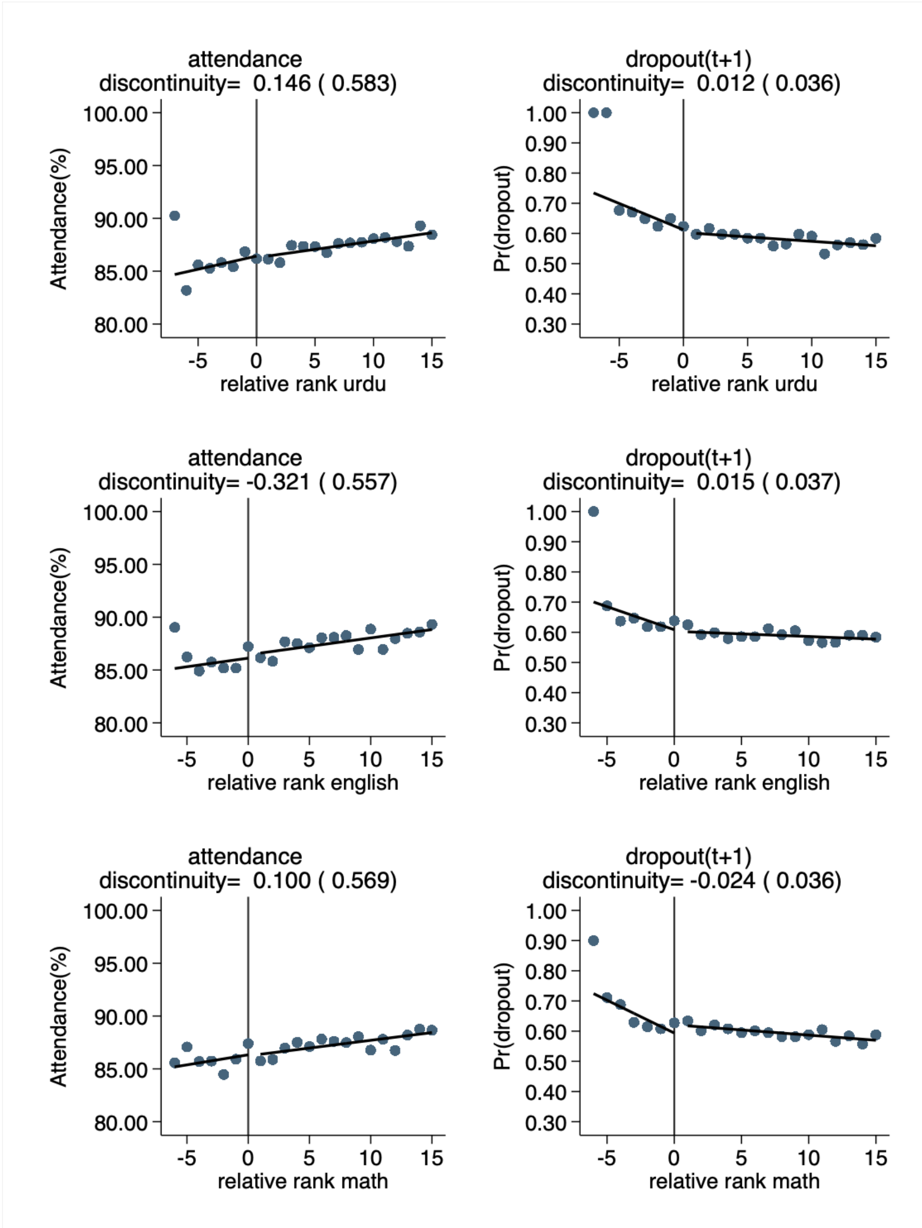
Notes- The table corresponds to figure A.1 and presents the coefficients from the regressions of pre-remediation covariates on the eligibility dummy, controlling for relative ranks and the interaction between the eligibility dummy and relative ranks. The table shows no significant discontinuities in student outcomes at the cutoff. We consider a bandwidth of -7 to +15 in these regressions.

Figure A.3: Impact of Remediation by Marginal Student



Notes- The figure plots the coefficients (along with a 90% confidence interval) of the effects of remediation by the position of the marginal student (the student at the cutoff) in the initial test score distribution. The plots show that remediation was largely equally effective regardless of the location of the cutoff in the initial score distribution.

