Free to choose: Introducing Technology for Foundational Learning in Pakistan

Tahir Andrabi[*], Juan Baron[†], Isabel Macdonald[‡], Zainab Qureshi[§]

April 13, 2025

Abstract

Prior research has found varying returns to education technology (EdTech). These studies almost universally mandate the use of technology when it is introduced. In this study, we find that the approach to introducing EdTech has a large impact on learning outcomes. We explore these dynamics in the context of a large-scale foundational learning program in public primary schools in Pakistan. The novel curricular program incorporated both targeted instruction and structured pedagogy, and some teachers were also randomly assigned access to a low-cost, smartphone application (tech tool) designed to reduce the administrative burden of classroom management. Test results show that the 8-week program raised student learning overall by 0.12 SD across math, English, and Urdu. Learning gains, especially among low-performing students, were highest when the tech tool was offered to teachers but framed as optional. In contrast, mandating the use of the tech tool (0.10 SD) had significantly lower student learning relative to the Optional tech (0.19 SD) version of the program. The differences are driven by schools with female teachers and students (schools are gender segregated), where the Mandated treatment arm performed even worse (0.02 SD) relative to the Optional arm (0.17 SD). These findings suggest substantial heterogeneity in teacher returns to EdTech, and support a self-selection approach for introducing new technology.

Keywords: Education Policy, EdTech, Targeted Instruction, Structured Pedagogy, Treatment Effects

[0] Maleeha Hameed, Mariam Raheem, Angela Tran, and Tomoya Murakawa provided outstanding program management and research assistance.

[*] Pomona College and CERP

[†] World Bank

[‡] Irrational Labs

[§] BEAJ

1 Introduction

Prior research has found varying returns to education technology (GEEAP Report: Akyeampong et al. [2023], Beg et al. [2022], Beuermann et al. [2015], Muralidharan, Singh and Ganimian [2019]). The polices in question almost always mandate the use of technology when it is introduced. This is particularly so in public schooling systems where standardization is enforced across most elements of education delivery. This mandated usage can be a significant imposition in large schooling systems in LMICs where there are considerable differences in student learning levels. Schools are spread over rural and urban areas, and teacher demographics and educational attainment vary as well. While we can observe many dimensions of this heterogeneity, there remain significant unobservable aspects of a particular schooling environment, teacher capability, and comfort with technology that could impact the effective utilization of technology innovation.

In this study, we investigate how learning is impacted when technology is offered, but teachers are given an individual choice of whether to use it or not. We explore this question among a sample of 2,920 teachers and 24,730 fourth grade students at 916 public primary schools in Khyber Pakhtunkhwa (KP), Pakistan. Schools in the sample were randomly assigned to implement a novel, low-cost foundational learning program called Targeted Instruction in Pakistan (TIP), which was designed to fill gaps in basic literacy and numeracy caused by COVID-19 school closures. The program included two hours of daily instruction across math, English, and Urdu for an 8 week period. It began with a diagnostic assessment, which was then used to sort students across grades 1-5 into peer learning groups based on actual learning rather than age. During TIP classes, students were shuffled into their appropriate peer groups, and teachers delivered a structured pedagogy curriculum. During the program, the classes replaced the standard math, English, and Urdu classes, so no additional teaching hours or teaching staff were involved in the delivery.

Among schools assigned to complete the program, some were also randomly given access to a low-cost technology tool (tech tool) designed to provide assistance and reduce the administrative burden for teachers. The tech tool enables teachers to enter student test scores, track student and classroom learning, sort students into ability-based learning groups, access and check off lesson plans, and watch program and pedagogy train-

ing videos. Teachers could use the tool on their personal smartphone, tablet, or computer, or log in to their account on a school or another teacher's device. Once downloaded, the tool could function offline (essential for rural schools), with only an occasional internet connection to sync and backup data. In addition to the tech tool itself, the treatment included access to WhatsApp groups where teachers could share tips with each other and receive help, encouragement, and reminders from program mentors.

Schools in the sample were randomized into four treatment groups: Control, which did not implement the TIP program; Paper, where teachers delivered the full TIP program, but used paper-tracking tools in lieu of the tech tool; Mandatory, where teachers were given access to the tech tool and required by the education department to use it regularly; and Optional, where teachers were given access to the tool but informed that each teacher could decide whether to use it or use paper tracking and support tools instead. All teachers in the Paper, Mandatory, and Optional received 4 days of initial training and 1 day of refresher training on the program and tech tool, and were visited throughout the program by mentors who provided additional support, collected data, and ensured program and treatment compliance.

Our core results show that on average, all versions of the program succeeded in raising student test scores relative to the control group, but learning was highest in the Optional technology treatment. When using paper tools (without technology), the program led to a 0.12 SD gain in endline test scores across the three subjects (math, Urdu, and English), which aligns with impacts from other no-tech targeted instruction programs like Teaching at the Right Level (Banerjee et al. 2007). The extra impact of teacher-support technology depended on the approach by which technology was introduced. When teacher technology use was mandated and enforced by education officials, student learning showed a gain of only 0.10 SD relative to the no-program control group. When technology was made accessible but optional to teachers, student learning was 0.19 SD, the highest of any treatment group and statistically different from the Mandated tech treatment.

When we separate these impacts by baseline student test scores, we find that students throughout the distribution had higher learning in the Optional treatment, whereas only higher performing students at baseline

showed test score gains in the Paper and Mandatory groups. This result may be driven by the specific design of the tech tool, which aimed to bring greater attention to low-performing students and encourage teachers to give extra support to help these students catch up.

Schools in KP are almost entirely gender-segregated. In addition, only female teachers are paired with female students and male teachers with male students. When we separate the results by gender, we find distinct patterns across male and female schools. For female schools, endline learning results in the Optional group (0.16 SD gain relative to control) are considerably higher than the Paper and Mandatory groups (0.07 SD and 0.02 SD). In fact, only the Optional treatment produced a statistically significant gain in student learning among females. For male schools, there were substantial learning gains across all treatment groups, and the differences between treatments were not statistically significant. However, the point estimate for male schools was also highest for the Optional treatment. These findings suggest substantial heterogeneity in the returns to technology, and support a self-selection model of adoption, especially for women.

To further understand these gender patterns, we conduct exploratory analyses on which teachers took up the tech tool in the Optional group, and how impacts varied by usage and teacher characteristics. [1]. First, we see that the take up of the tech tool is only weakly predictive by teacher and school characteristics (gender, rural/urban, school size, teacher qualification, age, year of joining). Second, we find that for both genders, teacher tech usage is highly correlated with others at their school. Despite frequent reminders that each teacher should make an individual choice, 79.2% of schools in the Optional group (79.1% for male and 79.4% for female schools) showed a universal decision to completely use tech or not at all. (median teachers at a school is 5). Among the rest of the schools, which had only some teachers with tech tool accounts, the median ratio of teachers with tech tool account was 58.6% for male and 57.1% for female schools. These data suggest that teacher decision-making and behavior around technology and program adoption is highly interconnected with other teachers, and models of group or organization-level decision-making may be more representative versus thinking of take-up as an individual choice. In addition, we observed difference in

[1] As technology choice within the Optional group is endogenous, this analysis cannot be interpreted as causal

school-level mean learning gains by the school's tech tool status in female schools. There is evidence that schools which have some teachers using the tech tool and others not are performing better than schools in which all use the tech tool or those in which no teacher adopts the tech tool. However, as this choice is endogenous, we cannot say anything more causally about this.

This study contributes to several areas of literature on education and technology. First, our most novel contribution is to show that the method of introducing technology in an education system (and more broadly, an organization) has a tremendous impact on outcomes. This study shows that technology implementation is effective if teachers can individually decide to opt in given their personal and school constraints. Mandating them to do so can yield less positive outcomes.

Second, this study contributes to the growing literature on the potential for technology to improve educational outcomes. Past literature such as Barrera-Osorio and Linden [2009] and Cristia et al. [2017] suggests that investment in hardware alone or only giving instructions for how to use the hardware itself will not lead to positive educational outcomes. On the other hand, recent literature suggests that technology can enhance personalized learning (Muralidharan, Singh and Ganimian [2019], Wang et al. [2024]) and assist teacher's classroom teaching (Beg et al. [2022]) when the implementation is carefully carried out. In most cases, the technology in question is shown directly to students, either by students using the device individually or by teachers showing a video or content on a device to students. In our context, we show that the technology that is designed solely for teacher use, in order to improve their pedagogy and reduce the administrative burden of new programs ca also improve student learning.

Third, this study also demonstrates how intentional design of a tech tool can lead to specific behavior change. In this case, the tech tool was specifically designed to encourage teachers to focus on students who were struggling, such as by putting low performing students at the top of the list, highlighting them in salient colors, and providing a feature to pair these students with high performing peer mentors. The results show that teachers who were permitted to use the tool as they liked succeeded in raising learning at the bottom

of the baseline distribution, whereas other treatments only raised student learning among students who were already performing well. This outcome is likely attributable to the product design, though multiple designs of the tool were not tested to provide causal proof.

