NBER WORKING PAPER SERIES

# ADAPTING FOR SCALE:
# EXPERIMENTAL EVIDENCE ON TECHNOLOGY-AIDED INSTRUCTION IN INDIA

Karthik Muralidharan
Abhijeet Singh

Working Paper 34205
http://www.nber.org/papers/w34205

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
September 2025

This project was executed in partnership with the Government of Rajasthan and Educational Initiatives. We thank Ramamurthy Sripada, Ankit Agarwal, Nawar Al-Ebadi, Petter Berg, Urmi Bhattacharya, Aditi Gautam, Aditya Jahagirdar, Jalnidh Kaur, Melitine Malezieux and Archana Prabhakar for excellent field administration and research assistance. Sridhar Rajagopalan, Pranav Kothari, Kashi Nath Jha, and Raghav Rohatgi at Educational Initiatives provided invaluable operational support for the evaluation. We are grateful for comments from Julie Cullen, Andy de Barros, Ajinkya Keskar, Ofer Malamud, Lant Pritchett, Imran Rasul, Mauricio Romero, and seminar participants at various institutions. The fieldwork in this paper was funded by the RISE Programme, which was funded by the UK's Foreign, Commonwealth and Development Office (FCDO), Australia's Department of Foreign Affairs and Trade (DFAT) and the Gates Foundation. Educational Initiatives received funding from the Global Innovation Fund for developing the in-school model we evaluate in this study. This study was registered on the AEA RCT Registry (AEARCTR-0002546). All errors are our own. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Adapting for scale: Experimental Evidence on Technology-aided Instruction in India
Karthik Muralidharan and Abhijeet Singh
NBER Working Paper No. 34205
September 2025
JEL No. C93, I21, O15

## ABSTRACT

Many interventions that "work" in small-scale trials often fail at scale, highlighting the centrality of effective scaling for realizing the promise of evidence-based policy. We study the scaling of a personalized adaptive learning (PAL) software that was highly effective in a small-scale trial. We adapt the PAL implementation for scalability by integrating it into public school schedules, and experimentally evaluate this adaptation in a more representative sample over 20 times larger than the original study. After 18 months, treated students scored 0.22  higher in Mathematics and 0.20 higher in Hindi, a 50–66% productivity increase over the control group. Learning gains were proportional to student time on the platform, providing a simple, low-cost metric for monitoring implementation quality in future scale-ups. The adaptation was cost effective, and its key design features make it widely scalable across diverse settings.

Karthik Muralidharan
University of California, San Diego
Department of Economics
and NBER
kamurali@ucsd.edu

Abhijeet Singh
Stockholm School of Economics
Department of Economics
abhijeetsingh1@gmail.com

A randomized controlled trials registry entry is available at
https://www.socialscienceregistry.org/trials/2546

# 1 Introduction

The rapid growth of field experiments in economics has been largely driven by the desire to improve social welfare by identifying interventions that 'work' and scaling them up. Yet, globally, many interventions shown to be effective in small-scale efficacy trials have been found to be ineffective at larger scales, even within the same geographical and institutional context.[1] Thus, delivering on the promise of 'evidence-based' policy requires as much (or more) attention to adapting successful interventions for scalability and testing them at larger scales, as finding evidence of interventions that 'work' under highly-controlled implementation at small scales.[2]

In this paper, we study scaling in the context of a computer-based personalized adaptive learning (PAL) software (called *Mindspark*), which had proven highly effective in a proof-of-concept trial (Muralidharan, Singh, and Ganimian (2019)). This study found ITT effects of 0.23 and 0.37 standard deviations ($\sigma$) in just 4.5 months of program exposure, making it one of the most effective education interventions evaluated to date. However, while this was a highly-promising efficacy trial, there were several reasons for why these positive effects may not be sustained at scale.

First, it was conducted at a modest scale of 314 treatment students. Second, while treatment was randomized at the student level, the study universe comprised a non-representative self-selected sample of students who expressed interest in the program. Third, the program was delivered in out-of-school learning centers dedicated to delivering Mindspark, which ensured high-quality implementation. Fourth, it *supplemented* regular school instruction rather than substitute it and hence did not disrupt the school day. Finally, while the program was cost-effective compared to business-as-usual instruction, it was expensive in absolute terms, making the model that was studied difficult to scale.

[1]In the US, Bhargava and Manoli (2015) find substantial increases in EITC claim rates from mailing information and reminders to EITC-eligible individuals; but larger RCTs with more representative samples found no effect (Linos et al. (2022)). More generally, DellaVigna and Linos (2022) document lower efficacy at scale in 123 RCTs carried out by government nudge units, compared to academic research. In low- and middle-income settings Mitchell et al. (2023) report that a large-scale migration loan program in Bangladesh failed to replicate pilot results (Bryan et al., 2014); Kerwin and Thornton (2021) find that a reduced-cost version of a highly effective mother tongue literacy program had sharply negative effects; and Banerjee et al. (2017) show that the effectiveness of "Teaching at the Right Level" programs were sensitive to whether they were implemented by community volunteers or public school teachers.

[2]Further, the combination of publication bias towards significant results (see DellaVigna and Linos (2022); Camerer et al. (2016); Andrews and Kasy (2019)) and increased donor funding for programs found to be 'effective' create incentives for both researchers and implementing organizations to invest in high-quality implementation and evaluation of interventions at small scales, which may not be sustainable at larger scales. As noted by List (2024), "The result is that we [the research community] are essentially performing efficacy tests on steroids without telling outsiders." On a related note, Al-Ubaydli, Lee, List, Mackevicius, and Suskind (2021), observe that: "...the chain connecting initial research discovery to the ultimate policy enacted has as its most susceptible link an understanding of the science of scaling."

The most promising way to deliver the benefits of PAL at scale and reach underprivileged children is through the public schooling system. We, therefore, partnered with Mindspark's developers, and the Government of the Indian state of Rajasthan (GoRJ) to design an implementation protocol for public schools. This involved setting up computer labs in treated schools, and modifying the timetable to *replace* 25-50% of weekly math and language (Hindi) instructional time with a "computer lab" period, where students studied the same subjects on the Mindspark platform. Instruction in the labs was personalized to each student's learning level, unlike classroom instruction that typically focused on the textbook. While the platform delivered instruction, the regular teacher was expected to accompany students to the lab to ensure adherence, answer questions, and manage the class. The intervention provided a modestly-paid, locally-hired lab-in-charge (LIC) who was responsible for the maintenance and functionality of the computer equipment. Finally, when the number of students exceeded the available computers, students were paired on a computer. Taken together, even though the software was the same, this scalable model was a *substantially different* intervention from the one evaluated in the efficacy trial.

We evaluate this model over 18 months using a cluster-randomized trial that treated 40 schools and $\sim$6,500 students (with 40 control schools) across both rural and urban areas in 4 districts of Rajasthan. We study program impacts using independent measurement of learning outcomes, designed to span the full range of student learning levels. We also collected data on software usage (in treatment schools), measured classroom practices using direct observations, and interviewed students and teachers.

We present five main results. First, the dynamic computerized diagnostic test allows us to characterize both mean learning levels in a grade, and also the *distribution* of learning levels within a grade. We find that average math skills progress at roughly *half* the pace of the curriculum and textbooks, with the average 8th grade student performing at a 4th grade level (Figure 1). We also document striking variation within grades with students in 8th grade ranging from 2nd to 8th grade skill levels. This is a pattern that is likely to replicate in several other education systems as well.[3] This fact also illuminates the enormous challenge teachers face in accommodating such wide variation with a common instruction protocol, and highlights the potential for personalized instruction using PAL software to improve the productivity of instructional time.

Second, after 18 months, the intervention improved learning outcomes by $0.22\sigma$ in math and $0.2\sigma$ in Hindi relative to control schools. These additional gains are around half of the control group's total learning gains in math and two-thirds in Hindi over this same period. Since the program was delivered during school hours, these effects can be interpreted

[3]See the Appendix to Muralidharan et al. (2019) for details.

as a 50-66% improvement in the productivity of schooling time. These treatment effects rank around the 90th percentile of effect sizes found across all education RCTs with large samples (N>5000) in LMICs (Evans and Yuan, 2022).

Third, learning gains were broad-based, with little evidence of heterogeneity. We find no differential effects by gender, socioeconomic status or baseline scores, and find similar effects across primary and middle school grades. However, academically weaker students progressed more slowly in the control group, likely because they were further away from grade-level instruction. Thus, while *absolute* treatment effects are similar across students, gains *relative* to the counterfactual are higher for weaker students. We also find that the average gap between student learning and curricular standards narrows in the treated group over time. The pattern of gains is also consistent with personalized instruction: treated students with higher baseline scores improved more on "difficult" questions, while those with lower scores progressed more on "easy" questions.

Fourth, we find no evidence of improvement on school exams, on which treatment effects are statistically insignificant, with small negative point estimates. This likely reflects the fact that Mindspark instruction was targeted at students' actual learning level, which was often several years below grade standards. Even meaningful increases in learning from this low base are unlikely to be captured by grade-level school exams. Further, the 25-50% reduction in class-time available for grade-level instruction, could have also contributed to the lack of gains on grade-level school tests.

Fifth, based on direct classroom observations, we find no evidence of disruptions to regular classroom practices due to time lost to computer-aided instruction. Two years into the program, teachers in treated schools had acclimatized to it, and report adapting to reduced regular instruction time by covering material faster and reducing revision time. Importantly for long-term acceptance and scalability, both teachers and students in treated schools found the computer-aided instruction very useful, with little opposition to continuing the program.

While the treatment effects were lower than in the efficacy trial, this scalable in-school model was twice as cost effective (in learning-gains-per-dollar) than the out-of-school model studied in the efficacy trial over a similar time period. Sources of efficiency gains include (a) improved utilization rates of the computers, (b) and lower facility and staff costs (since existing school resources were used).

Following the end of the main study after 18 months, the intervention was iterated to further reduce costs and evaluated for another year. Since the main *flow* cost of the program was the LIC, the iterated design reduced LIC coverage from one per school to a single LIC shared between 3-4 schools. We continued to collect process and outcome data in the third year of the program to monitor the impacts of this further iterated model.

Using non-experimental value-added estimates, we find that the treatment effects in the third year were still positive but lower than in the second year.[4] However, using student-level Mindspark usage data, we find that the correlation of test score gains with Mindspark usage time (i.e. the dose-response relationship) did not decline between Year 2 and 3. Rather, student time spent on Mindspark declined in Year 3, suggesting that the LIC's regular presence in the computer-lab may have been important for student "time on task". Thus, the time spent on the platform (which is easily observable) provides a ready metric for assessing implementation quality for further scale ups, and for personalized education technology interventions more generally.

Our first contribution is to the science of scaling (List, 2022). In particular, our study illustrates that, for many interventions, the challenge of scaling is not primarily one of taking "what works" in an efficacy trial, and ensuring fidelity to the original implementation protocols while delivering it at larger scales. Rather, it often requires *adapting* the intervention design in ways that build on the insights and *principles* illustrated by initial efficacy trials, but may lead to substantially different implementation protocols to accommodate new constraints that bind when implementing at scale.[5] This process of adapting for scale is similar to the approach of "problem driven iterative adaptation" (PDIA) advocated by Andrews et al. (2013). However, while PDIA is often framed as a substitute for RCTs (Nadel and Pritchett, 2016), we show that they are in fact complements: the process of effective scaling benefits from both an adaptive approach to discovery for intervention design, and experimentation at larger scales for testing and validation (Al-Ubaydli et al., 2017; Muralidharan and Niehaus, 2017).

Second, we contribute to the literature on the effective use of technology in education (EdTech). Policy interest in this area has grown rapidly, as has research evidence.[6] Yet,

[4] While we also present experimental ITT estimates of the cumulative 3-year effect, this confounds the effects of the main study's base implementation model (Y1 and Y2) and the lower-cost reduced-staffing model (Y3). Estimating the effects of the modified protocol in Y3, and comparing it to the base model, requires value-added methods (see Section 4.7).

[5] A useful contrast is with medical efficacy trials — approvals from the Federal Drug Administration are for specific drug combinations, dosages, and implementation protocols; effectiveness trials then investigate adherence to protocol in real-world settings. In contrast, many social policy interventions need to change the bundle of treatment, the intensity and the implementation for scaling. Thus, the efficacy trial is important to validate the core "theory of change", but does not by itself provide a guide to large-scale implementation. Developing new protocols that can be deployed at scale should, therefore, be considered an essential part of the "science" of scaling, requiring deep *intellectual* engagement that extends beyond logistical considerations.

[6] See Bulman and Fairlie (2016), Escueta et al. (2020) and Rodriguez-Segura (2022) for recent reviews. While evidence on EdTech remains mixed, some clear themes have emerged, including the lack of impact of hardware-only interventions and the promise of computer-aided interventions that supplement instructional

we still know remarkably little about how to effectively integrate EdTech into regular classroom instruction. In particular, *substituting* classroom teaching with computer-aided instruction has typically not been effective: Linden (2008) finds negative effects of doing so in India, and Ferman et al. (2019) find no significant effects on average (with negative point estimates) in evaluating a large-scale implementation of *Khan Academy* in Brazil.[7] These results are disappointing since integration in public schooling remains the most direct route to scaling the potential of EdTech and reaching disadvantaged students.

Our contribution to this literature is three-fold: (i) we demonstrate how a process of adaptation led to the creation of a PAL implementation protocol that is scalable across a wide variety of settings; (ii) we show experimentally that this protocol delivered significant productivity gains over a sustained period, across a wider range of grades than previous trials, while being implemented in public schools *during* the school day; and (iii) we show that student-level PAL usage data can serve as a low-cost continuous measure of implementation quality, which is especially important for scaling.[8] These contributions have also directly shaped practice with the scalable implementation protocol developed and tested above now being deployed in over 2,000 public schools and reaching over 250,000 students (see Section 5).

Third, our results also contribute to the global literature on tutoring in K-12 education. In the US, meta-analyses have identified high-dosage tutoring as highly effective for improving student learning (see, e.g., Fryer Jr (2017); Nickow et al. (2024)), especially in adolescence (Guryan et al., 2023). Yet, scaling has been difficult since effective tutors are costly and hard to find in adequate numbers, and the effectiveness of tutoring programs has been shown to decline sharply with scale (Kraft et al. (2024)).[9] We contribute to

time (Rodriguez-Segura, 2022).

[7]Similarly, Barrera-Osorio and Linden (2009) report null effects of computer-aided teaching in Colombia due to difficulties in integrating it with regular subject instruction, and de Barros (2023) documents negative effects of a technology-enabled blended learning program in India. A positive exception is Beg et al. (2022) who document positive effects of a video-led intervention in Pakistan, but they find positive effects only when the program also trained and engaged teachers and had negative effects when it did not, which highlights the sensitivity of impacts to the implementation protocol.

[8]Angrist and Hull (2023) and Angrist and Meager (2023) show that variation in experimental program effects can often be explained by differences in take-up and implementation quality. However, measuring implementation quality is often challenging in education systems. System-generated usage data from PAL programs can address this challenge and generate real-time data on implementation quality that can serve as a proxy for impact at larger scales (Athey et al., 2019; Budish et al., 2015). It can also provide an early warning to implementers about "voltage loss" in implementation quality (List, 2022). Digitally-generated user logs also reduce manipulation and misreporting, which undermines other standard sources of education data, such as teacher-reported test scores, in these settings (Singh, 2024).

[9]"High dosage" tutoring is defined by Dobbie and Fryer Jr (2013) as being tutored for at least 4 days per week in groups of 6 or fewer, making it difficult to sustain such high "voltage" at scale. Similarly, while online and phone-based tutoring programs have been shown to be effective, especially during the COVID-19 pandemic (see Carlana and La Ferrara (2024); Angrist et al. (2023a) but also Kraft et al. (2022)), scaling is again

this literature by showing that high-quality educational software providing personalized academic content (akin to personalized tutoring) can deliver meaningful learning gains even within the school day, without adding instructional time. Since computers (especially tablets) are much cheaper than tutors in higher-income countries, a model combining a single tutor for a group of students (to ensure adherence and engagement) and individual computers providing personalized instruction in a lab-like model similar to the one we study, may be a promising way of delivering the benefits of tutoring at scale.[10]

Finally, we contribute to a broader literature that aims to address the "global learning crisis" in low- and middle-income countries (Glewwe and Muralidharan, 2016; World Bank, 2017). While evidence on effective interventions has expanded in recent decades (see Angrist et al. (2023b) for a review), there is limited evidence on how to improve learning in middle-school grades. This is a critical gap since progression rates to middle school grades have increased sharply; yet student learning levels remain very low, and there are very few evidence-backed scalable interventions to improve middle school learning outcomes in LMICs. Our results suggest that technology-enabled PAL programs can be a promising approach to address this challenge. Further, EdTech is politically popular around the world and attracting growing amounts of funding, which makes our results timely for informing how these funds can be spent effectively to improve learning outcomes at scale. We discuss policy implications further in the Conclusion.

# 2 Intervention and Experiment Design

## 2.1 Background

Despite promising results, the after-school model of Mindspark evaluated in Muralidharan et al. (2019) was not viable for scaling, for the reasons discussed above. The primary goal of this study, therefore, was to develop and evaluate an implementation protocol for delivering Mindspark within the public schooling system in a form that could scale. Such adaptation was challenging for several reasons: (i) to fit within the school day, it would have to *replace* regular instruction time rather than adding to it, (ii) implementation would be overseen by regular government teachers rather than dedicated staff at after-school centers, (iii) public schools faced tighter resource constraints, requiring logistical adaptation and lower per-child costs, (iv) the scale of the intervention would be substantially larger and (v) it would include *all* students in a class, rather than a self-selected group attending

likely to be constrained by the supply of effective tutors, and the difficulty of sustaining parental engagement beyond the extreme COVID-19 shock.

[10] Bhatt et al. (2024) provide promising evidence along these lines by showing that adding a computer-aided learning (CAL) component to a tutoring program was able to deliver similar gains as a prior human-only tutoring program at 30% lower cost.

an after-school programs. Further, the intervention was broadened to include primary school grades, and cover all grades from 1-8. Thus, using the taxonomy of Muralidharan and Niehaus (2017) and Al-Ubaydli et al. (2021), the intervention required substantial modifications in the *scale*, *situation* and *population* relative to the original trial to adapt to conditions expected during real-world implementation at scale.

## 2.2 Intervention

Designing an implementation protocol suitable for scaling required careful adaptation across several dimensions.

**Setting:** Our intervention took place in Rajasthan, a large Indian state with a population of $\sim$ 84 million people in 2024 (see Fig. A.1). We worked in integrated public schools (called *Adarsh* schools) which span Grades 1 to 12. These integrated schools are larger and better resourced than stand-alone primary schools, but are common in Rajasthan and increasingly so across India.[11] These schools represent the type of public schools in India that are most likely to implement hardware-intensive EdTech interventions.

**Hardware:** Each treated school was provided a Mindspark lab, equipped with laptops with extended battery packs to avoid disruptions due to power cuts.[12] Treated schools were also provided a locally-hired laboratory in-charge (LIC), paid $\sim$ INR 10,000/month ($\sim$USD 150 in 2017), responsible for hardware maintenance, and helping students login. LICs were neither trained nor expected to provide any subject-specific instruction. As the role required no specialized skills, LICs can be easily recruited at scale.

**Scheduling Mindspark instruction:** The main design challenge for the scalable model was to integrate Mindspark instruction *within* the regular school day. The school schedule comprised six working days per week with eight periods of 35-40 minutes each day. To optimize the use of the hardware, the schedule allocated six "Mindspark Lab" periods a week to each of the 8 grades, split equally between Math and language (Hindi) instruction (see Figure A.3 for an illustrative time table).

This schedule represented a replacement of 12.5% of weekly classroom instruction (6 out of 48 periods), and an even larger share in targeted subjects. In primary grades (1-5), Mindspark replaced $\sim$ 25% of weekly Math and Hindi instructional time (3 out of 11-12

[11] The *Adarsh* schools program, started in 2014, consolidated smaller schools into larger units, aiming to have one in every village council (gram panchayat). The goal was to eliminate multi-grade teaching, and have the scale to offer better facilities, including computer labs. At the time of our study, Rajasthan had $\sim$ 10,000 such schools. Analogous efforts include the CM-RISE schools in Madhya Pradesh, and the national PM-SHRI program spanning over 14,000 schools across India.