Figure A.4: Impact of Remediation on Attendance and Dropouts



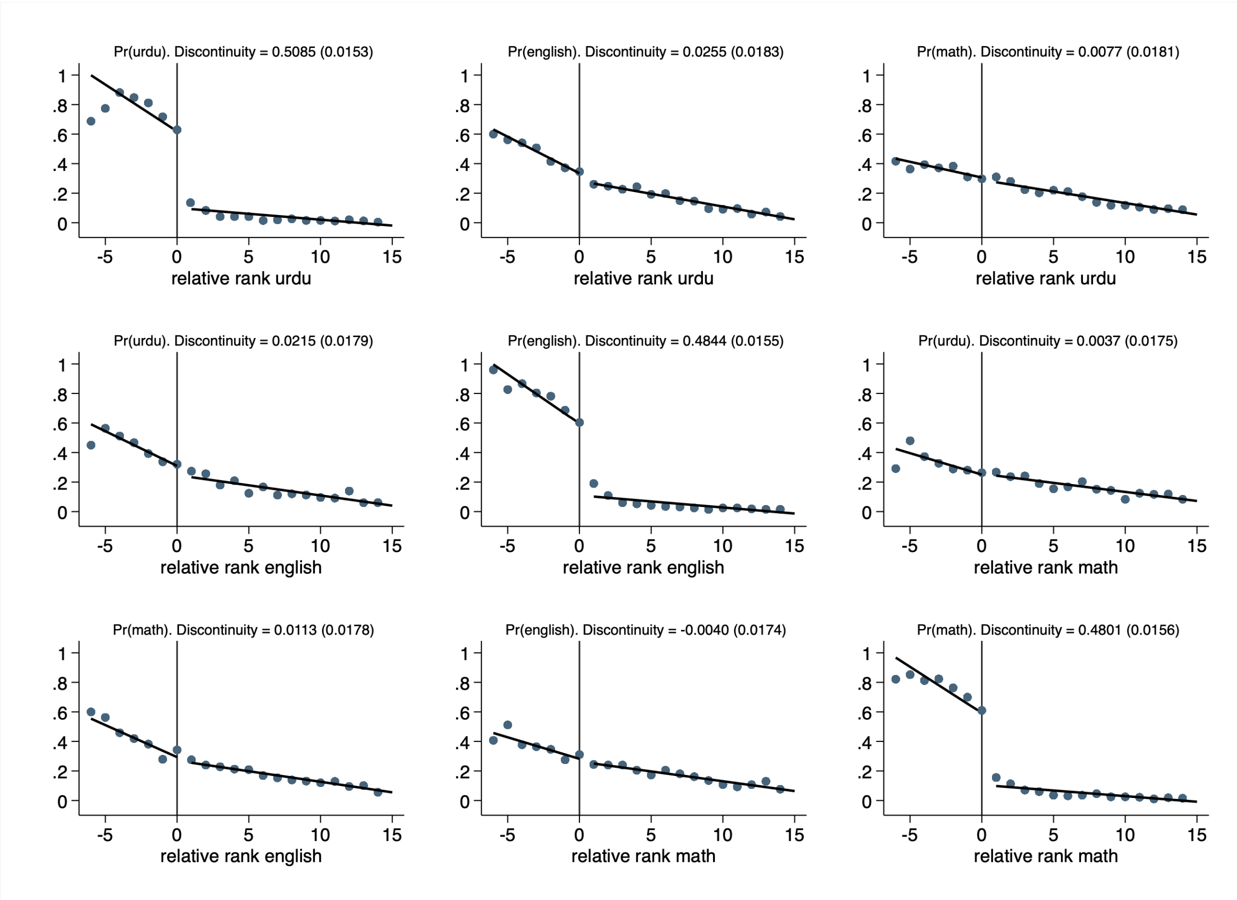
Notes- The figure presents the RD plots for the relationship between students' relative ranks and time-in-school variables of attendance rates and drop-out rates one term later. Each point represents a rank mean and the lines represent a linear fit on either side of the cutoff point. The graphs do not show any discontinuity in these variables.

Table A.2: Placebo Test

	(1) Math	(2) English
Remediation (D)	0.247**	0.345***
	(0.110)	(0.110)
D * Urdu Teacher	-0.0220	-0.114
	(0.0854)	(0.0811)
N	5,070	5,070

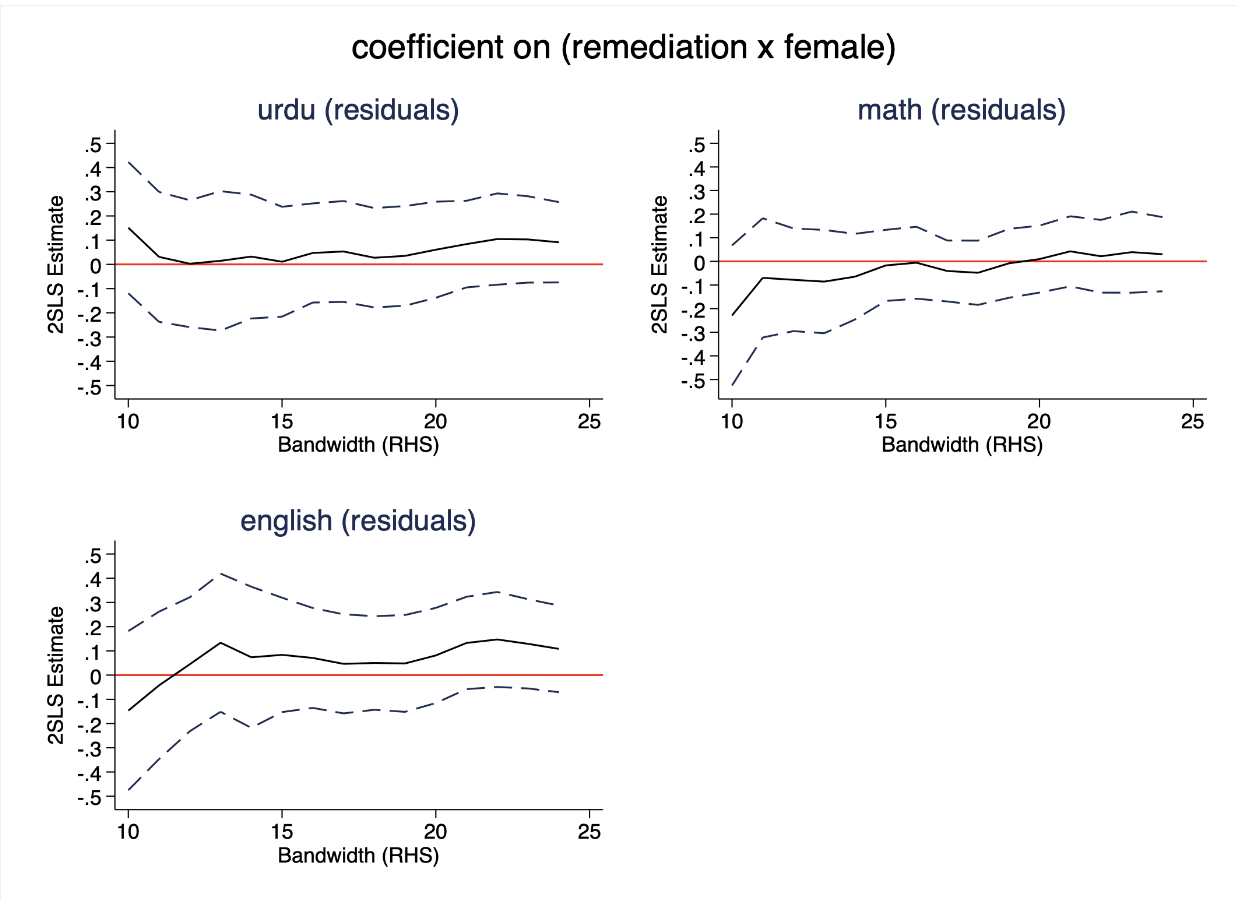
Notes- The table presents a placebo test for the impact of the Urdu teacher type on English and Math scores. It shows that the type of Urdu teacher does not impact test scores for English and Math

Figure A.5: First Stage extended.