Fourth, we demonstrate a potentially cheaper (and thus more scalable) method to deliver proven education interventions like targeted instruction and structured pedagogy. The 2023 report by the Global Education Evidence Advisory Panel (GEEAP) identified targeted instruction and structured pedagogy as two of the top three investments in education, given strong evidence of impact and cost effectiveness (GEEAP Report (2023)). Targeted instruction (i.e. tailoring content to actual student learning level rather than child age) has proven effective in different forms such as using volunteers to run in and after school hour programs, computer-enhanced individualized learning programs, intensive remedial camps, and student tracking in regular schools (Banerjee et al. [2007], Banerjee et al. [2010], Banerjee et al. [2016], Duflo, Dupas and Kremer [2011], Duflo, Kiessel and Lucas [2024]). Most programs yielded high educational returns, however, required additional inputs such as outside volunteers, extra work hours for teachers, or new hardware investments, which could pose a barrier to scale in low resource. The intervention studied here relied on regular public teachers during regular school hours and leveraged existing tech devices, establishing the effectiveness of targeted instruction in a new and highly scalable format.

Lastly, we also innovate on a cheaper way to deliver structured pedagogy interventions. Prior work has shown that printed teaching guides with detailed lesson plans can improve educational outcomes (Brunette et al. [2019], Cilliers et al. [2020], Piper et al. [2018]), as well as providing tablets with standardized lesson guides and school administration tools (Gray-Lobe et al. [2022]). We build on these findings by offering lesson plans via an app that teachers can use on their own smartphones, with no new devices required. We note that in the intervention, we also provided teachers with printed versions of the app-based lesson plans, however, the success of the program suggests testing with only the app-based version would be warranted.

The remainder of the paper proceeds as follows. In section 2, we provide additional context about the lo-

cation, program, and technology intervention. In section 3, we describe the experimental design and baseline characteristics. In section 4, we provide the results on student learning outcomes and describe secondary analyses around teacher characteristics and technology take up. Section 5 concludes.

2 Context and Intervention

2.1 Context

The intervention was conducted in public primary schools in the two largest districts (Peshawar and Mardan) of the province of Khyber Pakhtunkhwa (KP) in Pakistan. In KP, primary schools are almost entirely gender-segregated for both students and teachers. In rural areas, it is sometimes the case that a small number of children of the opposite gender may be enrolled in each school, such as a few female students joining a male school if a female school is not located nearby. Male schools make up 57.8% of total schools in our sample and tend to have more students. The median male school has 237 students (171 students for grades 1-5) and 5 teachers, whereas the median female school has 189 students (130 students for grades 1-5) and 4 teachers.

Both districts contain a mix of urban schools in cities such as Peshawar (the capital of KP) as well as peri-urban, semi-rural, and extremely rural schools. Urban schools, which makes up 17.1% of the sample, are generally larger (median 8 teachers and 321 students) while rural schools in our sample are smaller (median 4 teachers and 198 students). Schools are led by a head teacher who typically runs a classroom of their own, but in large schools, the head teacher may not have teaching duties.

The intervention began in November 2022 at a time when schools were coming off of multiple years of intermittent school closures due to the COVID-19 pandemic. UNESCO data shows that schools in Pakistan were fully closed for 37 weeks in 2020 and 2021 due to COVID-19, followed by 24 weeks of partial closures, where children may only come on alternate days (UNESCO [2024]). Additional school closures in 2022 were caused by political unrest and wide-scale flooding in KP in 2022.

2.2 Intervention

The Targeted Instruction in Pakistan (TIP) intervention aimed specifically at addressing learning losses from the COVID-19 pandemic and other school closures. The TIP intervention consisted of 40 days (8 weeks) of foundational learning instruction in math, English, and Urdu. Children in classes 1 through 5 were sorted across classrooms into four remedial groups (KG, Class 1, Class 2, or Class 3) based on their actual learning level rather than assigned grade or age. The TIP program was designed and coordinated by researchers at the Center for Education in Pakistan and Harvard University, with support from the Khyber Pakhtunkhwa Elementary and Secondary Education Department and the Ministry of Federal Education & Professional Training.

The TIP program included the following phases:

TIP teacher training (Dec 2021 - Jan 2022): all teachers in treatment schools (Paper, Mandatory, and Optional treatments) were required to attend a 4-day in-person training on the TIP methodology and lessons. This training was delivered in a cascade model by Education Department Master Trainers who had been trained by TIP program staff in August 2021. The training used a blended learning approach to deliver semi scripted lessons and guided activities like a micro teaching session where teachers practiced delivering a lesson and then gave each other feedback. All treated teachers received training on both the tech tool and paper trackers to ensure training was consistent across treatment groups.

Student baseline testing (Mar - May 2022): students in classes 1-5 in sample schools completed a written baseline assessment in math, Urdu, and English. The assessment covered content from the national curriculum for Classes 1, 2 and 3, and was designed to sort students into learning groups based on their actual learning level rather than assigned grade. There were four standardized versions of the assessment to reduce the likelihood of cheating, and the versions were then calibrated for difficulty as described in Appendix B. Students who performed poorly on even the lowest level questions (i.e. content from the Class 1 curriculum) were given an oral assessment in that subject, based on content from the national curriculum for Class KG, to determine if they should be included in Class 1 or Kindergarten level remedial groups. Teachers assisted

program staff and enumerators with baseline test grading, sorting students into learning groups, and administering the oral tests, with strict monitoring to prevent cheating. In special circumstances, teachers were able to "challenge" a student's remedial group assignment, for example, if a student was ill and performed worse than average on the test or if they saw a student cheating. Challenge cases were reviewed and approved by head teachers. During the program, 8.6% of assignments were changed.

Timetabling (Apr-May 2022 and Sep-Oct 2022): Once all students were sorted into remedial groups, mentors met with head teachers to discuss how teachers should be assigned across the groups and to work program lessons into the daily school timetable. Based on the number of classrooms and teachers available, remedial groups were sometimes merged, such as combining KG and remedial level 1 groups together.

Refresher training (Sep - Oct 2022): due to delays at the start of the program, teachers were given a one-day refresher training on lesson delivery, quiz administration, and tech tool usage during TIP classes. New teachers who were not present for the initial teacher training were administered the teacher baseline survey.

Intervention (Nov 2022 - Feb 2023): the TIP program was conducted over 40 days in intervention schools. TIP classes replaced the standard 40 minutes per day allocated to each subject (math, Urdu, and English) such that 2 hours were allotted to TIP classes on a daily basis. The standard curriculum in these subjects was adjusted to accommodate a 40-day pause for TIP classes. For each TIP lesson, teachers received a structured lesson plan that included a mix of lectures and small group activities. Some lessons also included Teaching and Learning Materials, such as flashcards or physical objects like straws used to teach counting and place value. After every 5-7 lessons, teachers conducted a short quiz to assess whether students had absorbed the material. Each quiz was then followed by a revision session where teachers were instructed to review the material with students in whichever way they preferred, with a special focus on children who had struggled on the quiz. Teachers were also encouraged to divide students into peer learning groups where children who had performed well on the quiz were matched with those who did poorly to work through review activities together. The KG level program did not have quizzes due to limited writing ability for students in this group,

which made administration and grading infeasible. During TIP administration, program mentors visited each school in person at least once to conduct at least three classroom observations, focus groups with teachers, collect monitoring data, and ensure compliance with treatment assignments. Mentors also called head teachers every other week to answer questions and boost compliance.

Endline testing and surveys (Feb - Mar 2023): Program staff collected endline tests from all students in the program with the help of a third-party survey firm. The same four versions of the assessment used at baseline were also used at endline, and later calibrated as described in Appendix B.

2.3 TIP Tech Tool

To support administration of the TIP program, the research team worked with software developers at the Centre for Economic Research in Pakistan to build a novel tech tool for teachers that they could utilize on their personal smartphones, tablets, or computers. The tool could be accessed via a browser, or downloaded as a mobile application to function offline in areas with low connectivity. Data could be synced across devices whenever teachers connected to the internet, generally through a mobile data plan. Teachers who lacked a device were encouraged to borrow another teacher's phone or a school tablet/computer where available.

As an alternative to the tech tool in Paper and Optional groups, teachers received a set of paper trackers, which could replace some but not all of the app's features. In addition to access to the tech tool, teachers were added to treatment-wise WhatsApp groups with teachers from their schools and other schools nearby. An average WhatsApp group had 15 teachers from 5 schools. The tech tool developers initially planned to include a chat feature in the mobile application, but this feature was moved to WhatsApp because many teachers had data plans that would allow them to receive messages for free, thereby enabling more consistent access. These groups are considered an extension of the tech tool itself.