[12] Hardware procurement and lab set-up were done by Educational Initiatives with funding from the Global Innovation Fund (GIF). However, similar infrastructure exists in integrated government schools, making this model broadly replicable.

periods). In middle school (grades 6-8), it replaced ∼40–50% of regular classroom time (3 out of 6-7 periods). Thus, integrating Mindspark lab periods into school schedules required substantial timetable modifications.

Head teachers were empowered to decide the details of how to implement these changes. In primary grades, Mindspark typically replaced classroom time in the same subject. In middle school, about half of the Mindspark time replaced scheduled classroom time in the same subjects (Math and Hindi), while the rest displaced time from non-targeted subjects, and remedial instruction.[13] Overall, the substitution of classroom time created significant adjustment costs for teachers, who had to cover the prescribed grade-level curriculum in much less time.

**Ensuring teacher support:** Interventions in the public sector, especially at scale, often fail due to the lack of support from frontline workers (Bold et al., 2018; Dhaliwal and Hanna, 2017). Thus, given the challenges posed to teachers from the substitution of instructional time (as noted above), a key element of adapting Mindspark for scale was to obtain the support of teachers.

Thus, the program design provided a central role to teachers. They were expected to accompany students to Mindspark labs during the lab period, answer student queries, and maintain time-on-task. Program communication emphasized that the role of Mindspark was to complement rather than substitute the teachers' role, and help by delivering differentiated instruction. Teachers received an orientation to the program, and access to a teacher-specific dashboard that summarized student achievement, progress, and learning gaps. In response to teacher feedback, the adapted version of Mindspark incorporated grade-specific worksheets for (optional) teacher use. Finally, the addition of LICs helped to assure teachers that they would not face additional administrative and logistical burdens.

**Student experience:** Mindspark is designed to offer personalized instruction to each student. However, despite adding extra hardware, it was infeasible – due to budget and space constraints – to provide one device per student in a government school setting.[14] These constraints required another important adaptation relative to the pilot, whereby two students were paired on one device where needed.

Students received individual login credentials, and completed a diagnostic test to set the starting level for personalized Mindspark instruction. Where the number of students exceeded the number of computers, students were paired by gender and similar diagnostic

[13]Based on comparing timetables across treatment and control schools (see Section 4.5 and Table A.2).

[14]Classrooms designated for use as computer labs were typically not large enough to accommodate the 40-50 computers needed for each student to have their own device. Further, government EdTech budgets would not be enough to provide those many computers at scale.

scores.[15] Paired students had individual headphones, but shared a mouse and keyboard, and were encouraged to discuss answers before entering them. This model was refined in the first year and remained unchanged thereafter.[16]

### 2.3 Study sample and experiment design

Our study was conducted in four districts of Rajasthan – Churu, Jhunjhunun, Udaipur and Dungarpur – spanning both northern and southern regions of the state (see Fig A.1). Educational Initiatives (EI) identified 80 *Adarsh* schools across rural and urban areas of these districts, that had the infrastructure to set up Mindspark labs. These schools constitute our study population.

We stratified schools into within-district pairs based on middle school enrollment in 2016-17. One school in each pair was randomly assigned to treatment, and the other to a control group. All analyses control for stratum fixed effects, and cluster standard errors at the stratum level (following De Chaisemartin and Ramirez-Cuellar (2024)).

We collected baseline data on school characteristics, student test scores, and socioeconomic status in October 2017 – after randomization but before Mindspark instruction started. Treatment and control schools were balanced on school characteristics, baseline test scores, and socioeconomic status (Table 1). The only exception was a small significant difference in the proportion of girls; so all regressions control for gender.[17] These covariates – except gender – remain balanced in later rounds (Table A.1). At baseline, 41% of students were in primary school (Grades 1-5), and the rest in middle school (Grades 6-8). There was no differential student attrition across treatment and control groups at the end of either Year 1 (Y1) or Year 2 (Y2).

## 3 Data

Our analysis is based mainly on primary data on student learning outcomes collected by the research team between 2017 and 2020, supplemented by administrative data from schools, and system data on usage from EI.

### 3.1 Student learning outcomes

Our primary outcome is student achievement, which we measure using independently designed and administered tests in Math and Hindi. We measure these four times: at

[15] Hardware constraints varied by school and grade based on enrollment. When only some students needed to share devices, weaker students were prioritized for access to their own device.

[16] While students were paired based on initial diagnostic scores, adherence to the assigned pairing was only partial since students sometimes paired up with their friends. We therefore focus on ITT effects based on random assignment of schools to treatment.

[17] 79 out of 80 schools were co-educational; the sole girls-only school was in the treatment group.

baseline (October 2017), and close to the end of each school year (February-March of 2018, 2019 and 2020). In the first three testing rounds, we tested all students in Grades 1-8 who were present in school on the date of the student assessment. In the final endline in February 2020, we also aimed to test students who were absent on the day of the school-level testing by visiting them at their household.

To capture the full distribution of student achievement and minimize ceiling and floor effects, we designed separate test booklets for each grade/subject/round combination, with difficulty increasing by grade. To reduce floor effects from many students scoring zero, we varied the testing mode by grade: tests for grades 1-2 had only oral questions; grades 3-5 included both oral and written items; grades 6-8 used written test booklets. A key measurement challenge, given the span from Grades 1-8, was to express student test-scores on a common scale across our full sample. We addressed this by including a subset of common test questions across adjacent grades and testing rounds, which enables us to use Item Response Theory (IRT) models to generate test scores on a comparable scale for all students and testing rounds.[18] For our main results, test scores are standardized to have a mean of zero and standard deviation of one in Grade 5 at baseline. Further details on test content and psychometric properties are in Appendix B.

We also use administrative data on Grade 5 and 8 exam scores in 2018-19, 18 months after the program began, to study effects on school exams on targeted subjects and potential spillovers to other subjects.[19]

## 3.2 Mindspark software data

The Mindspark software logs detailed usage data for each user and session, as users login with individual IDs. This includes session duration (with date and time), questions attempted and answered. The system also records the initial learning level assessed by the diagnostic test, which is then used to personalize the content provided to students.[20] This diagnostic test was implemented at the beginning of the intervention ($\sim$ November 2017), and then again at the beginning of each academic year (in July-August of 2018 and 2019).

## 3.3 Classroom observations and teacher interviews

In 2019, we collected time-use data for teachers and students in classrooms (in treated and control schools) and in Mindspark labs (treated schools only). Enumerators

[18]Linking items were administered in the same format (oral or written) across grades.

[19]Only Grades 5 and 8 have standardized exams across schools. End-of-year school exams were canceled in 2019-20 due to COVID-19 related school closures in March 2020.

[20]This test was also used to pair students at similar learning levels to share a computer, when needed. When paired, both students had to login, and usage was recorded for each of them.

recorded snapshots of activities of students/teachers/lab in-charges at regular intervals using an adapted version of the structured Stallings classroom observation protocol. We also administered surveys to teachers and students in treated schools to understand their subjective experience of Mindspark.

### 3.4 Other school level data

In 2018-19, the first full year of program implementation, we also collected school time tables to understand how scheduled class time adapted to the program. Time tables were obtained from 71 schools for middle grades and 63 for primary grades (out of 80). We also transcribed official attendance records to collect monthly student attendance for all students in 2018-19.

## 4 Results

### 4.1 Learning levels and variation under the status-quo

The Mindspark diagnostic assessment measures students' actual learning level regardless of the grade they are enrolled in. We use data from the first diagnostic test in 2017 to characterize learning levels, gaps, and heterogeneity among the students in our sample (prior to any Mindspark instruction). Figure 1 presents the joint distribution of students' enrolled grade, and their assessed grade-level at the start of treatment.

The figure highlights three key patterns. First, student learning levels in this setting are substantially below grade-level standards — for instance, the average 8th grade student has around a 4th grade level of math proficiency. Second, while learning improves in higher grades, the rate of progress is much flatter than the line of equality between curricular standards and actual achievement. Thus, by grade 8, students are (on average) 4 grades behind in math and 2 grades behind in Hindi. Third, the use of a dynamic computer-adaptive diagnostic test allows us to document the striking dispersion across student learning levels *within* the same grade. For instance, 8th grade students span *seven* grade-levels of learning in math (from grade 2 to grade 8). These patterns appear in both subjects but are more pronounced in Math than in Hindi.

Figure 1 reinforces the findings in Muralidharan et al. (2019) in three key ways. First, despite changes in geography and population (from a self-selected sample to all students in schools), we document nearly identical patterns in middle schools. This confirms the generalizability of a key fact about education in India, which is the large mismatch between curricular standards and students' actual learning levels. Second, we now track this gap across the *full span* of compulsory schooling, and show that learning deficits emerge early in primary schooling and widen over time, highlighting the need for early remediation. Third, the variation in within-grade learning levels emerges early and grows over time. While Figure 1 pools data from all treated schools, a variance decomposition shows that in Grade 8, 79% and 89% of the variation in math and Hindi learning levels are *within classrooms*; the corresponding figures for Grade 5 are 67% and 86%.[21]

The patterns in Figure 1 may be explained in part by the "no detention" policy implemented under India's Right to Education (RtE) Act, under which students are automatically promoted to the next grade regardless of whether they meet the standards of their current grade (Muralidharan and Singh, 2021). While well-intentioned and intended to reduce school dropout rates, it may have made teaching much more challenging, because even qualified and motivated teachers would struggle to cater to such wide variation in student preparation. It also highlights why technology-enabled personalized adaptive learning (PAL) could be highly effective in this setting, even when it *displaces* undifferentiated instruction based on following textbooks aligned to curricula standards.

## 4.2 Effects on Learning Outcomes

We estimate treatment effects on learning outcomes at the end of Years 1 and 2 (Y1 and Y2), and focus on Y2 results after 18 months of treatment. We present Intent-to-treat (ITT) effects estimated using the following specification:

$$Y^j_{igspt} = \beta_1.Treatment_s + \theta_p + \beta_2.X_{igsp} + \epsilon_{igspt} \tag{1}$$

where $Y^j_{igspt}$ is the test score in subject $j$ for student $i$ in grade $g$, school $s$, stratum $p$, at time $t$; Treatment is an indicator variable for being in a program school; $\theta_p$ are stratum (school-pair) fixed effects. $X_{igsp}$ includes student gender and baseline test-scores (from Oct 2017). For students absent during baseline testing (including new cohorts entering in the second year), we assign the average grade-subject score in the school in lieu of a baseline score.[22]

We find positive treatment effects of $0.15\sigma$ in math and $0.11\sigma$ in Hindi after 4-5 months of implementation (Table 2, Panel A, Columns 1 and 4). After 1.5 years, these rose to $0.22\sigma$ and $0.2\sigma$ respectively (Panel B). Over this period, control-group students gained $0.47\sigma$ in

[21] While we are not aware of similar measurement of within-grade variation in student learning in other settings, these patterns are likely to replicate in many other LMICs, and potentially in higher-income countries as well. See the Appendix to Muralidharan et al. (2019) for a more detailed discussion.

[22] This approach allows the benefits of improved precision from conditioning on baseline scores — including those who were absent at baseline — without introducing bias (Altonji and Mansfield, 2018). Our results are almost identical in specifications where we only control for randomization strata fixed effects (Table A.3).

math and 0.31$\sigma$ in Hindi on the same metric.[23] Thus, treatment effects equal roughly *half* of the business-as-usual learning gains in math, and about *two-thirds* in Hindi, over this period (see last two rows of Table 2). These effect sizes rank around the 90th percentile of those documented in large sample (N>5000) RCTs in LMICs (Evans and Yuan, 2022).

Compared to influential studies in primary education, our 18-month effects are similar to 18-month effects of tracking in Kenya (Duflo et al., 2011) and to two-year effects of teacher performance-pay (Muralidharan and Sundararaman, 2011) and remedial instruction by community volunteers in India (Banerjee et al., 2007). Notably, there is little evidence of scalable, effective interventions in public middle schools, and to our knowledge, these are the largest experimental treatment effects in public middle schools that we are aware of to date, that has been delivered at a large scale (N>5000 students).[24]

We find significant test-score gains in both subjects across primary and middle school grades (Table 2, Columns 2-3, 5-6). After 18 months, students in treated schools scored 0.15$\sigma$ higher in math in primary grades and 0.25$\sigma$ higher in middle school. Gains in Hindi were 0.2$\sigma$ and 0.15$\sigma$ respectively. We cannot reject equality of treatment effects across primary and middle school grades in either subject. This suggests that the Mindspark PAL system improved productivity of school instructional time across the full span of elementary school grades, which is new evidence beyond the efficacy trial, which only covered middle school grades.

## 4.3 Heterogeneity and personalization

### 4.3.1 Heterogeneity by student characteristics

Consistent with the personalized nature of Mindspark instruction, we find broad-based gains across all initial learning levels. Figure 2 illustrates this non-parametrically, and presents local polynomial regressions of Y1 and Y2 student test-scores on their (within-grade) baseline percentiles, separately for treatment and control groups. In both years and subjects, the conditional expectation function shifts upward for the treatment group, indicating broad-based gains across the learning distribution.

To probe this further, we classify students into within-grade quintiles of baseline achievement and allow treatment effects to vary by quintile. Point estimates for interaction

[23] This is the average within-student change in test scores for control-group students tested at baseline and 18-months later in 2019. IRT-linking of test scores enables us to express student scores on a common scale across *all* rounds and grades.

[24] The only study showing larger effect sizes on middle-school grades that we are aware of is Gray-Lobe et al. (2022) in Kenya. However, these are the effects of attending *private* New Globe schools and reflect a bundled intervention including pedagogy, management, peer effects, and staffing, rather than a specific intervention in public middle schools.

effects are typically small and insignificant, and we do not reject the null that they are jointly different from zero, though Y2 effects in Hindi appear larger for lower-scoring students (Table 3). We also examine heterogeneity using a standard linear interaction model and find limited evidence of heterogeneity by baseline test-scores, except for Y2 Hindi scores (Table A.4). We also find no heterogeneity by gender or socioeconomic status (Table A.5).

Table 3 also offers important insights on learning progress in the control group. Relative to the lowest quintile (omitted category), students in higher quintiles show significantly faster learning progress across both subjects and years.[25] Remarkably, learning progress is monotonically increasing by quintile of initial achievement in both subjects, highlighting how weaker students get progressively left behind under default classroom instruction focused on grade-level standards.[26] While Figure 1 presents a cross-sectional snapshot of learning dispersion within grades, Table 3 sheds light on how that divergence emerges over time. A key implication of Figure 2 and Table 3 is that while absolute treatment effects are comparable across students, the effects *relative* to progress in the counterfactual are much greater for weaker students, because their regular rate of progress is significantly lower.[27]

#### 4.3.2 Heterogeneity by question characteristics

We also test for personalization by examining heterogeneity by question difficulty. We classify questions as "easy" if the rate of correct responses in the control group was in the top third of questions, and "hard" if they were in the bottom third. This classification is specific to each grade/round/subject combination.[28] We then estimate program effects on percentage correct for "easy" and "hard" items separately, allowing this effect to be heterogeneous across within-grade terciles of student achievement at baseline.

We see treatment effects consistent with personalization. In both Math and Hindi, and in both years, we see that students in the bottom tercile have significantly larger achievement gains for "easy" test questions than students in the top tercile (Cols 1-2, Table

[25] Note that, while current and lagged test scores are measured on a common IRT scale, quintiles are defined *within* grade and subject. Thus, students with the same IRT-score can be in different quintiles if they are in different grades, which serves as a proxy for distance from curricular standards.

[26] This pattern also holds when comparing value-added in one subject across quintiles defined by baseline test scores in the *other* subject and controlling for baseline scores in both subjects (Table A.6). Following Jerrim and Vignoles (2013), this suggests that greater progress for initially-high-scoring students is not driven by measurement error in baseline scores.

[27] Note that we cannot quantify this ratio precisely because the rate of progress in the omitted category (lowest quintile) is not identified, and that has to be added to the quintile interaction terms to calculate absolute rates of counterfactual progress in each quintile. However, dividing a constant treatment effect with increasing values of counterfactual progress at higher quintiles implies that the relative treatment effect is mechanically higher for weaker students. This point is strengthened by the negative coefficients on the linear interaction in Table A.4, implying slightly higher absolute treatment effects for weaker students.

[28] Recall that a subset of items are common across grades and rounds. So, a given test item may be "hard" for, say, Grade 5 students but "easy" for Grade 8 students.

4). Conversely, in both years in math, students in the top tercile have larger treatment effects for "hard" items (although the evidence in Hindi is more mixed).[29]

### 4.4 Insights from Mindspark system data

All treatment effects reported above are based on independently designed and administered tests in both treatment and control groups. We now present additional insights from the Mindspark system data, available only for the treatment group.

Mindspark conducts a diagnostic test at the start of each school year to personalize the content it delivers. This test provides a summary assessment of each students' actual grade level, and allows us to examine progress relative to curricular standards. Since the test is conducted at the start of each school year, it reflects learning up to the end of the previous year. We therefore refer to the diagnostic test at the start of year 2 (and 3) as Y1 (and Y2), for consistency with the terminology used for treatment effects above and reflect the same duration of treatment.

Figure 3 plots students' assessed grade level at Y1 and Y2 by their assessed level at baseline.[30] The key result is an upward shift relative to the line of equality in both subjects over time. Averaged across grades, treated students with 18-months of program exposure (proxied by being present for both Y0 and Y2 diagnostic tests) gained an average of 1.7 and 2.1 grade levels in Math and Hindi between Y0 and Y2, implying that treated students gained around a year's curricular standards of learning per year of school.[31]

This is an important result in light of Figure 1, which shows that typical annual learning progress is far below curricular expectations. The treatment appears to raise the productivity of a year in school to align learning gains with curricular expectations for progress, and suggests that learning gaps relative to grade-level curricular standards could fall over time for treated students. Using the Mindspark diagnostic tests at the start of each year, we find exactly this pattern: the gap between students' assessed learning levels and curricular standards narrows significantly over time (Figure 4), and Table 5 shows a rising slope in mean learning levels by grade. Since treated students' annual learning gains now match curricular expectations (Figure 3), starting personalized

[29]We classify students into terciles of baseline achievement (rather than quintiles, as in Table 3) for greater power, since we are now examining heterogeneity across two dimensions (question difficulty and initial learning levels). For completeness, Table A.7 presents the analogous table, classifying students into quintiles. Results are very similar, but less precisely estimated.

[30]The histograms plot the distribution of assessed grade level (regardless of enrolled grade level), which is typically well below the enrolled grade level (Figure 1)

[31]The mean learning gains for each grade-level of Y0 scores can be calculated from the estimates of the slopes and intercepts presented in (Table A.8).

instruction early could prevent such gaps from emerging in the first place, or at least sharply reduce them. This is an important area for future research.[32]

### 4.5 Distinguishing productivity gains from additional instruction

One caveat to interpreting the positive treatment effects on test-scores as *solely* reflecting increased productivity of instructional time, is that treated schools often adjusted their timetables to partly make up for lost classroom time in math and Hindi due to the substitution with Mindspark lab time. As a result, total scheduled instruction time in targeted subjects (classroom plus Mindspark lab time) was an insignificant ∼6.5% higher in treated schools in primary grades and was a significant ∼25% higher in middle school grades (Table A.2).[33] We provide two pieces of evidence suggesting that Mindspark time was more productive than classroom instruction.

First, in primary grades, treatment effects are 53% of control group learning gains in Hindi, and 22% in Math (2, last row of columns 2 and 5), substantially exceeding the ∼6.5% increase in subject-specific instructional time. In middle grades, treatment effects are 65% of control gains in Hindi, and 100% in Math (2, columns 3 and 6), far exceeding the ∼25% increase in instructional time. In the absence of productivity differences between classroom and Mindspark instruction, we would expect gains to be proportional to added instructional time.

Second, we use 2018-19 school time-tables to identify the subset of treated grades in our sample where total subject-specific instructional time (classroom plus Mindspark lab) was equal to the scheduled classroom time in the same grade and subject in the paired control school in the same randomization stratum. Treatment effects in this restricted sample are nearly identical to those in the full sample (Table 6). These results suggest that treatment effects primarily reflect increased productivity of school time, and not simply added instructional time by displacing other subjects.