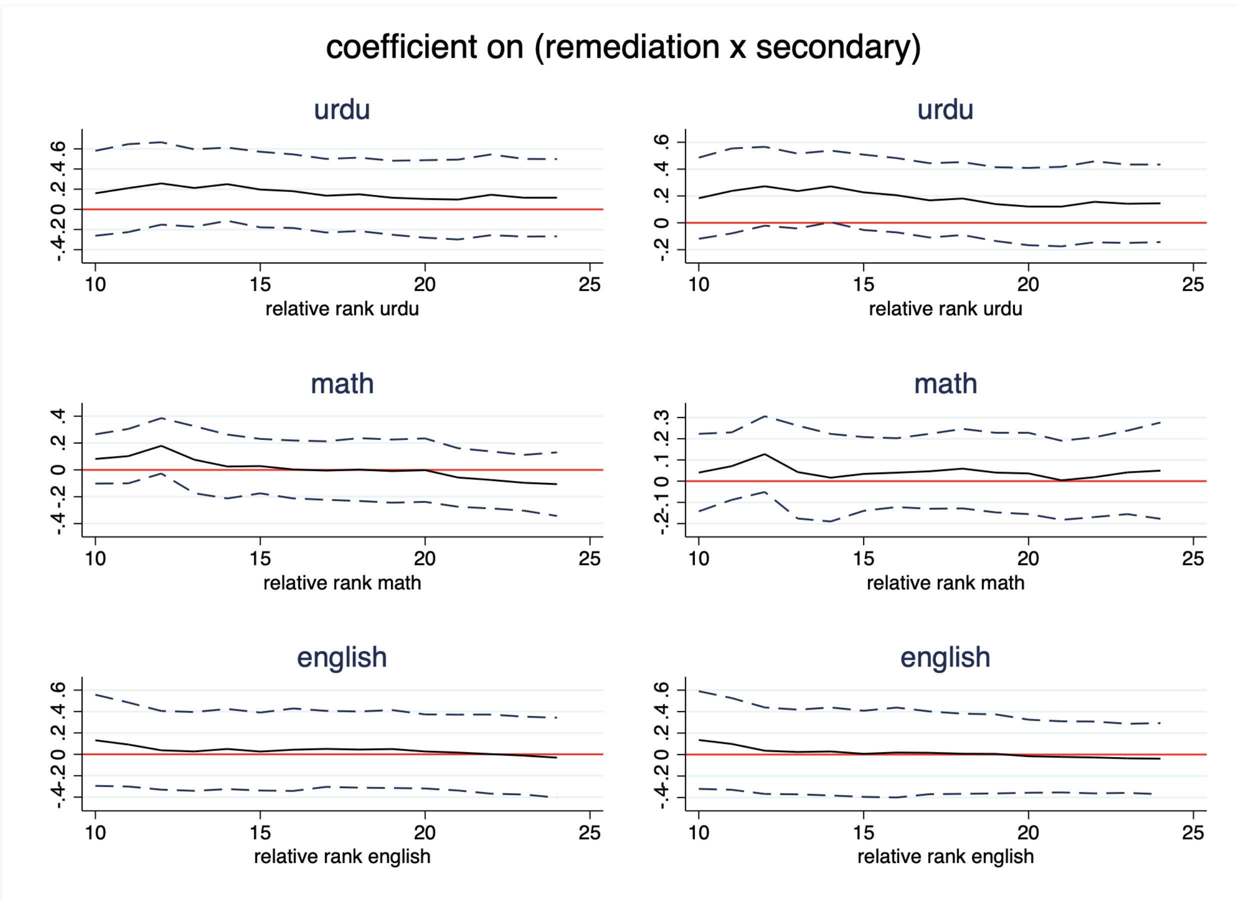
These graphs plot the relationsip between the relative ranks and the probability of remediation for English, Urdu, and math. These graphs show that while relative rank for subject i has a discontinuous relationship with the probability of remediation for subject i at the cutoff, it does has a smooth relationship with the probability of remediation for subject j where $j \neq i$

Figure A.6: Robustness to bandwidth- Gender.



These graphs test the robustness of the coefficient on the interaction between gender and remediation across bandwidths.

Figure A.7: Robustness to bandwidth- Grade Level.



These graphs test the robustness of the coefficient on the interaction between grade level and remediation across bandwidths.

A.1 Second order polynomial

This subsection test the robustness of the main estimates of the paper to a second-order polynomial of the relative ranks.

Table A.3: Estimates using a second-order polynomial of relative ranks

	Urdu	English	Math
First-stage	0.371***	0.374***	0.378***
	(0.021)	(0.021)	(0.021)
Panel A: Post-Pre			
Remediation	0.147	0.266**	0.423***
	(0.196)	(0.130)	(0.152)
Panel B: Residualized Score			
Remediation	0.151	0.230**	0.344***
	(0.157)	(0.115)	(0.115)
N	5169	5562	5584

This table presents the first-stage and 2SLS estimates using a second order polynomial of the realtive-ranks. It shows that the estimates are largely similar to the ones presented in the main text.

A.2 Limited Sample

This subsection test the robustness of the main estimates of the paper by limiting the sample to only those schools that followed the 5-student rule in the remediation classes. While the estimates are noisier due to smaller sample sizes, they remain similar to the ones presented in the main text.