In the WhatsApp groups, mentors posted weekly messages like reminders to administer quizzes, links to training videos for material taught that week, government directives on program timing, and general messages of encouragement. Messages also enforced treatment assignments, such as reminding teachers to enter quiz scores in the mobile application if they were in the Mandatory treatment and thus required to do so. Mentor messages were scripted so all groups in the same treatment received the same set of scheduled messages. Teachers could also post to the group with tips from their own experience, or pose questions to the mentor or other teachers.

3 Experimental Design, Methodology, and Balance

3.1 Experimental Design

Schools were randomly selected for inclusion in the study from the universe of public primary schools in Peshawar and Mardan. All schools in the district were eligible to participate provided that at least 30% of their students were from Classes 1-5. This restriction eliminated a small number of schools that were almost entirely Kindergarten students. An initial sample of 1,250 schools was randomly selected from the eligible list. Within the sample, treatment was assigned at the school level, with stratification to ensure that each treatment group had equivalent proportions of female and male schools, schools in Peshawar and Mardan, and small and large schools. A small school is defined as having three or fewer teachers. We use the number of teachers as a key stratifying variable because the TIP program requires at least four teachers in order to run a separate classroom for all the potential learning levels. If a school has three or fewer teachers, they would need to merge learning groups, such as teaching KG and Class 1 together in one classroom. This may impact program effectiveness, thus we ensure each treatment group has even numbers of schools in this category.

The selected sample schools were distributed evenly into five treatment groups: Control, which did not run the TIP program; Paper, where teachers conducted the program but did not have access to the TIP tech tool; Mandatory, where teachers had access to the tech tool and were required to use it; Optional, where teachers had access to the tech tool but were informed they could decide whether to use the tool or paper equivalents;

and Grace Period, where teachers were required to use the tool for the first two weeks and then could decide whether to keep using it or switch to paper tools. Unfortunately, the Grace Period group showed evidence of imbalance (baseline test scores were statistically lower than the other treatment groups) so we eliminated this group from analysis and focused on the control and three other treatments.

Additional details about the three treatment groups are as follows:

Paper Treatment: Teachers in this group received training on the tech tool during the initial teacher training in order to ensure that training was consistent across all treatments. However, when the program began, teachers were informed that they would not have access to the tech tool for the first cycle of the program and should use the paper equivalents instead.

Mandatory Treatment: Teachers in this group were informed that they were required by the education department to use the tech tool regularly, such as to record quiz scores and track student learning. Program mentors reiterated this message with each contact to the school, including in-person monitoring visits and biweekly calls to the head teacher. WhatsApp messages on the group thread also emphasized this requirement. If teachers did not comply, education department officials called Assistant District Education Officers (ASDEOs) were called in to enforce the requirement.

Optional Treatment: Teachers in this group were informed from the start of the program that they could decide individually whether to use the tech tool to record quiz scores, or record scores in a paper tracking tool. During each call to head teachers and in-person school visits, program mentors emphasized that each teacher should make their own decision about tech or paper usage. All teachers were added at the start to the WhatsApp groups, but frequent messages emphasized that they could leave or ignore the group if they wished and make their own decisions about using tech or paper tools. We note, however, that teachers still may have faced pressure from their head teacher or consulted and coordinated with other teachers to use one system or another. We discuss evidence of this pressure later in section 4.

The study was pre-registered with the American Economic Association (AEARCTR-0007469).

3.2 School Sample

As per the submitted pre-analysis plan, our focus for measuring the impact was on fourth grade students. Of the initial 1,000 schools that were randomized into the Control, Paper, Mandatory, and Optional treatment groups, 19 schools were dropped from the sample at various stages of the intervention due to closures for construction or COVID (5 schools), logistical complications during TIP administration (1 school), or inability to implement the program due to teacher refusal/retirement (13 schools). Among the remaining 981 schools, 13 schools did not have grade 4 students, and 52 schools did not have students in grade 4 who completed both baseline and endline testing. This leads to a final sample of 916 schools with 2,920 teachers and 24,730 students. The balance for baseline test score and school-level characteristics across treatment groups are described in appendix tables A1 and A2. Majority of school-level variables such as gender, district, number of teachers, availability of basic facilities (electricity, water, and toilet) and student enrollment balance across treatment groups. Proportion of urban schools did not balance between the control and Paper treatment (p-value <0.1) and proportion of schools with walls did not balance across all treatment groups (p-value <0.1). However, the proportion of schools with walls in each control, Paper, and Mandated are 100% as opposed to 98% in Optional, indicating that the differences are minimal. The school and teacher level summary statistics can be found in appendix tables A3 and A4.

3.3 Data Collection

The following data was collected to measure program impacts:

- **Tech tool usage:** Each teacher was assigned a unique login for the tech tool, and their activity was recorded on the backend. If teachers used the tool offline, their data was recorded whenever they synced their devices to the internet.
- **Student baseline and endline assessments:** The same assessment tool that was utilized to sort children into remedial groups also serves as the student learning assessment for baseline and endline. The written

assessment covers content from the national curriculum in math, Urdu, and English up to Class 3. Program staff and enumerators visited schools individually to help administer the assessment. Four versions of the assessment were administered to prevent cheating, and difficulty levels were calibrated using the procedure described in Appendix B [2].

- **Teacher baseline and endline surveys:** Baseline surveys were collected during TIP teacher training, or during baseline testing or refresher training for teachers who joined after the initial training. Endline surveys were collected during endline student testing visits to schools.

- **Teacher digital competency testing:** After endline testing, program staff visited schools one more time to administer an assessment of digital skills and conduct behavioral games.

- **Implementation monitoring:** Mentors visited schools in person once or twice during the intervention to collect monitoring data via interviews with the head teacher, conduct focus groups with teachers, and observe classrooms.

4 Results

4.1 Treatment Effects

As per the submitted pre-analysis plan, our focus of the impact evaluation was on fourth grade students. The core outcome is the endline average student theta score after IRT calibration, averaged across the three subjects (math, Urdu, and English). Table 1 shows the endline ITT result combining all treatments (paper, optional, and mandated). The regression controls for gender, whether school is located in an urban setting, school size (number of teacher larger than 5), district, and baseline test score. Overall, the TIP program yield a 0.12 SD learning gain with all treatments combined.

[2]The concern about teacher cheating was raised by education officials throughout the piloting and planning process. The research team was especially concerned about teachers' differing ability to cheat by treatment group. The technology treated schools had been added to WhatsApp groups with teachers at other schools, which made it much easier for them to share test questions with teachers at other schools who were scheduled for testing at a later date. To reduce the potential for cheating, multiple versions of the test were deployed, and then difficulty levels were calibrated using an Item Response Theory procedure established in Patel and Sandefur (2020).

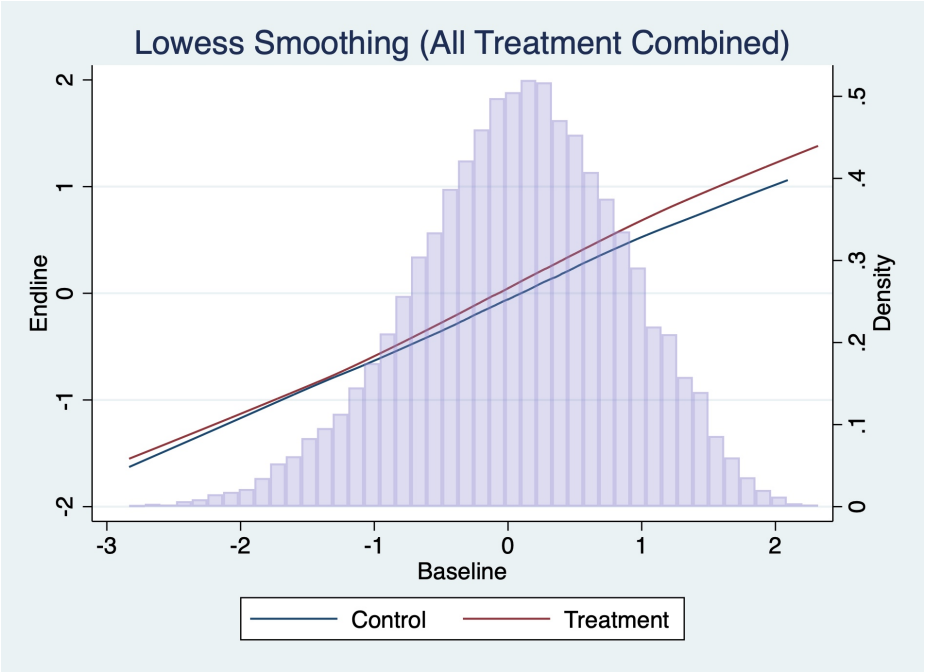
Table 1: Endline Regression: All Treatments Combined