### 4.6 Treatment effects on school examinations

Next, we examine impacts on treated students' performance in official school exams. The direction of this effect is *ex ante* ambiguous. On one hand, the gains on our independent tests should ideally also be seen in school exams. On the other hand, school exams narrowly assess grade-specific curricula. Since most students are several years behind

[32] We had intended to conduct longer-term follow-ups of study cohorts, but did not do so because its value was significantly reduced by COVID-related school closures of ∼18 months shortly after 3 years of treatment, making findings difficult to interpret.

[33] In primary grades, treated schools had 0.5-0.65 more weekly periods in math and Hindi over a base of ∼11 periods in control schools; in middle grades, they had 1.6 more over a base of 6.2. This extra time in middle school appears to have come from a 5-10% reduction in time for other subjects (Table A.2).

grade level, and Mindspark provides instruction targeted to actual learning levels, even substantial improvements in learning may not translate into better performance on grade-level exams. Moreover, the intervention *displaced* classroom time typically dedicated to grade-level instruction, including time for exam preparation and revision.

We investigate this for Grades 5 and 8, where standardized exams are administered across schools. We find no significant effects in Math or Hindi, with point estimates being negative in both grades (Table 7). We also test for spillovers on non-targeted subjects such as English, Science and Social Studies. Consistent with the potential displacement of class time from these subjects, we find small negative point estimates that are not statistically significant (Table A.10). Thus, despite sizable learning gains on assessments designed to capture the full range of true student learning, we find no impact on grade-level school exams.[34]

Overall, the absence of a significant negative effect on school exams can be seen as a positive result given the ∼25-50% reduction in grade-level instructional time, and reduced time spent on revision and cramming for the final exams. At the same time, the lack of positive effects on grade-level exams underscores the trade-off between teaching "at the right level" and "at the curricular level" within finite instructional time.

More broadly, it highlights the tension between the "sorting and screening" and "human capital formation" functions of education systems. The former focuses more on *identifying* high-achieving students, often through curricula and exams aimed at the top end of the distribution; the latter requires improving learning for *all* students, regardless of their starting point. The Indian education system has historically served the sorting function well (Muralidharan, 2024). However, as Figure 1 and Table 3 show, this has come at the cost of low effectiveness in human capital formation for the vast majority of students who fall behind curricular standards.[35]

In this context, Mindspark may have been especially effective because the status quo does not adequately serve students who fall behind curricular standards. Our results also highlights the promise of PAL systems to narrow learning gaps relative to curricular standards by starting in early grades (Figures 3 and 4). Doing so may reduce the tension between teaching "at the right level" and "at curricular standards" in later grades.[36]

[34] These results also underscore the importance of appropriate test design in evaluating education programs in settings where there is wide dispersion in student learning, and where impacts on learning are likely to happen at a considerably different level than those targeted by curricular school tests.

[35] This challenge has been exacerbated by rising enrollment of first-generation learners without adequate parental support for learning at home (Muralidharan and Singh, 2021). Further, the "no detention" policy under India's Right to Education (RtE) Act—though well-intentioned and intended to reduce school dropout rates—may have hurt learning by depriving weaker students additional time to reach grade-level standards before tackling harder material in higher grades.

[36] Early grades are also well suited to PAL and a focus on conceptual learning. In higher grades, all

## 4.7 Further adaptation for scaling: results from the third year

Beyond hardware, the largest recurring program cost was the dedicated lab-in-charge (LIC) in each treated school. To bring costs down further, program funders and the government sought implementation models that either eliminated the LIC role or spread it out across schools. Accordingly, in its third year, the program reduced the number of LICs, and assigned each LIC to a 'beat' of 3-4 schools. They were expected to rotate among them to ensure smooth functioning of systems with no technical challenges. Data collection continued in this third year.

Program usage declined sharply after the staff reduction July 2019, to about half the previous year's levels (Figure A.2). However, because usage data was *visible* to implementers, they received an early signal of reduced "voltage", and worked with schools to resolve challenges and improve usage. As a result, usage recovered to 2018-19 levels by November 2019. However, total usage was ~15% lower in Y3 compared to Y2.

The change in implementation protocols, and associated disruptions, complicates interpretation of 3-year ITT effects. We therefore treat Y2 effects as the primary *experimental* estimates of the original implementation protocol, and use non-experimental value-added models to evaluate the modified Y3 protocol. We also do this for Y2 to compare value-added in Y2 and Y3. We estimate:

$$Y^j_{igspt} = \theta_s + \mu_g + \beta_1.Treatment_s + \beta_2.Female_i + \lambda.Y^j_{igspt-1} + \epsilon_{igspt} \tag{2}$$

Here, $Y^j_{igspt}$ is the test score in subject $j$ at time $t$ for student $i$ in grade $g$ in school $s$; $\theta_s$ and $\mu_g$ are strata and grade fixed effects; the $Female_i$ indicator is included to account for baseline imbalance. $\beta_1$ identifies per-year program value-added (VA). Interpreting $\beta_1$ causally requires that, in the absence of the program in year $t$, test scores of students with the same lagged achievement ($Y_{t-1}$) would have evolved similarly in treated and control schools. Given random assignment of treatment, the main identifying assumption is that lagged scores act as a summary statistic of previous program effects.

Results, presented in Table 8, suggest that between Y2 and Y3, program VA declined from $0.14\sigma$ to $0.1\sigma$ in math and from $0.12\sigma$ to $0.07\sigma$ in Hindi, though we lack power to reject equality. This decline could reflect either reduced usage or lower Mindspark productivity in Y3 (e.g. if students were less focused on computer-aided instruction with reduced adult supervision).

stakeholders—parents, students, teachers, and administrators—place much greater emphasis on high-stakes external exams, increasing pressure to memorize grade-level content for exams.

To explore this, we use data from treatment schools to estimate the dose-response of Mindspark usage — i.e. the correlation between time spent on the platform with learning gains — using similar VA specifications that condition on the previous year's test score, school, and class fixed effects. Causal interpretation of this association requires assuming that lagged achievement and controls fully address selection into greater usage (for example, through student motivation).[37] We find no evidence that the dose-response of Mindspark deteriorated across the two years (Table 9). This suggests that the reduced usage of Mindspark accounts for the decline in program value added between Years 2 and 3, and highlights the value of usage data in EdTech platforms as a real-time metric of program implementation quality.[38]

Finally, for completeness, we present the ITT comparisons between treatment and control schools at the end of Year 3 (Table A.11). After ∼30 months, program effects are $0.24\sigma$ in math and $0.21\sigma$ in Hindi, and are broad-based and positive across grades. That these effects are not larger than after ∼18 months (Table 2), despite a positive value-added in Year 3, is explained by the impersistence of test scores over time. This issue is ubiquitous in panels and cautions against linearly extrapolating treatment effects over longer durations.[39]

## 4.8 Program implementation in steady state

The intervention aimed to develop an implementation protocol to integrate technology-enabled PAL into regular teaching. In Year 3 (Nov-Dec 2019), we conducted a round of data collection in treated schools, focused on grades 5 and 8, to directly observe (i) the implementation of Mindspark in treated schools, (ii) any effects on regular classroom instruction, and (iii) teacher and student opinions about Mindspark instruction, after two years of observation.

[37] Similar value-added models have been shown to substantially address selection in many settings including, e.g., teacher effects (Chetty et al., 2014; Bau and Das, 2020), school effects (Andrabi et al., 2011, 2025; Angrist et al., 2017; Singh, 2015), and years of schooling (Singh, 2020). Further, we showed in Muralidharan et al. (2019) that similar panel-based VA estimates were statistically indistinguishable from IV models using the (randomized) voucher offer as an instrument. While we cannot use the same IV strategy here (since treatment affects learning through both the individual-specific usage of Mindspark and the displacement of class time) the equivalence of VA and IV estimates in evaluating the same software provides additional confidence in interpreting the dose-response relationship causally.

[38] These results parallel those from recent multi-site RCTs in the US on scaling high-dosage tutoring, which find that dose-response remains unchanged, but treatment effects fall due to sharp dosage reductions at larger scales (Bhatt et al., 2025).

[39] As noted by Muralidharan (2012) in the context of multi-year experiments, "*the n-year 'net' treatment effect consists of the sum of each of the previous n-1 years' 'gross' treatment effects, the depreciation of these effects, and the n'th year 'gross' treatment effect.*" Our ITT and value-added estimates imply a ∼ 60-70% test-score persistence between years, consistent with the range of estimates reported by Andrabi et al. (2011) using four years of panel data on test scores in Pakistan.

Classroom and lab observations, along with teacher and student interviews, indicate that Mindspark was well integrated into regular school practice. In directly-observed lab periods, most students were actively engaged on Mindspark, and 90% of computers were being used for the assigned subject (Table A.12); despite teachers only being present for about 50% of lab observation snapshots (Table A.13).[40] Direct observations of teacher-led classroom periods showed no significant differences in regular in-person instruction (Table A.14). Nearly all teachers reported finding Mindspark useful, although primarily for conceptual understanding and student learning, but not for exam performance (Table 10). This suggests that teachers are aware of Mindspark's strengths and limitations, with their views being consistent with the results in Table 7. This broad acceptance of Mindspark, both in student use and among teachers, is important since the buy-in from teachers is key for sustaining interventions in public schools (Bold et al., 2018).

## 5 Cost Effectiveness and Policy Implications

As implemented, the program cost ∼INR 53 million over three years (∼USD 750,000 @ 1 USD = INR 70, the Y3 exchange rate). Roughly half comprised costs of hardware (INR 22 million) and of repairs and infrastructure for labs (INR 2.8 million). The other half comprised recurring costs, including program staff salaries, training, and ongoing engagement with teachers and principals. The recurring costs declined over time, from INR 12.6 million and INR 10.5 million in Y1 and Y2, to INR 4.9 million in Y3 after reducing LIC staffing.

To calculate annual costs, we assume that (i) hardware and lab repairs are depreciated over five years, (ii) software license fees would be USD 2 per child per year,[41] and (iii) 6500 students were treated annually, in line with enrollment in program schools. Under these assumptions, per-student annual costs were roughly INR 2903 in Y1, 2571 in Y2, and 1718 in Y3 (∼USD 41, 37 and 25).

The adapted in-school model was more cost-effective than the Delhi efficacy trial, which cost USD 180 per student annually and yielded ITT effects of $0.22\sigma$ in Hindi and $0.36\sigma$ in Math after half a year (Muralidharan et al., 2019). Over the same period, the scaled-up version improved scores by $\sim 0.15\sigma$ in Math and $0.11\sigma$ in Hindi. This is about half the original effect, but was achieved at under one-quarter the cost. This gain in

[40] This suggests that teachers may have often relied on the LIC to ensure usage, and used that time for other tasks (administrative work, class preparation) or leisure. It may also explain the sharp usage drop at the start of Y3 with reduced LIC presence (Figure A.2). However, the later recovery of usage suggests that awareness that usage reductions are being monitored and flagged may itself increase teacher engagement in the labs to maintain adequate usage (though we cannot directly test this).

[41] These fees were waived for this study. EI committed to capping software licensing fees, inclusive of cloud storage, at USD 2 per child per year for future government school scale ups.

cost-effectiveness occurred despite *substituting* instruction and serving a more diverse, non-self-selected population. Three factors explain this improvement: (i) hardware costs were spread over more students (especially when computers were shared) and labs were more fully utilized than after-school centers often running below capacity, (ii) there were no rental costs for out-of-school premises, and (iii) implementation within school hours by regular teachers, limiting additional salary costs to the LIC.

Our main experimental treatment effects are from Y2, and the gains of $0.22\sigma$ in math and $0.2\sigma$ in Hindi were achieved at a per-pupil cost of $\sim$USD 78 over Y1 and Y2. A key policy relevant benchmark is to compare these costs and benefits with status quo public education spending. In 2018-19, Rajasthan spent an average of INR 39,490 annually per student in government schools (Accountability Initiative, 2021). Scheduled instruction in math and Hindi accounted for $\sim$46% of the school day in primary grades and 25% in middle-school grades (Table A.2). Pro-rating per-pupil spending by this share yields annual expenditures of INR 17,700 in primary and INR 9,872 in middle school on these subjects. Thus, in middle schools, average program spending in the first two years (INR 2735) increased spending on these subjects by $\sim$27% but doubled productivity in math and raised it by 64% in Hindi (Table 2, last row). In primary school, spending rose by 15% but increased productivity by 22% in Math and 53% in Hindi. Thus, as implemented, the incremental spending on the intervention was 1.5 to 4 times more productive than business-as-usual spending in these schools.

Looking ahead, costs are likely to fall further due to falling hardware costs, and operational improvements that reduce LIC costs without reducing usage. In Y3 itself, usage fell in the transitional period when LIC staffing was reduced, but recovered to previous levels within a few months despite halving the recurring staff costs. If future annual costs match Year 3 with no usage reduction, then cumulated two-year program costs would be $\sim$50 USD rather than the 78 USD actually incurred. Finally, recent models of Mindspark deployment have featured 1 LIC-equivalent across 8 schools, suggesting that further reductions in personnel costs are feasible.

Moreover, many schooling systems, including growing parts of India's public education system, already have computer labs and equipment.[42] Where the hardware investments have already been made, the marginal cost of deploying Mindspark, or other PAL software, is much lower. Excluding pro-rated hardware costs, Y3 costs were $\sim$USD 11 per child.

While total cost-effectiveness calculations should include hardware costs, hardware-excluded figures also matter for decision-makers looking to scale evidence-backed education

[42] For example, the PM-SHRI program intends to equip 14,500 upgraded schools across India with computer labs. Similar state-level programs are also common.

interventions. Several studies show that hardware by itself typically has no impact on learning (Malamud and Pop-Eleches (2011), Cristia et al. (2017), Escueta et al. (2020)). Yet, governments in India spend much more on hardware than PAL software because hardware purchases are (i) visible and hence politically popular, and (ii) administratively easier because procurement rules are much simpler for standardized hardware than PAL software platforms that vary on several dimensions. Thus, if the fixed costs of hardware have already been incurred, the cost-effectiveness of PAL software on the margin will be even higher.[43]

The implementation protocols developed and tested in this study have already contributed to scaling up of PAL in public schools. As of 2024-25, Mindspark was operational in 2217 government schools, serving over 266,000 students, across 13 Indian states. These scale-ups directly build upon the process discovery and evaluation documented in this paper. The stable dose-response relationship in our data suggests that monitoring student-level usage could be a key metric for ensuring implementation quality and forecasting the effects of these scaled up deployments.

Finally, in return for receiving taxpayer funds through the Global Innovation Fund (GIF) for developing and evaluating the in-school Mindspark model, Educational Initiatives agreed to make the implementation protocols public and broadly accessible. This is a significant public good since most Ed-tech firms aim to keep their protocols proprietary. These protocols have since informed within-school implementation by other PAL providers in multiple states in recent years.[44]

# 6 Discussion and conclusion

This paper makes both methodological and substantive contributions.

Methodologically, it advances the science of scaling by showing how scaling challenges in social policy contexts differ fundamentally from those in medical contexts. In the latter, the challenge in moving from small-scale efficacy trials to effectiveness at scale is primarily one of maintaining compliance with the original clinically-tested protocol. In contrast, scaling social interventions requires considerable adaptation to take the core principles validated in smaller trials, and then iterating delivery models to account for the fiscal, logistical, organizational, and political economy challenges that arise at scale. Our study illustrates this process, using the example of personalized adaptive learning (PAL) software, and shows the value of continued experimental evaluation of delivery models adapted for scale.

[43]It may also be possible to improve cost effectiveness further by using existing hardware more intensely. For instance, computer labs could be used to deliver summer programs or after-school remediation programs.

[44]Examples include ongoing implementation and evaluations of Convegenius in Andhra Pradesh and Khan Academy in Uttar Pradesh. These evaluations may also aid scaling by facilitating public procurement, which is easier when there are multiple qualified bidders whose products have been credibly evaluated.

It also highlights the critical value of having a readily observable proxy indicator of implementation quality at scale – here, usage of the EdTech platform. The ability to continuously monitor usage enabled us to detect when implementation quality ("voltage") dropped following a reduction in LIC support in Year 3, and to drive prompt corrective action. Further, the strong dose-response relationship we document aligns with emerging evidence that variation in implementation quality is a key driver of differences in program impacts across contexts (Angrist and Meager, 2023). Thus, investing in proxy measures of implementation quality is likely to be a critical enabler of maintaining the "voltage" of interventions at scale, and is one of our most important recommendations to practitioners seeking to deliver impact at scale.

Beyond these methodological contributions, our results also yield several substantive insights. First, they contribute to the growing EdTech literature. We show that PAL can be effectively integrated into the core instructional timetable of public schools, improve the productivity of school time, and reach large numbers of disadvantaged students at modest cost. Moreover, we find that PAL can deliver broad-based gains with no evidence of differential impacts by initial learning, gender or socioeconomic status. These findings are globally relevant as governments and donors continue to invest billions of dollars annually in educational technology, and can inform both the design and management of PAL programs to improve learning at scale.

Second, we show how PAL can help mitigate critical systemic challenges in LMIC education systems. Reviews of global evidence highlight that weaknesses in pedagogy (low teacher subject knowledge, ineffective instructional practices, and poor matching of content to student learning levels), and governance (high teacher absence, low time on task, and limited accountability for effort or outcomes), are major obstacles to improving learning outcomes (Glewwe and Muralidharan, 2016). The PAL model addresses both. It provides personalized high-quality content that meets students where they are, keeps them engaged, provides rapid feedback, and adapts with student learning. At the same time, its granular usage data offers rare visibility into classroom processes, enabling more effective governance and accountability.

Third, compared to other promising education interventions in LMICs, technology-enabled PAL appears more feasible to scale. Interventions such as contract teachers and teacher performance pay have shown positive impacts but face political resistance or implementation challenges at scale.[45] Other promising reforms such as charter schools

[45] See Bold et al. (2018) on the political challenges of scaling contract teachers. While teacher performance-pay has proven effective in trials across India, East Africa, China and Mexico (see, e.g., Muralidharan and Sundararaman (2011); Behrman et al. (2015); Mbiti et al. (2019); Leaver et al. (2021)) but has proven difficult to scale outside of researcher-led pilots. Singh (2024) and Singh and Berg (2024) highlight the

(Romero et al., 2020; Gray-Lobe et al., 2022) face strong resistance from teacher unions.[46] Finally, effective primary-grade pedagogical innovations such as teacher or volunteer-led efforts to "teach at the tight level" (see, e.g., Banerjee et al. (2007, 2017); Duflo et al. (2024)) may not be feasible to scale in middle schools as the complexity of subject matter and range of students' initial learning levels is much greater. In contrast, technology-enabled PAL has proven politically attractive: it enjoys support from politicians, is popular with parents, and—crucially—was implemented with the endorsement of teachers. This broad coalition of support enhances its prospects for scalability within public systems.

More broadly, our results invite reflection on how PAL may enable a re-imagining of education itself. Historically, individualized instruction was an elite privilege, delivered through private tutoring, while public education relied on standardized classroom instruction. While this made mass education fiscally feasible, it came at the cost of curricular mismatch and limited personalization. Indeed, the ability to customize instruction may partly explain the success of high-dosage tutoring (Nickow et al., 2024), but scaling is constrained by tutor cost and availability. Our results suggest that PAL can enable differentiated instruction at scale, potentially reshaping how instruction is delivered in mass education. It also suggests an evolving role for teachers: less focused on uniform content delivery, and more on supporting student engagement with adaptive tools, fostering non-cognitive skills, and creating an environment where personalized learning can flourish. Such an approach can enable technology to complement rather than replace teachers in the classrooms of the future (Autor et al., 2003).

Finally, from a policy perspective, our study illustrates the value of a staged approach to funding both intervention development and evaluation at increasing scales. Advancing evidence-based policy requires more than identifying "what works"; it requires ongoing adaptation and evaluation to make "what works" work at scale. Funding models such as the Global Innovation Fund (GIF), AFD's Fund for Innovation in Development, and USAID's (former) Development Innovation Ventures (DIV) explicitly support such staged, adaptive scaling. Embedding this logic widely could greatly improve the prospects for better evidence enabling sustained impact at scale.

challenge of widespread test-score inflation in administrative data, which may be a key practical constraint in scaling teacher performance-pay. Unconditional pay increases, by contrast, *are* routinely implemented at scale, but have been shown to not improve learning (De Ree et al., 2018).