Figure A.8: First-Stage

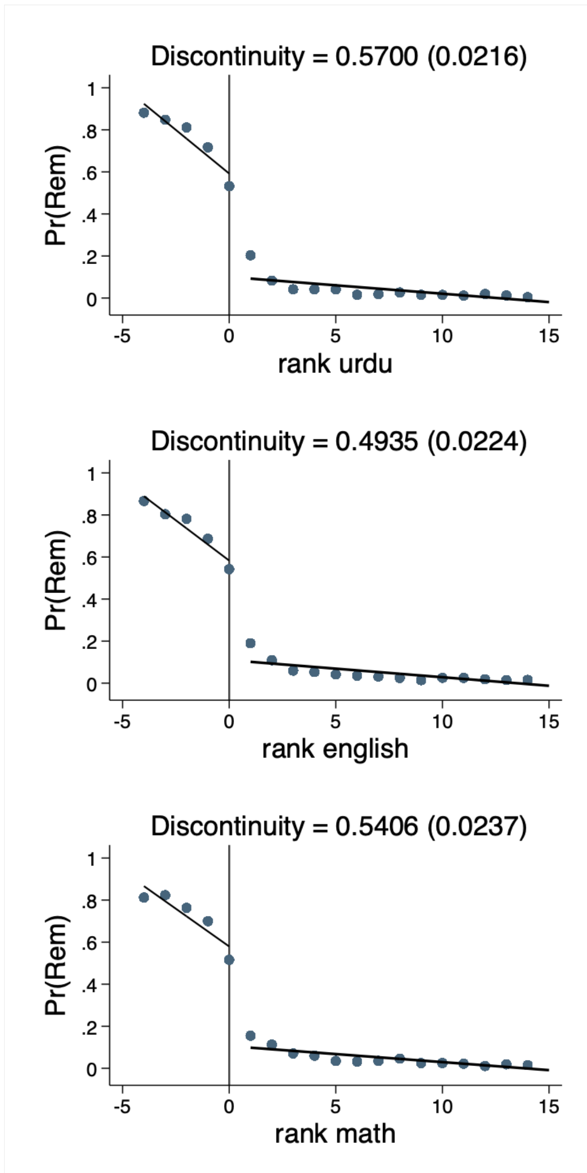


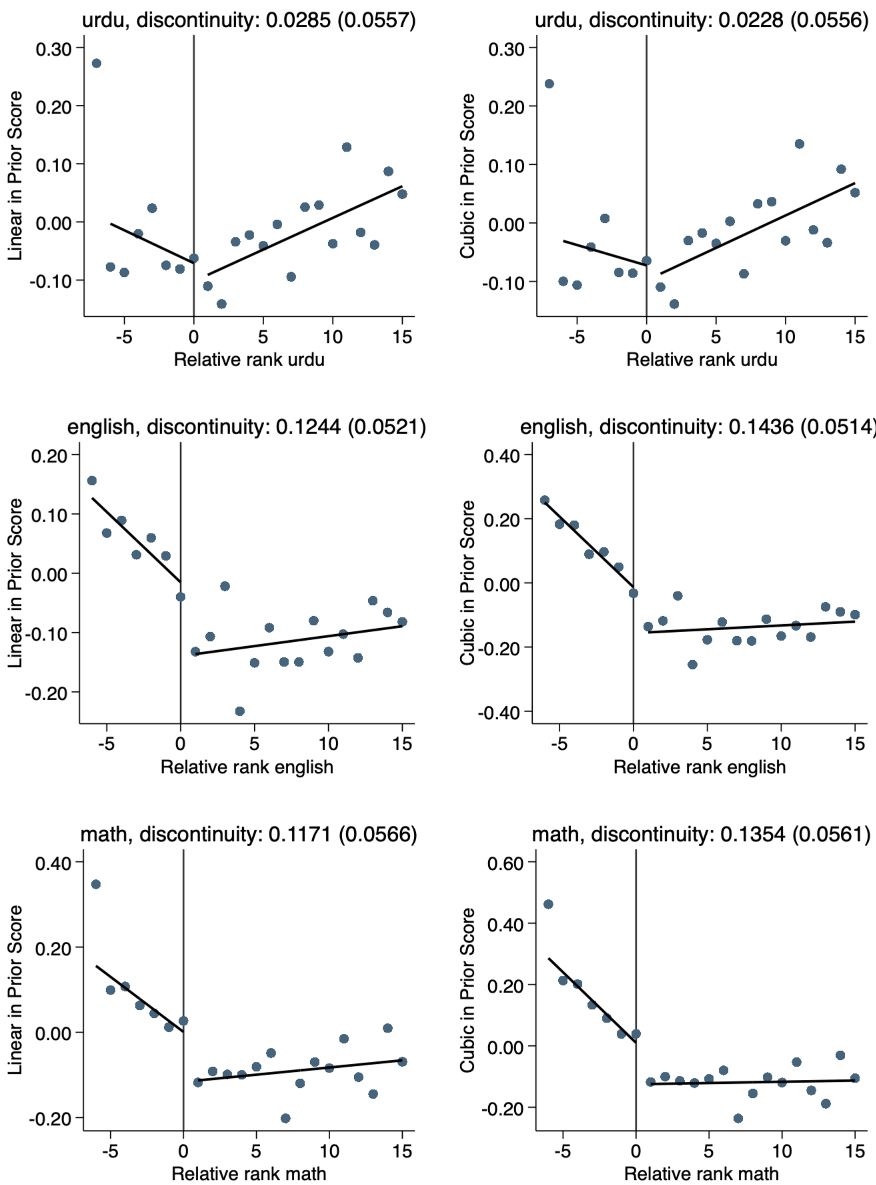
Table A.4: Limiting to schools with 5 student rule

	Urdu	English	Math
Panel A- Post-Pre			
Remediation	-0.0539	0.240*	0.217**
	(0.153)	(0.144)	(0.0869)
Panel B- Residualized Scores			
Remediation	-0.0621	0.228*	0.176**
	(0.114)	(0.138)	(0.0837)
N	2,246	2,499	2,360

A.3 Alternate definitions of the outcome variable

In this subsection, we consider various definitions of the outcome variables to test for the robustness of our results. We condier two outcome definitions. The first is where we obtain residualized scores from a regression of the post-remediation test scores on a linear function of the pre-remediation test scores and school fixed effects. The second definition is where we obtain the residuals from a regression of the post scores on a cubic function of the pre-remediation test scores. In both these cases, we find that the estimates are very similar to the main estimates using a quadratic function of the pre-remediation test scores presented in the main text.

Figure A.9: Alternate outcome variable definitions- RD plots



The graphs present the RD plots for the English, Math, and Urdu using the two definitions of the outcome variable defined above.

