NBER WORKING PAPER SERIES

HOW IN-SCHOOL SUPERVISED ED-TECH SUPPORT PRODUCES MASSIVE LEARNING GAINS: A KHAN ACADEMY FIELD EXPERIMENT IN INDIA

Philip Oreopoulos
Oliver Keyes-Krysakowski
Deepak Agarwal

Working Paper 34683
http://www.nber.org/papers/w34683

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
January 2026

This research was supported through the Jameel Poverty Action Lab and the Smith Richardson Foundation. We are especially grateful to Shri Prashant Kumar, Director of the Uttar Pradesh Social Welfare Department, and Shri J Ram, Deputy Director of the Uttar Pradesh Social Welfare Department, for their support and partnership in enabling this research. We also thank Nina Low and other RAs for their help, as well as the dozens of lab-in-charges who worked tirelessly to implement the program. Participants at numerous seminars and workshops provided helpful suggestions. Any errors or omissions 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.

How In-School Supervised Ed-Tech Support Produces Massive Learning Gains: A Khan Academy Field Experiment in India

Philip Oreopoulos, Oliver Keyes-Krysakowski, and Deepak Agarwal

NBER Working Paper No. 34683
January 2026
JEL No. I2, I25, I3, O2

ABSTRACT

Computer-assisted learning (CAL) platforms frequently underperform at scale not because the technology is ineffective, but because schools face substantial implementation frictions: teachers and administrators must overcome initial technical hurdles, reorganize instructional routines, manage competing scheduling pressures, and do so while uncertain about the technology's effectiveness—conditions that often lead to low and unproductive student engagement. This study explores whether strengthening implementation structure can raise both the quantity and quality of CAL usage in 83 residential government middle schools in Uttar Pradesh, India and, in turn, learning gains. All schools had access to Khan Academy, but randomly selected treatment schools received on-the-ground lab-in-charges whose sole responsibility was to ensure high-fidelity implementation by securing reliable connectivity, simplifying student rostering, protecting weekly practice time, supervising in-class use, coordinating content with teachers, and monitoring progress. The intervention increased platform usage from 7.2 to 47.4 minutes per week. Mathematics achievement rose by almost half a standard deviation over 31 weeks, with gains broad-based across achievement levels and question difficulty. These results show that the central constraint on effective and scalable CAL is not technology or content, but the presence of organizational structures that ensure sustained, productive instructional use.

Philip Oreopoulos
University of Toronto
Department of Economics
and NBER
philip.oreopoulos@utoronto.ca

Deepak Agarwal
Khan Academy
deepak@khanacademy.org

Oliver Keyes-Krysakowski
University of Toronto
Department of Economics
oliver.keyeskrysakowski@mail.utoronto.ca

A randomized controlled trials registry entry is available at
https://www.socialscienceregistry.org/trials/13921

I. Introduction

A fundamental challenge in education is that students are different. Students arrive at each grade with vastly different skill levels and progress at different rates. For decades, managing this variation in student needs has been regularly reported by teachers as one of the most difficult challenges of teaching (Guryan et al., 2023). In classroom settings, teachers cannot easily provide individualized attention. As a result, some students move on to other topics before clearly grasping earlier ones while other students could be learning more advanced topics faster. This mismatch between instruction and individual needs creates a dynamic whereby early differences in skill acquisition lead to widening achievement gaps as students progress through school (Stanovich, 1986; Duncan et al., 2007). The average fifth-grade class contains students working at levels ranging from third grade to eighth grade, with this performance variance increasing over time (Peters et al., 2017; Cascio and Staiger, 2012; Nielsen, 2023).

One approach to addressing differences in learning rates across students is tutoring, which allows students to progress more at their own pace and receive immediate, individualized feedback (Beck, 2007). Dietrichson et al. (2017) concludes that tutoring is the most effective academic intervention for elementary and secondary students (generating an average learning gain impact of 0.36 standard deviations (SD)). High dosage tutoring experiments often find effective results across subject areas, grade levels, and educational contexts (Nickow, Oreopoulos and Quan, 2020, 2024; Guryan et al., 2023). However, tutoring alone remains prohibitively expensive and difficult to scale. Effective tutoring programs, with 2-3 sessions a week over the school year, cost thousands of dollars per student annually and face significant scaling challenges, particularly in ensuring consistent student participation (Oreopoulos et al., 2024; Strassberger and Condliffe, 2024; White, Groom-Thomas and Loeb, 2023). The pandemic highlighted these constraints as many districts struggled to deliver tutoring at scale despite unprecedented funding (Fahle et al., 2024; Guryan and Ludwig, 2023).

Motivated by these challenges, computer-assisted learning (CAL) has emerged as a potentially more scalable approach for delivering personalized instruction. Several well-implemented ran-

domized controlled trials of CAL platforms have demonstrated positive impacts on learning outcomes, with standardized assessment scores improving by 0.20 SD (Escueta et al., 2020). CAL platforms also allow students to progress through material at their own pace, receive immediate feedback, and work at levels matched to their actual skills rather than their enrolled grade, potentially mimicking benefits of tutoring at far lower cost. Research on "teaching at the right level" has demonstrated that targeting instruction to students' actual learning levels can generate substantial gains and directly address heterogeneity by filling learning gaps (Banerjee et al., 2007, 2016; Duflo, Dupas and Kremer, 2011). This approach appears particularly promising in developing countries, where classroom heterogeneity is often more pronounced (Duflo, Dupas and Kremer, 2011; Rodriguez-Segura, 2022). A meta-analysis of self-led educational technology interventions in developing countries found a median effect size of 0.29 standard deviations (Rodriguez-Segura, 2022).

CAL effectiveness, however, depends on more than just its platform quality. How it is set up in schools and how it is motivated and monitored also matter. Students often lack motivation to work on a computer by themselves, compared to working with a human tutor. CAL programs produce dramatically different outcomes depending on how they are structured. Some studies find negative or null effects, often citing implementation challenges (Morgan and Ritter, 2002; Pane et al., 2010, 2014). Other studies find modest positive effects when implementation fidelity was high (Barrow, Markman and Rouse, 2009; Roschelle et al., 2016), while others found smaller effects when teachers faced integration challenges (Copeland et al., 2023) or null effects when teachers struggled to balance CAL with regular curriculum demands (Phillips et al., 2020). Even within studies, average treatment effects belie enormous variation in class level fidelity, with some teachers integrating CAL into their lessons very little while others embracing its potential determined to follow instructions and embrace new methods. These patterns reveal a crucial insight: the technology alone is not the only factor in CAL success or failure. As Hill and Erickson (2021) demonstrate, "low fidelity increases the likelihood of weak student outcomes," while "moderate fidelity may be enough to yield positive program outcomes." The fundamental challenge is not whether CAL

can work in principle, but how schools can facilitate sufficient dosage and quality engagement to realize its potential benefits.

Evidence suggests that successful implementation hinges on structural support and supervision. As Fancsali et al. (2016) conclude, learner efficiency depends on whether teachers actively support students in turning engaged time into productive learning time. Studies consistently show clear dosage-response differences: classrooms using CAL platforms averaging 50-90+ minutes per week are associated with substantially higher learning gains than classrooms averaging under 20 minutes per week (Oreopoulos et al., 2024; Canbolat and Arndt, 2024). Of course, time alone is insufficient to predict gains. Students can engage in "wheel-spinning," practicing extensively without making progress if they lack appropriate support (Beck and Gong, 2013). Achieving both adequate quantity and quality of engagement requires implementation structures that ensure consistent, educationally productive use. Variation across classrooms in CAL effectiveness is driven primarily by differences in implementation conditions and the extent to which CAL practice is integrated into their curriculum (Oreopoulos et al., 2024).

Recent evidence from India illustrates both the promise and challenge of scaling personalized learning through technology. A recent study of Mindspark, a computer-based adaptive learning platform, achieved gains of 0.22 standard deviations in mathematics over 18 months when implemented in public schools with dedicated computer labs, modified timetables, and lab-in-charges ensuring smooth operations (Muralidharan and Singh, 2025). But achieving this implementation fidelity is difficult and requires thoughtful design and implementation, as our study setting in the Indian state of Uttar Pradesh illustrates. There, Khan Academy began a partnership with 105 government boarding schools in 2022, with schools encouraged to dedicate one to two of their 6 weekly math sessions to the platform during the first year. Despite best efforts to promote 120 minutes of weekly practice, only 44 percent of registered students used the platform even once during the entire year, and very few students actually achieved usage targets. Schools lacked dedicated time in schedules for platform use, with sessions happening once monthly or less. In many schools, it was difficult to allocate time for Khan Academy in the time table amid competing demands and,

in other schools, regular sessions could not be conducted due to intermittent internet connection and electricity access.

This paper presents results from an experiment in which more intensive, on-the-ground support was provided to these schools to better ensure recommended practice sessions actually occur. With administrative approval from the Uttar Pradesh Department of Social Welfare, we provided government schools with on-the-ground support through lab-in-charges, whose sole responsibility was ensuring high-fidelity implementation. Lab-in-charges guaranteed two Khan Academy sessions per week for each grade-section, troubleshooted technical and program related challenges, monitored student progress data, and worked with school leadership to integrate Khan Academy practice as mandatory curriculum time. We randomly assigned 24 schools to receive this support and 50 schools to a control condition over 31 weeks, covering 5,535 students in grades 6-8.

The impact from treatment was substantial. Students in treatment schools scored almost half a standard deviation higher on independently administered mathematics assessments, representing one of the most effective interventions reported in the education policy literature. The mechanism driving these gains corresponds to treatment schools practicing 47.4 minutes of Khan Academy practice per student per week compared to just 7.2 minutes in control schools, a 6.6-fold increase in usage. This dramatic difference highlights the importance of implementation quality and dedicated staff with clear ownership of the program for ensuring effective gains from CAL. While much attention focuses on developing better platforms or identifying optimal configurations, our results suggest that the quality and intensity of implementation support are important complements to the technology itself.

The remaining part of this paper is organized as follows: Section II provides background on computer-assisted learning in India, reviewing evidence on implementation challenges and the role of support structures in determining program success. Section III describes the setting, detailing the 2023-24 Khan Academy implementation that preceded our experiment and the design of our randomized controlled trial with dedicated lab-in-charges. Section IV presents our data sources, including baseline and endline mathematics assessments and Khan Academy platform us-

age tracking. Section V presents preliminary analysis including sample construction and balance tests across treatment and control conditions. Section VI examines implementation fidelity and platform engagement patterns across treatment and control schools. Section VII presents our main results on learning outcomes, including effects across question difficulty levels and student subgroups. Section VIII discusses implications for cost-effectiveness and scalability, acknowledges limitations, and section IX concludes with lessons for implementing computer-assisted learning at scale.

II. Background

The challenge of student heterogeneity is particularly acute in developing countries. Teachers often manage 40 pupils with limited capacity to address wide achievement gaps (Duflo, Dupas and Kremer, 2011), while two-thirds lack minimum proficiency in their subject content (Rodriguez-Segura, 2022). In India, these constraints can compound to severe learning deficits. In Uttar Pradesh, where our study takes place, only 27.9 percent of third-grade students in government schools can read second-grade-level text, and only 31.6 percent can perform basic subtraction (ASER Centre, 2025). These foundational gaps persist through later grades, with diagnostic assessments showing that eighth graders perform, on average, four grades below their enrolled level (Muralidharan and Singh, 2025). Fundamentally, Indian school systems often face constraints that limit their ability to adapt infrastructure and curricula to students' needs (Banerjee et al., 2007), resulting in widespread student variation that has made India a particularly important context for testing whether personalized learning approaches can address these challenges.