[46]Voucher policies have also not appeared promising for improving achievement levels. They have been shown to not improve math and local language scores (Muralidharan and Sundararaman, 2015) and, at scale, also appear to severely constrained in effectiveness due to self-selection in take-up (Romero and Singh, 2024).

# References


Accountability Initiative (2021): *School Education Finances: An Overview of Eight States (At a Glance)*, Accountability Initiative, New Delhi.

Al-Ubaydli, O., M. S. Lee, J. A. List, C. L. Mackevicius, and D. Suskind (2021): "How can experiments play a greater role in public policy? Twelve proposals from an economic model of scaling," *Behavioural Public Policy*, 5, 2–49.

Al-Ubaydli, O., J. A. List, and D. L. Suskind (2017): "What can we learn from experiments? Understanding the threats to the scalability of experimental results," *American Economic Review*, 107, 282–86.

Altonji, J. G. and R. K. Mansfield (2018): "Estimating group effects using averages of observables to control for sorting on unobservables: School and neighborhood effects," *American Economic Review*, 108, 2902–2946.

Andrabi, T., N. Bau, J. Das, and A. I. Khwaja (2025): "Heterogeneity in School Value Added and the Private Premium," *American Economic Review*, 115, 147–182.

Andrabi, T., J. Das, A. Ijaz Khwaja, and T. Zajonc (2011): "Do value-added estimates add value? Accounting for learning dynamics," *American Economic Journal: Applied Economics*, 3, 29–54.

Andrews, I. and M. Kasy (2019): "Identification of and correction for publication bias," *American Economic Review*, 109, 2766–2794.

Andrews, M., L. Pritchett, and M. Woolcock (2013): "Escaping capability traps through problem driven iterative adaptation (PDIA)," *World Development*, 51, 234–244.

Angrist, J. D. and P. Hull (2023): "Instrumental variables methods reconcile intention-to-screen effects across pragmatic cancer screening trials," *Proceedings of the National Academy of Sciences*, 120, e2311556120.

Angrist, J. D., P. D. Hull, P. A. Pathak, and C. R. Walters (2017): "Leveraging lotteries for school value-added: Testing and estimation," *The Quarterly Journal of Economics*, 132, 871–919.

Angrist, N., M. Ainomugisha, S. P. Bathena, P. Bergman, C. Crossley, C. Cullen, T. Letsomo, M. Matsheng, R. M. Panti, S. Sabarwal, et al. (2023a): "Building Resilient Education Systems: Evidence from Large-Scale Randomized Trials in Five Countries," Tech. rep., National Bureau of Economic Research, Inc.

Angrist, N., D. K. Evans, D. Filmer, R. Glennerster, F. H. Rogers, and S. Sabarwal (2023b): "How to Improve Education Outcomes Most Efficiently? A Review of the Evidence Using a Unified Metric," *A Review of the Evidence Using a Unified Metric*.

Angrist, N. and R. Meager (2023): "Implementation matters: Generalizing treatment effects in education," *Available at SSRN 4487496*.

Athey, S., R. Chetty, G. W. Imbens, and H. Kang (2019): "The surrogate index: Combining short-term proxies to estimate long-term treatment effects more rapidly and precisely," Tech. rep., National Bureau of Economic Research.

Autor, D. H., F. Levy, and R. J. Murnane (2003): "The skill content of recent technological change: An empirical exploration," *The Quarterly Journal of Economics*, 118, 1279–1333.

Banerjee, A., R. Banerji, J. Berry, E. Duflo, H. Kannan, S. Mukerji, M. Shotland, and M. Walton (2017): "From proof of concept to scalable policies: Challenges and solutions, with an application," *Journal of Economic Perspectives*, 31, 73–102.

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, 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*.

Bau, N. and J. Das (2020): "Teacher value added in a low-income country," *American Economic Journal: Economic Policy*, 12, 62–96.

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, 61–90.

Behrman, J. R., S. W. Parker, P. E. Todd, and K. I. Wolpin (2015): "Aligning learning incentives of students and teachers: Results from a social experiment in Mexican high schools," *Journal of Political Economy*, 123, 325–364.

Bertling, M., A. Singh, and K. Muralidharan (2025): "Psychometric Quality of Measures of Learning Outcomes in Low and Middle Income Countries," in *Handbook of Experimental Development Economics*, Cheltenham, UK: Edward Elgar, 250 – 280.

Bhargava, S. and D. Manoli (2015): "Psychological frictions and the incomplete take-up of social benefits: Evidence from an IRS field experiment," *American Economic Review*, 105, 3489–3529.

Bhatt, M., T. Chau, B. Condliffe, R. Davis, J. Grossman, J. Guryan, J. Ludwig, M. Magnaricotte, S. Mattera, F. Momeni, P. Oreopolous, and G. Stoddard (2025): "Personalized Learning Initiative Interim Report: Findings from 2023-24," Tech. rep., University of Chicago (Education Lab).

Bhatt, M. P., J. Guryan, S. A. Khan, M. LaForest-Tucker, and B. Mishra (2024): "Can Technology Facilitate Scale? Evidence from a Randomized Evaluation of High Dosage Tutoring," Tech. rep., National Bureau of Economic Research, Inc.

Bold, T., M. Kimenyi, G. Mwabu, J. Sandefur, et al. (2018): "Experimental evidence on scaling up education reforms in Kenya," *Journal of Public Economics*, 168, 1–20.

Bryan, G., S. Chowdhury, and A. M. Mobarak (2014): "Underinvestment in a profitable technology: The case of seasonal migration in Bangladesh," *Econometrica*, 82, 1671–1748.

Budish, E., B. N. Roin, and H. Williams (2015): "Do firms underinvest in long-term research? Evidence from cancer clinical trials," *American Economic Review*, 105, 2044–2085.

Bulman, G. and R. W. Fairlie (2016): "Technology and education: Computers, software, and the internet," in *Handbook of the Economics of Education*, Elsevier, vol. 5, 239–280.

Camerer, C. F., A. Dreber, E. Forsell, T.-H. Ho, J. Huber, M. Johannesson, M. Kirchler, J. Almenberg, A. Altmejd, T. Chan, et al. (2016): "Evaluating replicability of laboratory experiments in economics," *Science*, 351, 1433–1436.

Carlana, M. and E. La Ferrara (2024): "Apart But Connected: Online Tutoring, Cognitive Outcomes, and Soft Skills," *NBER Working Paper*.

Chetty, R., J. N. Friedman, and J. E. Rockoff (2014): "Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates," *American Economic Review*, 104, 2593–2632.

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, 295–320.

de Barros, A. (2023): "Explaining the Productivity Paradox: Experimental Evidence from Educational Technology. EdWorkingPaper No. 23-853." *Annenberg Institute for School Reform at Brown University*.

De Chaisemartin, C. and J. Ramirez-Cuellar (2024): "At what level should one cluster standard errors in paired and small-strata experiments?" *American Economic Journal: Applied Economics*, 16, 193–212.

De Ree, J., K. Muralidharan, M. Pradhan, and H. Rogers (2018): "Double for nothing? Experimental evidence on an unconditional teacher salary increase in Indonesia," *The Quarterly Journal of Economics*, 133, 993–1039.

DellaVigna, S. and E. Linos (2022): "RCTs to scale: Comprehensive evidence from two nudge units," *Econometrica*, 90, 81–116.

Dhaliwal, I. and R. Hanna (2017): "The devil is in the details: The successes and limitations of bureaucratic reform in India," *Journal of Development Economics*, 124, 1–21.

Dobbie, W. and R. G. Fryer Jr (2013): "Getting beneath the veil of effective schools: Evidence from New York City," *American Economic Journal: Applied Economics*, 5, 28–60.

Duflo, A., J. Kiessel, and A. M. Lucas (2024): "Experimental Evidence on Four Policies to Increase Learning at Scale," *The Economic Journal*, ueae003.

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, 1739–74.

Escueta, M., A. J. Nickow, P. Oreopoulos, and V. Quan (2020): "Upgrading education with technology: Insights from experimental research," *Journal of Economic Literature*, 58, 897–996.

Evans, D. K. and F. Yuan (2022): "How big are effect sizes in international education studies?" *Educational Evaluation and Policy Analysis*, 01623737221079646.

Ferman, B., L. Finamor, and L. Lima (2019): "Are Public Schools Ready to Integrate Math Classes with Khan Academy?" Tech. rep., University Library of Munich, Germany.

Fryer Jr, R. G. (2017): "The production of human capital in developed countries: Evidence from 196 randomized field experiments," in *Handbook of economic field experiments*, Elsevier, vol. 2, 95–322.

Glewwe, P. and K. Muralidharan (2016): "Improving education outcomes in developing countries: Evidence, knowledge gaps, and policy implications," in *Handbook of the Economics of Education*, Elsevier, vol. 5, 653–743.

Gray-Lobe, G., A. Keats, M. Kremer, I. Mbiti, and O. W. Ozier (2022): "Can education be standardized? Evidence from Kenya," *Evidence from Kenya (September 16, 2022). University of Chicago, Becker Friedman Institute for Economics Working Paper*.

Guryan, J., J. Ludwig, M. P. Bhatt, P. J. Cook, J. M. Davis, K. Dodge, G. Farkas, R. G. Fryer Jr, S. Mayer, H. Pollack, et al. (2023): "Not too late: Improving academic outcomes among adolescents," *American Economic Review*, 113, 738–765.

Jerrim, J. and A. Vignoles (2013): "Social mobility, regression to the mean and the cognitive development of high ability children from disadvantaged homes," *Journal of the Royal Statistical Society Series A: Statistics in Society*, 176, 887–906.

Kerwin, J. T. and R. L. Thornton (2021): "Making the grade: The sensitivity of education program effectiveness to input choices and outcome measures," *Review of Economics and Statistics*, 103, 251–264.

Kraft, M. A., J. A. List, J. A. Livingston, and S. Sadoff (2022): "Online tutoring by college volunteers: Experimental evidence from a pilot program," in *AEA Papers and Proceedings*, American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203, vol. 112, 614–618.

Kraft, M. A., B. E. Schueler, and G. Falken (2024): "What Impacts Should We Expect from Tutoring at Scale? Exploring Meta-Analytic Generalizability. EdWorkingPaper No. 24-1031." *Annenberg Institute for School Reform at Brown University*.

Leaver, C., O. Ozier, P. Serneels, and A. Zeitlin (2021): "Recruitment, effort, and retention effects of performance contracts for civil servants: Experimental evidence from Rwandan primary schools," *American Economic Review*, 111, 2213–2246.

Linden, L. L. (2008): *Complement or substitute?: The effect of technology on student achievement in India*, Citeseer.

LINOS, E., A. PROHOFSKY, A. RAMESH, J. ROTHSTEIN, AND M. UNRATH (2022): "Can nudges increase take-up of the EITC? Evidence from multiple field experiments," *American Economic Journal: Economic Policy*, 14, 432–452.

LIST, J. A. (2022): *The voltage effect: How to make good ideas great and great ideas scale*, Currency.

——— (2024): "Optimally generate policy-based evidence before scaling," *Nature*, 626, 491–499.

MALAMUD, O. AND C. POP-ELECHES (2011): "Home computer use and the development of human capital," *The Quarterly journal of economics*, 126, 987–1027.

MBITI, I., K. MURALIDHARAN, M. ROMERO, Y. SCHIPPER, C. MANDA, AND R. RAJANI (2019): "Inputs, incentives, and complementarities in education: Experimental evidence from Tanzania," *The Quarterly Journal of Economics*, 134, 1627–1673.

MITCHELL, H., A. M. MOBARAK, K. NAGUIB, M. E. REIMÃO, AND A. SHENOY (2023): "Delegation risk and implementation at scale: Evidence from a migration loan program in Bangladesh," *Unpublished manuscript. https://ashishenoy. github. io/Website/Paper_NLS_Evaluation. pdf*.

MURALIDHARAN, K. (2012): "Long-Term Effects of Teacher Performance Pay: Experimental Evidence from India," *Unpublished manuscript*.

——— (2024): *Accelerating India's Development: A State-led Roadmap for Effective Governance*, Penguin Viking.

MURALIDHARAN, K. AND P. NIEHAUS (2017): "Experimentation at scale," *Journal of Economic Perspectives*, 31, 103–24.

MURALIDHARAN, K. AND A. SINGH (2021): "India's new national education policy: Evidence and challenges," *Science*, 372, 36–38.

MURALIDHARAN, K., A. SINGH, AND A. J. GANIMIAN (2019): "Disrupting education? Experimental evidence on technology-aided instruction in India," *American Economic Review*, 109, 1426–60.

MURALIDHARAN, K. AND V. SUNDARARAMAN (2011): "Teacher performance pay: Experimental evidence from India," *Journal of Political Economy*, 119, 39–77.

——— (2015): "The Aggregate Effect of School Choice: Evidence from a Two-Stage Experiment in India *," *The Quarterly Journal of Economics*, 130, 1011–1066.

NADEL, S. AND L. PRITCHETT (2016): "Searching for the Devil in the Details: Learning about Development Program Design," *Center for Global Development working paper*.

NICKOW, A., P. OREOPOULOS, AND V. QUAN (2024): "The promise of tutoring for PreK–12 learning: A systematic review and meta-analysis of the experimental evidence," *American Educational Research Journal*, 61, 74–107.

Rodriguez-Segura, D. (2022): "EdTech in developing countries: A review of the evidence," *The World Bank Research Observer*, 37, 171–203.

Romero, M., J. Sandefur, and W. A. Sandholtz (2020): "Outsourcing education: Experimental evidence from Liberia," *American Economic Review*, 110, 364–400.

Romero, M. and A. Singh (2024): "The incidence of affirmative action: Evidence from quotas in private schools in India," *Working Paper*.

Singh, A. (2015): "Private school effects in urban and rural India: Panel estimates at primary and secondary school ages," *Journal of Development Economics*, 113, 16–32.

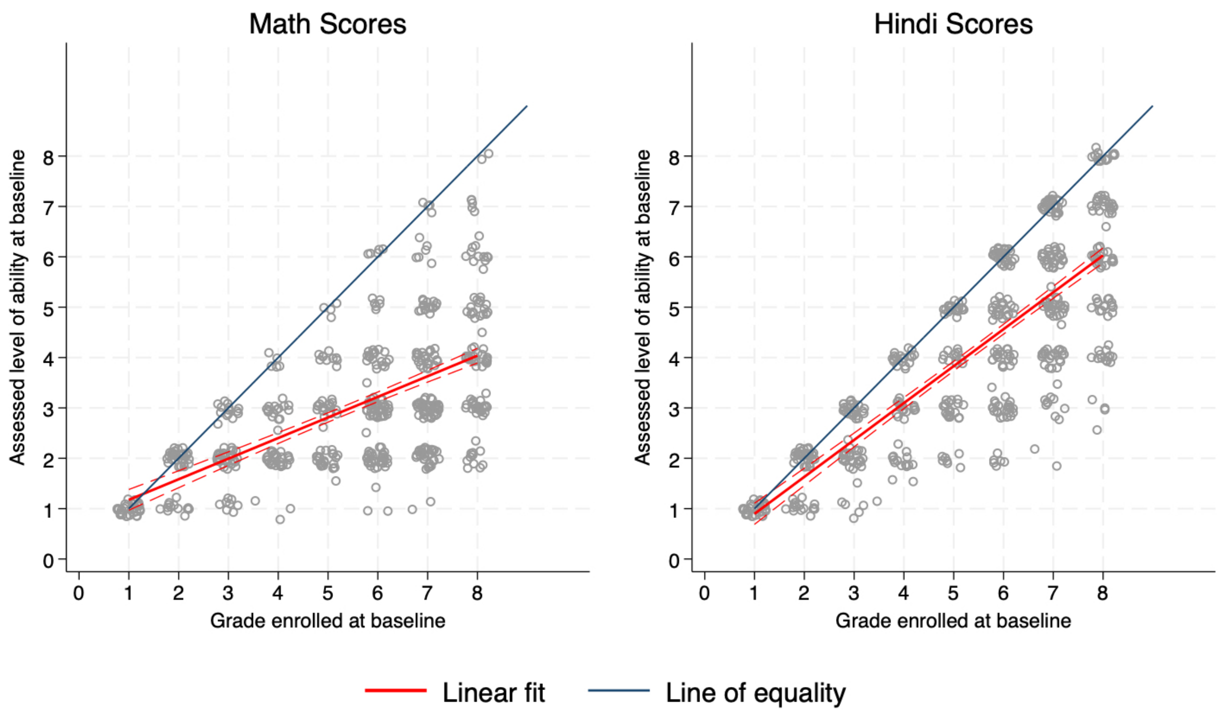
——— (2020): "Learning more with every year: School year productivity and international learning divergence," *Journal of the European Economic Association*, 18, 1770–1813.

——— (2024): "Improving administrative data at scale: Experimental evidence on digital testing in Indian schools," *The Economic Journal*, 134, 2207–2223.

Singh, A. and P. Berg (2024): "Myths of official measurement: Limits to test-based education reforms with weak governance," *Journal of Public Economics*, 239, 105246.

World Bank (2017): *World Development Report 2018: Learning to realize education's promise*, The World Bank.

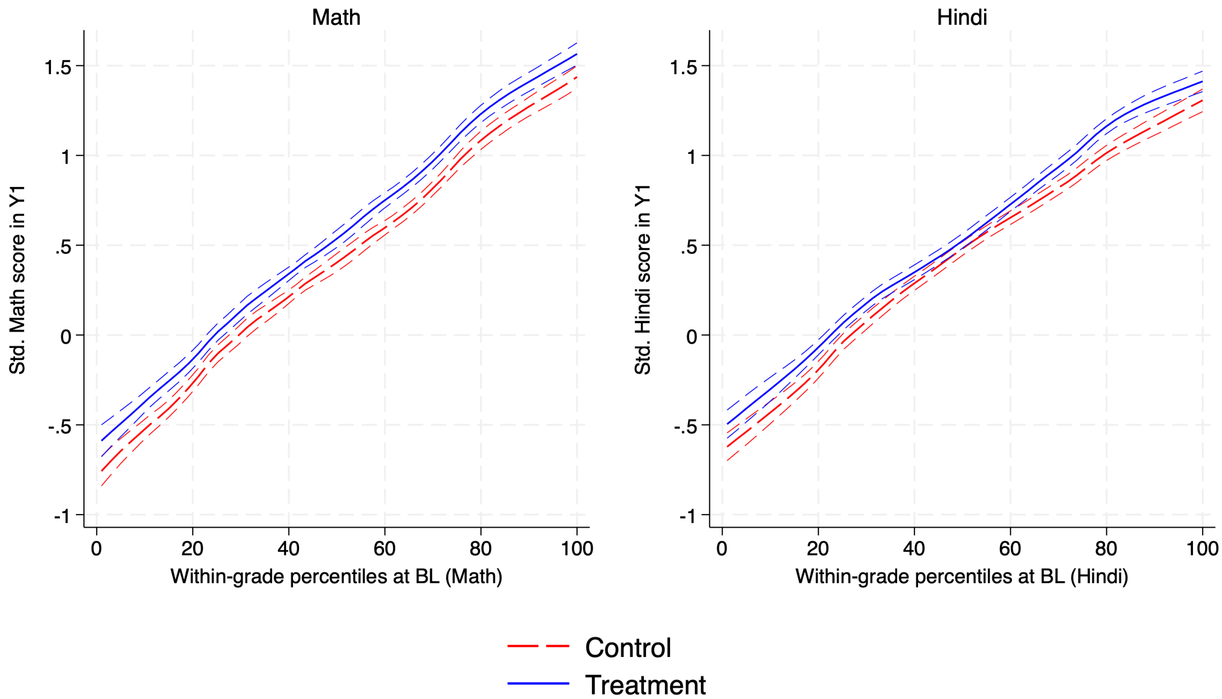
Figure 1: Assessed levels of student achievement versus grade enrolled in at baseline



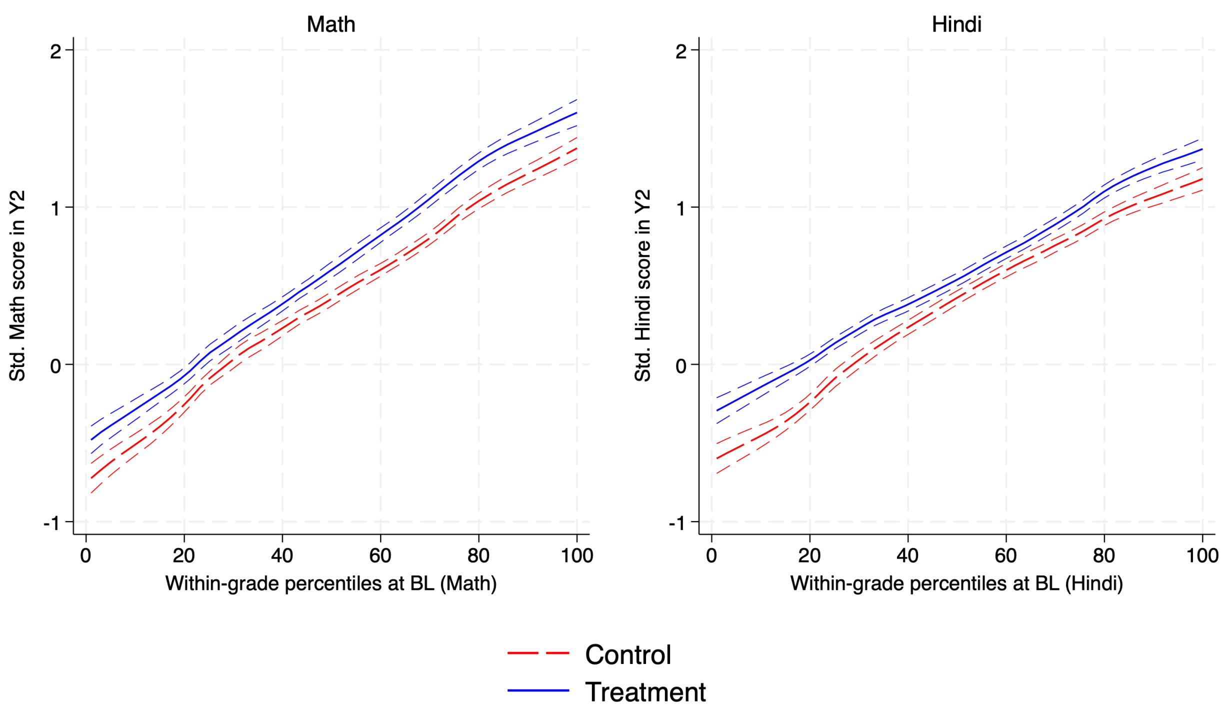

*Note:* This figure shows, for treatment group, the estimated level of student achievement (determined by the Mindspark CAL program) plotted against the grade they are enrolled in. These data are from the initial diagnostic test, and do not reflect any instruction provided by Mindspark. Each marker represents 10 students — markers have been jittered for legibility. In both subjects, we find three main patterns: (i) there is a general deficit between average attainment and grade-expected norms; (ii) this deficit is larger in later grades; and (iii) within each grade, there is a wide dispersion of student achievement. Deficits appear to be larger in mathematics than Hindi.