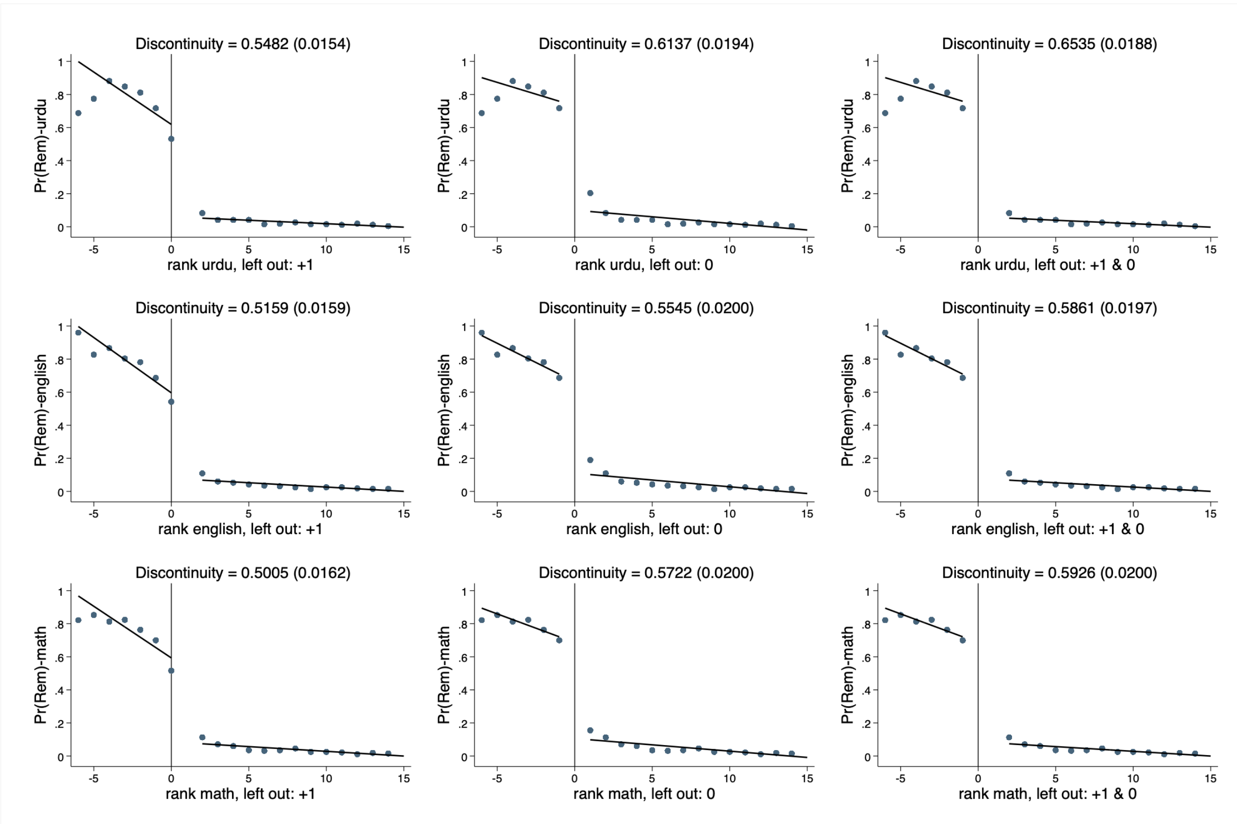
Table A.5: IV Estimates: Alternative Definition of the outcome

	(1) No Controls	(2) Student Controls	(3) All Controls
Panel A: Linear in pre-remediation score			
Urdu	0.0543	0.0472	0.0403
	(0.106)	(0.106)	(0.0963)
English	0.257**	0.269**	0.268***
	(0.108)	(0.108)	(0.0952)
Math	0.240**	0.228**	0.219**
	(0.116)	(0.116)	(0.108)
Panel B: Cubic in pre-remediation score			
Urdu	0.0435	0.0365	-0.0390
	(0.106)	(0.106)	(0.0967)
English	0.296***	0.308***	0.277***
	(0.106)	(0.106)	(0.0953)
Math	0.277**	0.268**	0.212**
	(0.115)	(0.115)	(0.106)
Observations	5,169	5,562	5,584

A.4 Donut-hole RD

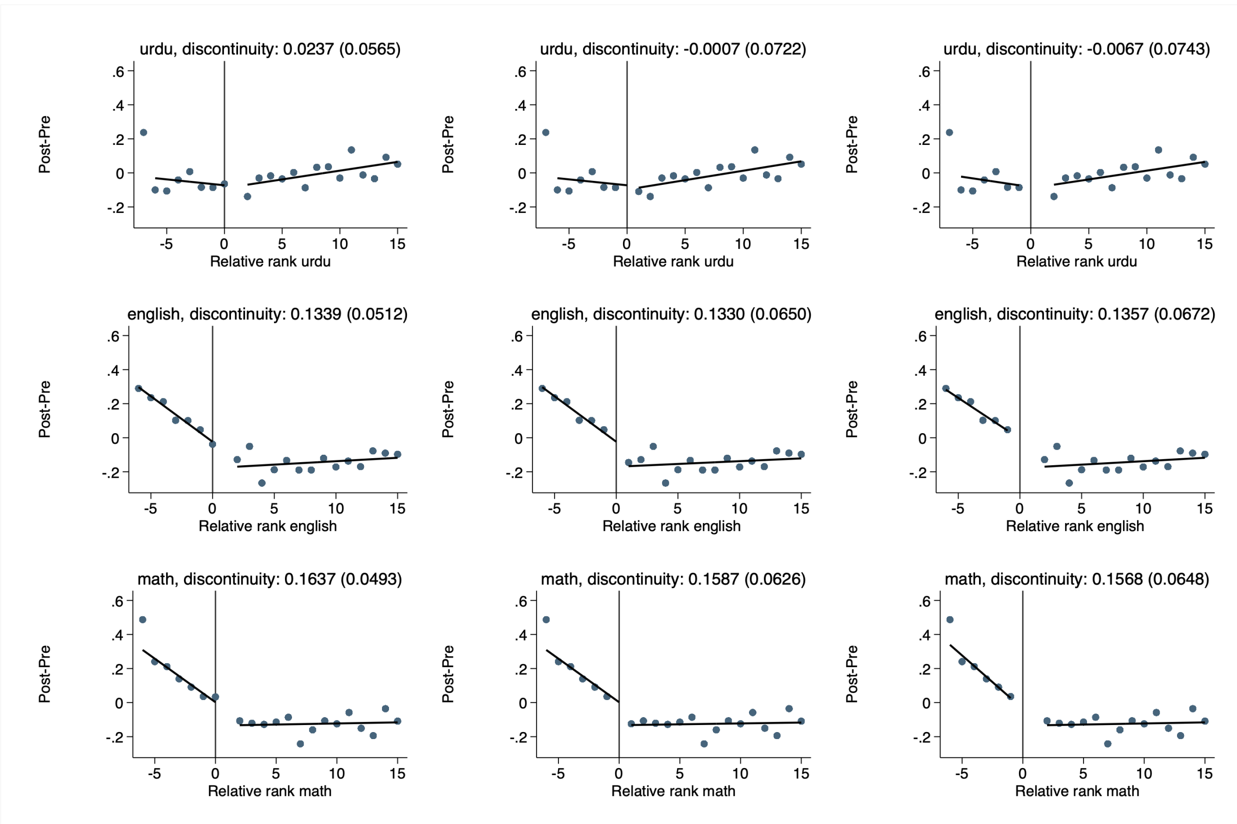
This subsection test for the robustness of the estimates using a Donut-hole RD approach that leave out observations extremely close to the cutoff point (M. D. Cattaneo et al., 2019)