Evidence from India demonstrates that computer-assisted learning can generate substantial gains when implementation is properly structured, though results depend critically on program design and support. Banerjee et al. (2007) found that a Pratham CAL program providing two hours of shared computer time weekly increased Grade 4 math scores by 0.47 standard deviations, with the key lesson being that "it is not the number of teachers that seems to matter but how they are deployed, and what they do." Muralidharan, Singh and Ganimian (2019) evaluated

Mindspark adaptive learning software in after-school centers, finding effects of 0.36 standard deviations in math, noting results reflected "the extent to which using technology increased the productivity of an instructor, as opposed to technology by itself." However, control students received no after-school instruction, meaning the observed effects could reflect additional practice time rather than computer-assisted learning specifically. These supplementary programs, which added instructional time rather than replacing regular instruction, have typically shown positive effects (Singh, 2025). However, translating success into scalable government school models that substitute for regular instruction has proven more challenging. Linden (2008) found that while supplementary computer-assisted learning generated positive effects, "pulling out" students from regular instruction for computer-based learning produced "very large negative effects" even in high-performing NGO schools, suggesting technology can be actively worse than regular classroom instruction when poorly integrated (Singh, 2025). Similarly, a Khan Academy evaluation in Brazilian schools that replaced one weekly classroom period found null effects, with researchers hypothesizing implementation problems (Ferman, Finamor and Lima, 2019; Singh, 2025). These contrasting results underscore that successful substitution requires more than training and monitoring alone.

More recent evidence demonstrates that substitution models can succeed with adequate structural adaptation. When Mindspark was adapted for government schools with modified timetables displacing 25-50 percent of instruction time and lab-in-charges for hardware support, it achieved 0.22 standard deviation gains after 18 months (Muralidharan and Singh, 2025). This substitution design provided a clearer assessment of computer-assisted learning effectiveness, as control schools continued regular mathematics instruction rather than receiving no instruction. Critically, "learning gains were proportional to student time on the platform," and when lab-in-charge presence decreased, "usage declined sharply to about half the previous year's levels," demonstrating that achieving gains through substitution "will require very careful consideration of the complementarities between adult supervision and computer use" (Singh, 2025; Muralidharan and Singh, 2025). Similar patterns emerge across contexts: field staff supporting Andhra Pradesh schools generated 0.43 standard deviation gains (Kremer et al., 2025), phone-based targeted tutoring across

five developing countries, including India, achieved 0.30-0.35 standard deviations with similar effects from government teachers and NGO instructors (Angrist et al., 2023), and computer-assisted learning in El Salvador providing additional instruction outperformed traditional teacher-led lessons (Büchel et al., 2020). Research also reveals limits to simply increasing dosage, with Bettinger et al. (2023) finding positive effects at 45 minutes weekly but similar rather than larger effects when doubling to 90 minutes, which parallels the findings of "wheel-spinning" by Beck and Gong (2013). A meta-analysis confirms that effectiveness "rests on the thoughtful customization of the EdTech solution to the policy constraints at hand" (Rodriguez-Segura, 2022). These findings suggest that the binding constraint for scaling personalized learning may not be technology quality or general support alone, but rather the intensity and comprehensiveness of implementation structures. This hypothesis is particularly relevant in contexts like India where previous computer-assisted learning attempts with only training and monitoring have struggled to achieve consistent platform usage.

The 2022-2024 Khan Academy implementation across 105 Uttar Pradesh government schools involved training and encouragement to utilize the CAL platform for 120 minutes a week. Despite this, students practiced little, possibly due to poor internet connectivity, a lengthy CAL rostering process, divided teacher priorities across multiple programs, or lack of teacher buy-in. These bottlenecks all provide a useful setting to test the benefits from more on-the-ground implementation support while holding CAL usage targets the same. Our study introduces dedicated lab-in-charges with comprehensive implementation management responsibilities. These lab-in-charges ensure that scheduled sessions actually occur, troubleshoot technical challenges, monitor student progress, and work with school leadership to integrate platform usage as mandatory curriculum time. Importantly, control schools retained full access to the Khan Academy platform and guidance to use it for 120 minutes, allowing us to isolate the impact of overcoming implementation bottlenecks rather than differences in technology on its own. This approach addresses a critical gap identified across the literature. While studies have demonstrated that supplementary computer-assisted learning can work with adequate support, and that substitution models show promise when carefully adapted

with structural modifications and adult supervision, few have tested whether intensive implementation capacity can enable successful substitution using freely available platforms in contexts where previous attempts with lighter-touch support have failed. Our contribution is isolating implementation structure as a potentially crucial variable. We demonstrate how schools with dedicated staff whose sole responsibility is ensuring technology use can massively increase CAL usage and, in turn, achieve higher learning gains compared to those without, thereby informing pathways for scaling personalized learning in resource-constrained government school systems.

III. Setting

Our study takes place within the residential school system operated by the Department of Social Welfare in Uttar Pradesh, India's most populous state. In November 2022, the Department entered into partnership with Khan Academy India to launch a mathematics improvement program across 105 schools serving grades 6-12. The 2023-24 program followed conventional computer-assisted learning approaches. Teachers were instructed to use Khan Academy for lesson-aligned practice and remediation, assigning syllabus-matched exercises and videos to students, tracking completion of assigned tasks, and using the platform's dashboard to diagnose learning levels and guide individualized progression. Students were expected to complete these tasks using individual devices (computers or tablets) in online mode during designated timetable slots. The recommended implementation called for one to two Khan Academy sessions per week with teacher supervision, targeting 120 minutes of platform practice per student monthly.

The program included support across multiple dimensions. Khan Academy provided capacity building through two in-person training sessions for mathematics teachers, supplemented by online refresher training. The partnership established continued support through dedicated WhatsApp channels with district-level coordinators and a helpline for technical troubleshooting. An eight-month "Shikshak SuperStars Programme" served as an engagement campaign, recognizing high-performing schools quarterly with awards and magazine features. The program implemented data tracking and review through weekly performance data shared with coordinator groups and monthly

newsletters delivering actionable insights on active learner metrics.

Despite the above mentioned support, the program adoption remained weak. First, while these schools were equipped with sufficient tablets, access to reliable internet connectivity and electricity was occasionally subject to intermittent disruptions. Second, the Khan Academy program required teachers to complete a one-time student rostering process, which could be considered as time-intensive or not always intuitive. For teachers unfamiliar with the platform, these initial setup costs may have discouraged adoption or led to implementation difficulties. Third, and likely the main reason for weak program adoption, was an overall lack of teacher buy-in. In the absence of any convincing evidence about the benefits of the program and divided priorities across multiple objectives, Khan Academy sessions often got deprioritized or deferred. Schools failed to consistently allocate dedicated time for Khan Academy within schedules, with practice sessions occurring once monthly or less as teachers prioritized multiple competing programs. Of the 27,309 registered students, only 44 percent accessed Khan Academy even once during the entire academic year.

This implementation shortfall created the setting from which our experiment emerged. Teachers, facing competing demands and lacking dedicated support, perpetually deferred Khan Academy implementation in favor of more immediately pressing obligations. Our research question thus became whether intensive implementation support through dedicated supervisory personnel could enable successful computer-assisted learning where conventional support had proven insufficient. Following the 2023-24 academic year, we designed a randomized controlled trial to test this hypothesis directly.

We randomly assigned schools to treatment or control conditions. Treatment schools received dedicated lab-in-charges (LICs) whose sole responsibility was ensuring high-fidelity Khan Academy implementation. This intervention ran for 31 weeks from August 2024 through February 2025, focusing on students in grades 6-8, with baseline assessments conducted at the intervention start and endline assessments administered upon completion. The intervention directly addressed each barrier identified in the 2023-24 implementation shortfall. LICs were recruited primarily through

local newspaper advertisements and selected through multi-stage interviews. Throughout the intervention period, LICs received ongoing support through weekly online sessions, follow-up training modules, and instructional materials via WhatsApp. This equipped LICs to conduct two Khan Academy sessions weekly for each grade-section, formally integrating these sessions into school timetables and transforming platform usage from optional supplement to mandatory curriculum time. During sessions, LICs trained students on basic digital literacy including operating tablets and navigating the platform. They actively monitored student behavior to prevent diversion to games or other applications and escalated electricity and internet connectivity issues to school administrators and the research team for immediate troubleshooting. LICs assigned mathematical content to students, initially focusing on units teachers had recently completed but later pivoting to bridge courses for approximately six weeks to address foundational gaps. Following bridge courses, LICs implemented sequential unit progression strategies where students attempted unit tests and practiced units where performance was weak. Beyond core implementation, LICs supported program enhancements including a "streaks" campaign that incentivized consistent weekly practice with certificates, badges, medals, and earphones, and facilitated student webinars and at-home usage initiatives during winter breaks.

Control schools retained full access to the Khan Academy program, received initial training and encouragement to implement 120 minutes of weekly practice, but some of the other support mechanisms from the 2023-24 program were not strictly maintained. They could continue using Khan Academy independently and had access to monitoring systems and usage data, though the Khan Academy program was not actively promoted in control schools during the intervention period. The time table for Treated and Control schools remained identical across all subjects. The key difference was the presence of dedicated personnel in treatment schools whose full-time responsibility was ensuring implementation fidelity. This experimental design isolates implementation capacity as the crucial variable determining computer-assisted learning success. Critically, our control condition represents a baseline that reflects the setting of similar computer-assisted learning studies done in India. Our experiment thus provides a test for whether on-the-ground external

support, endorsed by school administrators can make the difference for raising CAL practice to targeted levels, and what learning gains result from such practice.

IV. Data

Our paper encompasses 83 schools from the Uttar Pradesh Social Welfare Department's residential school system. The initial sample consisted of 102 schools, from which we excluded 11 tribal schools operating under a distinct administrative structure, 7 schools which self-reported insufficient digital infrastructure for program implementation, and 1 school which was discontinued as a Social Welfare school. The remaining 83 schools span over 50 districts across the region. To ensure geographic balance, we clustered schools into geographic groups and randomly assigned one school to treatment in groups with four or fewer schools, and two schools in larger groups. Schools were assigned to treatment using a random number generator, sorted by geographic cluster and random number in ascending order, with the required number from the top of each cluster selected for treatment. This yielded our final sample of 83 schools with 28 assigned to treatment and 55 to control.

We measure learning outcomes through baseline and endline mathematics assessments administered to students in grades 6-8. Teach For Learning, an independent assessment agency, developed grade-specific tests aligned with both CBSE (Central Board of Secondary Education, India's national curriculum) and UP Board (Uttar Pradesh state curriculum) curricula and administered in Hindi. The baseline assessment occurred in August-September 2024 using a digital format hosted on SurveyMonkey. Tests incorporated anti-cheating features, randomizing question and answer option order across students. External proctors from Datamation administered all assessments, with comprehensive tracking of timestamps, completion times, and proctor identifiers enabling quality control. The endline assessment, administered in February-March 2025, employed a pen-and-paper format with optical mark recognition sheets to address logistical constraints encountered during baseline administration. There were four sets of test papers for each grade. Though all the four sets consisted of the same questions, question and option order were randomized across them.

Each grade received a 60-minute assessment, with external proctors again ensuring standardized administration across all schools.

The Khan Academy platform automatically tracked all student activity through unique student identifiers. Usage data captured total practice time in minutes, time allocation across school sessions versus at-home practice, skills worked on, skills achieving proficiency or mastery, and net skill level improvements. Our primary outcome variable is the standardized endline mathematics test score, normalized to mean zero and standard deviation one within the control group by grade. Key covariates include baseline test scores, school-level characteristics obtained by the Social Welfare Department (curriculum type, all-girls versus all-boys composition, and enrollment size), the school's United Nations' Human Development Index (according to the school's local region), and student grade level.

V. Sample Construction & Balance Tests

Table 1 presents our sample construction process from the 83 randomized schools. We construct two analytic samples to address different data quality concerns. Both samples begin by removing three control schools that could not administer endline tests. These three schools served as board centres, an administrative designation that predated treatment assignment and was unrelated to the intervention. These schools had conducted internal examinations for students earlier in the academic year, after which students were sent back home. By the scheduled endline assessment in February-March 2025, conducting examinations in these schools was not possible. These three schools were distributed across three different district groups, each of which retained multiple control schools after their removal. Appendix Table A1 confirms that including these schools in balance tests does not alter the pattern of successful randomization on pre-treatment characteristics. The Endline Only Sample retains all remaining 80 schools (28 treated and 52 control) and 7,888 students, including six schools with suspected baseline cheating identified through unusually brief test completion times and attempts outside designated testing windows. We retain these schools since we do not condition on baseline scores with this sample. The endline assessment

switched to pen-and-paper OMR format with enhanced supervision, and no such irregularities were detected. Five treatment schools discontinued participation during the intervention due to technical barriers including lack of digital devices, internet unavailability, and insufficient support. These schools remain in the Endline Only Sample based on their original treatment assignment to preserve the integrity of randomization, as removing them could introduce selection bias. As such, we include in the results section both Intent to Treat (ITT) and Treatment on the Treated (TOT) effect estimates.