Figure 2: Nonparametric investigation of treatment effect by baseline percentiles

**Year 1**

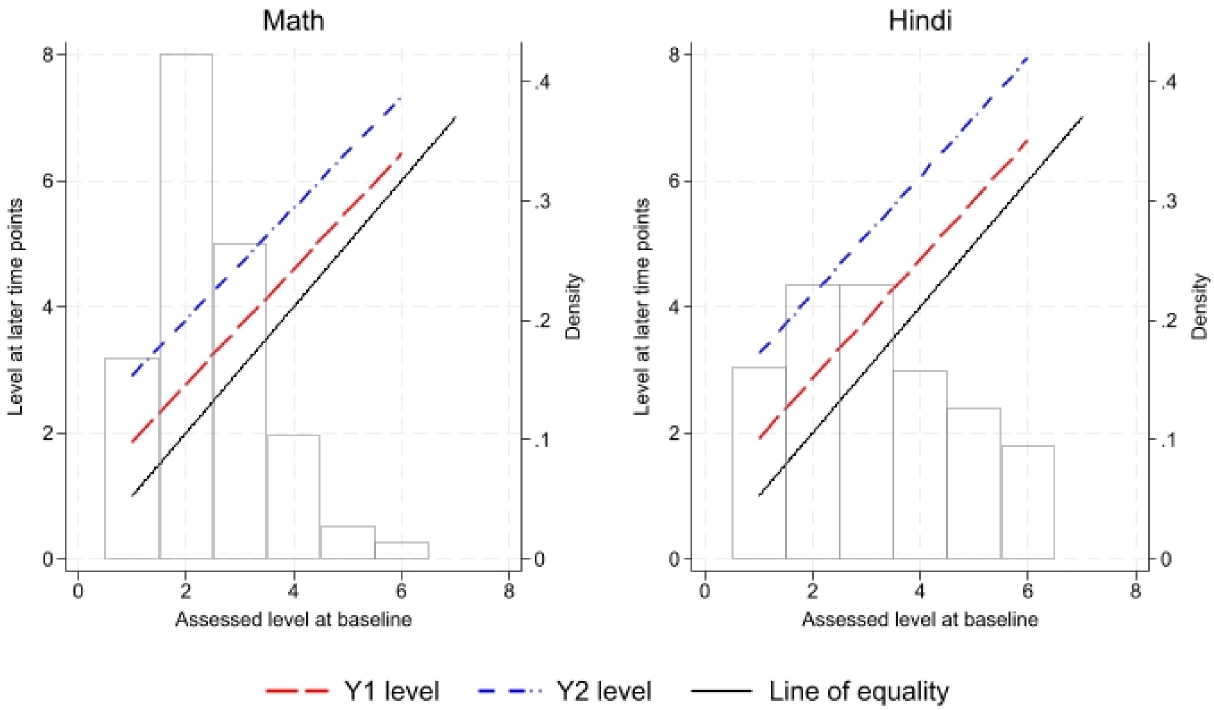



**Year 2**



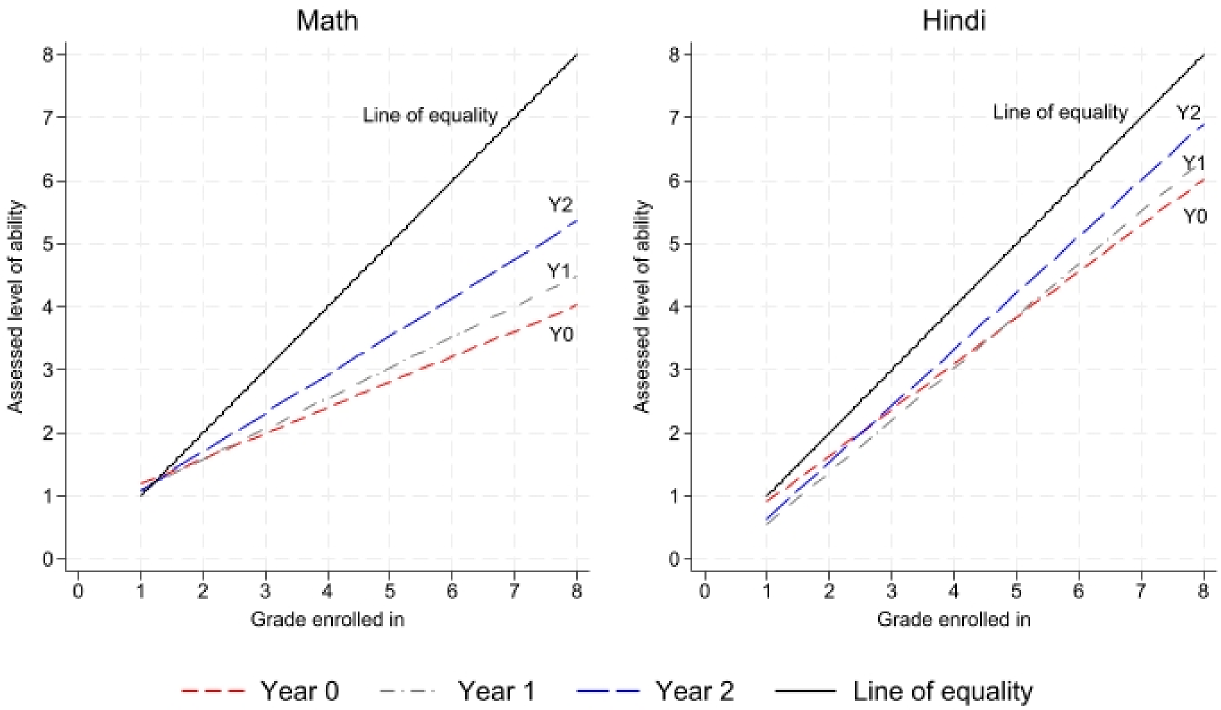

*Note:* The figures present local polynomial regressions which relate endline test scores to within-grade percentiles of baseline achievement, separately for the treatment and control groups, alongside 95 percent confidence intervals. Across the achievement distribution, treatment group students score higher in the endline tests than the control group.

Figure 3: Learning progress in treated schools after 2 years




*Note:* The figures present, for students in the treatment group, the line of best fit between their academic ability, as assessed by the Mindspark system at baseline (Nov 2017, X-axis) and at the beginning of the subsequent academic years (Jul-Aug 2018 and 2019, Y-axis). The sample includes only those students who were observed in all three years. The histogram shows the distribution of assessed grade level at baseline for this sample (using the Y-axis on the right).

Figure 4: Change in academic mismatch in treated schools after 2 years




*Note:* This figure shows, for the treatment group, a linear fit of the estimated level of student achievement (determined by the Mindspark CAL program) plotted against the grade they are enrolled in. These tests were administered to all primary and middle school students in treated schools in each academic year. Y1 refers to the baseline diagnostic assessment (Nov 2017), Y2 to early in the second year (July 2018) and Y3 to early in the third year (Aug 2019). The principal result is that academic mismatch declines in treated schools with the line of best fit pivoting closer to the pace of progress expected by the curriculum (the line of equality) in later years.

Table 1: Balance at baseline on school and student characteristics

| Variable | (1) Treatment Mean/(SE) | (2) Control Mean/(SE) | (1)-(2) p-value |
|---|---|---|---|
| | **Panel A: School Characteristics** | | |
| Enrolment: Primary School (Grades 1-5) | 68.28 | 72.10 | 0.75 |
| | (6.06) | (6.15) | |
| Enrolment: Middle School (Grades 6-8) | 93.22 | 90.00 | 0.70 |
| | (6.53) | (7.31) | |
| Total Enrollment (Grades 1-12) | 367.50 | 392.55 | 0.56 |
| | (27.26) | (26.97) | |
| Total Teachers | 14.75 | 15.95 | 0.29 |
| | (0.61) | (0.45) | |
| Number of observations | 40 | 40 | 80 |
| Number of clusters | 40 | 40 | 40 |
| | **Panel B: Student Characteristics** | | |
| Baseline score (Math) † | 0.10 | 0.10 | 0.88 |
| | (0.06) | (0.08) | |
| Baseline score (Hindi)† | 0.20 | 0.17 | 0.67 |
| | (0.05) | (0.06) | |
| Female | 0.47 | 0.42 | 0.02** |
| | (0.02) | (0.03) | |
| Socioeconomic Status Index (PCA) | -0.01 | -0.00 | 0.93 |
| | (0.06) | (0.07) | |
| Enrolled in primary grade | 0.41 | 0.41 | 0.59 |
| | (0.02) | (0.02) | |
| Enrolled in middle school grade | 0.59 | 0.59 | 0.59 |
| | (0.02) | (0.02) | |
| Follow-up rate: BL to Y1 | 0.67 | 0.69 | 0.32 |
| | (0.03) | (0.02) | |
| Follow-up rate: BL to Y2 ‡ | 0.65 | 0.64 | 0.91 |
| | (0.02) | (0.02) | |
| Number of observations | 4783 | 4803 | 9586 |
| Number of clusters | 40 | 40 | 40 |

*Notes*: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Standard errors are clustered at the stratum level.
School characteristics are measured using pre-program administrative data, student characteristics are measured using independently-collected baseline data.
†Test scores are standardized to have $\mu$=0, $\sigma$=1 in grade 5 in the control group. SES scores are standardized to $\mu$=0, $\sigma$=1 at baseline in the control group. Regressions underlying column (3) include stratum fixed effects.
‡Follow-up rates in Y2 exclude students enrolled in Grade 8 in baseline (since they exit the study in subsequent years).

Table 2: Effects on student test scores

| | Math | | | Hindi | | |
|---|---|---|---|---|---|---|
| | Pooled Gr. 1-8 | Primary Gr. 1-5 | Middle Gr. 6-8 | Pooled Gr. 1-8 | Primary Gr. 1-5 | Middle Gr. 6-8 |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| | | | **Panel A: Year 1** | | | |
| Treatment | 0.15*** | 0.11* | 0.15*** | 0.11*** | 0.09* | 0.11*** |
| | (0.03) | (0.06) | (0.03) | (0.03) | (0.05) | (0.02) |
| Baseline score | 0.67*** | 0.65*** | 0.65*** | 0.72*** | 0.64*** | 0.70*** |
| | (0.01) | (0.02) | (0.01) | (0.01) | (0.02) | (0.01) |
| Observations | 7994 | 3329 | 4665 | 7994 | 3329 | 4665 |
| R-squared | 0.53 | 0.40 | 0.53 | 0.51 | 0.36 | 0.50 |
| | | | **Panel B: Year 2** | | | |
| Treatment | 0.22*** | 0.15*** | 0.25*** | 0.20*** | 0.20*** | 0.15*** |
| | (0.036) | (0.054) | (0.038) | (0.032) | (0.043) | (0.036) |
| Baseline score | 0.59*** | 0.39*** | 0.66*** | 0.65*** | 0.41*** | 0.62*** |
| | (0.016) | (0.039) | (0.021) | (0.015) | (0.033) | (0.023) |
| Observations | 8733 | 3716 | 5017 | 8733 | 3716 | 5017 |
| R-squared | 0.35 | 0.19 | 0.36 | 0.38 | 0.20 | 0.37 |
| ΔY in control | 0.47 | 0.67 | 0.25 | 0.31 | 0.38 | 0.23 |
| Effect/ΔY in control | 0.47 | 0.22 | 1 | 0.65 | 0.53 | 0.64 |

*Notes*: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Standard errors are clustered at the stratum level.
This table presents Intention-to-treat treatment effects of the program at the end of ∼6 months of treatment (EL Y1) and 18 months (EL Y2). Test scores are based on independent tests conducted in class for all students present on the day of testing at baseline (July 2017) and each end-of-year assessment (in February of each academic year). All regressions control for gender and strata (school-pair) fixed effects. Test scores are linked across rounds and across grades using Item Response Theory models, and standardized to have a mean of zero and a standard deviation of one in the baseline in grade 5 in the control group. For students who were absent on the day of the baseline test, we replace the baseline score with the classroom average. ΔY refers to the within-person change in test scores for students between baseline and end-of-year assessments in Year 2.

Table 3: Heterogeneity by within-grade quintiles of student achievement

| | Math | | Hindi | |
|---|---|---|---|---|
| | Year 1 | Year 2 | Year 1 | Year 2 |
| Treatment | 0.13** | 0.19** | 0.15** | 0.28*** |
| | (0.053) | (0.076) | (0.057) | (0.064) |
| Treatment x Q2 | 0.04 | -0.01 | -0.04 | -0.07 |
| | (0.056) | (0.078) | (0.070) | (0.065) |
| Treatment x Q3 | 0.02 | 0.00 | -0.08 | -0.14* |
| | (0.055) | (0.079) | (0.066) | (0.076) |
| Treatment x Q4 | 0.00 | 0.08 | -0.04 | -0.12 |
| | (0.056) | (0.089) | (0.066) | (0.073) |
| Treatment x Q5 | -0.00 | 0.04 | -0.07 | -0.20** |
| | (0.068) | (0.089) | (0.073) | (0.093) |
| Quintile 2 | 0.23*** | 0.20*** | 0.33*** | 0.25*** |
| | (0.051) | (0.050) | (0.046) | (0.044) |
| Quintile 3 | 0.48*** | 0.43*** | 0.58*** | 0.54*** |
| | (0.060) | (0.054) | (0.047) | (0.053) |
| Quintile 4 | 0.72*** | 0.64*** | 0.85*** | 0.78*** |
| | (0.076) | (0.066) | (0.047) | (0.048) |
| Quintile 5 | 0.90*** | 0.83*** | 1.02*** | 1.05*** |
| | (0.092) | (0.079) | (0.062) | (0.071) |
| F-test that interaction terms equal zero (p-val) | .93 | .76 | .76 | .21 |
| Observations | 6539 | 4825 | 6539 | 4825 |
| R-squared | 0.643 | 0.549 | 0.610 | 0.577 |

*Notes*: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Standard errors are clustered at the stratum level. Quintiles of student achievement are based on their baseline test scores. We exclude students who did not take the baseline test. Test scores are standardized to have a mean of zero and a standard deviation of one in the baseline in grade 5 in the control group. All regressions control for baseline test scores, student gender and fixed effects for randomization strata and the grade enrolled in. The F-test reported tests whether each of the interaction terms is different from each other and from zero.

Table 4: Heterogeneity in effect on hardest/easiest items by within-grade BL score terciles

| | Easy items | | Hard items | |
|---|---|---|---|---|
| | Year 1 | Year 2 | Year 1 | Year 2 |
| **Math** | | | | |
| Treatment | 0.05*** | 0.06*** | 0.02 | 0.04*** |
| | (0.014) | (0.016) | (0.010) | (0.011) |
| Treatment x middle tercile | -0.03** | -0.02 | 0.02 | 0.03** |
| | (0.011) | (0.016) | (0.011) | (0.012) |
| Treatment x top tercile | -0.04** | -0.04** | 0.03* | 0.05*** |
| | (0.016) | (0.016) | (0.014) | (0.014) |
| Middle tercile | 0.16*** | 0.14*** | 0.02* | 0.01 |
| | (0.011) | (0.014) | (0.011) | (0.011) |
| Top tercile | 0.16*** | 0.18*** | 0.13*** | 0.08*** |
| | (0.014) | (0.016) | (0.016) | (0.013) |
| Mean percentage correct in bottom tercile | .64 | .56 | .16 | .18 |
| Observations | 6538 | 4825 | 6537 | 4825 |
| R-squared | 0.426 | 0.464 | 0.363 | 0.278 |
| **Hindi** | | | | |
| Treatment | 0.06*** | 0.10*** | 0.02** | 0.03*** |
| | (0.015) | (0.018) | (0.011) | (0.011) |
| Treatment x middle tercile | -0.04** | -0.06*** | 0.02 | -0.00 |
| | (0.016) | (0.017) | (0.014) | (0.013) |
| Treatment x top tercile | -0.05*** | -0.08*** | 0.02 | 0.00 |
| | (0.017) | (0.020) | (0.013) | (0.015) |
| Middle tercile | 0.12*** | 0.16*** | 0.11*** | 0.05*** |
| | (0.011) | (0.015) | (0.012) | (0.012) |
| Top tercile | 0.10*** | 0.18*** | 0.27*** | 0.18*** |
| | (0.015) | (0.020) | (0.013) | (0.014) |
| Mean percentage correct in bottom tercile | .74 | .67 | .28 | .22 |
| Observations | 6535 | 4825 | 6534 | 4825 |
| R-squared | 0.374 | 0.400 | 0.547 | 0.380 |

*Notes*: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Standard errors are clustered at the stratum level.
Terciles of question difficulty are defined at the grade-round level, and terciles of student achievement on their baseline scores. We exclude students who did not take the baseline test. The dependent variables is the proportion of questions correctly answered in the hardest/easiest terciles in a given round of testing. All regressions control for baseline scores, gender, and fixed effects for randomization strata and grade.