	(1) Endline Theta Score
Treatment Combined	0.117***
	(0.04)
Boy	-0.196***
	(0.03)
Urban	-0.0423
	(0.04)
≥5 Teachers	0.0961***
	(0.03)
PESHAWAR	0.0133
	(0.03)
Baseline Mean	0.604***
	(0.01)
Constant	-0.00418
	(0.04)
Sample	Class 4
Obs	23,438
R2	0.39
Model	OLS
SE Clustered	school

Standard errors in parentheses
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

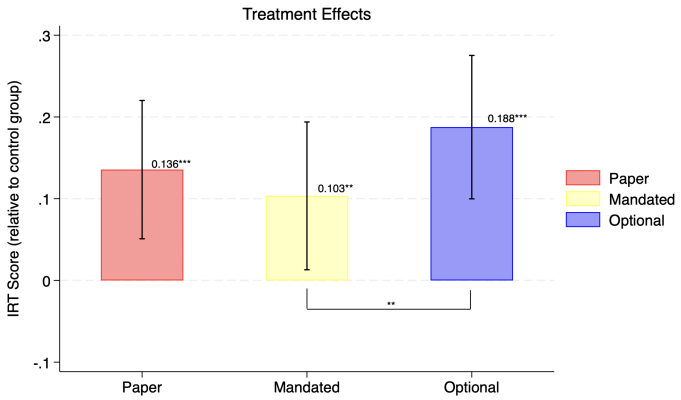
Figure 1 shows the treatment effect across the baseline ability spectrum. The figure shows that the TIP treatment worked across the student's ability spectrum, including the very bottom of the distribution.

Figure 1: Lowess Endline - Baseline Treatments Combined



Across the three treatment groups – Paper, Optional technology, and Mandated technology – Optional technology treatment (0.19 SD) outperformed Paper (0.14 SD) and Mandated (0.08 SD) treatment. Furthermore, the F-statistic suggests that the difference in the impact of the Optional and Mandated treatment arms was statistically significant at the 5% level. The gain for each treatment separately can be found in Figure 2[3]. One explanation for the results is that giving a choice allows teachers to opt into either technology or paper based on their preferences of technology, private information on their comfort in using technology, and constraints, resulting in better performance. In addition, the low magnitude of the result for the Mandated technology arm suggests that a top-down approach to implementing technology may not benefit students.

Figure 2: Treatment Effects for Each Treatment Groups



Note: The sample size drops from 24,730 to 23,438 since some student's test scores at the top and bottom of the distribution do not convert to a common scale when applying the "Rosetta Stone" calibration procedure proposed in Patel and Sandefur (2020). The regression controls for covariates of location (urban/rural), school size (3 teachers or less / 4 or more teachers), gender of school, district (Mardan / Peshawar), and baseline test score for each student. Standard errors are clustered at the school level.

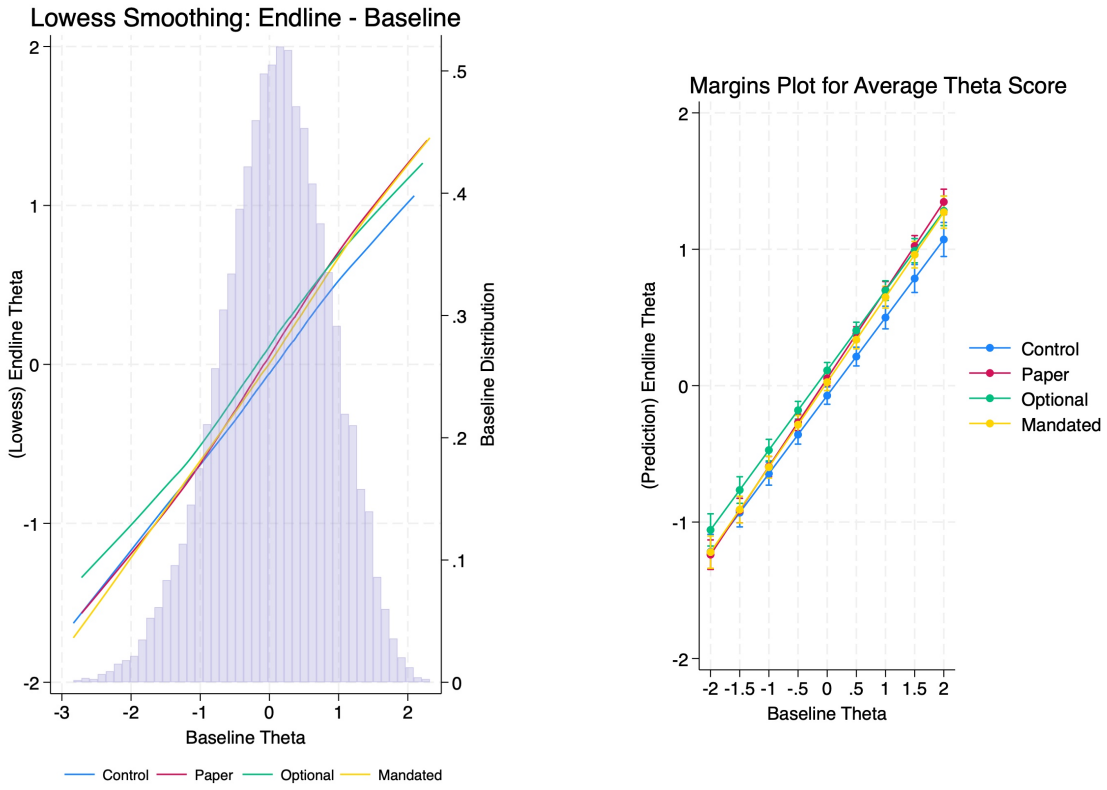
4.2 Heterogeneity in Treatment Effects

To examine the distribution of learning gains across the ability spectrum, we present the Lowess smoothing plot and margins plot for endline theta scores in comparison to the baseline theta scores in Figure 3. The majority of learning gains for the Paper (and Mandated) treatment come from students who were performing

[3]Details of the results can be found in the full regression table presented in Table A5

well at the baseline. On the other hand, the Optional treatment group benefits the student who are performing worse at the baseline with similar magnitude as those in the middle and top tercile. The margins plot analysis shows that the slope of the graph is significantly different between Optional and the other treatments. The design of the tech tool was meant to make the performance of the low performing student more salient to the teachers. This could have potentially contributed to the teacher focusing on these students, leading to their better performance.

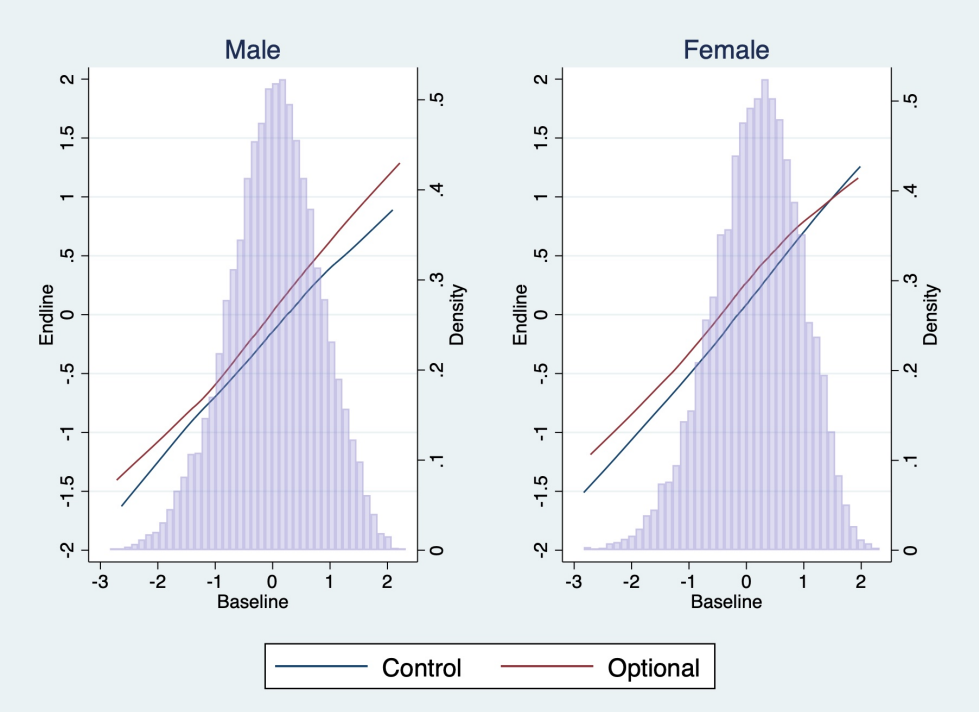
Figure 3: Difference in Treatment Gain by Baseline Ability Level



Note: The regression used to produce the margins plot controls for covariates of location (urban/rural), school size (3 teachers or less / 4 or more teachers), gender of school, district (Mardan / Peshawar), and baseline test score for each student. Standard errors are clustered at the school level.