The Baseline + Endline Sample excludes the six schools suspected of baseline cheating. This allows us to more confidently consider treatment effects with and without conditioning on baseline scores. Four of the five treated schools that discontinued participation early in the experiment were also schools suspected of baseline cheating. As such, estimated ITT effects are closer to TOT effects, since this sample includes only one school with discontinued treatment. The resulting Baseline + Endline Sample contains 74 schools (24 treated and 50 control) and 5,535 students, all with complete baseline and endline data enabling value-added specifications.

Table 2 shows school-level balance tests across our observable characteristics between treatment and control schools. Schools are balanced on district 2011 Human Development Index scores (mean 0.597, a United Nations composite measure of health, education, and income on a zero-to-one scale where higher values indicate greater development), gender composition (approximately 29 percent all-girls schools and 71 percent all-boys schools), curriculum type (46-48 percent CBSE and 52-54 percent UP Board)[1], and enrollment size. School-level attrition rates, measured as the fraction of registered students who did not complete testing, also do not differ significantly between treatment and control schools, indicating that sample loss was balanced at the school level. All regressions control for district group fixed effects to account for geographic clustering in the randomization process. None of the treatment-control differences are statistically significant at the 5 percent level.

[1]CBSE (Central Board of Secondary Education) is India's standardized national curriculum overseen by the central government, while UP Board is the Uttar Pradesh state curriculum designed to reflect regional educational priorities and cultural context. These curricular differences influence the scope and rigor of material covered and may affect baseline student preparation and learning approaches.

Table 3 documents student-level attrition patterns. Attrition rates are calculated relative to estimated school enrollment based on administrative registration data collected from participating schools. Baseline data is missing for approximately 24 percent of students in both groups among those who completed endline assessments, likely due to absenteeism. In the Endline Only Sample, 41.5 percent of treatment students and 46.2 percent of control students are missing endline data, likely due to students leaving the boarding schools. The Baseline + Endline Sample, which by construction requires that students completed both baseline and endline tests, exhibits higher overall attrition rates. In both samples, control students exhibit slightly higher attrition rates than treatment students. As shown in Table 2, these differences in school-level attrition rates are statistically insignificant, indicating balanced sample loss across treatment and control schools.

Student-level baseline test scores are also generally balanced. Table 4 shows that treatment and control students exhibit similar baseline mathematics achievement. In the Endline Only Sample, the treatment-control difference is 0.098 standard deviations with a standard error of 0.126, while in the Baseline + Endline Sample the difference is 0.009 standard deviations with a standard error of 0.101. Figure 1 displays the baseline score distribution for the Baseline + Endline Sample (Kolmogorov-Smirnov test p-value = 0.168)[2].

VI. Implementation Fidelity & Platform Engagement

The intervention's effectiveness depends on whether lab-in-charge support successfully increased platform engagement. We first establish implementation fidelity by examining platform usage patterns and skill development across treatment and control schools. Figures 2 and 3 compare total platform usage over the 31-week intervention period for our two core samples. In the Baseline + Endline Sample, treatment students averaged 47.4 minutes per week (1,470 total minutes) while control students averaged just 7.2 minutes per week (222 total minutes), a 6.6-fold difference. The Endline Only Sample shows a similar pattern with treatment students averaging 39.6 minutes per week (1,229 total minutes) and control students averaging 7.1 minutes per week

[2]The corresponding baseline score distribution for the Endline Only Sample is presented in Appendix Figure A1 (Kolmogorov-Smirnov test p-value = 0.0285).

(219 total minutes), a 5.6-fold difference. Control students had identical platform access but lacked dedicated support, and their usage remained concentrated near zero. Breaking down this total platform usage to active learning time on math content including watching videos, doing exercises, and reading articles, Table 5 quantifies the treatment effect on practice time after controlling for district fixed effects. The intervention increased total mathematics practice by 37.1 minutes per week (1,150 total minutes over the intervention, Panel A, $p < 0.01$). This additional practice occurred primarily during school hours, which increased by 22.0 minutes per week (682 total minutes, Panel C, $p < 0.01$), but also extended to at-home practice during holidays and after-school hours, which increased by 15.1 minutes per week (469 total minutes, Panel B, $p < 0.01$). Lab-in-charges thus established regular school routines while encouraging independent practice beyond formal instruction. Figures 4 and 5 show this engagement sustained throughout the seven-month intervention, with practice time rising quickly to 50-60 minutes per week and remaining elevated despite a temporary dip during extended holidays, while control usage stayed consistently low at 5-15 minutes per week.

Table 6 reveals this increased platform time translated into productive learning activity. Treatment students worked on 101 additional mathematics skills and achieved proficiency in 63 more skills than control students (Panels A and B, $p < 0.01$). Panel D shows treatment students mastered 0.96 more skills per hour of practice ($p < 0.01$), suggesting that lab-in-charges helped students navigate the platform effectively rather than accumulating unproductive time. Panel C shows 276 more skill-level improvements ($p < 0.01$), indicating progression through increasingly difficult material. This combination of increased practice time, sustained engagement, and improved efficiency suggests that lab-in-charge support successfully converted platform access into meaningful skill development.

Figures 6 and 7 show the raw dosage-response relationship at school and student levels. At both levels, treatment observations cluster toward higher usage and larger gains while control observations concentrate near the origin. Table 7 and Figure 8 present model predictions from this relationship. While not for direct causal interpretation, it is noteworthy that linear models predict

higher gains of 0.465 standard deviations for schools with 47.4 minutes of practice each week (the treatment school's average practice level), compared to students and schools with no KA usage. At the student level, the linear model predicts 0.406 standard deviations gains for students with 47.4 minutes of practice per week.

VII. Results

Table 8 presents intention-to-treat (ITT) and treatment-on-treated (TOT) estimates of the intervention's impact on standardized mathematics test scores. Using the Endline Only Sample, the ITT effect is 0.326 standard deviations (SE = 0.111, $p < 0.01$), while the TOT effect is 0.400 standard deviations (SE = 0.131, $p < 0.01$). The Baseline + Endline Sample without baseline controls yields similar ITT and TOT estimates of 0.445 (SE = 0.122, $p < 0.01$) and 0.467 standard deviations (SE = 0.126, $p < 0.01$), respectively. Adding baseline controls produces an ITT effect of 0.440 standard deviations (SE = 0.100, $p < 0.01$) and a TOT effect of 0.462 standard deviations (SE = 0.102, $p < 0.01$). These estimates remain remarkably consistent across specifications, ranging from 0.326 to 0.467 standard deviations with all effects significant at the one percent level. To contextualize these magnitudes, a 0.44 to 0.47 standard deviation gain represents moving an average student from the 50th percentile to approximately the 67th to 68th percentile. Using estimates from low- and middle-income countries, this effect size is equivalent to approximately two to three years of learning in business-as-usual schooling (Evans and Yuan, 2019).

Figures 9 and 10 reveal that treatment shifted the entire achievement distribution rightward rather than affecting particular performance levels. Treatment group means of 0.352 and 0.488 standard deviations in the two samples closely approximate our regression estimates (that also include stratified group fixed effects), and the distributions maintain similar shapes across groups, suggesting benefits throughout the performance spectrum.

Table 9 confirms these comprehensive gains extended uniformly across question difficulty levels on the endline assessment. Treatment students improved 0.107 standard deviations on questions at their own grade level and 0.088 to 0.147 standard deviations on questions one, two, and three

grades below (Panel C, all $p < 0.01$). F-tests fail to reject equality across difficulty levels ($p = 0.168$), confirming that the total 0.438 standard deviation effect reflects broad-based learning gains rather than concentration at particular skill levels.

Table 10 examines whether these gains were broadly distributed or concentrated among specific subgroups. Most interaction terms across school gender composition, district development levels, curriculum type, student grade, and baseline performance are statistically insignificant, suggesting treatment benefits extended throughout the study population. Panel E reveals patterns by baseline achievement: students in the highest performance tercile gained 0.521 standard deviations (SE = 0.150, $p < 0.01$), while students in the medium and low terciles gained 0.414 and 0.382 standard deviations respectively. Although point estimates suggest larger gains for initially high-performing students, these differences are not statistically significant. Thus, lab-in-charge support substantially benefited both struggling and more advanced students. The intervention similarly produced comparable effects across all-boys and all-girls schools, high and low HDI districts, and grade levels. This consistency across diverse populations and question difficulty levels indicates that dedicated implementation support generated comprehensive learning gains rather than narrow, context-specific improvements.

These results demonstrate that lab-in-charge support led to large and sustained dosage of the Khan Academy platform that generated substantial learning gains approaching half a standard deviation across the achievement spectrum, grade levels, and student subgroups. The implementation fidelity ensured by lab-in-charges facilitated an environment where increased platform usage translated into improved skill mastery and ultimately test score gains. Importantly, these gains reflect not simply usage alone, but effective implementation support that made such usage productive, successfully converting platform access into meaningful mathematical learning.

VIII. Discussion

Our results demonstrate the importance of facilitating recommended high dosage levels of computer-assisted-learning in order to achieve impressive classroom learning gains. The jump in

CAL usage for treated schools, from 7.2 to 47.4 minutes per week, drove learning gains of almost half a standard deviation. Critically, this increased time translated into productive learning rather than unproductive practice. Treatment students mastered 0.96 more skills per hour of practice than control students, demonstrating that lab-in-charges ensured both sustained participation and quality engagement. Third-party, in-school implementation support enabled high productive dosage, which facilitated skill development, which generated test score gains. This validates the implementation fidelity hypothesis established in prior literature showing that "low fidelity increases the likelihood of weak student outcomes" while adequate support structures predict program success (Hill and Erickson, 2021; Fancsali et al., 2016).

The key component to our intervention was the presence of dedicated personnel whose sole responsibility was ensuring implementation fidelity. While we tested this principle through dedicated lab-in-charges, the generalizable insight is the implementation structure itself: dedicated responsibility for platform use, protected curriculum time where computer-assisted learning substitutes for rather than supplements regular instruction, active troubleshooting capacity for technical and pedagogical challenges, quality monitoring that distinguishes productive engagement from unproductive use, extended engagement beyond school hours, and motivational systems that sustain participation. Various staffing models could deliver these features. Schools might reallocate existing teachers, hire part-time personnel, train community members, or deploy paraprofessionals. The key is that someone must be responsible and accountable for implementation, transforming computer-assisted learning from an optional supplement that gets deferred perpetually to mandatory curriculum time with dedicated support.

Our effect sizes of 0.326 to 0.467 standard deviations in seven months exceed comparable interventions while demonstrating superior cost-effectiveness. Mindspark's implementation in Uttar Pradesh government schools achieved 0.22 standard deviations over 18 months using proprietary adaptive software and lab-in-charges for hardware support (Muralidharan and Singh, 2025), meaning we achieved double their effect in less than half the time. One explanation is from higher quality practice time, facilitated in our study by more dedicated class time and comprehensive lab-

in-charge support. Another possibility is that Khan Academy was integrated more closely with the classroom curriculum. Mindspark's own experience confirms the critical role of implementation support: when they reduced lab-in-charge presence in their third year, usage initially dropped by approximately 50 percent before recovering after several months of intensive monitoring efforts (Muralidharan and Singh, 2025). Banerjee et al. (2007) achieved 0.47 standard deviations in fourth grade mathematics by providing two hours of weekly computer-assisted learning time with dedicated supervision, demonstrating that similar magnitudes are achievable with simpler platforms when implementation ensures consistent practice. Our gains also exceed the 0.36 standard deviation effect size from high-dosage tutoring meta-analyses (Dietrichson et al., 2017). The critical insight is that properly implemented computer-assisted learning can deliver tutoring-like personalization, but requires a dedicated support structure to ensure implementation fidelity. The binding constraint for scaling personalized learning appears to be not the sophistication of adaptive algorithms or the quality of digital content, but rather whether educational systems possess the implementation capacity to ensure technology translates into sustained, productive student engagement.