Figure A.10: First-stage: Donut RD



These graphs present the first-stage estimates for the Donut-Hole RD approach leaving out either the student with a relative rank of +1 (column 1), relative rank of 0 (column 2), or both +1 and 0 (column 3)

Figure A.11: Donut-hole RD plots



These graphs present the RD plot for the Donut-Hole RD approach leaving out either the student with a relative rank of +1 (column 1), relative rank of 0 (column 2), or both +1 and 0 (column 3). The outcome here is a simple post-pre difference in scores

Table A.6: Donut-hole RD: IV estimates

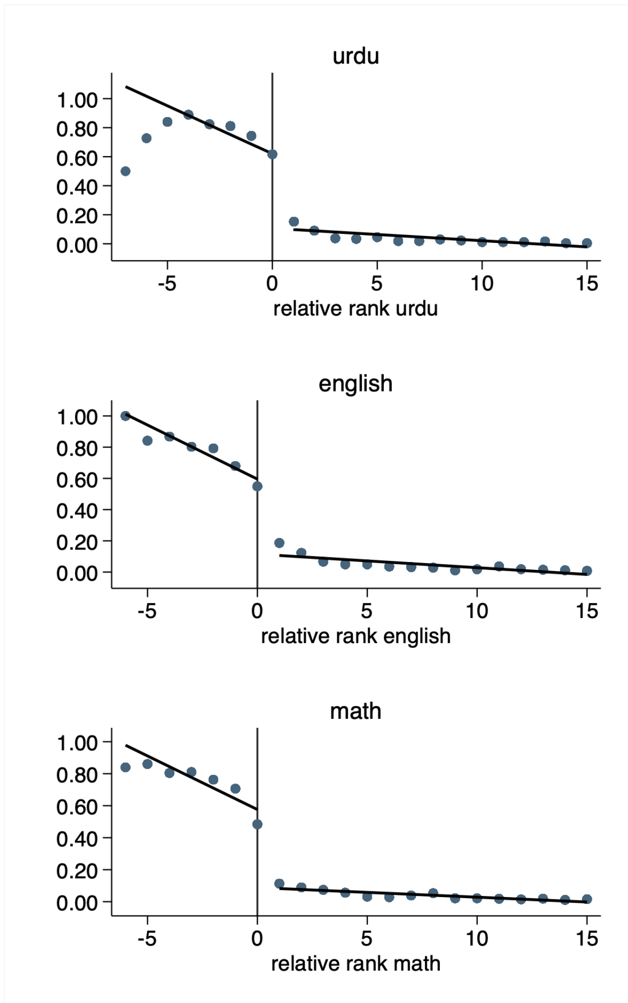
Urdu	English	Math	Rank(s) left out
-0.00133	0.247***	0.299***	+1
(0.106)	(0.0916)	(0.0577)	
-0.0321	0.233***	0.245***	0
(0.116)	(0.0892)	(0.0658)	
-0.0453	0.218**	0.237***	0 & +1
(0.113)	(0.0953)	(0.0660)	

This table presents the IV estimates for the Donut-Hole RD approach leaving out either the student with a relative rank of +1 (column 1), relative rank of 0 (column 2), or both +1 and 0 (column 3). The outcome here is a simple post-pre difference in scores