Figure 4: Lowess Endline - Baseline by Subgroups

(a) by Gender



(b) by School Size

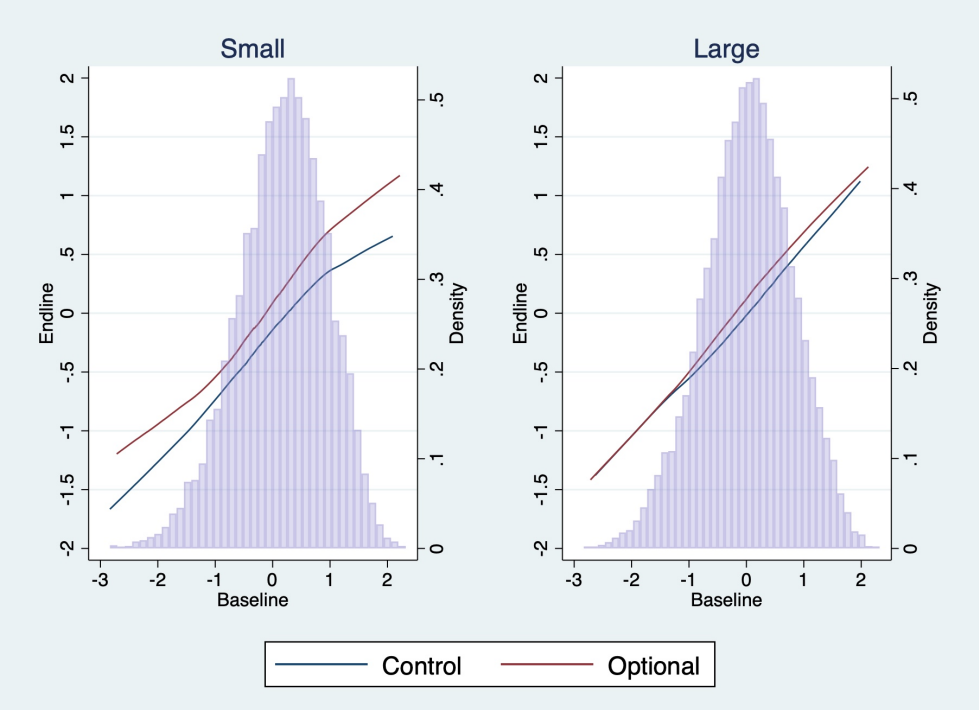
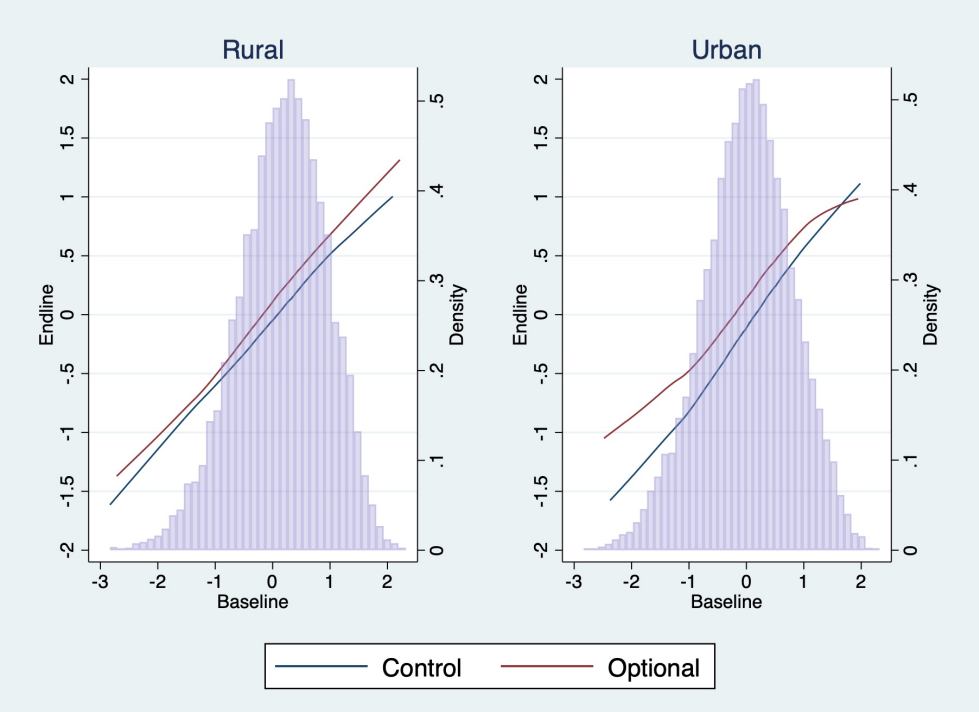
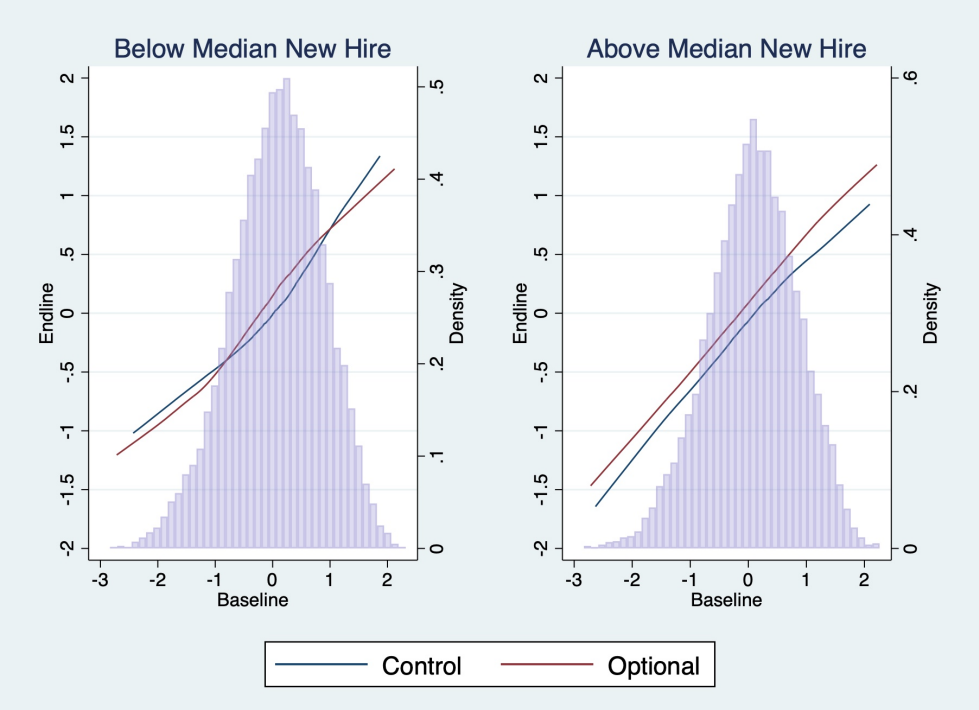


Figure 4: Lowess Endline - Baseline by Subgroups *(continued)*

(c) by Urban/Rural



(d) by ratio of Teachers Hired in New Scheme



From figure 3, we see that optional treatment works well across the baseline ability spectrum. Figures 4a–4d shows the treatment gain by baseline ability separating for gender (male/female), school size (number of teachers above or below five), urban/rural, and by whether the ratio of teachers hired under the new scheme is above or below median (hiring scheme changed in 2014). Notably, the gain from optional treatment is constantly large across the baseline ability spectrum for small schools compared to large schools. Since students in larger schools perform about 0.15 SD better, we see that the optional treatment closed the gap between large and small schools.

4.3 Next Steps

In this section, we discuss the next steps of analyses planned to uncover the particularly large gain from the optional treatment group.

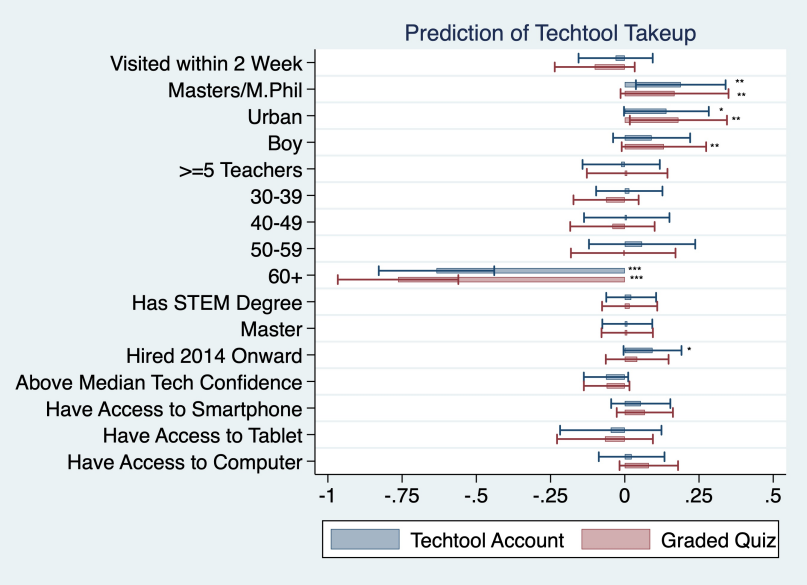
Teacher's Techtool Takeup In the Optional treatment group, the decision to take up the tech tool or not was given to each individual teachers. We determine teacher's techtool takeup in two ways:

(1) The teacher created the techtool account

(2) The teacher graded one or more of their students' quizzes using the techtool (it is possible to grade quizzes using other teacher's techtool account without opening their own)

Below, we regress various teacher characteristics as well as the mentor visit timing on teacher's techtool takeup status and observe how well it predicts the teacher's techtool take up decision. Schools were visited by a mentor whom assisted teachers with using the techtool several times over the course of eight weeks of the program.