A. Cost-Effectiveness and Scalability

At $14 (USD) per student over seven months, equivalent to approximately $24 (USD) annually, our intervention demonstrates cost-effectiveness through part-time implementation labor and a freely available platform. The cost structure allocates 61 percent to lab-in-charge salaries and training, 22 percent to program management, and 17 percent to monitoring and motivational campaigns. Lab-in-charges were part-time contractors dedicating two to four days weekly rather than full-time employees, while the Khan Academy platform is completely free with no licensing fees[3]. This cost advantage stems from two factors. First, using a freely available platform eliminates licensing fees that proprietary software requires. Second, our part-time lab-in-charge model reduced

[3]For comparison, Mindspark's scaled implementation in Uttar Pradesh cost approximately $39 (USD) per student annually during Years 1 and 2 when they measured their 0.22 standard deviation effect (Muralidharan and Singh, 2025), meaning we achieved roughly double their effect size at 60 percent of their cost within the same Indian context.

labor costs while still providing comprehensive implementation support. Compared to high-dosage tutoring programs costing thousands per student annually (Oreopoulos et al., 2024; Strassberger and Condliffe, 2024; White, Groom-Thomas and Loeb, 2023) and generating effect sizes of 0.36 standard deviations (Dietrichson et al., 2017), we produced larger learning gains at substantially lower cost.

While our absolute costs reflect Indian wage structures and will be higher in higher-wage countries, the cost structure reveals that effective implementation depends on labor for supervision and support rather than expensive technology or proprietary software. The primary concern for schools considering this approach is not the technology cost but rather the investment in implementation personnel. However, the key lesson from our experiment is that lab-in-charges were not the necessary variable for these large effects. Rather, the implementation structure of our intervention enabled high productive dosage with computer-assisted learning, which led to substantial learning gains. While staffing arrangements and costs will vary across contexts, the fundamental tradeoff remains: intensive implementation support costs substantially more than laissez-faire technology adoption but vastly less than high-dosage tutoring while generating comparable or larger learning gains.

B. Limitations

Several limitations warrant acknowledgment in interpreting these findings. The intervention lasted 31 weeks. Whether learning gains persist, fade, or compound with continued exposure remains a topic for future research. Mathematics learning was measured exclusively through standardized test scores. Effects on other subjects, problem-solving skills, or broader competencies are unknown, though the personalized learning principles validated here should generalize across content areas if implementation structures remain intact and the time table for both Treatment and Control schools remains the same across all subjects. The study occurred in boarding schools within Uttar Pradesh's Social Welfare system. This context may facilitate implementation because students are present throughout the day and schools face fewer competing demands from families

or extracurricular activities compared to day schools. Results may not generalize to other countries, different grade levels beyond 6-8, or contexts where family, community, and other factors play larger roles in educational processes. Intensive monitoring and attention in treatment schools may have generated Hawthorne effects that boosted performance beyond what the implementation structure alone would produce. However, sustained usage throughout the full 31-week period rather than declining after initial novelty suggests genuine engagement rather than temporary enthusiasm.

Even with intensive support, implementation faced barriers that constrain generalizability. Five of 28 treatment schools discontinued participation during the intervention due to technical barriers including insufficient digital devices, unavailable internet connectivity, and inadequate infrastructure support. These schools remain in intention-to-treat estimates to preserve randomization integrity, but their experience demonstrates that even dedicated implementation support cannot overcome all infrastructure constraints. Among treatment schools, the achieved dosage still varied substantially. Weekly practice time ranged from approximately 25 minutes per week, on average, in some schools to over 65 minutes per week in others, reflecting variation in schools' capacity to integrate CAL sessions consistently despite similar lab-in-charge support and practice targets. However, the difference compared to other CAL studies is that, despite this variation, even the lowest practice schools maintained significant dosage levels during the school-year.

The study required schools to possess basic digital infrastructure of tablets, internet access, and reasonably reliable electricity. Implementation challenges differ substantially in contexts lacking this foundation, where providing technology access represents a prerequisite rather than a given. Attrition rates of approximately 55 to 60 percent relative to estimated enrollment, though balanced across treatment and control conditions, mean results describe effects for students who completed assessments rather than all enrolled students. Whether non-completers would have experienced similar gains remains unknown.

IX. Conclusion

Computer-assisted learning has long promised to facilitate personalized instruction, yet implementations have often struggled to deliver on this potential. Our experience from this study suggests that the primary constraint is not technology but implementation capacity. When dedicated personnel ensured platform use through protected curriculum time and active support, the same Khan Academy platform that produced minimal results elsewhere generated learning gains approaching half a standard deviation. Implementation support enabled productive dosage that facilitated skill development and test score improvements, achieving these gains at substantially lower cost than alternative interventions. The implication is that scaling personalized learning requires investing in organizational structures that guarantee implementation fidelity rather than simply providing better platforms or more comprehensive training. Looking forward, emerging AI-powered tutoring systems may offer additional pedagogical benefits beyond the solutions and videos available in current platforms, potentially enhancing learning effectiveness when combined with strong implementation support. However, further research is needed to determine how best to integrate such technologies within effective implementation structures. Ultimately, the challenge ahead remains clear: building CAL programs with the organizational capacity to translate platform access into sustained productive learning.

References

Angrist, Noam, Micheal Ainomugisha, Sai Pramod Bathena, Peter Bergman, Colin Crossley, Claire Cullen, Thato Letsomo, Moitshepi Matsheng, Rene Marlon Panti, Shwetlena Sabarwal, and Tim Sullivan. 2023. "Building Resilient Education Systems: Evidence from Large-Scale Randomized Trials in Five Countries." National Bureau of Economic Research Working Paper 31208.

ASER Centre. 2025. "Annual Status of Education Report (Rural) 2024 (Provisional)." ASER Centre, New Delhi.

Banerjee, Abhijit, Rukmini Banerji, James Berry, Esther Duflo, Harini Kannan, Shobhini Mukherji, Marc Shotland, and Michael Walton. 2016. "Mainstreaming an Effective Intervention: Evidence from Randomized Evaluations of 'Teaching at the Right Level' in India." National Bureau of Economic Research Working Paper 22746.

Banerjee, Abhijit V., Shawn Cole, Esther Duflo, and Leigh Linden. 2007. "Remedying Education: Evidence from Two Randomized Experiments in India." *Quarterly Journal of Economics*, 122(3): 1235–1264.

Barrow, Lisa, Lisa Markman, and Cecilia Elena Rouse. 2009. "Technology's Edge: The Educational Benefits of Computer-Aided Instruction." *American Economic Journal: Economic Policy*, 1(1): 52–74.

Beck, Joseph E., and Yue Gong. 2013. "Wheel-Spinning: Students Who Fail to Master a Skill." In *Artificial Intelligence in Education. AIED 2013*. Vol. 7926 of *Lecture Notes in Computer Science*, , ed. H. Chad Lane, Kalina Yacef, Jack Mostow and Philip Pavlik, 431–440. Berlin:Springer.

Beck, Robert J. 2007. "Towards a Pedagogy of the Oxford Tutorial." Lawrence University.

Bettinger, Eric, Robert Fairlie, Anastasia Kapuza, Elena Kardanova, Prashant Loyalka, and Andrey Zakharov. 2023. "Diminishing Marginal Returns to Computer-Assisted Learning." *Journal of Policy Analysis and Management*, 42: 552–570.

Büchel, Konstantin, Martina Jakob, Christoph Kühnhanss, Daniel Steffen, and Aymo Brunetti. 2020. "The Relative Effectiveness of Teachers and Learning Software: Evidence from a Field Experiment in El Salvador." University of Bern, Department of Social Sciences Social Sciences Working Paper 36.

Canbolat, Yusuf, and Rebeca Arndt. 2024. "Can Computer-Assisted Instruction Help Schools to Close the Achievement Gap: Evaluation of a District-Wide Reading Intervention." Annenberg Institute at Brown University EdWorkingPaper 24-938.

Cascio, Elizabeth, and Douglas Staiger. 2012. "Knowledge, Tests, and Fadeout in Education Interventions." National Bureau of Economic Research Working Paper 18038.

Copeland, Susan, Michael A. Cook, Ashley A. Grant, and Steven M. Ross. 2023. "Randomized-Control Efficacy Study of IXL Math in Holland Public Schools." Johns Hopkins Center for Research and Reform in Education, Baltimore, MD.

Dietrichson, Jens, Martin Bøg, Trine Filges, and Anne-Marie Klint Jørgensen. 2017. "Academic Interventions for Elementary and Middle School Students with Low Socioeconomic Status: A Systematic Review and Meta-Analysis." *Review of Educational Research*, 87(2): 243–282.

Duflo, Esther, Pascaline Dupas, and Michael Kremer. 2011. "Peer Effects, Teacher Incentives, and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya." *American Economic Review*, 101: 1739–1744.

Duncan, Greg J., Chantelle J. Dowsett, Amy Claessens, Katherine Magnuson, Aletha C. Huston, Pamela Klebanov, Linda S. Pagani, Leon Feinstein, Mimi Engel, Jeanne Brooks-Gunn, Holly Sexton, Kathryn Duckworth, and Christa Japel. 2007. "School Readiness and Later Achievement." *Developmental Psychology*, 43(6): 1428–1446.

Escueta, Maya, Andre J. Nickow, Philip Oreopoulos, and Vincent Quan. 2020. "Upgrading Education with Technology: Insights from Experimental Research." *Journal of Economic Literature*, 58(4): 897–996.

Evans, David, and Fei Yuan. 2019. "Equivalent Years of Schooling: A Metric to Communicate Learning Gains in Concrete Terms." World Bank Group Policy Research Working Paper WPS 8752, Washington, D.C.

Fahle, Erin, Thomas J. Kane, Sean F. Reardon, and Douglas Staiger. 2024. "The First Year of Pandemic Recovery: A District-Level Analysis." Center for Education Policy Research, Harvard University.

Fancsali, Stephen E., Steven Ritter, Michael Yudelson, Michael Sandbothe, and Susan R. Berman. 2016. "Implementation Factors and Outcomes for Intelligent Tutoring Systems: A Case Study of Time and Efficiency with Cognitive Tutor Algebra." 473–478.

Ferman, Bruno, Lucas Finamor, and Lycia Lima. 2019. "Are Public Schools Ready to Integrate Math Classes with Khan Academy?" Munich Personal RePEc Archive Paper 94736.

Guryan, Jonathan, and Jens Ludwig. 2023. "Overcoming Pandemic-Induced Learning Loss." In *Building a More Resilient US Economy*. , ed. Melissa S. Kearney, Justin Schardin and Luke Pardue. Washington, DC:Aspen Institute.

Guryan, Jonathan, Jens Ludwig, Manasi P. Bhatt, Philip J. Cook, Jonathan M. V. Davis, Kenneth Dodge, George Farkas, Roland G. Fryer, Susan Mayer, Harold Pollack, Laurence Steinberg, and Greg Stoddard. 2023. "Not Too Late: Improving Academic Outcomes among Adolescents." *American Economic Review*, 113(3): 738–765.

Hill, Heather C., and Anna Erickson. 2021. "Using Implementation Fidelity to Aid in Interpreting Program Impacts: A Brief Review." Annenberg Institute at Brown University EdWorkingPaper 21-414.

Kremer, Michael, Alex Eble, Guthrie Gray-Lobe, Saloni Gupta, Sabareesh Ramachandran, and Wendy Wong. 2025. "Results from Randomised Evaluation of Personalised Adaptive Learning (PAL) in Andhra Pradesh, India (2023-2025)." Development Innovation Lab, University of Chicago.

Linden, Leigh. 2008. "Complement or Substitute? The Effect of Technology on Student Achievement in India." World Bank Working Paper 44863.

Morgan, Pat, and Steven Ritter. 2002. "An Experimental Study of the Effects of Cognitive Tutor Algebra I on Student Knowledge and Attitude." Carnegie Learning, Inc., Pittsburgh, PA.

Muralidharan, Karthik, Abhijeet Singh, and Alejandro J. Ganimian. 2019. "Disrupting Education? Experimental Evidence on Technology-Aided Instruction in India." *American Economic Review*, 109(4): 1426–1460.