A.5 Appendix: Term 1 Class Size as Threshold

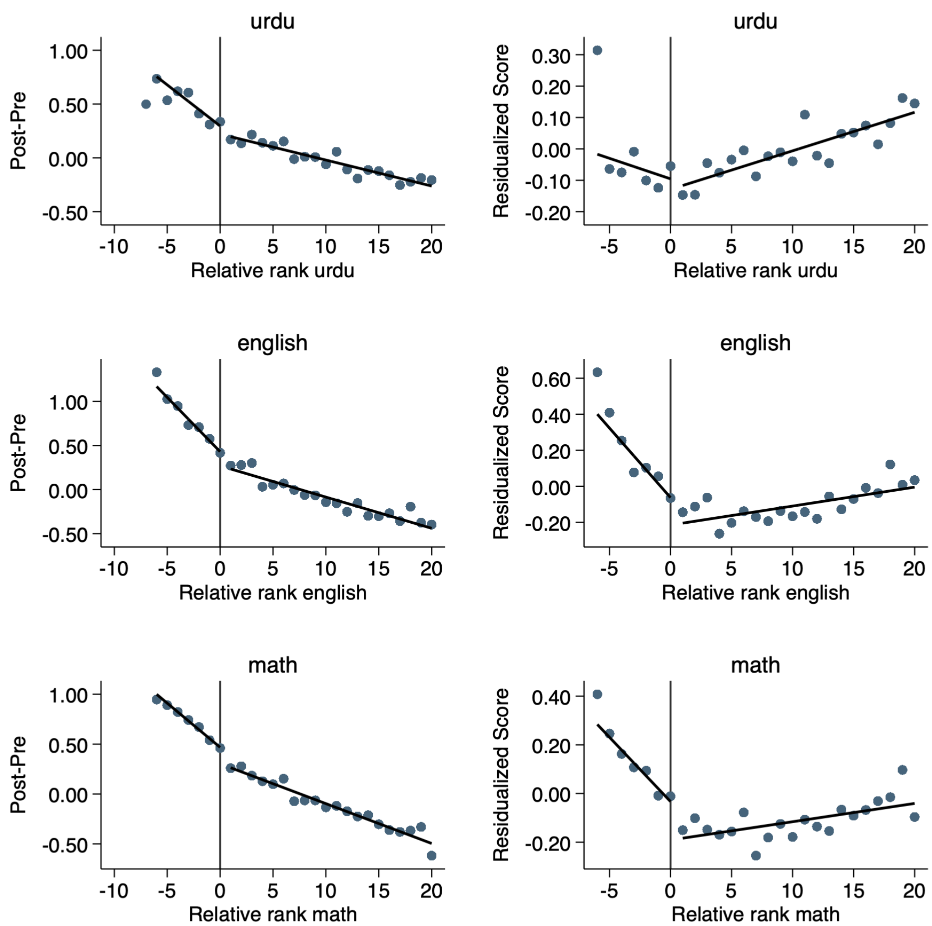
This section repeats the main analysis but uses Term 1's remediation class sizes as thresholds for Term 2's remediation classes for each of Urdu, English, and Math to check for the robustness of the main estimates.

Figure A.12: First Stage



Notes-The figure shows the first stage relationship between the probability of remediation and the relative ranks for each of Urdu, English, and Math

Figure A.13: RD- Plots



Notes- The figure show the RD plots for three outcomes. The first column shows the reduced form relationship between relative ranks and the difference in post- and pre-remediation scores. The second column plots the relationship between relative ranks and residualized scores formed by regressing post remediation scores on a quadratic function of pre-remediation score and school fixed effects.

Table A.7: Reduced Form Results

	(1) First- Stage	(2) No Controls	(3) Student Controls	(4) Student Controls + prior score	(5) All Controls + FEs
Panel A: Post-Pre					
Urdu Remediation	0.514***	0.0675	0.0670	0.0164	0.0117
	(0.0153)	(0.0536)	(0.0533)	(0.0520)	(0.0521)
English Remediation	0.480***	0.115**	0.121**	0.0840*	0.0843*
	(0.0154)	(0.0486)	(0.0485)	(0.0465)	(0.0465)
Math Remediation	0.467***	0.161***	0.160***	0.120***	0.118***
	(0.0156)	(0.0463)	(0.0462)	(0.0444)	(0.0443)
Panel B: Residualized Scores					
Urdu Remediation	-	0.0254	0.0246	-0.0115	0.00315
		(0.0563)	(0.0561)	(0.0556)	(0.0525)
English Remediation	-	0.108**	0.113**	0.0958*	0.101**
		(0.0504)	(0.0503)	(0.0499)	(0.0463)
Math Remediation	-	0.129**	0.128**	0.101*	0.106**
		(0.0554)	(0.0554)	(0.0547)	(0.0528)

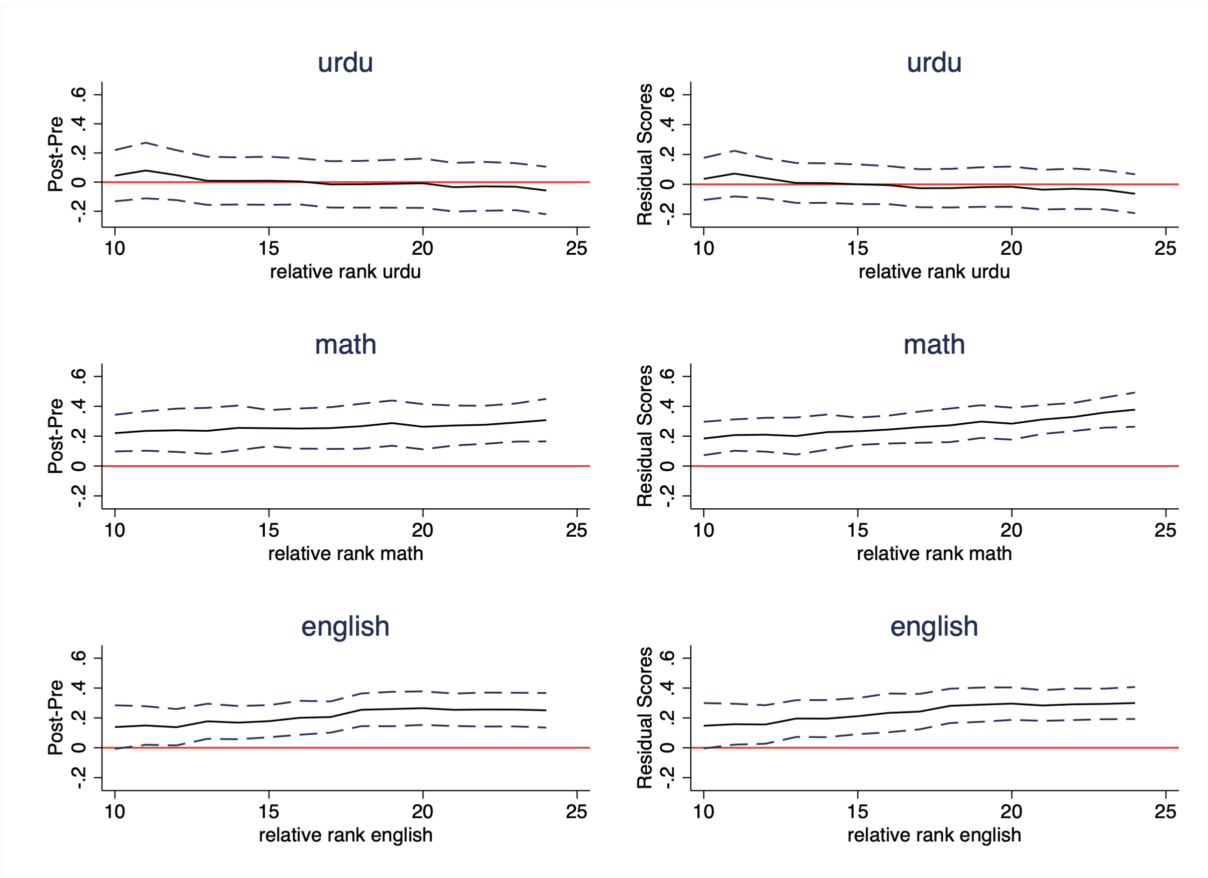
Notes- The table presents the reduced form estimates that correspond to the RD plots presented above. Panel A considers the simple difference between Post and Pre remediation scores as an outcome. Panel B considers residualized scores, formed by regressing the post-remediation scores on a quadratic function of prior scores and school FE, as the outcome. Panel B does not include school FE but Panel A does include school FE.