Figure 5: Prediction of Techtool Takeup Decision



The results suggest that for both definition for techtool use, whether the mentor has a Master degree, school located in urban settings, and teacher being older than the age of 60 were predictive of tech tool take up. For the quiz definition, male teachers were more likely to takeup the techtool. The R^2 of the regression result was 0.07 and 0.09 for each definitions, and overall, the results are suggestive that the take up of techtool is only weakly predictive by teacher characteristics.

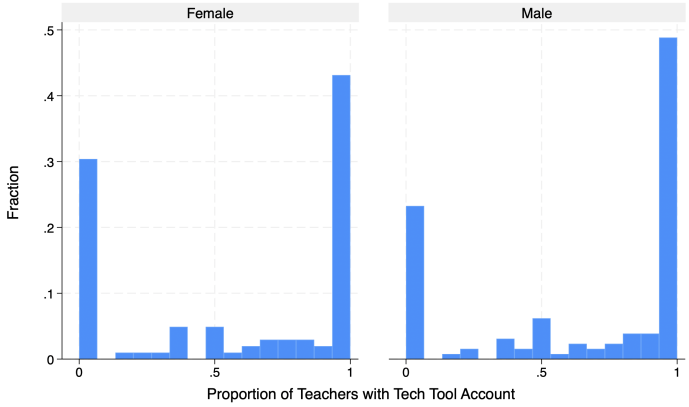
Organizational vs Individual Takeup In addition to the weak predictability of teacher's techtool takeup decision, we observe that the decision is potentially made collectively at the school level in many cases. Below in table 2, we show whether all teachers, any teacher, or none of the teachers in a school took up the techtool based on the two definitions explained in the paragraph above. The table presents the breakdown by school gender as well. Based on the techtool account definition, 27.7% of schools had no teachers opening a techtool account, 28.6% of schools with some teachers opening an account, and 43.7% of schools with all teachers opening an account. The breakdown was polarized for the quiz grading definition where only 2.6% of schools had some teacher grade quiz while 35.1% of schools had no teachers grade quiz and 62.3% of schools had all teachers grade quiz. Male schools were more likely to have a higher takeup of the techtool than female schools in both definitions.

Table 2: School-level Techtool Take-up Status

*showing column percentage points	All		Male		Female	
	Techtool account	Graded Quiz	Techtool account	Graded Quiz	Techtool account	Graded Quiz
No Teachers with techtool account / graded quiz	27.7	35.1	24.0	27.9	32.4	44.1
Any Teachers with techtool account / graded quiz	28.6	2.6	31.0	2.3	25.5	2.9
All Teachers with techtool account / graded quiz	43.7	62.3	45.0	69.8	42.2	52.9

In addition, figure 6 shows shows the full distribution of the proportion of teachers who took up the techtool in each school by gender. For the schools with any teachers with tech tool accounts, the median percentage of teachers with accounts was 56.3%.

Figure 6: Distribution of School-level Tech Tool Take-up Ratio

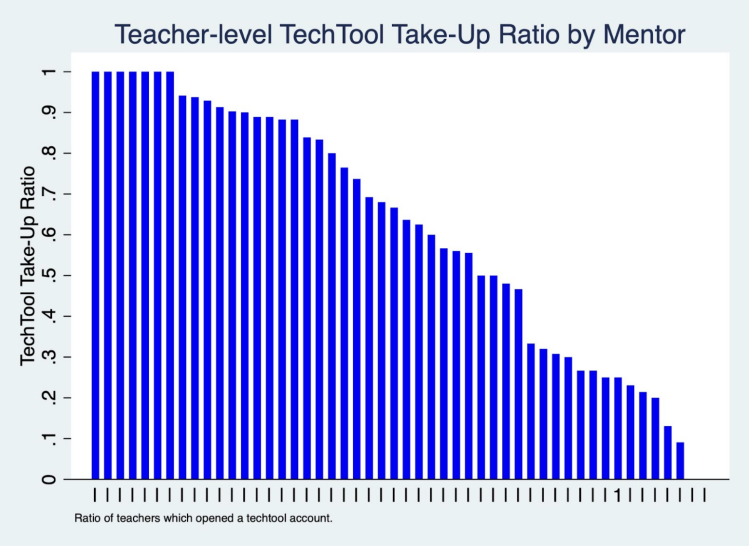


Measuring Techtool Effectiveness Since the choice of taking up techtool within the optional treatment group is confounded with various factors such as teacher's gender and age, simply comparing the student performance among teachers who used the techtool with those who didn't would yield a biased estimate. To identify the causal effect of techtool usage on student achievement, we consider applying a "judge leniency IV" approach seen in various empirical work recently (Dobbie, Goldin and Yang [2018], Doyle Jr et al. [2015]).

Each treatment school received a visit from a techtool mentor during the 40-day TIP program. With only 50 mentors assigned to several hundred schools, visit timing was largely quasi-random and plausibly exogenous to both school and teacher level characteristics. In addition, although only weakly predictive, receiving a mentor earlier in the program was associated with a higher takeup of the techtool, and more importantly,

techtool take-up varied substantially across mentors (Figure 7). Hence, the mentor and its visited timing can be considered as an instrumental variable. Estimating a two-stage least square model taking the mentor variable as an instrument would allow us to obtain an unbiased estimate of the effect of techtool adoption on student test scores.

Figure 7: Takeup Ratio by Mentor



Prediction of the Optional Treatment Group's self-selection and empowering effect In addition, we consider that the optional treatment has a strength over the other two treatments (paper and mandated) through two channels: allowing teachers to decide whether to use the techtool based on their own benefits (***self-selection effect***) as well as providing teachers with a sense of ownership in delivering the remedial program (***empowering effect***). We aim to identify the causal effect of self-selection effect as well as the empowering effect on the student's test score.

We leveraging a machine learning approach such as random forest algorithms for the identification of the effects. We first train a model that predicts techtool takeup using the optional treatment group teacher data, then fit the model to the mandated treatment group teachers using their teacher and school level information.

In this way, we would be able to build counterfactual groups of "teachers in the mandated techtool treatment group who would have taken up the techtool had they been given the choice" and "teachers in the mandated techtool treatment group who wouldn't have taken up techtool if not mandated" (similar classification can be done at the school-level as well) within the mandated techtool treatment group. Here, the comparison of the counterfactual group of "teachers in the mandated techtool treatment group who would have taken up the techtool had they been given the choice" with the entire mandated techtool treatment would provide a measurement of the self-selection effect. In addition, the comparison of the counterfactual group of "teachers in the mandated techtool treatment group who would have taken up the techtool had they been given the choice" with the optional treatment group would show the empowering effect of giving options to teachers.

5 Conclusion

Prior research has found varying returns to education technology (EdTech). These studies almost universally mandate the use of technology when it is introduced. In this study, we find that the approach to introducing EdTech has a large impact on learning outcomes. We explore these dynamics in the context of a large-scale foundational learning program in public primary schools in Pakistan. The novel curricular program incorporated both targeted instruction and structured pedagogy, and some teachers were also randomly assigned access to a low-cost, smartphone application (tech tool) designed to reduce the administrative burden of classroom management. Test results show that the 8-week program raised student learning overall by 0.12 SD across math, English, and Urdu. Learning gains, especially among low-performing students, were highest when the tech tool was offered to teachers but framed as optional. In contrast, mandating the use of the tech tool (0.10 SD) had significantly lower student learning relative to the Optional tech (0.19 SD) version of the program. The differences are driven by schools with female teachers and students (schools are gender segregated), where the Mandated treatment arm performed even worse (0.02 SD) relative to the Optional arm (0.17 SD). These findings suggest substantial heterogeneity in teacher returns to EdTech, and support a self-selection approach for introducing new technology.