Muralidharan, Karthik, and Abhijeet Singh. 2025. "Adapting for Scale: Experimental Evidence on Computer-Aided Instruction in India." Working paper.

Nickow, Andre, Philip Oreopoulos, and Vincent Quan. 2020. "The Impressive Effects of Tutoring on PreK-12 Learning: A Systematic Review and Meta-Analysis of the Experimental Evidence." National Bureau of Economic Research Working Paper 27476.

Nickow, Andre, Philip Oreopoulos, and Vincent 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(1): 74–107.

Nielsen, Eric. 2023. "The Variance of Achievement Increases During Childhood." Federal Reserve Board of Governors Working paper.

Oreopoulos, Philip, Chloe Gibbs, Michael Jensen, and Joseph Price. 2024. "Teaching Teachers to Use Computer Assisted Learning Effectively: Experimental and Quasi-Experimental Evidence." National Bureau of Economic Research Working Paper 32388.

Pane, John F., Beth Ann Griffin, Daniel F. McCaffrey, and Rita Karam. 2014. "Effectiveness of Cognitive Tutor Algebra I at Scale." *Educational Evaluation and Policy Analysis*, 36(2): 127–144.

Pane, John F., Daniel F. McCaffrey, Mary Ellen Slaughter, Jennifer L. Steele, and Gina S. Ikemoto. 2010. "An Experiment to Evaluate the Efficacy of Cognitive Tutor Geometry." *Journal of Research on Educational Effectiveness*, 3: 254–281.

Peters, Scott J., Karen Rambo-Hernandez, Matthew C. Makel, Michael S. Matthews, and Jonathan A. Plucker. 2017. "Should Millions of Students Take a Gap Year? Large Numbers of Students Start the School Year above Grade Level." *Gifted Child Quarterly*, 61(3): 229–238.

Phillips, Andrea, John F. Pane, Rebecca Reumann-Moore, and Oluwatosin Shenbanjo. 2020. "Implementing an Adaptive Intelligent Tutoring System as an Instructional Supplement." *Educational Technology Research and Development*, 68: 1409–1437.

Rodriguez-Segura, Daniel. 2022. "EdTech in Developing Countries: A Review of the Evidence." *The World Bank Research Observer*, 37(2): 171–203.

Roschelle, Jeremy, Mingyu Feng, Robert F. Murphy, and Craig A. Mason. 2016. "Online Mathematics Homework Increases Student Achievement." *AERA Open*, 2(4).

Singh, Abhijeet. 2025. "Education Technology in Low- and Middle-Income Countries: Experimental Evidence on Computer-Aided Instruction in India." Working paper, Stockholm School of Economics.

Stanovich, Keith E. 1986. "Matthew Effects in Reading: Some Consequences of Individual Differences in the Acquisition of Literacy." *Reading Research Quarterly*, 21(4): 360–407.

Strassberger, Marissa, and Barbara Condliffe. 2024. "How to Build It and Ensure They Will Come: Educators' Advice on High-Dosage Tutoring Programs." MDRC, New York.

White, Sara, Leiah Groom-Thomas, and Susanna Loeb. 2023. "A Systematic Review of Research on Tutoring Implementation: Considerations when Undertaking Complex Instructional Supports for Students." Annenberg Institute at Brown University EdWorkingPaper 22-652.

Table 1: Sample Construction

	Schools			Removed Schools		
Sample	**Total**	**Treated**	**Control**	**Total**	**Treated**	**Control**
Endline Only						
Initial Sample Size	83	28	55	—	—	—
Remove: Schools without endline data	80	28	52	3	0	3
Remaining: Discontinued treatment schools	—	5	0	—	—	—
Final Sample Size	*80*	*28*	*52*	—	—	—
Baseline + Endline						
Initial Sample Size	83	28	55	—	—	—
Remove: Schools without endline data	80	28	52	3	0	3
Remove: Baseline cheating schools	74	24	50	6	4	2
Remaining: Discontinued treatment schools	—	1	0	—	—	—
Final Sample Size	*74*	*24*	*50*	—	—	—

Notes: Samples constructed from 83 schools (28 treatment, 55 control) randomized to treatment intervention. Endline Only Sample focuses on endline analysis only, excluding schools without endline tests. Baseline + Endline Sample enables baseline and endline analysis by additionally removing schools with suspected baseline cheating. Discontinued treatment schools are those where the treatment could not continue due to technical barriers such as lack of digital devices, internet unavailability, or insufficient support. Final sample sizes: 7,888 students for Endline Only Sample, 5,535 students for Baseline + Endline Sample.

Table 2: School-Level Balance Tests

	Endline Only (N=80)		Baseline + Endline (N=74)	
Variable	**Control Mean [SD]**	**T-C Diff. (SE)**	**Control Mean [SD]**	**T-C Diff. (SE)**
District HDI Score	0.597	0.005	0.596	0.004
	[0.051]	(0.009)	[0.052]	(0.010)
All-Girls School	0.288	0.009	0.280	0.009
		(0.106)		(0.113)
CBSE Curriculum Adoption	0.462	0.054	0.480	0.061
		(0.119)		(0.131)
School Size	94.519	11.624	71.040	11.458
	[38.652]	(8.517)	[29.599]	(7.853)
Missing Endline Survey	0.467	-0.049	0.597	-0.048
	[0.169]	(0.040)	[0.133]	(0.038)
Missing Baseline Survey	0.238	0.019	—	—
	[0.104]	(0.026)		
Regression specifications				
District Group Fixed Effects		✓		✓

Notes: School-level balance tests regressing each school characteristic on treatment assignment. Each cell shows control group means with standard deviations in square brackets for continuous variables and treatment-control differences with standard errors in parentheses. District group fixed effects control for geographic clustering in treatment assignment. District HDI Score is Human Development Index (0-1 scale). All-Girls School is proportion of all-girls schools versus all-boys schools (binary variable). CBSE Curriculum Adoption is proportion using CBSE (Central Board of Secondary Education) versus UP (Uttar Pradesh) Board curriculum (binary variable). School size is number of students per school. Missing Endline Survey is the fraction of registered students who did not complete endline testing (out-of-sample attrition). Missing Baseline Survey is the fraction of sample students missing baseline test scores and applies only to Endline Only Sample (Baseline + Endline Sample requires complete baseline data). Endline Only Sample size: 80 schools (28 treatment, 52 control). Baseline + Endline Sample size: 74 schools (24 treatment, 50 control). No significance stars appear as all differences are non-significant at conventional levels, indicating successful randomization and balanced attrition. Significance levels: *** $p \leq 0.01$, ** $p \leq 0.05$, * $p \leq 0.10$.

Table 3: Student-Level Attrition Summary

Sample	Est. Registered	Sample Endline	Missing Endline	Sample Baseline	Missing Baseline
Endline Only					
Treated	5,078	2,973	0.415	2,248	0.244
Control	9,128	4,915	0.462	3,732	0.241
Total	14,206	7,888	0.445	5,980	0.242
Baseline + Endline					
Treated	4,356	1,983	0.545	1,983	0.000
Control	8,775	3,552	0.595	3,552	0.000
Total	13,131	5,535	0.578	5,535	0.000

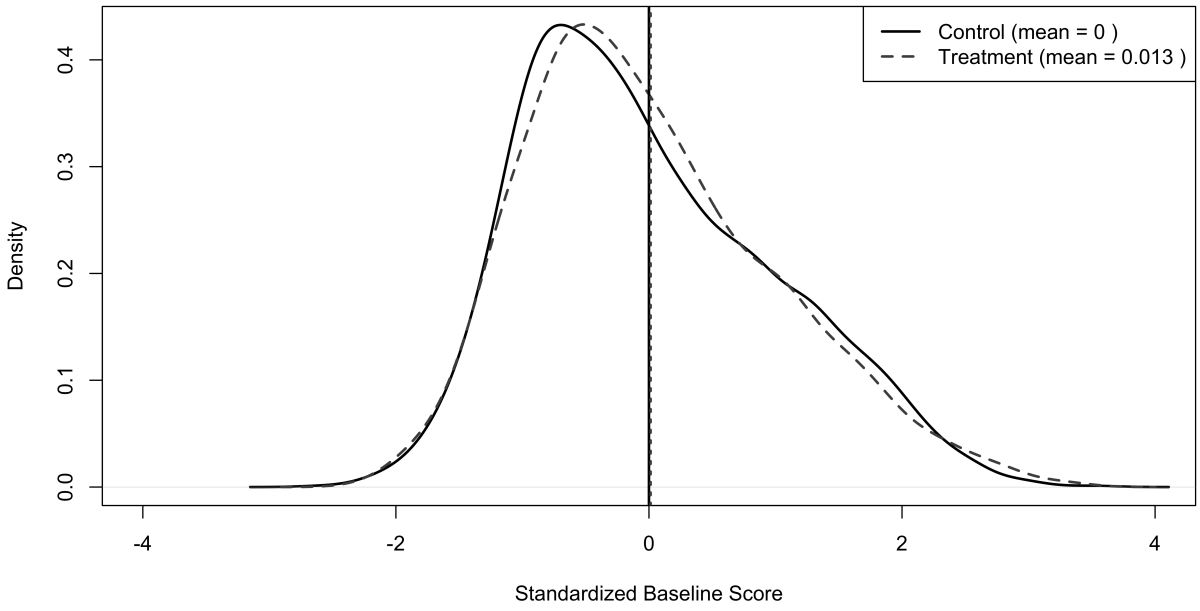
Notes: Student-level attrition summary showing sample composition and missingness patterns across treatment groups. Est. Registered represents estimated total school enrollment based on administrative registration data collected from participating schools. Sample Endline shows the number of students with completed endline test scores who are included in the analysis samples. Missing Endline represents the proportion of estimated registered students who did not complete endline testing. Sample Baseline shows the number of students in the sample who have baseline test scores available. Missing Baseline represents the proportion of students in the sample who lack baseline test scores. Endline Only Sample includes all students with endline test data regardless of baseline test availability. Baseline + Endline Sample restricts to students with both baseline and endline test scores, resulting in zero baseline missingness. Sample sizes: Endline Only Sample contains 7,888 students from 80 schools, Baseline + Endline Sample contains 5,535 students from 74 schools.

Table 4: Student-Level Baseline Balance Tests

Sample	**Control Mean [SD]**	**T-C Diff. (SE)**
Endline Only	0.000	0.098
	[0.9997]	(0.126)
Baseline + Endline	0.000	0.009
	[0.9997]	(0.101)
Regression specifications		
District Group Fixed Effects		✓
School Level Clustered SEs		✓

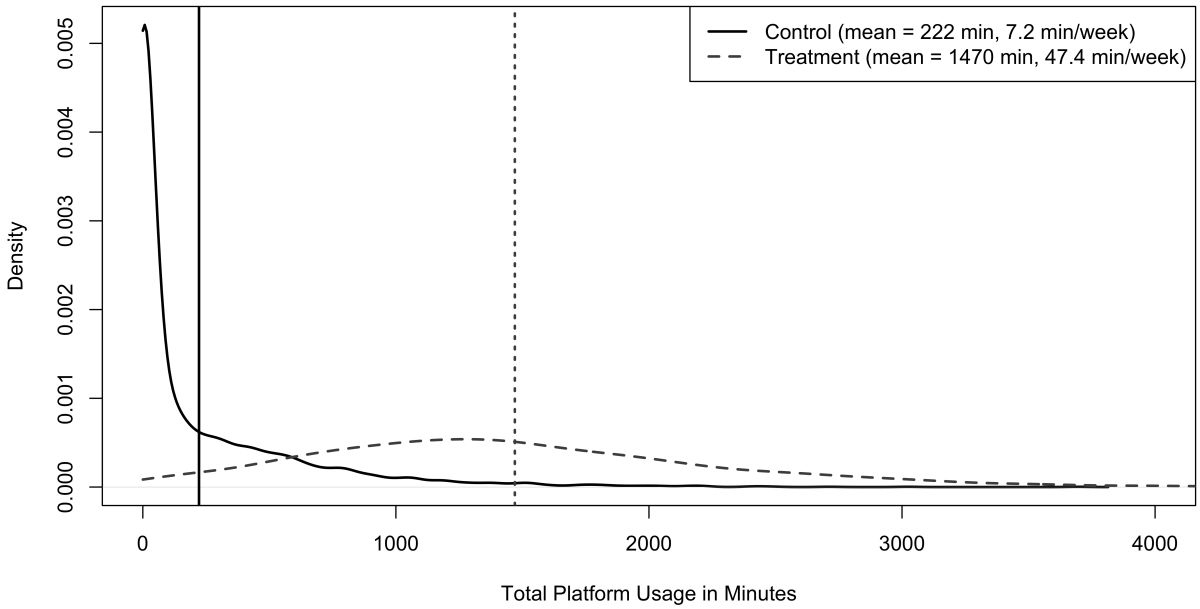
Notes: Student-level baseline balance tests regressing standardized baseline test scores on treatment assignment with school-level clustering and district fixed effects. T-C Diff. shows treatment-control differences with school-clustered standard errors in parentheses. Control group means and standard deviations in square brackets reflect standardized baseline scores. District group fixed effects control for geographic clustering in treatment assignment. Endline Only Sample includes 5,980 students with baseline data from 7,888 total students with endline data. Baseline + Endline Sample includes 5,535 students with complete baseline and endline data. No significance stars appear as all differences are non-significant at conventional levels ($p > 0.10$), indicating successful randomization. Significance levels: *** $p \leq 0.01$, ** $p \leq 0.05$, * $p \leq 0.10$.