Table A.8: 2SLS Estimates

	(1) No Controls	(2) Student Controls	(3) All Controls + FE
Panel A: Post-Pre			
Urdu Remediation	0.116	0.0163	0.00922
	(0.105)	(0.101)	(0.100)
English Remediation	0.242***	0.177***	0.178***
	(0.0698)	(0.0633)	(0.0651)
Math Remediation	0.345***	0.257***	0.254***
	(0.0789)	(0.0732)	(0.0737)
Panel B: Residualized Scores			
Urdu Remediation	0.0434	-0.0259	-0.00262
	(0.0814)	(0.0775)	(0.0833)
English Remediation	0.225***	0.201***	0.228***
	(0.0769)	(0.0746)	(0.0736)
Math Remediation	0.279***	0.223***	0.229***
	(0.0562)	(0.0572)	(0.0532)

Notes- The table presents the 2SLS estimates that correspond to the RD plots presented above. Panel A considers the simple difference between Post and Pre remediation scores as an outcome. Panel B considers residualized scores, formed by regressing the post-remediation scores on a quadratic function of prior scores and school FE, as the outcome. Panel B does not include school FE but Panel A does include school FE

Figure A.14: Testing Robustness of Estimates to Bandwidth



Notes- The figure checks for the robustness of the 2SLS estimates to variations in the bandwidth. The figure shows that the estimates are robust to bandwidth.

Table A.9: Comparing Class Sizes Between Terms

class ID	math term1	math term2	diff_math	english term1	english term2	diff_eng	urdu term1	urdu term2	diff_urdu
1	5	5	0	5	5	0	5	5	0
2	5	5	0	5	5	0	5	5	0
3	5	5	0	5	5	0	5	5	0
4	5	5	0	5	5	0	5	5	0
5	5	6	-1	5	5	0	5	5	0
6				4	6	-2	4	6	-2
7	4	6	-2	4	5	-1	4	6	-2
8	4	5	-1	4	5	-1	4	5	-1
9	5	5	0	5	5	0	5	5	0
10	6	7	-1	6	7	-1	6	7	-1
11	4	4	0	4	5	-1	4	4	0
12	5	5	0	5			5		
13	5	5	0	5			5	3	2
14	5	4	1	5	4	1	5	4	1
15	5	5	0	5			5	5	0
16	5			5			5		
17	4	5	-1	4	5	-1	4	5	-1
18	5	6	-1	5	6	-1	5	6	-1
19	6	5	1	6	5	1	5	5	0
20	4			4			4	5	-1
21	5	5	0	5	5	0	5	5	0
22	6	5	1	5	5	0	6		
23	7	5	2	6			5		
24	5			5	6	-1	5	6	-1
25	4	5	-1	4			4		
26	4	4	0	4	4	0	4	4	0
27	4	4	0	4			4	4	0
28	4	4	0	4			4	4	0
29	6	6	0	6	6	0	6	6	0
30	6	6	0	6	6	0	6	6	0
31	6	6	0	6			6	6	0
32	7	6	1	4	5	-1	4	4	0
33	4	4	0	4	4	0	4	4	0
34	5			5	5	0	5	5	0
35	5			5	7	-2	5	7	-2
36	5	5	0	5	5	0	5	5	0
37	5	5	0	5	5	0	5	5	0
38	4	3	1	4			4		
39	5	4	1	5			5		
40	4	3	1	4			4	5	-1
41	4	3	1	4			4	3	1
42	5			5	3	2	5		
43	5	3	2	5	3	2	5	3	2
44	4			5			5		
45	5	5	0	5	5	0	5	5	0
46	5	5	0	5	5	0	5	5	0
47	5	5	0		4		4	3	1
48	5	4	1		4		4	4	0
49	4	5	-1	4	5	-1	4	5	-1
50	4	5	-1	4	5	-1	4	5	-1
51	5	7	-2	5	7	-2	5	7	-2
52	5	4	1	5	6	-1	5		
53	6	4	2	6	3	3	6		
54	6	4	2	6	6	0	6		
55	5	5	0	5	4	1	5		
56	5	5	0	5	5	0	6		
57	5	4	1	5	3	2	4	3	1
58	5	4	1	5	4	1	5	4	1
59	7	5	2	4			5		
60	7	5	2	4			5		
61	6	5	1		6				
62	4	5	-1	5	4	1	4	4	0
63	5	4	1	5	4	1	5	4	1
64	5			5	3	2	4		
65	5	5	0	5	5	0			
66	5	4	1	5	4	1			
67	4			4	4	0	4	4	0
68	4	7	-3	4	7	-3	4	7	-3

Notes- The table compares remediation class sizes between term 1 and term 2 for a given grade-school. It shows that for many classes, the number of students at for given remediation classes does not vary from term to term suggesting that teachers have a fixed class size that they populate with poor-performing students. It also suggests that the number of students across subject within a class remains similar The missing values represent classes for which we did not find data due to poor record keeping by schools.