References

Akyeampong, K., T. Andrabi, A. Banerjee, R. Banerji, S. Dynarski, R. Glennerster, S. Grantham-McGregor, K. Muralidharan, B. Piper, S. Ruto, J. Saavedra, S. Schmelkes, and H. Yoshikawa. 2023. *Cost-Effective Approaches to Improve Global Learning: What Does Recent Evidence Tell Us Are "Smart Buys" for Improving Learning in Low- and Middle-Income Countries?* London, Washington D.C., New York:FCDO, the World Bank, UNICEF, and USAID.

Banerjee, A., R. Banerji, J. Berry, E. Duflo, H. Kannan, S. Mukherji, and M. Walton. 2016. "Mainstreaming an effective intervention: Evidence from randomized evaluations of "Teaching at the Right Level" in India." National Bureau of Economic Research w22746.

Banerjee, A. V., R. Banerji, E. Duflo, R. Glennerster, and S. Khemani. 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. V., S. Cole, E. Duflo, and L. Linden. 2007. "Remedying education: Evidence from two randomized experiments in India." *The Quarterly Journal of Economics*, 122(3): 1235–1264.

Barrera-Osorio, F., and L. L. Linden. 2009. "The use and misuse of computers in education: Evidence from a randomized experiment in Colombia." World Bank Policy Research Working Paper 4836.

Beg, S., W. Halim, A. M. Lucas, and U. Saif. 2022. "Engaging teachers with technology increased achievement, bypassing teachers did not." *American Economic Journal: Economic Policy*, 14(2): 61–90.

Beuermann, D. W., J. Cristia, S. Cueto, O. Malamud, and Y. Cruz-Aguayo. 2015. "One laptop per child at home: Short-term impacts from a randomized experiment in Peru." *American Economic Journal: Applied Economics*, 7(2): 53–80.

Brunette, T., B. Piper, R. Jordan, S. King, and R. Nabacwa. 2019. "The impact of mother tongue reading instruction in twelve Ugandan languages and the role of language complexity, socioeconomic factors, and program implementation." *Comparative Education Review*, 63(4): 591–612.

Cilliers, J., B. Fleisch, C. Prinsloo, and S. Taylor. 2020. "How to improve teaching practice? An experimental comparison of centralized training and in-classroom coaching." *Journal of Human Resources*, 55(3): 926–962.

Cristia, J., P. Ibarrarán, S. Cueto, A. Santiago, and E. Severín. 2017. "Technology and child development: Evidence from the one laptop per child program." *American Economic Journal: Applied Economics*, 9(3): 295–320.

Dobbie, Will, Jacob Goldin, and Crystal S Yang. 2018. "The effects of pre-trial detention on conviction, future crime, and employment: Evidence from randomly assigned judges." *American Economic Review*, 108(2): 201–240.

Doyle Jr, Joseph J, John A Graves, Jonathan Gruber, and Samuel A Kleiner. 2015. "Measuring returns to hospital care: Evidence from ambulance referral patterns." *Journal of Political Economy*, 123(1): 170–214.

Duflo, A., J. Kiessel, and A. M. Lucas. 2024. "Experimental Evidence on Four Policies to Increase Learning at Scale." *The Economic Journal*, 134(661): 1985–2008.

Duflo, E., P. Dupas, and M. Kremer. 2011. "Peer effects, teacher incentives, and the impact of tracking: Evidence from a randomized evaluation in Kenya." *American Economic Review*, 101(5): 1739–1774.

Gray-Lobe, G., A. Keats, M. Kremer, I. Mbiti, and O. W. Ozier. 2022. "Can education be standardized? Evidence from Kenya." University of Chicago, Becker Friedman Institute for Economics Working Paper 2022-68.

Muralidharan, K., A. Singh, and A. J. Ganimian. 2019. "Disrupting education? Experimental evidence on technology-aided instruction in India." *American Economic Review*, 109(4): 1426–1460.

Piper, B., J. Destefano, E. M. Kinyanjui, and S. Ong'ele. 2018. "Scaling up successfully: Lessons from Kenya's Tusome national literacy program." *Journal of Educational Change*, 19: 293–321.

UNESCO. 2024. "COVID-19 Educational Response." *https://covid19.uis.unesco.org/global-monitoring-school-closures-covid19/country-dashboard/*, Retrieved from https://covid19.uis.unesco.org/

global-monitoring-school-closures-covid19/country-dashboard/. Accessed Dec 24, 2024.

Wang, L. C., M. Vlassopoulos, A. Islam, and H. Hassan. 2024. "Delivering remote learning using a low-tech solution: Evidence from a randomized controlled trial in Bangladesh." *Journal of Political Economy Microeconomics*, 2(3): 562–601.

A Tables

Table A1: Baseline Test Score Balance

	(1) Baseline Test Score
Paper	0.0112 (0.21)
Optional	-0.0328 (-0.65)
Mandated	-0.0597 (-1.18)
Constant	0.113^{***} (3.09)
Observations	23438

$^{*}\ p < 0.1$, $^{**}\ p < 0.05$, $^{***}\ p < 0.01$

Table A2: Balance Table

	(1) Control		(2) Paper		(3) Optional Tech		(4) Mandated Tech		F-test for balance across all groups		(1)-(2)		(1)-(3)		(1)-(4)		(2)-(3)		(2)-(4)		(3)-(4)	
Variable	N	Mean/(SE)	N	Mean/(SE)	N	Mean/(SE)	N	Mean/(SE)	N	F-stat/P-value	N	Mean difference	N	Mean difference	N	Mean difference	Pairwise t-test N	Mean difference	N	Mean difference	N	Mean difference
Prop. Male Schools	230	0.55 (0.03)	230	0.57 (0.03)	231	0.56 (0.03)	225	0.56 (0.03)	916	0.11 0.95	460	-0.03	461	-0.01	455	-0.02	461	0.02	455	0.01	456	-0.01
Prop. Peshawar Schools	230	0.43 (0.03)	230	0.41 (0.03)	231	0.42 (0.03)	225	0.40 (0.03)	916	0.11 0.95	460	0.02	461	0.01	455	0.03	461	-0.00	455	0.01	456	0.01
Prop. Urban Schools	230	0.13 (0.02)	230	0.21 (0.03)	231	0.17 (0.02)	225	0.17 (0.03)	916	1.67 0.17	460	-0.08*	461	-0.03	455	-0.03	461	0.04	455	0.04	456	-0.00
Prop. Electricty Availability	230	0.98 (0.01)	230	0.96 (0.01)	231	0.98 (0.01)	225	0.97 (0.01)	916	1.04 0.37	460	0.02	461	-0.00	455	0.01	461	-0.02	455	-0.01	456	0.01
Prop. Water Availability	230	1.00 (0.00)	230	0.99 (0.01)	231	0.99 (0.01)	225	1.00 (0.00)	916	1.79 0.15	460	0.01	461	0.01	.n	.n	461	-0.00	455	-0.01	456	-0.01
Prop. Toilet Availability	230	1.00 (0.00)	230	0.99 (0.01)	231	1.00 (0.00)	225	1.00 (0.00)	916	1.27 0.28	460	0.01	461	-0.00	455	0.00	461	-0.01	455	-0.01	456	0.00
Prop. School with Walls	230	1.00 (0.00)	230	1.00 (0.00)	231	0.98 (0.01)	225	1.00 (0.00)	916	2.86* 0.04	.n	.n	461	0.02*	455	0.00	461	0.02*	455	0.00	456	-0.01
Mean Number of Teachers	230	5.57 (0.24)	230	5.99 (0.27)	231	5.99 (0.25)	225	5.75 (0.26)	916	0.63 0.59	460	-0.42	461	-0.42	455	-0.18	461	-0.00	455	0.24	456	0.24
Mean School Enrollment	230	254.62 (11.15)	230	271.84 (13.31)	231	267.04 (12.31)	225	259.31 (12.71)	916	0.39 0.76	460	-17.23	461	-12.42	455	-4.69	461	4.80	455	12.54	456	7.73
Mean School Enrollment (Grades 1 - 5)	230	186.09 (8.82)	230	200.53 (10.32)	231	196.21 (9.51)	225	189.52 (9.87)	916	0.46 0.71	460	-14.44	461	-10.12	455	-3.43	461	4.32	455	11.01	456	6.69

Table A3: Descriptive Statistics: School

Variable	Standard Error in Brackets
School-Level Variables (N = 916)	
Prop. Male Schools	0.56 (0.50)
Prop. Peshawar Schools	0.42 (0.49)
Prop. Urban Schools	0.17 (0.38)
Prop. Electricity Availability	0.97 (0.16)
Prop. Water Availability	0.99 (0.07)
Prop. Toilet Availability	0.99 (0.07)
Prop. School with Walls	0.99 (0.07)
Median Number of Teachers	5.0 (3.87)
Median School Enrollment	216.5 (187.33)
Median School Enrollment (Grades 1–5)	152.5 (145.83)

Table A4: Descriptive Statistics: Teacher

Variable	Standard Error in Brackets
Teacher-Level Variables (N = 2,881)	
Prop. Male Teachers	0.59 (0.49)
Median Teacher Age	37.0 (8.68)
Prop. Married Teachers	0.87 (0.34)
Prop. Teacher New Hire (>2013)	0.55 (0.50)
Prop. Teachers with STEM degree	0.26 (0.44)
Prop. Teachers below 40 years old	0.62 (0.49)
Prop. Pre-University Degree	0.09 (0.28)
Prop. Bachelor's Degree	0.29 (0.45)
Prop. Master's Degree	0.63 (0.48)

Note: Missing teacher roster survey data for 39 teachers.