Figure 1: Baseline Test Scores - Baseline + Endline Sample



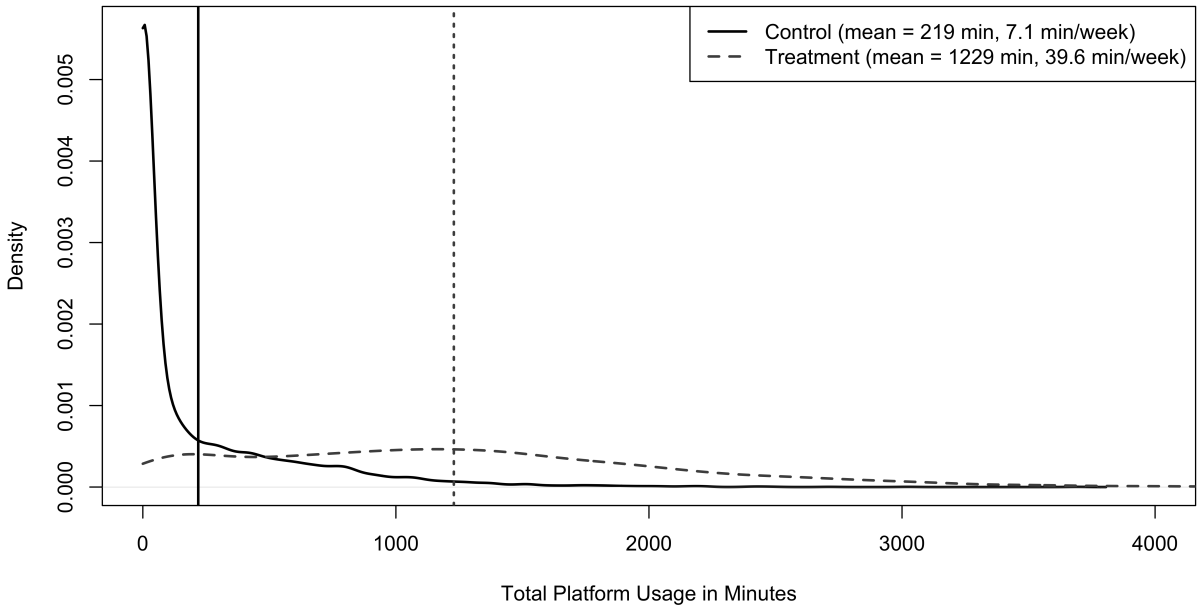
Notes: Kernel density plots comparing standardized baseline test score distributions between treatment and control groups. Solid line shows control group distribution, dashed line shows treatment group distribution. Vertical lines indicate group means (control mean = 0, treatment mean = 0.013). Baseline + Endline Sample includes 5,535 students with available baseline data.

Figure 2: Platform Usage - Baseline + Endline Sample



Notes: Kernel density plots comparing total platform usage distributions between treatment and control groups. Solid line shows control group distribution, dashed line shows treatment group distribution. Vertical lines indicate group means (control mean = 222 minutes, treatment mean = 1470 minutes). Platform usage measured in total minutes over the 31-week intervention period. Weekly averages: control 7.2 minutes/week, treatment 47.4 minutes/week. Baseline + Endline Sample includes 5,460 students with available usage data.

Figure 3: Platform Usage - Endline Only Sample



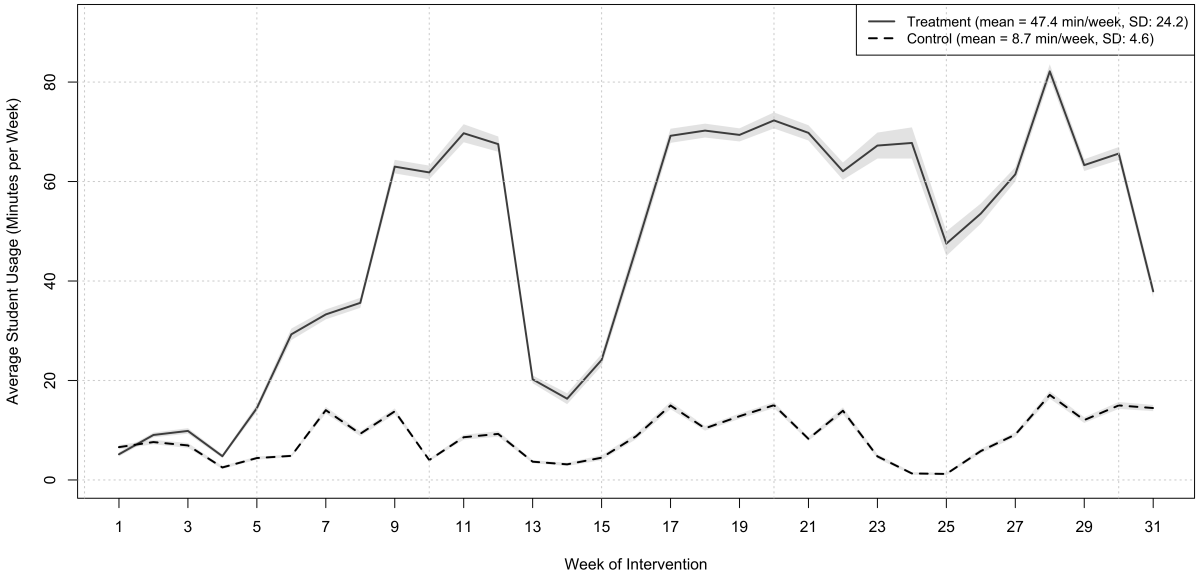
Notes: Kernel density plots comparing total platform usage distributions between treatment and control groups. Solid line shows control group distribution, dashed line shows treatment group distribution. Vertical lines indicate group means (control mean = 219 minutes, treatment mean = 1229 minutes). Platform usage measured in total minutes over the 31-week intervention period. Weekly averages: control 7.1 minutes/week, treatment 39.6 minutes/week. Endline Only Sample includes 7,764 students with available usage data.

Table 5: Treatment Effects on Practice Time

	Endline Only	**Baseline + Endline**	
	(1)	(2)	(3)
Panel A: Total Practice Time			
Treatment Effect	920.99*** (92.11)	1150.17*** (77.49)	1150.41*** (76.90)
Control Mean [SD]	235.40 [329.67]	236.36 [330.09]	236.36 [330.09]
Panel B: At-Home Practice			
Treatment Effect	379.96*** (47.06)	468.66*** (49.41)	468.78*** (49.35)
Control Mean [SD]	43.44 [106.37]	45.18 [113.06]	45.18 [113.06]
Panel C: At-School Practice			
Treatment Effect	541.02*** (63.49)	681.51*** (56.84)	681.63*** (56.50)
Control Mean [SD]	191.96 [272.58]	191.18 [270.77]	191.18 [270.77]
Panel D: Holiday Practice			
Treatment Effect	226.26*** (26.59)	277.65*** (28.70)	277.72*** (28.91)
Control Mean [SD]	20.54 [63.02]	21.71 [66.41]	21.71 [66.41]
Panel E: After-School Practice			
Treatment Effect	262.79*** (39.34)	324.03*** (42.90)	324.11*** (42.90)
Control Mean [SD]	28.78 [84.46]	29.62 [90.73]	29.62 [90.73]
Regression specifications			
Baseline Controls			✓
District Group Fixed Effects	✓	✓	✓
School Level Clustered SEs	✓	✓	✓

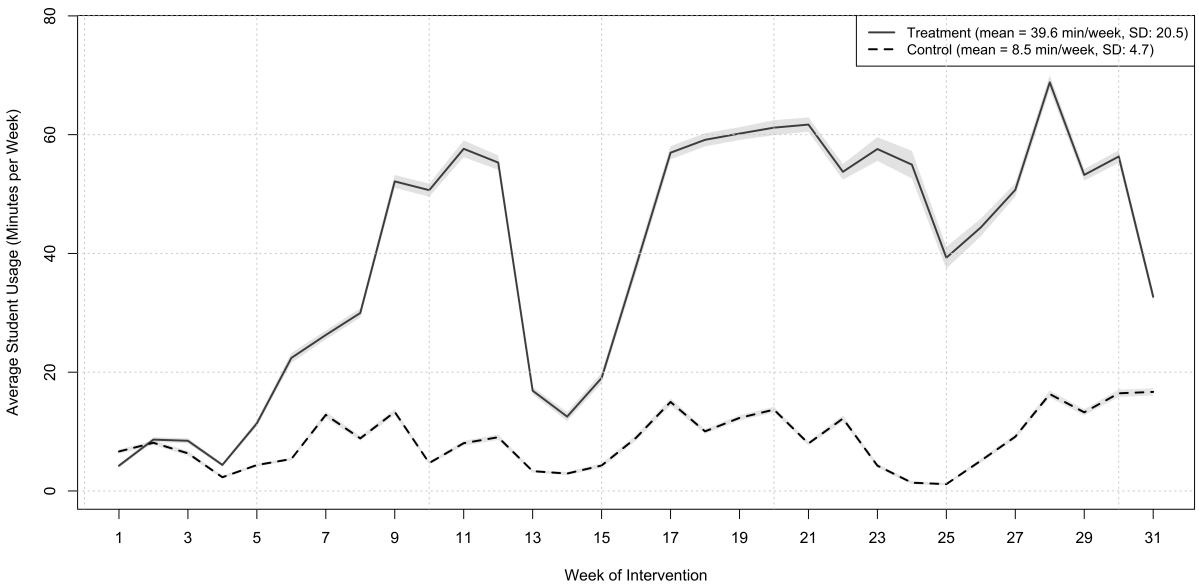
Notes: Treatment effects on math practice time over the 31-week intervention period. Each panel shows results from separate regressions with school-clustered standard errors in parentheses and control means with standard deviations in square brackets. All practice time measured in minutes. Total Practice Time (Panel A) measures active learning time on math content including watching videos, doing exercises, and reading articles. At-Home Practice (Panel B) combines usage during holidays and after-school hours. At-School Practice (Panel C) equals Total Practice Time minus At-Home Practice. Holiday Practice (Panel D) includes usage during any school holidays excluding Sundays. After-School Practice (Panel E) includes usage during after-school hours on school days and all day on Sundays. All regressions include district group fixed effects and school-level clustering. Baseline controls (column 3 only) include standardized baseline test scores. Students without usage identifiers excluded from analysis. Endline Only Sample: 6,923 students. Baseline + Endline Sample: 4,853 students. Significance levels: *** p $\leq$ 0.01, ** p $\leq$ 0.05, * p $\leq$ 0.10.

Figure 4: Weekly CAL Usage Over Time - Baseline + Endline Sample



Notes: Weekly average CAL usage in minutes per week over the 31-week intervention period (August 2024 to February 2025). Solid line represents treatment group average, dashed line represents control group average. Shaded regions represent standard error bands. Analysis includes 4,857 students (Treatment: 1,955, Control: 2,902) with valid usage identifiers from Baseline + Endline Sample. Students without usage identifiers (678 students, 12.2% of sample) were excluded as their usage data could not be linked to treatment assignment. Weekly usage calculated as zero for students with valid identifiers but no recorded usage in a given week.