Table 5: Reduction in academic mismatch in the treatment group

| | *Dep var:* Assessed level of achievement | | | | | |
|---|---|---|---|---|---|---|
| | Math | Hindi | Math | Hindi | Math | Hindi |
| Enrolled grade | 0.41*** | 0.73*** | 0.41*** | 0.74*** | | |
| | (0.02) | (0.02) | (0.02) | (0.02) | | |
| Enrolled grade x Y1 | 0.08*** | 0.09*** | 0.09*** | 0.10*** | 0.10*** | 0.10*** |
| | (0.01) | (0.02) | (0.01) | (0.02) | (0.01) | (0.02) |
| Enrolled grade x Y2 | 0.20*** | 0.16*** | 0.21*** | 0.17*** | 0.22*** | 0.17*** |
| | (0.02) | (0.01) | (0.02) | (0.01) | (0.02) | (0.02) |
| Year FE | Y | Y | Y | Y | Y | Y |
| School FE | | | Y | Y | Y | Y |
| Grade FE | | | | | Y | Y |
| Observations | 16,956 | 17,116 | 16,956 | 17,116 | 16,956 | 17,116 |
| R-squared | 0.50 | 0.70 | 0.54 | 0.72 | 0.54 | 0.72 |

*Notes*: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Standard errors are clustered at the school level. This table presents the regression analog of Figure 4. Pooling data from the diagnostic Mindspark assessment from each of three academic years, we examine whether academic mismatch in treated schools declines over time. In Y1 and Y2, the slope between assessed and enrolled grade is significantly steeper. This indicates that students came closer to curricular levels across the sample. This relationship is robust to adding year, school and grade fixed effects.

Table 6: Treatment effects in restricted sample with no change in instructional time

| | Math | | Hindi | |
|---|---|---|---|---|
| | Year 1 | Year 2 | Year 1 | Year 2 |
| Treatment | 0.193*** | 0.213*** | 0.147*** | 0.176** |
| | (0.0522) | (0.0730) | (0.0429) | (0.0656) |
| Baseline score | 0.636*** | 0.637*** | 0.724*** | 0.636*** |
| | (0.0208) | (0.0357) | (0.0262) | (0.0349) |
| Observations | 1956 | 2120 | 2056 | 2399 |
| R-squared | 0.555 | 0.372 | 0.541 | 0.434 |

*Notes*: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Standard errors are clustered at the stratum level. This table presents intention-to-treat estimates of the treatment effect, restricted to only those grade×stratum pairs where the program did not increase scheduled instruction in the relevant targeted subject. The specification is identical to Table 2. All regressions include baseline achievement, strata (school-pair) fixed effects and student gender. Standard errors are clustered at the stratum level.

Table 7: Treatment effect on school examinations, math and Hindi (year 2)

| | Half-year examinations | | Board examinations | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | Grade A and above | | Grade B and above | | Grade C and above | |
| | Math | Hindi | Math | Hindi | Math | Hindi | Math | Hindi |
| **Grade 8** | | | | | | | | |
| Treatment | 0.95 | 0.35 | -0.02 | -0.06 | -0.04 | -0.06 | -0.07 | -0.02 |
| | (2.651) | (2.279) | (0.041) | (0.040) | (0.056) | (0.036) | (0.042) | (0.025) |
| Baseline score | 2.35*** | 2.60*** | 0.10*** | 0.21*** | 0.15*** | 0.25*** | 0.10*** | 0.12*** |
| | (0.531) | (0.562) | (0.014) | (0.020) | (0.013) | (0.017) | (0.018) | (0.021) |
| Mean score | 88.66 | 89.7 | .13 | .29 | .35 | .62 | .76 | .9 |
| Observations | 1010 | 1040 | 1496 | 1496 | 1496 | 1496 | 1496 | 1496 |
| R-squared | 0.615 | 0.660 | 0.279 | 0.235 | 0.335 | 0.288 | 0.366 | 0.235 |
| **Grade 5** | | | | | | | | |
| Treatment | | | 0.00 | -0.08 | -0.06* | -0.03 | -0.01 | -0.01 |
| | | | (0.079) | (0.057) | (0.034) | (0.046) | (0.005) | (0.013) |
| Mean score | | | .63 | .56 | .93 | .83 | 1 | .98 |
| Observations | | | 1189 | 1192 | 1189 | 1192 | 1189 | 1192 |
| R-squared | | | 0.403 | 0.321 | 0.257 | 0.372 | 0.052 | 0.122 |

*Notes*: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Standard errors are clustered at the stratum level.
For half-year examinations, the dependent variable is the score obtained at the half year school examinations (scores between 0 and 100). For board examinations, the dependent variable is a dummy variable for obtaining the grade of interest or above. Stratum fixed effects are included in all regressions. Gender and baseline test scores are controlled for in Grade 8 but not Grade 5 due to lower match rates from the administrative rosters to our data.

Table 8: Value-added estimates of program effectiveness in Years 2 and 3

| Variable | Math | | Hindi | |
|---|---|---|---|---|
| | Year 2 | Year 3 | Year 2 | Year 3 |
| Treatment | 0.14*** | 0.10** | 0.12*** | 0.07 |
| | (0.035) | (0.045) | (0.031) | (0.041) |
| Constant | 0.25*** | 0.24*** | 0.21*** | 0.21*** |
| | (0.018) | (0.026) | (0.018) | (0.023) |
| Grade FE | Yes | Yes | Yes | Yes |
| Stratum FE | Yes | Yes | Yes | Yes |
| Observations | 8131 | 10297 | 8131 | 10297 |
| R-squared | 0.353 | 0.357 | 0.397 | 0.265 |

*Notes*: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Standard errors are clustered at the stratum level. This table shows treatment effects when controlling for previous year's test score. Test scores are standardized to have a mean of zero and a standard deviation of one in the baseline in grade 5 in the control group. If a child's lagged test score is missing, it is replaced by the average of the class she would have attended in the previous round (average score in the same school, in the grade preceding her actual grade, in the previous round).

Table 9: Comparing dose response of Mindspark usage across Years 2 and 3

| Variable | Math | | Hindi | |
|---|---|---|---|---|
| | Year 2 | Year 3 | Year 2 | Year 3 |
| Mindspark Usage (10 hrs) | 0.07*** | 0.07*** | 0.06*** | 0.08*** |
| | (0.015) | (0.017) | (0.014) | (0.018) |
| Lagged test scores | Yes | Yes | Yes | Yes |
| Stratum FE | Yes | Yes | Yes | Yes |
| Class FE | Yes | Yes | Yes | Yes |
| Average Mindspark usage (10 hrs) | 3.16 | 2.93 | 2.71 | 2.18 |
| Observations | 1942 | 2027 | 1939 | 2026 |
| R-squared | 0.572 | 0.617 | 0.596 | 0.486 |

*Notes*: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Standard errors are clustered at the school level.
This table shows the association between a child's usage of Mindspark, measured in units of 10 hours, and their end-of-year scores, controlling for previous year's test score, gender, grade fixed effects and school fixed effects. Students whose test scores are not observed in the previous year are excluded from the regression. Test scores are standardized to have a mean of zero and a standard deviation of one in the baseline in grade 5 in the control group. The results do not suggest a weakening of the effect of Mindspark usage across years. For year 3, we only include students who took the exams in school.

Table 10: Teacher opinions about Mindspark usefulness

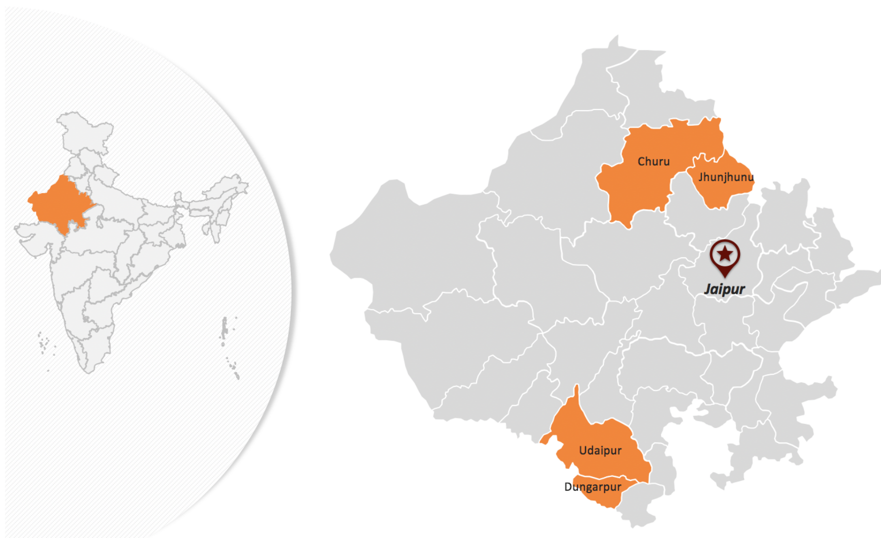
| | Subject-wise | | Grade-wise | | Aggregate |
|---|---|---|---|---|---|
| Variable | Hindi | Math | Grade 5 | Grade 8 | |
| Mindspark useful | 1 | 0.97 | 0.99 | 0.99 | 0.99 |
| | [0] | [0.02] | [0.01] | [0.01] | [0.01] |
| **If yes, why was it useful?** | | | | | |
| Student learning has improved substantially | 0.88 | 0.89 | 0.87 | 0.9 | 0.88 |
| | [0.04] | [0.04] | [0.04] | [0.04] | [0.03] |
| Students understand things better | 0.81 | 0.71 | 0.81 | 0.71 | 0.76 |
| | [0.05] | [0.05] | [0.05] | [0.05] | [0.04] |
| Students perform well in exams | 0.26 | 0.39 | 0.22 | 0.43 | 0.32 |
| | [0.05] | [0.06] | [0.05] | [0.06] | [0.04] |
| I do not know why | 0 | 0.01 | 0 | 0.01 | 0.01 |
| | [0] | [0.01] | [0] | [0.01] | [0.01] |
| Other | .07 | 0.07 | 0.09 | 0.06 | 0.07 |
| | [0.03] | [0.03] | [0.03] | [0.03] | [0.02] |
| N | 69 | 72 | 70 | 71 | 141 |

*Notes*: This table displays means and standard errors in parentheses. It presents summary statistics about teacher opinions about Mindspark two years after the program was introduced in the treatment schools. These survey questions were administered to mathematics and Hindi teachers in Grades 5 and 8 in treated schools. Teachers report finding Mindspark useful, primarily for increasing student knowledge and comprehension. The proportion reporting improvement in school exams is much lower, likely reflecting that the Mindspark program typically presented instructional material that was much below grade level to students.

# Appendix

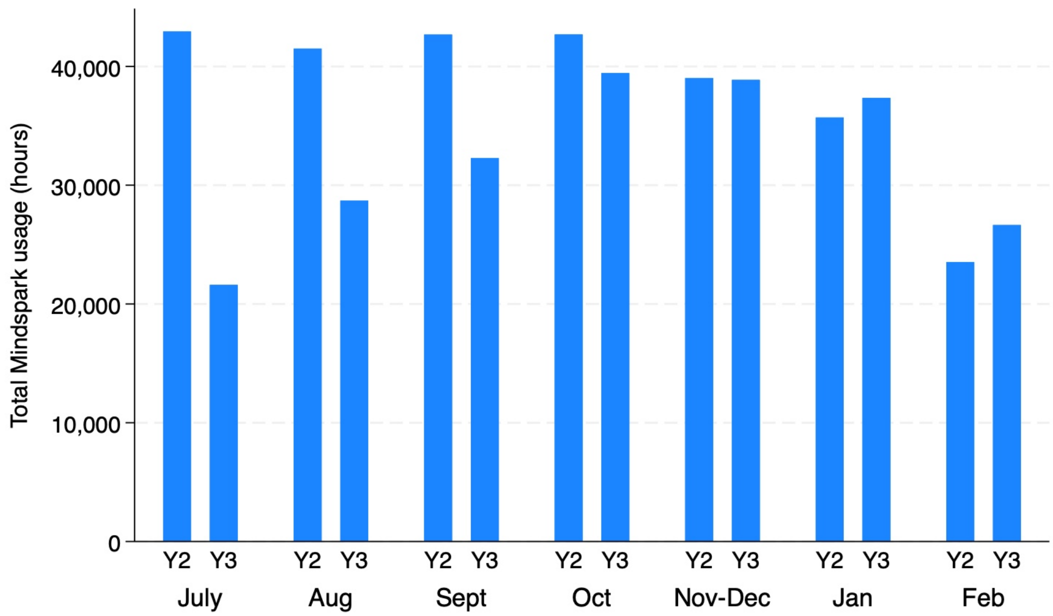
## A Further Tables and Figures

Figure A.1: Map of study districts



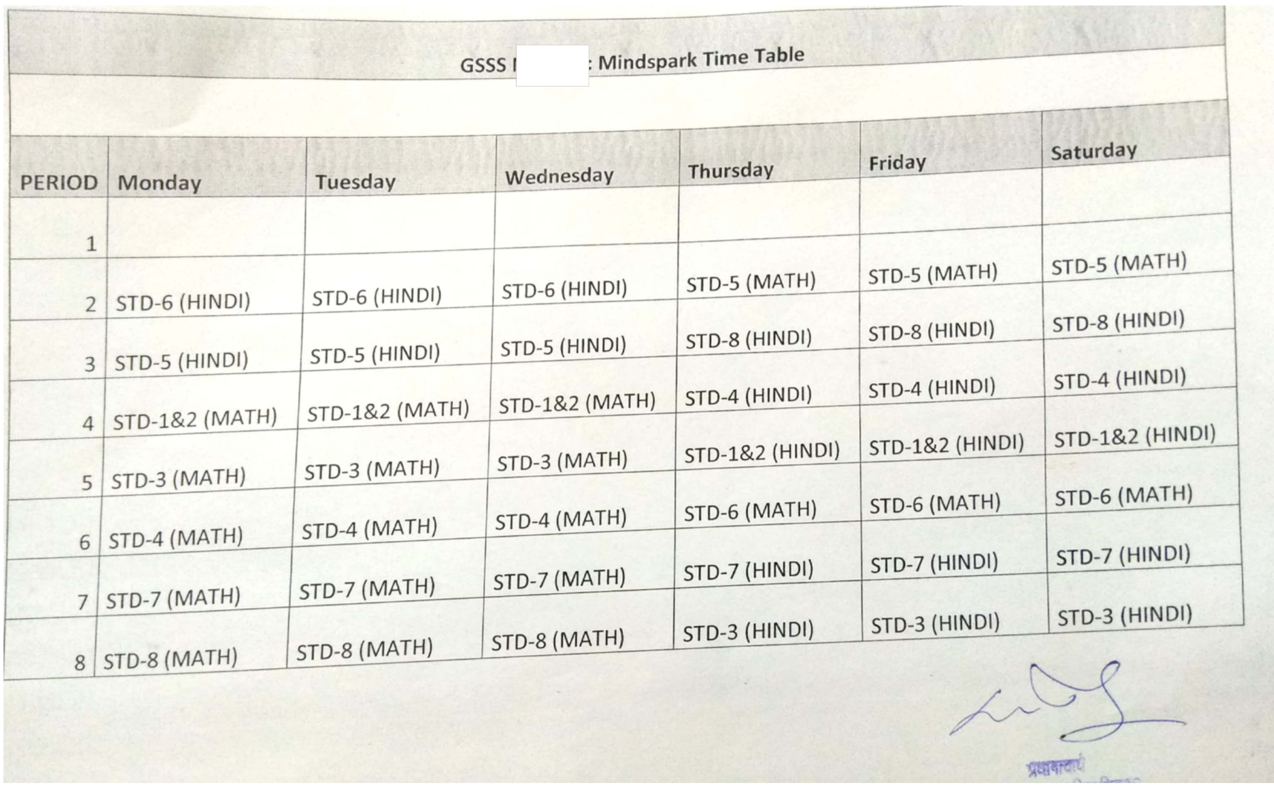

*Note:* The study was conducted in 80 schools spread across 4 districts — Churu and Jhunjhunun in northern Rajasthan, and Udaipur and Dungarpur in southern Rajasthan — which are highlighted in this map.

Figure A.2: Total usage of Mindspark in treated schools in Years 2 and 3




*Notes:* This graph presents total monthly usage of Mindspark, cumulated over both subjects and all students in treated schools, in Y2 and Y3. November and December usage is combined due to the timing of school holidays across years. Usage in February is only included until February 15 (around the time of our end-of-year data collection).

Figure A.3: Sample time-table of Mindspark lab



GSSS : Mindspark Time Table

| PERIOD | Monday | Tuesday | Wednesday | Thursday | Friday | Saturday |
|---|---|---|---|---|---|---|
| 1 | | | | | | |
| 2 | STD-6 (HINDI) | STD-6 (HINDI) | STD-6 (HINDI) | STD-5 (MATH) | STD-5 (MATH) | STD-5 (MATH) |
| 3 | STD-5 (HINDI) | STD-5 (HINDI) | STD-5 (HINDI) | STD-8 (HINDI) | STD-8 (HINDI) | STD-8 (HINDI) |
| 4 | STD-1&2 (MATH) | STD-1&2 (MATH) | STD-1&2 (MATH) | STD-4 (HINDI) | STD-4 (HINDI) | STD-4 (HINDI) |
| 5 | STD-3 (MATH) | STD-3 (MATH) | STD-3 (MATH) | STD-1&2 (HINDI) | STD-1&2 (HINDI) | STD-1&2 (HINDI) |
| 6 | STD-4 (MATH) | STD-4 (MATH) | STD-4 (MATH) | STD-6 (MATH) | STD-6 (MATH) | STD-6 (MATH) |
| 7 | STD-7 (MATH) | STD-7 (MATH) | STD-7 (MATH) | STD-7 (HINDI) | STD-7 (HINDI) | STD-7 (HINDI) |
| 8 | STD-8 (MATH) | STD-8 (MATH) | STD-8 (MATH) | STD-3 (HINDI) | STD-3 (HINDI) | STD-3 (HINDI) |

प्रधानाचार्य

*Notes:* This is the schedule for usage of the Mindspark lab in one sample school. Since enrollment in grades 1 and 2 were typically lower in the Adarsh schools, they were often scheduled jointly in the Mindspark Lab period. In the example above, the first period was the home room period, after which the Mindspark lab was used fully for a total of 42 periods each week.

Table A.1: Balance on observed student characteristics, by round

| | Year 1 | | Year 2 | | Year 3 | |
|---|---|---|---|---|---|---|
| | Control mean (1) | Treatment difference (2) | Control mean (3) | Treatment difference (4) | Control mean (5) | Treatment difference (6) |
| Baseline score Math | 0.25 | 0.02 | -0.04 | 0.03 | -0.36 | 0.03 |
| | | (0.09) | | (0.09) | | (0.08) |
| Baseline score Hindi | 0.30 | 0.03 | 0.11 | 0.03 | -0.17 | 0.03 |
| | | (0.07) | | (0.08) | | (0.07) |
| Female | 0.44 | 0.07*** | 0.44 | 0.08*** | 0.44 | 0.07** |
| | | (0.03) | | (0.03) | | (0.03) |
| SES score | 0.08 | -0.01 | 0.03 | 0.02 | 0.02 | 0.04 |
| | | (0.05) | | (0.05) | | (0.06) |
| Enrollment in primary school | 0.40 | -0.01 | 0.39 | -0.03 | 0.36 | -0.04 |
| | | (0.03) | | (0.03) | | (0.02) |
| Enrollment in middle school | 0.60 | 0.01 | 0.61 | 0.03 | 0.64 | 0.04* |
| | | (0.03) | | (0.03) | | (0.02) |
| Observations | 3333 | 6539 | 2392 | 4825 | 1760 | 3520 |

*Notes*: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Standard errors are clustered at the school level and shown in parentheses. Stratum fixed effects are included in treatment effect regressions.
This table shows, for students who took the baseline test in 2017, the observed characteristics in end-of-year tests in February 2018, 2019 and 2020. Note that new cohorts enter study schools, especially in Grades 1 and Grade 6, while students who move to Grade 9 or other schools exit the study. Test scores are standardized to have a mean of zero and a standard deviation of one in the baseline in grade 5 in the control group.