Table A5: Endline Regression: Treatment Effects

	(1) Endline Theta Score	(2) Endline Theta Score	(3) Endline Theta Score	(4) Endline Theta Score
Paper	0.120**	0.113***	0.142***	0.136***
	(0.05)	(0.04)	(0.05)	(0.04)
Optional	0.142***	0.163***	0.186***	0.188***
	(0.05)	(0.05)	(0.05)	(0.04)
Mandated	0.0426	0.0795*	0.0764	0.103**
	(0.05)	(0.05)	(0.05)	(0.05)
Baseline Theta Score		0.618***		0.608***
		(0.01)		(0.01)
Urban			0.101**	-0.0263
			(0.04)	(0.04)
Large School			0.105***	0.0550
			(0.04)	(0.04)
PESHAWAR			0.0673**	0.00757
			(0.03)	(0.03)
Male			-0.260***	-0.189***
			(0.04)	(0.03)
Constant	0.0102	-0.0598*	0.0198	0.00535
	(0.04)	(0.03)	(0.06)	(0.05)
Sample	Class 4	Class 4	Class 4	Class 4
Obs	23,438	23,438	23,438	23,438
R2	0.01	0.38	0.04	0.39
Model	OLS	OLS	OLS	OLS
F-Stat(P=O)	0.19	1.41	0.89	1.73
F-Stat(P=M)	2.23	0.62	1.88	0.59
F-Stat(O=M)	3.90**	3.50*	5.41**	3.87**
SE Clustered	school	school	school	school

Standard errors in parentheses
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table A6: Endline Regression: Treatment Effects by Gender

	Female				Male			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Endline Theta Score	Endline Theta Score	Endline Theta Score	Endline Theta Score	Endline Theta Score	Endline Theta Score	Endline Theta Score	Endline Theta Score
Paper	0.109	0.0658	0.0631	0.0695	0.131**	0.147***	0.171***	0.175***
	(0.08)	(0.06)	(0.08)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)
Optional	0.153*	0.157**	0.168**	0.167**	0.144**	0.173***	0.186***	0.200***
	(0.08)	(0.06)	(0.08)	(0.07)	(0.06)	(0.06)	(0.06)	(0.06)
Mandated	0.00358	0.00707	0.0446	0.0238	0.0721	0.128**	0.113*	0.161***
	(0.09)	(0.07)	(0.08)	(0.07)	(0.06)	(0.06)	(0.06)	(0.06)
Baseline Theta Score		0.588***		0.573***		0.623***		0.623***
		(0.02)		(0.02)		(0.02)		(0.02)
Urban			0.275***	0.0189			-0.0137	-0.0456
			(0.06)	(0.05)			(0.05)	(0.05)
Large School			0.0883	0.0313			0.108**	0.0708
			(0.07)	(0.06)			(0.05)	(0.04)
PESHAWAR			0.194***	0.0835*			-0.00197	-0.0367
			(0.06)	(0.05)			(0.04)	(0.04)
Constant	0.185***	0.0928**	-0.0390	0.0241	-0.100**	-0.154***	-0.202***	-0.199***
	(0.06)	(0.05)	(0.08)	(0.07)	(0.04)	(0.04)	(0.06)	(0.06)
Sample	Class 4	Class 4	Class 4	Class 4	Class 4	Class 4	Class 4	Class 4
Obs	8,831	8,831	8,831	8,831	14,607	14,607	14,607	14,607
R2	0.01	0.36	0.06	0.36	0.01	0.38	0.01	0.38
Model	OLS	OLS	OLS	OLS	OLS	OLS	OLS	OLS
F-Stat(P=O)	0.31	2.29	2.02	2.53	0.04	0.24	0.07	0.24
F-Stat(P=M)	1.49	0.73	0.05	0.44	1.01	0.13	0.98	0.08
F-Stat(O=M)	3.29*	4.28**	2.34	3.89**	1.47	0.69	1.65	0.55
SE Clustered	school	school	school	school	school	school	school	school

Standard errors in parentheses
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table A7: Tech Tool Take-up Channel

	(1)
	Teacher Opened Tech Tool Account (0/1)
Male	0.0603
	(0.06)
Bachelor's	0.0480
	(0.07)
Master's	0.0466
	(0.07)
Teacher Has STEM Degree	0.0251
	(0.04)
Urban	0.119*
	(0.07)
Teacher First Hired >2013	0.117***
	(0.04)
PESHAWAR	0.0999*
	(0.06)
30-40	0.0275
	(0.05)
40-50	0.0503
	(0.07)
50 or above	0.122
	(0.08)
Constant	0.384***
	(0.10)
Obs	858
R2	0.04
SE Clustered	School
Method	OLS
Sample	Optional Treatment Group Teachers

Standard errors in parentheses
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: Regression results for whether the teacher created a tech account or not for the 858 teachers in the Optional treatment group (19 teachers not part of the regression due to some missing data in teacher survey).

B Calibration

As a requirement of the KP education department, four different versions of the baseline test were created and schools were randomly given one of the four versions. The tests had similar question form and structure (based on Student Learning Outcomes of grades 1-3) and had the same number of items. The four tests were re-randomized at the school level at the endline. While the tests had similar form and structure, there were no overlapping items. In order to create a common scale, the "Rosetta Stone" methodology (Patel and Sandefur, 2020) was utilized. 3 fresh composite tests were created out of the four original tests (details below) and administered one year later to students in grades 3-5 from a representative sample of 4,137 students from 50 schools (not in the study sample) in the same two districts.

Each original test has 30 question items per subject. The calibration test has 15 items from each original test per subject (a total of 30 items) in the main part. Specifically, we create 3 versions of the calibration test namely tests A, B, and C where each is a combination of test version 1 + 3, version 2 + 3, and version 4 + 3 accordingly. Additionally, the test has an appendix of 5 items each from the target test and the source test (not used in the analyses, collected just in case IRT not converging only with the main items). All items are graded using 2PL IRT. Different item selection designs were evaluated by measuring the MSEs of student ability (theta) scores produced using each design against the theta scores produced using the full baseline test. Among the designs evaluated, the optimal item selection design was to have all odd-numbered items from the target test (version 1, 2, 4), all even-numbered items from the source test (version 3) in the main part (30 test items) since it yielded one of the least MSEs as well as preserving the original test structure.

For the appendix items, from the source test (version 3) we selected the 5 most discriminating items from the items not used in the main item selection (selecting the 5 most discriminating items in the odd-numbered items of the source test). Symmetrically, from the target test (versions 1, 2, 4) we selected the 5 most discriminating items from the items not used in the main item selection (selecting the 5 most discriminating items in the even-numbered items of the target test). The math exam was constructed following this procedure exactly, while for English and Urdu, we had a small change as there were a few grouped questions that had

to come together. The calibration of the original baseline test score followed the procedure described below (calibration for endline test score followed the exact same procedure). One note for the calibration procedure is that some student's scores at the top and bottom of the distribution did not get converted if an appropriate counterpart did not exist in the calibration test data.

(i) **Obtain fixed item estimates** ... Obtain fixed question item parameters (difficulty and discrimination parameters) by applying IRT on each source baseline test (version 1, 2, 3, and 4) data. Here, all baseline test data including those that did not match with endline test data were used for estimation (there was almost no difference between using data that matched with endline test data and including all baseline data). Through this procedure, we obtain difficulty and discrimination parameters for every test item in source tests version 1 through 4.

(ii) **Estimate theta scores of the calibration test data** ... By restricting the item parameters with the values obtained in step 1, estimate the theta scores for the calibration test data in both directions. For instance, for test version A (combining original test version 1 + 3), we first restrict the difficulty and discrimination parameters of items brought from version 1 and estimate the theta scores (items brought from version 3 are freely estimated) which provides theta scores in version 1 scale. Next, we restrict the difficulty and discrimination parameters of items brought from version 3 and estimate the theta scores, which provide the theta scores in the version 3 scale. As a result, we will obtain two theta scores in the scale of each source test for the calibration test data.

(iii) **Convert the baseline theta scores** ... Using the 2 theta scores obtained for each calibration test A, B, and C, we run a local linear regression on each corresponding theta score and apply the conversion scale to items in the baseline test. As a result, all baseline test theta scores will be in the original test version 3 scale, making it comparable across different test versions.