Figure 5: Weekly CAL Usage Over Time - Endline Only Sample



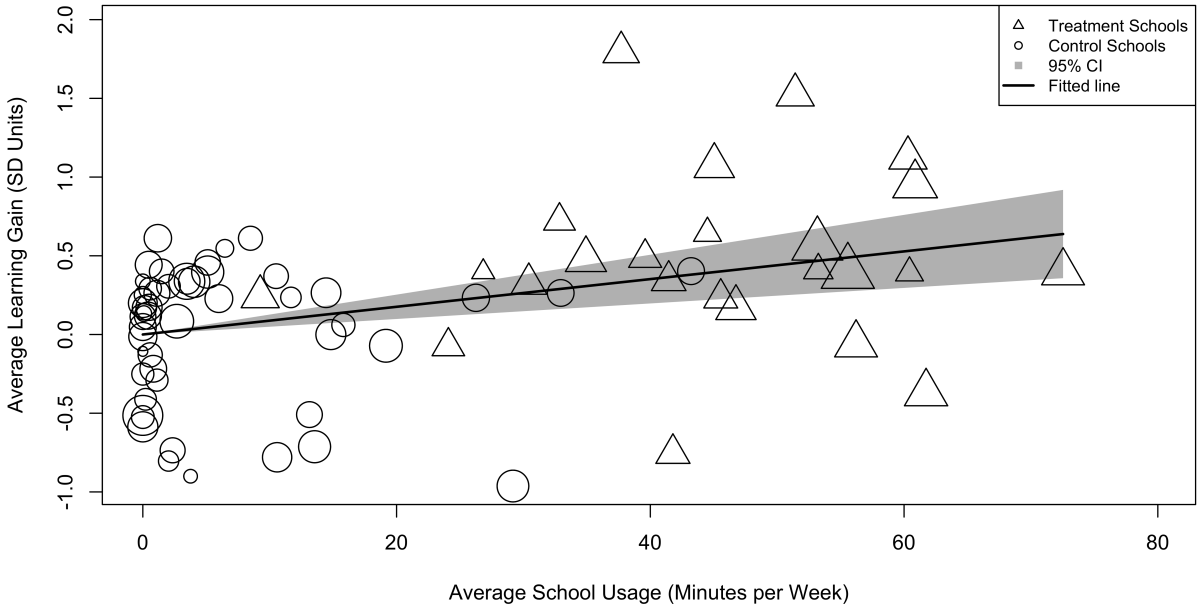
Notes: Weekly average CAL usage in minutes per week over the 31-week intervention period (August 2024 to February 2025). Solid line represents treatment group average, dashed line represents control group average. Shaded regions represent standard error bands. Analysis includes 6,926 students (Treatment: 2,916, Control: 4,010) with valid usage identifiers from Endline Only Sample. Students without usage identifiers (961 students, 12.2% of sample) were excluded as their usage data could not be linked to treatment assignment. Weekly usage calculated as zero for students with valid identifiers but no recorded usage in a given week.

Table 6: Treatment Effects on Skills Mastery

	Endline Only	**Baseline + Endline**	
	(1)	(2)	(3)
Panel A: Skills Worked On			
Treatment Effect	82.25*** (8.03)	101.33*** (6.58)	101.35*** (6.56)
Control Mean [SD]	20.24 [31.05]	19.84 [30.72]	19.84 [30.72]
Panel B: Skills Proficient			
Treatment Effect	49.56*** (6.01)	63.46*** (5.64)	63.49*** (5.52)
Control Mean [SD]	9.83 [19.63]	9.42 [19.52]	9.42 [19.52]
Panel C: Net Skill Upgrades			
Treatment Effect	218.40*** (23.83)	275.53*** (21.24)	275.64*** (20.90)
Control Mean [SD]	47.30 [83.02]	45.96 [82.85]	45.96 [82.85]
Panel D: Skills Proficient per Hour			
Treatment Effect	0.62*** (0.22)	0.95*** (0.21)	0.96*** (0.21)
Control Mean [SD]	2.16 [2.60]	2.06 [2.57]	2.06 [2.57]
Regression specifications			
Baseline Controls			✓
District Group Fixed Effects	✓	✓	✓
School Level Clustered SEs	✓	✓	✓

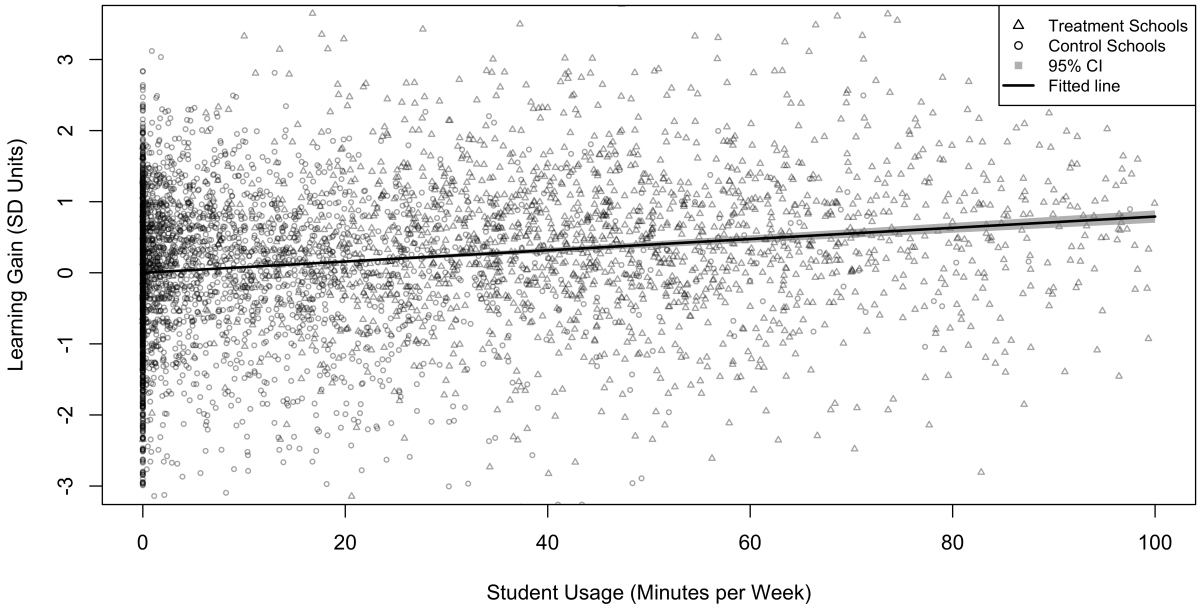
Notes: Treatment effects on math skills mastery over the 31-week intervention period. Each panel shows results from separate regressions with school-clustered standard errors in parentheses and control means with standard deviations in square brackets. Skills Worked On (Panel A) counts unique math skills students engaged with during the intervention. Skills Proficient (Panel B) counts skills where students achieved proficiency level or higher. Net Skill Upgrades (Panel C) measures total skill level improvements across all mathematics skills. Skills Proficient per Hour (Panel D) measures learning efficiency as the ratio of skills achieving proficiency to total practice hours, restricted to students with positive practice time. All regressions include district group fixed effects and school-level clustering. Baseline controls (column 3 only) include standardized baseline test scores. Students without usage/skill identifiers excluded from analysis. Endline Only Sample: 6,923 students for Panels A-C, 5,535 students for Panel D. Baseline + Endline Sample: 4,853 students for Panels A-C, 3,954 students for Panel D. Significance levels: *** $p \leq 0.01$, ** $p \leq 0.05$, * $p \leq 0.10$.

Figure 6: School-Level Dosage-Response Relationship



Notes: School-level relationship between average platform usage and learning gains. Learning gain measured as difference between standardized endline and baseline test scores. Triangles represent treatment schools, circles represent control schools, with point sizes proportional to number of students per school. Fitted line represents linear regression of average school learning gains on average school weekly usage minutes, forced through the origin, using all schools (treatment and control). Gray shaded area represents 95% confidence interval. Analysis includes schools from Sample 2 with complete baseline, endline, and usage data.

Figure 7: Student-Level Dosage-Response Relationship



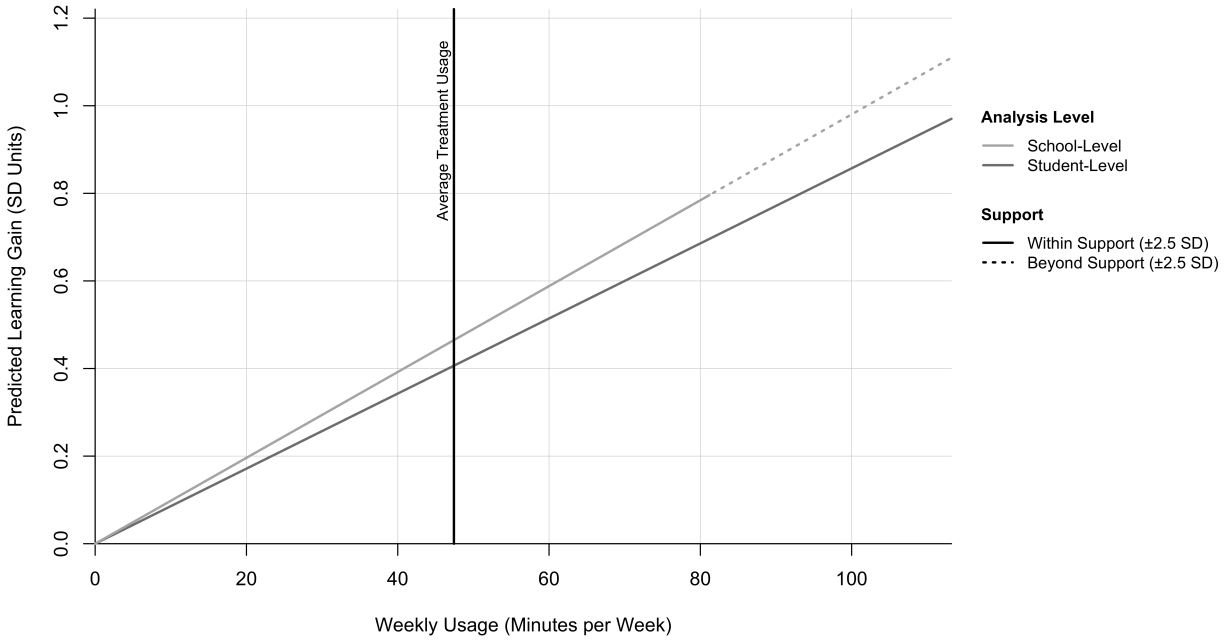
Notes: Student-level relationship between individual platform usage and learning gains. Learning gain measured as difference between standardized endline and baseline test scores. Triangles represent treatment students, circles represent control students. Fitted line represents linear regression of individual student learning gains on individual student weekly usage minutes, forced through the origin, using all students (treatment and control). Gray shaded area represents 95% confidence interval. Usage truncated at 100 minutes per week for display. Analysis includes students from Sample 2 with complete baseline, endline, and usage data.

Table 7: Predicted Learning Gains by Usage and Model

	Average Treatment Usage	
Model	**Student-Level**	**School-Level**
Linear	+0.406 SD	+0.465 SD
Quadratic	+0.493 SD	+0.524 SD
Cubic	+0.514 SD	+0.538 SD

Notes: Predicted learning gains from dosage-response models evaluated at average treatment group usage (47.4 minutes per week, corresponding to 1,470 total minutes over the 31-week intervention period). All models regress learning gains (difference between standardized endline and baseline test scores) on weekly usage minutes, forced through the origin, using treatment group data only. Linear model includes usage as single predictor. Quadratic model adds squared usage term. Cubic model adds both squared and cubed usage terms. Student-level models use individual student data; school-level models use school-aggregated average gains and usage. Analysis uses Sample 2 treatment group students and schools with complete baseline, endline, and usage data.