Table A.2: Schedule of control and treated schools (2018-19)

| | Primary Grades | | Middle Grades | |
|---|---|---|---|---|
| Subject | Control mean | Treatment difference | Control mean | Treatment difference |
| **Targeted subjects** | | | | |
| Math (all) | 11.07 | 0.65 | 6.27 | 1.60*** |
| | | (0.66) | | (0.27) |
| Math (in class) | 11.05 | -2.51*** | 6.27 | -1.45*** |
| | | (0.65) | | (0.28) |
| Hindi (all) | 11.34 | 0.49 | 6.11 | 1.57*** |
| | | (0.54) | | (0.30) |
| Hindi (in class) | 11.33 | -2.69*** | 6.11 | -1.42*** |
| | | (0.60) | | (0.25) |
| **Untargeted curricular subjects** | | | | |
| English | 8.4 | -0.97 | 6.4 | -0.64*** |
| | | (0.65) | | (0.24) |
| Environmental Science | 7.84 | -0.43 | | |
| | | (0.73) | | |
| Science | | | 6.15 | -0.58*** |
| | | | | (0.20) |
| Social Studies | | | 6.09 | -0.25* |
| | | | | (0.14) |
| Sanskrit | | | 6.05 | -0.48*** |
| | | | | (0.14) |
| **Other subjects** | | | | |
| Library | .08 | 0.13 | .69 | -0.01 |
| | | (0.22) | | (0.20) |
| Art Education | 2.19 | -0.26 | 1.78 | -0.30 |
| | | (0.47) | | (0.24) |
| Health Education | 2.85 | -0.32 | 1.8 | -0.44** |
| | | (0.57) | | (0.20) |
| SUPW (Craft) | 2.21 | -0.46 | 1.8 | -0.17 |
| | | (0.46) | | (0.25) |
| Remedial Education | | | 3.45 | -0.44 |
| | | | | (0.50) |
| Observations | 175 | 140 | 108 | 105 |

*Notes*: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Standard errors are clustered at the stratum level. Each observation is at school $\times$ grade level.

Table A.3: Effects on student test scores (without covariate adjustment)

| | Math | | | Hindi | | |
|---|---|---|---|---|---|---|
| | Pooled Gr. 1-8 | Primary Gr. 1-5 | Middle Gr. 6-8 | Pooled Gr. 1-8 | Primary Gr. 1-5 | Middle Gr. 6-8 |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| **Panel A: Year 1** | | | | | | |
| Treatment | 0.15** | 0.06 | 0.17** | 0.14* | 0.05 | 0.16** |
| | (0.07) | (0.09) | (0.07) | (0.07) | (0.09) | (0.06) |
| Constant | 0.36*** | -0.11** | 0.72*** | 0.40*** | -0.03 | 0.73*** |
| | (0.04) | (0.04) | (0.03) | (0.03) | (0.04) | (0.03) |
| Observations | 7998 | 3333 | 4665 | 7998 | 3333 | 4665 |
| R-squared | 0.046 | 0.064 | 0.095 | 0.041 | 0.058 | 0.080 |
| **Panel B: Year 2** | | | | | | |
| Treatment | 0.24*** | 0.11 | 0.27*** | 0.22*** | 0.14** | 0.22*** |
| | (0.076) | (0.087) | (0.072) | (0.069) | (0.065) | (0.071) |
| Constant | 0.25*** | -0.15*** | 0.57*** | 0.24*** | -0.21*** | 0.60*** |
| | (0.039) | (0.044) | (0.038) | (0.036) | (0.033) | (0.037) |
| Observations | 8772 | 3725 | 5047 | 8772 | 3725 | 5047 |
| R-squared | 0.060 | 0.078 | 0.098 | 0.059 | 0.085 | 0.093 |

*Notes*: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Standard errors are clustered at the stratum level.
This table presents Intention-to-treat treatment effects of the program at the end of ∼6 months of treatment (EL Y1) and 18 months (EL Y2). Test scores are based on independent tests conducted in class for all students present on the day of testing at baseline (July 2017) and each end-of-year assessment (in February of each academic year). All regressions control for strata fixed effects but no other covariates. Test scores are linked across rounds and across grades using Item Response Theory models, and standardized to have a mean of zero and a standard deviation of one in the baseline in grade 5 in the control group.

Table A.4: Heterogeneity in treatment effects by baseline achievement

| | Standardized IRT scores (endline) | | | |
|---|---|---|---|---|
| | Year 1 | | Year 2 | |
| | Math | Hindi | Math | Hindi |
| | (1) | (2) | (3) | (4) |
| Treatment | 0.15*** | 0.11*** | 0.21*** | 0.18*** |
| | (0.036) | (0.031) | (0.036) | (0.031) |
| Interaction | -0.02 | -0.02 | -0.01 | -0.07** |
| | (0.023) | (0.025) | (0.029) | (0.028) |
| Baseline score | 0.61*** | 0.65*** | 0.54*** | 0.54*** |
| | (0.021) | (0.022) | (0.024) | (0.027) |
| Observations | 7994 | 7994 | 8733 | 8733 |

*Notes*: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Standard errors are clustered at the stratum level. The dependent variable is the IRT subject test score standardized to have mean zero and standard deviation 1 in Grade 5 at baseline. All regressions include stratum fixed effects and gender, as well as class fixed effects.

Table A.5: Heterogeneity in treatment effect by gender and socioeconomic status

| | Standardized IRT scores (endline) | | | |
|---|---|---|---|---|
| | Year 1 | | Year 2 | |
| | Math (1) | Hindi (2) | Math (3) | Hindi (4) |
| **Female** | | | | |
| Treatment | 0.17*** | 0.12*** | 0.23*** | 0.18*** |
| | (0.036) | (0.031) | (0.036) | (0.034) |
| Covariate | 0.05* | 0.10*** | 0.04 | 0.13*** |
| | (0.025) | (0.021) | (0.030) | (0.029) |
| Interaction | -0.06 | -0.02 | -0.04 | -0.00 |
| | (0.041) | (0.031) | (0.047) | (0.039) |
| Observations | 7994 | 7994 | 8733 | 8733 |
| **Socioeconomic status Index** | | | | |
| Treatment | 0.14*** | 0.10*** | 0.21*** | 0.18*** |
| | (0.030) | (0.026) | (0.036) | (0.030) |
| Covariate | 0.03*** | 0.02*** | 0.06*** | 0.06*** |
| | (0.008) | (0.008) | (0.011) | (0.009) |
| Interaction | -0.01 | -0.00 | -0.00 | -0.01 |
| | (0.012) | (0.010) | (0.014) | (0.014) |
| Observations | 7147 | 7147 | 8733 | 8733 |

*Notes*: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Standard errors are clustered at the stratum level. The dependent variable is the IRT subject test score standardized with mean zero and standard deviation 1 in Grade 5 at baseline. All regressions include stratum and class fixed effects and controls for the baseline individual/classroom mean test score and gender. The SES score is generated using Principal Components Analysis based on household ownership of 14 consumer durables (elicited in student surveys).

Table A.6: Heterogeneity by within-grade quintiles (Robustness)

| | Math | | Hindi | |
|---|---|---|---|---|
| | Year 1 | Year 2 | Year 1 | Year 2 |
| Treatment | 0.17*** | 0.18*** | 0.07 | 0.24*** |
| | (0.043) | (0.066) | (0.055) | (0.064) |
| Treatment x Q2 | -0.07 | 0.02 | 0.05 | -0.03 |
| | (0.045) | (0.058) | (0.059) | (0.063) |
| Treatment x Q3 | -0.10* | -0.06 | 0.06 | -0.07 |
| | (0.048) | (0.071) | (0.066) | (0.073) |
| Treatment x Q4 | -0.03 | 0.06 | 0.02 | -0.12 |
| | (0.053) | (0.075) | (0.068) | (0.088) |
| Treatment x Q5 | -0.02 | 0.11 | 0.05 | -0.13* |
| | (0.061) | (0.092) | (0.075) | (0.079) |
| Quintile 2 | 0.22*** | 0.14*** | 0.12** | 0.14** |
| | (0.044) | (0.046) | (0.051) | (0.054) |
| Quintile 3 | 0.36*** | 0.28*** | 0.20*** | 0.28*** |
| | (0.053) | (0.051) | (0.062) | (0.063) |
| Quintile 4 | 0.46*** | 0.42*** | 0.31*** | 0.41*** |
| | (0.065) | (0.065) | (0.068) | (0.075) |
| Quintile 5 | 0.57*** | 0.55*** | 0.31*** | 0.43*** |
| | (0.069) | (0.090) | (0.091) | (0.092) |
| F-test of equality of interaction terms (p-val) | .12 | .14 | .85 | .49 |
| Observations | 6539 | 4825 | 6539 | 4825 |
| R-squared | 0.669 | 0.574 | 0.612 | 0.594 |

*Notes*: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Standard errors are clustered at the stratum level. This table repeats the specification presented in Table 3 with two differences: (i) quintiles of baseline achievement are defined based on the *other* subject and (ii) we condition on both math and Hindi test scores from baseline. This procedure, inspired by Jerrim and Vignoles (2013), aims to assess the sensitivity of our conclusions to mean reversion induced by measurement error. The main patterns of Table 3 — that control group value-added is greater for students in upper quintiles while treatment effects are similar — is not sensitive to these changes.

Table A.7: Heterogeneity in effect on hardest/easiest items by within-grade BL score quintiles

| | Easy items | | Hard items | |
|---|---|---|---|---|
| | Year 1 | Year 2 | Year 1 | Year 2 |
| **Math** | | | | |
| Treatment | 0.04* | 0.06*** | 0.01 | 0.04** |
| | (0.02) | (0.02) | (0.01) | (0.01) |
| Treatment x Q2 | 0.01 | -0.01 | 0.02 | 0.02 |
| | (0.02) | (0.03) | (0.01) | (0.02) |
| Treatment x Q3 | -0.01 | -0.02 | 0.02 | 0.03* |
| | (0.02) | (0.02) | (0.01) | (0.02) |
| Treatment x Q4 | -0.02 | -0.04 | 0.02 | 0.06*** |
| | (0.02) | (0.02) | (0.01) | (0.02) |
| Treatment x Q5 | -0.03 | -0.05** | 0.04** | 0.04* |
| | (0.02) | (0.02) | (0.02) | (0.02) |
| Quintile 2 | 0.11*** | 0.09*** | -0.01 | 0.00 |
| | (0.01) | (0.01) | (0.01) | (0.02) |
| Quintile 3 | 0.19*** | 0.17*** | 0.04*** | 0.03** |
| | (0.01) | (0.02) | (0.02) | (0.01) |
| Quintile 4 | 0.21*** | 0.22*** | 0.10*** | 0.06*** |
| | (0.02) | (0.02) | (0.02) | (0.01) |
| Quintile 5 | 0.19*** | 0.22*** | 0.21*** | 0.13*** |
| | (0.02) | (0.02) | (0.02) | (0.02) |
| Mean percentage correct in bottom quintile | .6 | .52 | .15 | .16 |
| Observations | 6538 | 4825 | 6537 | 4825 |
| R-squared | 0.43 | 0.47 | 0.38 | 0.28 |
| **Hindi** | | | | |
| Treatment | 0.08*** | 0.11*** | 0.01 | 0.02* |
| | (0.02) | (0.02) | (0.01) | (0.01) |
| Treatment x Q2 | -0.04** | -0.04* | 0.03 | 0.01 |
| | (0.02) | (0.02) | (0.02) | (0.01) |
| Treatment x Q3 | -0.06*** | -0.07*** | 0.04* | 0.01 |
| | (0.02) | (0.02) | (0.02) | (0.01) |
| Treatment x Q4 | -0.07*** | -0.08*** | 0.03 | 0.03 |
| | (0.02) | (0.02) | (0.02) | (0.02) |
| Treatment x Q5 | -0.07*** | -0.09*** | 0.02 | -0.01 |
| | (0.02) | (0.02) | (0.02) | (0.02) |
| Quintile 2 | 0.16*** | 0.12*** | 0.02* | 0.02** |
| | (0.01) | (0.02) | (0.01) | (0.01) |
| Quintile 3 | 0.20*** | 0.21*** | 0.12*** | 0.08*** |
| | (0.01) | (0.02) | (0.01) | (0.01) |
| Quintile 4 | 0.21*** | 0.25*** | 0.25*** | 0.15*** |
| | (0.01) | (0.02) | (0.01) | (0.01) |
| Quintile 5 | 0.18*** | 0.24*** | 0.35*** | 0.27*** |
| | (0.02) | (0.02) | (0.02) | (0.02) |
| Mean percentage correct in bottom quintile | .64 | .6 | .2 | .19 |
| Observations | 6535 | 4825 | 6534 | 4825 |
| R-squared | 0.40 | 0.41 | 0.56 | 0.39 |

*Notes*: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Standard errors are clustered at the stratum level. Terciles of question difficulty are defined at the grade-round level, and quintiles of student achievement on their baseline scores. We exclude students who did not take the baseline test. The dependent variables is the proportion of questions correctly answered in the hardest/easiest terciles in a given round of testing. All regressions control for baseline scores, gender, and fixed effects for randomization strata and grade.

Table A.8: Progress in diagnosed achievement in treated schools

| | (1) Math Y1 | (2) Math Y2 | (3) Hindi Y1 | (4) Hindi Y2 |
|---|---|---|---|---|
| Assessed level at baseline | 0.91*** | 0.89*** | 0.92*** | 0.94*** |
| | (0.02) | (0.07) | (0.02) | (0.02) |
| Constant | 1.05*** | 2.02*** | 1.12*** | 2.31*** |
| | (0.06) | (0.14) | (0.06) | (0.08) |
| Observations | 4,122 | 2,705 | 4,223 | 2,710 |
| R-squared | 0.61 | 0.37 | 0.69 | 0.59 |

*Notes*: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Standard errors are clustered at the school level.
This table presents the regression analog of Figure 3. The dependent variable is the student achievement level in Math/Hindi, as assessed by the *Mindspark* software, after the first and second academic year of achievement.

Table A.9: Effects by grade, accounting for potential increase in instructional time

| | Year 1 | | Year 2 | |
|---|---|---|---|---|
| | Primary | Middle | Primary | Middle |
| **Math** | | | | |
| Treatment | 0.096 | 0.225*** | 0.086 | 0.247** |
| | (0.0623) | (0.0713) | (0.0933) | (0.0915) |
| Observations | 648 | 1308 | 655 | 1465 |
| R-squared | 0.481 | 0.553 | 0.302 | 0.383 |
| **Hindi** | | | | |
| Treatment | 0.171* | 0.143*** | 0.134 | 0.188** |
| | (0.0945) | (0.0418) | (0.0788) | (0.0895) |
| Observations | 738 | 1318 | 832 | 1567 |
| R-squared | 0.372 | 0.538 | 0.226 | 0.401 |

*Notes*: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Standard errors are clustered at the stratum level. This table presents Intention-to-treat treatment effects of the program at the end of ∼6 months of treatment (EL Y1) and 18 months (EL Y2), restricted to only those grade×stratum pairs where the program did not increase scheduled instruction in the relevant targeted subject. The specification is identical to Table 2. Test scores are standardized to have a mean of zero and a standard deviation of one in the baseline in grade 5 in the control group. All regressions include strata (school-pair) fixed effects and control for baseline scores and gender.

Table A.10: Treatment effect on school examinations, other subjects (year 2)

| | Half-year examinations | | | Board examinations | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | Grade A and above | | | Grade B and above | | | Grade C and above | | |
| | Science | Soc. Sc. | English | Science | Soc. Sc. | English | Science | Soc. Sc. | English | Science | Soc. Sc. | English |
| **Grade 8** | | | | | | | | | | | | |
| Treatment | -0.96 | 1.21 | 2.47 | 0.02 | -0.00 | -0.02 | 0.07 | -0.03 | -0.06 | -0.04 | -0.06 | -0.05 |
| | (2.151) | (1.851) | (2.014) | (0.038) | (0.037) | (0.062) | (0.068) | (0.059) | (0.056) | (0.042) | (0.052) | (0.030) |
| Mean score | 87.27 | 89.49 | 88.11 | .29 | .15 | .12 | .56 | .36 | .37 | .88 | .79 | .79 |
| Observations | 1764 | 1767 | 1713 | 2568 | 2566 | 2566 | 2568 | 2566 | 2566 | 2568 | 2566 | 2566 |
| R-squared | 0.578 | 0.605 | 0.638 | 0.078 | 0.115 | 0.119 | 0.124 | 0.157 | 0.121 | 0.196 | 0.190 | 0.192 |

*Notes*: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Standard errors are clustered at the stratum level.
For half-year examinations, the dependent variable is the score obtained at the half year school examinations (scores between 0 and 100). For board examinations, the dependent variable is a dummy variable for obtaining the grade of interest or above. ”Soc. Sc.” stands for Social Science, and ”Env. Science” stands for environmental science. In all regressions, stratum fixed effects are included, but gender and baseline test scores are not controlled for.

Table A.11: Cumulated effects on student achievement, year 3

| | Math | | | Hindi | | |
|---|---|---|---|---|---|---|
| | Pooled | Primary | Middle | Pooled | Primary | Middle |
| Treatment | 0.24*** | 0.19*** | 0.26*** | 0.21*** | 0.20*** | 0.19*** |
| | (0.057) | (0.067) | (0.066) | (0.050) | (0.072) | (0.068) |
| Baseline score | 0.59*** | 0.37*** | 0.64*** | 0.65*** | 0.42*** | 0.60*** |
| | (0.030) | (0.051) | (0.027) | (0.033) | (0.054) | (0.038) |
| Observations | 10971 | 4760 | 6211 | 10971 | 4760 | 6211 |
| R-squared | 0.261 | 0.122 | 0.264 | 0.210 | 0.128 | 0.184 |

*Notes*: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Standard errors are clustered at the stratum level.
This table presents Intention-to-treat treatment effects of the program at the end of 30 months (EL Y3). Test scores are based on independent tests conducted in class for all students present on the day of testing at baseline (July 2017) and end-of-year assessment (in February 2020). In the end round, students who were absent on the day of the test were tracked to households for testing. For students who were absent on the day of testing at baseline, we replace the baseline score with the classroom average. Test scores are linked across rounds and across grades using Item Response Theory models. Test scores are standardized to have a mean of zero and a standard deviation of one in the baseline in grade 5 in the control group. All regressions include strata (school-pair) fixed effects and student gender.

Table A.12: Student time use in observed Mindspark lab periods

| Variable | (1)<br>Pooled<br>Mean/(SD) | (2)<br>Grade 5<br>Mean/(SD) | (3)<br>Grade 8<br>Mean/(SD) |
|---|---|---|---|
| *What are students doing?* | | | |
| On Mindspark working actively | 0.75 | 0.76 | 0.74 |
| | (0.21) | (0.19) | (0.24) |
| On Mindspark guessing/clicking randomly | 0.11 | 0.10 | 0.12 |
| | (0.13) | (0.11) | (0.16) |
| Talking to the teacher asking questions | 0.02 | 0.03 | 0.02 |
| | (0.08) | (0.09) | (0.07) |
| Talking casually (Other than the partner) | 0.03 | 0.03 | 0.04 |
| | (0.05) | (0.04) | (0.06) |
| Sitting idle | 0.01 | 0.01 | 0.01 |
| | (0.02) | (0.02) | (0.02) |
| Out of their seats | 0.01 | 0.02 | 0.01 |
| | (0.05) | (0.07) | (0.02) |
| Not able to work due to technical problem | 0.01 | 0.02 | 0.01 |
| | (0.04) | (0.05) | (0.02) |
| Not in the lab | 0.05 | 0.04 | 0.06 |
| | (0.11) | (0.09) | (0.13) |
| *What do you observe on the laptop screens?* | | | |
| Assigned subject questions on Mindspark | 0.90 | 0.90 | 0.90 |
| | (0.14) | (0.13) | (0.15) |
| Other subject questions on Mindspark | 0.01 | 0.01 | 0.01 |
| | (0.05) | (0.03) | (0.07) |
| Login screen on Mindspark | 0.05 | 0.05 | 0.04 |
| | (0.07) | (0.07) | (0.07) |
| Student progress report | 0.01 | 0.01 | 0.01 |
| | (0.03) | (0.02) | (0.03) |
| Blank screen | 0.03 | 0.02 | 0.04 |
| | (0.08) | (0.07) | (0.10) |
| Other | 0.00 | 0.01 | 0.00 |
| | (0.02) | (0.02) | (0.02) |
| Number of observations | 95 | 46 | 49 |

Table A.13: Teacher time use in observed Mindspark lab periods

| Variable | (1)<br>Total<br>Mean/(SD) | (2)<br>5<br>Mean/(SD) | (3)<br>8<br>Mean/(SD) |
|---|---|---|---|
| Is the LIC in the lab? | 0.30 | 0.34 | 0.27 |
| | (0.44) | (0.46) | (0.43) |
| Is the teacher present in the lab? | 0.52 | 0.48 | 0.55 |
| | (0.42) | (0.44) | (0.40) |
| *What is the teacher doing?* | | | |
| Helping students settle down | 0.09 | 0.06 | 0.12 |
| | (0.16) | (0.16) | (0.16) |
| Helping students with questions | 0.19 | 0.22 | 0.15 |
| | (0.28) | (0.31) | (0.24) |
| Helping students with technical difficulties | 0.02 | 0.01 | 0.02 |
| | (0.05) | (0.04) | (0.06) |
| Doing rounds in the lab | 0.04 | 0.03 | 0.05 |
| | (0.11) | (0.09) | (0.13) |
| Sitting in the lab doing admin work | 0.06 | 0.08 | 0.05 |
| | (0.16) | (0.18) | (0.14) |
| Sitting in the lab doing corrections | 0.02 | 0.02 | 0.02 |
| | (0.10) | (0.08) | (0.11) |
| Sitting in the lab at leisure | 0.06 | 0.03 | 0.10 |
| | (0.14) | (0.09) | (0.17) |
| Talking to other staff in the lab | 0.03 | 0.02 | 0.03 |
| | (0.09) | (0.09) | (0.09) |
| Standing outside the lab | 0.00 | 0.00 | 0.01 |
| | (0.03) | (0.02) | (0.03) |
| Teacher is not present | 0.48 | 0.52 | 0.45 |
| | (0.42) | (0.44) | (0.40) |
| Number of observations | 95 | 46 | 49 |

Table A.14: Effects on classroom practice

| Variable | Grade 5 | | Grade 8 | |
|---|---|---|---|---|
| | Control mean | Treatment difference | Control mean | Treatment difference |
| Is the teacher present? | 0.97 | -0.02 | 0.94 | 0.03 |
| | | (0.02) | | (0.02) |
| **If yes, what is the teacher doing?** | | | | |
| Teaching the full class | 0.90 | -0.03 | 0.91 | -0.01 |
| | | (0.04) | | (0.05) |
| Teaching students in small groups/one by one | 0.32 | -0.12 | 0.11 | 0.01 |
| | | (0.08) | | (0.04) |
| Reading out loud | 0.45 | -0.01 | 0.41 | 0.13* |
| | | (0.07) | | (0.07) |
| Reading but students repeat after teacher | 0.52 | -0.17*** | 0.31 | 0.03 |
| | | (0.06) | | (0.05) |
| Sitting idle in the chair | 0.06 | -0.01 | 0.01 | 0.02 |
| | | (0.03) | | (0.02) |
| Correcting notebooks/test papers | 0.06 | -0.01 | 0.06 | -0.03 |
| | | (0.02) | | (0.03) |
| Filling records/Data/Taking attendance | 0.03 | 0.02 | 0.01 | 0.01 |
| | | (0.02) | | (0.01) |
| Talking casually | 0.09 | 0.03 | 0.02 | 0.06* |
| | | (0.05) | | (0.03) |
| Total observations | 147 | | 141 | |

*Notes*: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Standard errors are clustered at the stratum level and shown in parentheses. Stratum fixed effects are included in all regressions.
This table is based on classroom observations that were carried out in mathematics and Hindi lessons in Grades 5 and 8, both in treatment and control schools. Teacher time use was captured through structured snapshots recorded every 5 minutes. While we see some differences being statistically significant between the treatment and control groups during the observed period, there does not appear to be any consistent difference in the practices observed by enumerators during classroom visits.

# B Test score construction and validation

## B.1 Overview

We measured student achievement, which is the main outcome for our evaluation, using independent assessments in math and Hindi. These tests were administered under the supervision of the research team at baseline (Oct 2017) and close to the end of the school year in each of three years (February-to-March of 2018, 2019 and 2020). Here we present details about the test content and development, administration, and scoring.

## B.2 Objectives of test design

Our test design reflects three objectives, closely following principles from the efficacy trial and recommended best practices (Muralidharan et al., 2019; Bertling et al., 2025).

First, in each grade, the test should be informative over the full range of achievement.

Second, the tests should allow us to express student scores on a common scale, both across rounds and across grades. This challenge was more severe in the current project than the efficacy trial given the span over the full range of primary and middle schooling.

Third, the test should be a fair benchmark to judge the actual skill acquisition of students. Reflecting this need, tests were administered using pen-and-paper rather than on computers so that they do not conflate increments in actual achievement with greater familiarity with computers in the treatment group. Further, the items were taken from a wide range of independent sources, and selected by the research team without consultation with Education Initiatives, to ensure that the selection of items was not prone to "teaching to the test" in the intervention.

## B.3 Test booklets and administration

We assembled grade-specific tests in math and Hindi for each round of testing.

We aimed to avoid ceiling and floor effects in each grade. Recognizing that students may be much below grade-appropriate levels of achievement, test booklets included items ranging from very basic competences to harder items which are closer to grade-appropriate standards. Test booklets included between 20-30 items per subject in each round for Grades 3-8 and between 11-16 items in Grades 1-2.[47] Booklets administered to middle school students (Grades 6-8) included only written questions; for students in Grades 1-2, only orally-administered items; and, for Grades 3-5, a combination of orally-administered and written items. These items were sourced from previous research studies in India and state-level textbooks and exams. To ensure the integrity of assessments, all tests were administered and proctored independently by surveyors.[48]

The test booklets were designed to partially overlap across grades — in each grade, typically about a third of items were common with the preceding grade and a third

[47]The smaller number of test items in Grades 1-2 reflect a smaller number of competences to be tested as well as the need to reduce survey burden for young children who were administered these items individually with oral stimuli and responses.

[48]See Singh and Berg (2024) and **?** for evidence of test score manipulation in teacher-administered tests in Indian public schools in multiple states.

with the grade above. This allowed us to increase the difficulty of test booklets across grades while preserving sufficient overlap of items to link test scores using Item Response Theory models and express them on a common scale. Similarly, in each round, a substantial share of items were retained in common from the previous assessments to ensure that scores could be linked across survey rounds. Our final Item Bank includes 149 unique test items in Math and 122 test items in Hindi. All items were scored dichotomously (correct or incorrect) and scored using a 3-parameter logistic model estimated using a dataset that pooled all rounds and grades.

## B.4 Psychometric validation

We validate our test scores in three successive exercises.

### B.4.1 Empirical distribution of test scores

First, we present the distribution of the percentage of correctly answered questions in each round in the sample separately for primary and middle school students. Since test booklets vary both across grades and across survey rounds, these distributions are not comparable to each other — they are presented only as a diagnostic of whether a large proportions of students answer all questions incorrectly (floor effects) or all questions correctly (ceiling effects); the presence of either would indicate a censoring problem induced by our test instruments being unable to pick up the range of achievement in the sample. Percentage correct scores are smoothly distributed without evidence of large floor or ceiling effects (Figures A.4 and A.5).

Figure A.4: Distribution of percentage correct by test round: Math

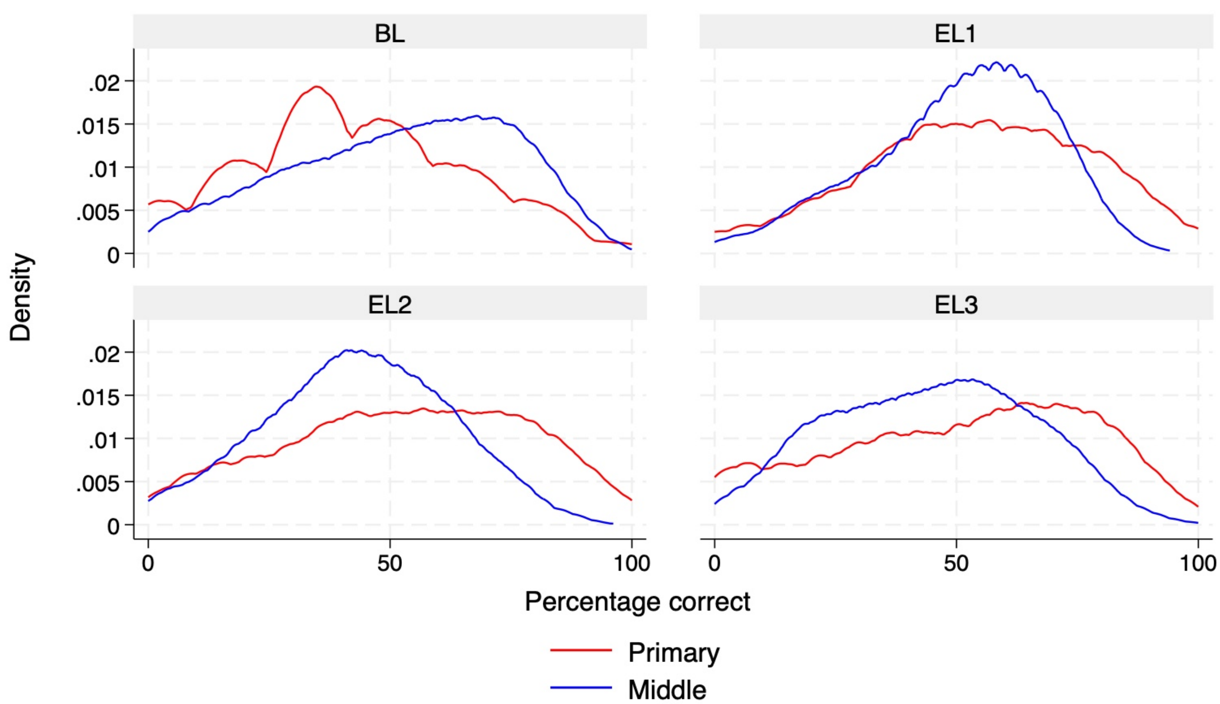
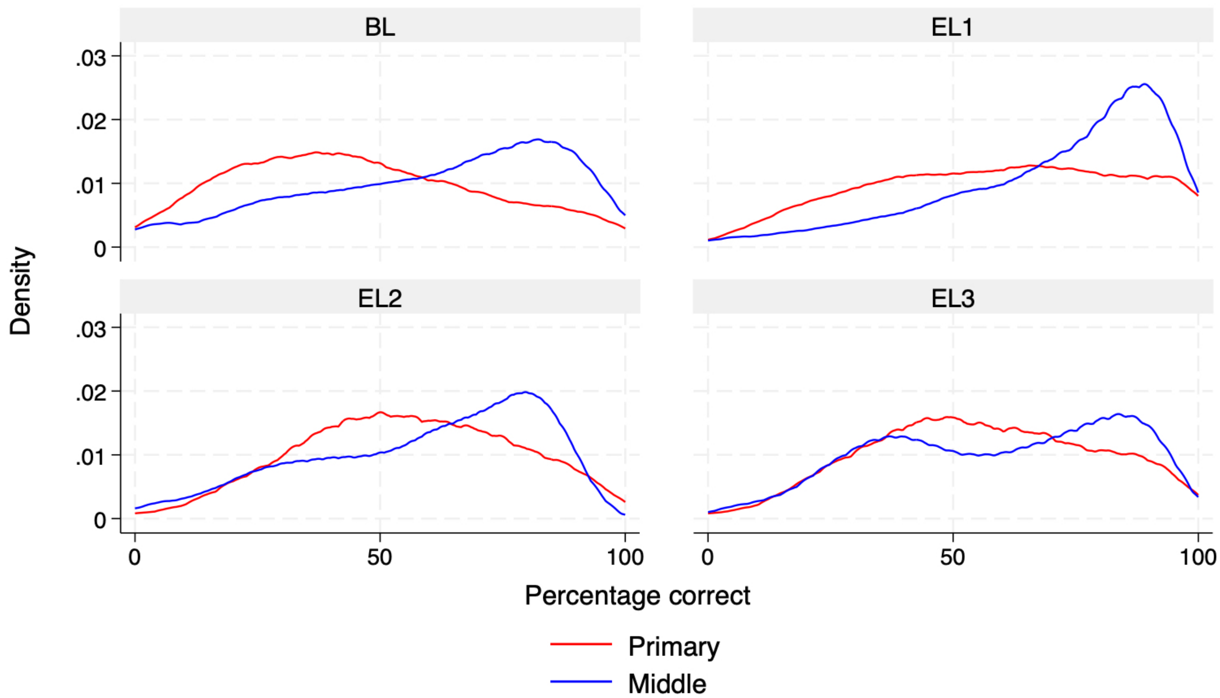


Figure A.5: Distribution of percentage correct by test round: Hindi




Our outcome measures in this paper are IRT linked scores, not the percentage of questions answered correctly. We present the distribution of these IRT scores in Figures A.6 and A.7. Reassuringly, we find that the distributions are well-distributed in both math and Hindi, in each survey round, for both primary and middle school students. Note that, as a consequence of the IRT linking using common ("anchor") items, these distributions are comparable to each other across grades and survey rounds. In both math and Hindi, the distribution is shifted substantially to the right for middle school students; this indicates that, as expected, students in middle school grades have greater subject knowledge than those in primary grades (and that our test design with overlapping booklets is sufficiently sensitive to pick up these differences in student learning).

Figure A.6: Distribution of IRT scores: Math

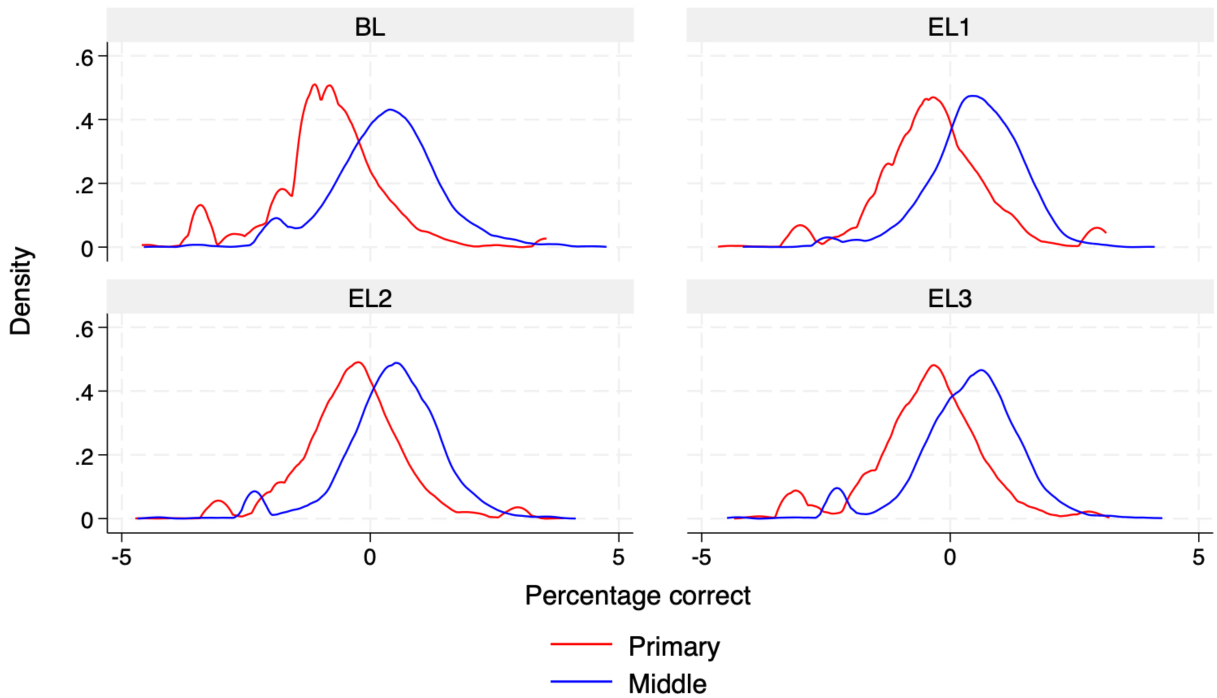
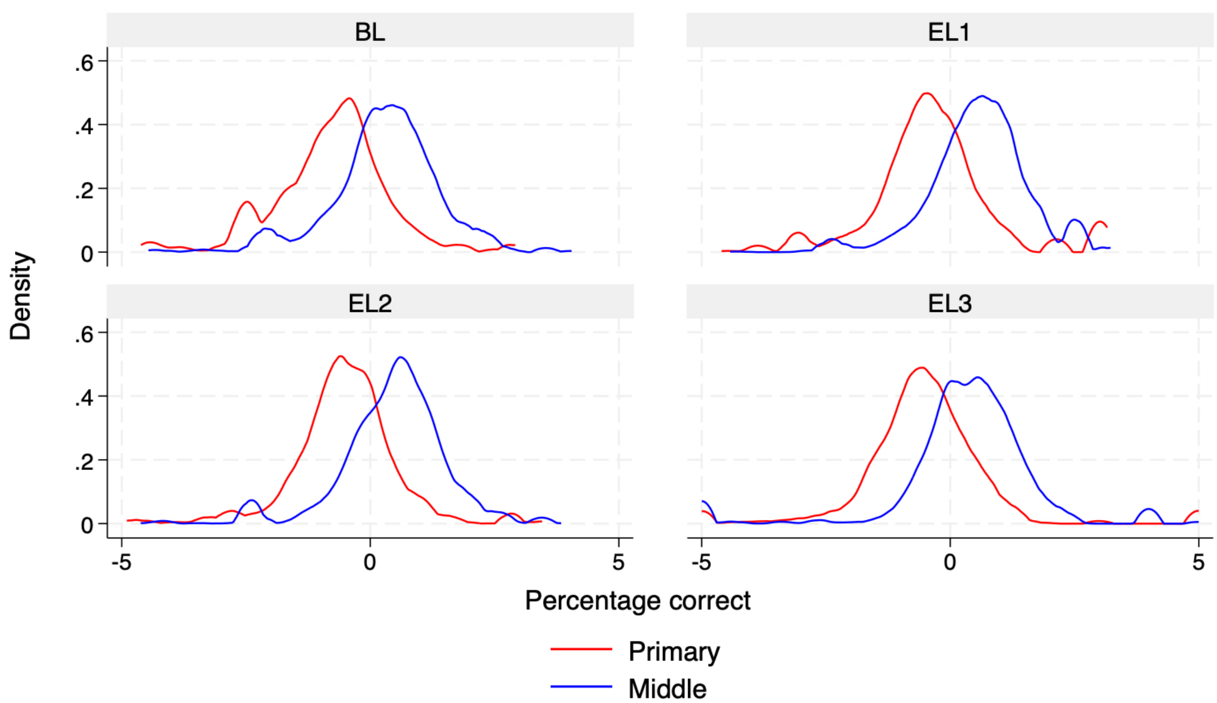



Figure A.7: Distribution of IRT scores: Hindi



### B.4.2 Classical Test Theory validation

Next, we present the Cronbach's alpha, a standard measure of scale reliability in classical test theory, for each test booklet (Table A.15). Values above 0.7-0.8 are considered precise for measuring group differences, which are comfortably exceeded for all our booklets.

Table A.15: Cronbach's alpha for test booklets

| | **Math** | | | **Hindi** | | |
|---|---|---|---|---|---|---|
| | Endline Y1 | Endline Y2 | Endline Y3 | Endline Y1 | Endline Y2 | Endline Y3 |
| Grade 1 | 0.78 | 0.88 | 0.87 | 0.88 | 0.80 | 0.79 |
| Grade 2 | 0.74 | 0.88 | 0.88 | 0.89 | 0.83 | 0.81 |
| Grade 3 | 0.83 | 0.83 | 0.85 | 0.83 | 0.83 | 0.83 |
| Grade 4 | 0.85 | 0.86 | 0.83 | 0.88 | 0.87 | 0.88 |
| Grade 5 | 0.91 | 0.88 | 0.90 | 0.90 | 0.89 | 0.91 |
| Grade 6 | 0.89 | 0.88 | 0.88 | 0.91 | 0.91 | 0.91 |
| Grade 7 | 0.85 | 0.77 | 0.84 | 0.90 | 0.89 | 0.90 |
| Grade 8 | 0.77 | 0.82 | 0.86 | 0.90 | 0.91 | 0.91 |

### B.4.3 Empirical fit to Item Characteristic Curves

Third, to assess the reliability of the IRT linking exercise, we examined item-by-item the empirical fit of the data to estimated item characteristics. We inspected the fit separately for potential Differential Item Functioning across (i) treatment and control groups and (ii) survey rounds. Given the large number of test items, and the multiple dimensions on which we inspected DIF, we do not present the full set of Item Characteristic Curves and empirical fits here. Figure A.8 shows examples of our visual inspection of DIF for a subset of Math items across survey rounds and treatment groups. The top sub-figure shows visual inspection of DIF across survey rounds, while the bottom sub-figure shows it across treatment and control group students. Similar graphs were generated and inspected for all assessment items in both subjects.

Figure A.8: Visual inspection of Differential Item Functioning