Figure 8: Predicted Dosage-Response Effects by Analysis Level



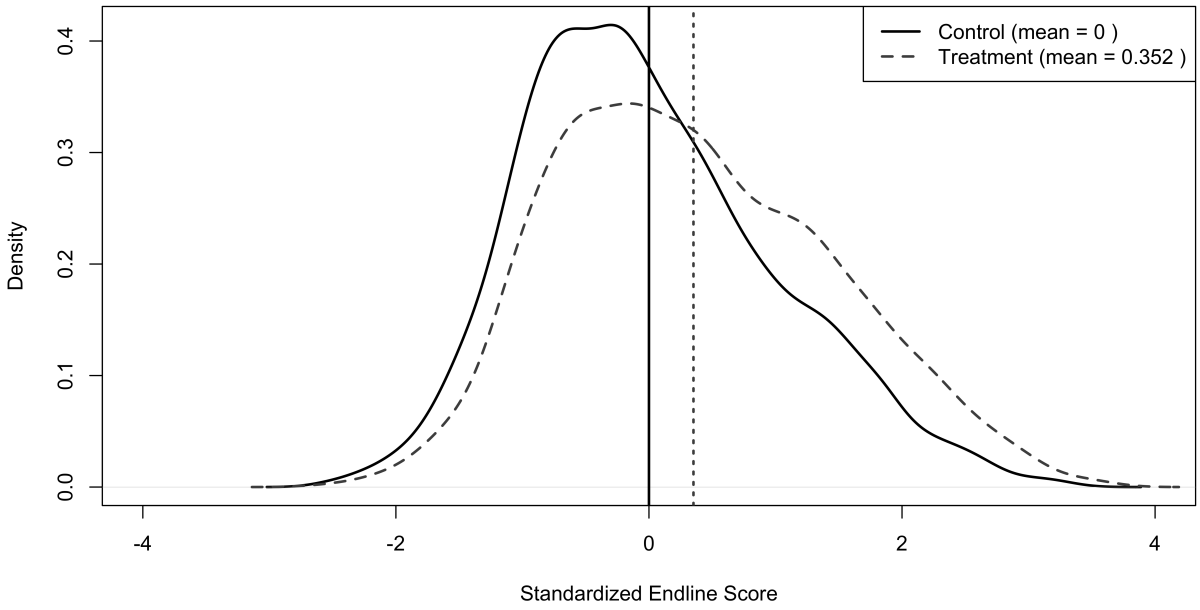
Notes: Predicted learning gains from linear dosage-response models at student and school analysis levels. Both models regress learning gains on weekly usage minutes, forced through the origin, using treatment group data only. School-level model uses school-aggregated average gains and usage; student-level model uses individual student data. Solid lines represent predictions within empirical support range (±2.5 standard deviations of observed usage distribution at each level); dotted lines show extrapolated predictions beyond empirical support. Vertical line indicates average treatment group usage (47.4 minutes per week). Student empirical support extends to 113.2 minutes/week; school empirical support extends to 81.3 minutes/week. Analysis uses Sample 2 treatment group data with complete baseline, endline, and usage information.

Table 8: Treatment Effects on Student Test Scores

Sample	Control Mean [SD]	ITT (SE)	TOT (SE)
Endline Only	0.000	0.326***	0.400***
	[0.9998]	(0.111)	(0.131)
Baseline + Endline	0.000	0.445***	0.467***
	[0.9997]	(0.122)	(0.126)
Baseline + Endline with Baseline Controls	0.000	0.440***	0.462***
	[0.9997]	(0.100)	(0.102)

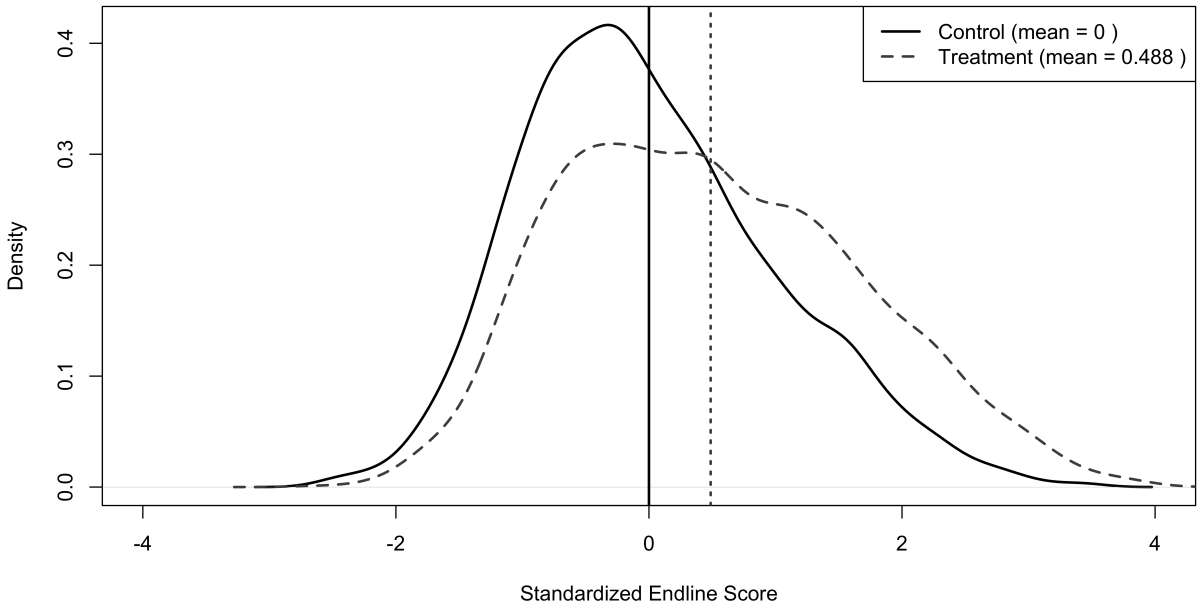
Notes: Treatment effects on standardized endline test scores. ITT shows intention-to-treat effects comparing all students in treatment vs control schools. TOT shows treatment-on-treated effects using instrumental variables estimation where treatment assignment instruments for actual school-level treatment receipt. Each row shows results from separate regressions with school-clustered standard errors in parentheses. Control group means and standard deviations in square brackets reflect standardized endline scores. Treatment receipt defined as schools assigned to treatment that did not discontinue participation due to implementation challenges. Compliance rates for TOT estimation: 81.6% (Endline Only), 95.2% (Baseline + Endline samples). First-stage F-statistics for TOT estimation: 2623 (Endline Only), 7706.1 (Baseline + Endline), 7254 (Baseline + Endline with Baseline Controls), all exceeding conventional thresholds indicating strong instruments. All regressions include district group fixed effects and school-level clustering. District group fixed effects control for geographic clustering in treatment assignment. Baseline controls include standardized baseline test scores for the Baseline + Endline with Baseline Controls specification only. Endline Only Sample: 7,888 students from 80 schools. Baseline + Endline Sample: 5,535 students from 74 schools. Significance levels: *** $p \leq 0.01$, ** $p \leq 0.05$, * $p \leq 0.10$.

Figure 9: Endline Test Scores - Endline Only Sample



Notes: Kernel density plots comparing standardized endline test score distributions between treatment and control groups. Solid line shows control group distribution, dashed line shows treatment group distribution. Vertical lines indicate group means (control mean = 0, treatment mean = 0.352). Endline Only Sample includes 7,888 students with complete endline data.

Figure 10: Endline Test Scores - Baseline + Endline Sample



Notes: Kernel density plots comparing standardized endline test score distributions between treatment and control groups. Solid line shows control group distribution, dashed line shows treatment group distribution. Vertical lines indicate group means (control mean = 0, treatment mean = 0.488). Baseline + Endline Sample includes 5,535 students with complete endline data.

Table 9: Treatment Effects on Endline Performance by Question Grade Level

Question Grade Level	Control Mean [SD]	T-C Diff. (SE)
Panel A: Endline Only Sample		
At Grade Level	0.000	0.092***
	[0.303]	(0.029)
1 Grade Below	0.000	0.138***
	[0.379]	(0.042)
2 Grades Below	0.000	0.091***
	[0.284]	(0.025)
3 Grades Below	0.000	0.086***
	[0.276]	(0.023)
Sum (Total)	0.000	0.406***
	[1.000]	(0.116)
F-test: Equality of effects	F = 1.93, p = 0.122	
Panel B: Baseline + Endline Sample		
At Grade Level	0.000	0.108***
	[0.305]	(0.030)
1 Grade Below	0.000	0.149***
	[0.378]	(0.044)
2 Grades Below	0.000	0.097***
	[0.286]	(0.026)
3 Grades Below	0.000	0.089***
	[0.275]	(0.023)
Sum (Total)	0.000	0.443***
	[1.000]	(0.121)
F-test: Equality of effects	F = 1.68, p = 0.168	
Panel C: Baseline + Endline Sample with Baseline Controls		
At Grade Level	0.000	0.107***
	[0.305]	(0.024)
1 Grade Below	0.000	0.147***
	[0.378]	(0.037)
2 Grades Below	0.000	0.096***
	[0.286]	(0.020)
3 Grades Below	0.000	0.088***
	[0.275]	(0.019)
Sum (Total)	0.000	0.438***
	[1.000]	(0.098)
F-test: Equality of effects	F = 1.68, p = 0.168	

Notes: Treatment effects on student performance across question groups targeting different grade levels. Scores are standardized using control group means and the total score standard deviation, ensuring level effects sum exactly to the total effect. T-C Diff. shows the treatment-control difference in standard deviations with school-clustered standard errors in parentheses. Control means and standard deviations (in brackets) are from the control group. Question grade level indicates the relative difficulty: At Grade Level tests students on their current grade questions, while 1, 2, and 3 Grades Below test on progressively easier material. Each level pools students across grades 6, 7, and 8 based on relative question difficulty. The Sum (Total) row shows the treatment effect on overall standardized test scores and equals the sum of the four level effects by construction. F-test examines whether treatment effects differ across the four grade levels; non-significant tests indicate uniform effects. All regressions include district group fixed effects and school-level clustering. Baseline controls in Panel C include standardized baseline test scores. Panel A: 7,281 students. Panels B & C: 5,527 students . Sum (Total) effects closely match main results, with some divergence due to smaller sample sizes. Significance levels: *** $p \leq 0.01$, ** $p \leq 0.05$, * $p \leq 0.10$.

Table 10: Heterogeneous Intent-to-Treat Effects on Endline Test Scores

	Endline Only	**Baseline + Endline**	
	(1)	(2)	(3)
Panel A: Gender			
Treatment Effect	0.279**	0.459***	0.420***
	(0.134)	(0.121)	(0.094)
Treatment × All-Girls	0.190	0.005	0.105
	(0.219)	(0.219)	(0.189)
Panel B: HDI			
Treatment Effect	0.505***	0.603***	0.567***
	(0.150)	(0.185)	(0.158)
Treatment × Above Median HDI	-0.351	-0.294	-0.241
	(0.236)	(0.251)	(0.208)
Panel C: Curriculum			
Treatment Effect	0.321**	0.373***	0.268***
	(0.126)	(0.134)	(0.092)
Treatment × CBSE	-0.015	0.123	0.318*
	(0.198)	(0.222)	(0.169)
Panel D: Grade			
Treatment Effect	0.257**	0.406***	0.398***
	(0.116)	(0.131)	(0.096)
Treatment × Grade 7	0.042	-0.030	0.002
	(0.133)	(0.153)	(0.147)
Treatment × Grade 8	0.153*	0.142	0.117
	(0.089)	(0.108)	(0.109)
Panel E: Baseline Performance			
Treatment Effect	—	0.504***	0.521***
		(0.147)	(0.150)
Treatment × Low Tercile	—	-0.119	-0.139
		(0.165)	(0.170)
Treatment × Medium Tercile	—	-0.086	-0.107
		(0.121)	(0.128)
Control Mean	0.000	0.000	0.000
[SD]	[0.9998]	[0.9997]	[0.9997]
Regression specifications			
Baseline Controls			✓
District Group Fixed Effects	✓	✓	✓
School Level Clustered SEs	✓	✓	✓

Notes: Heterogeneous intention-to-treat effects on standardized endline test scores examining differential impacts across student and school characteristics. Each panel shows results from separate regressions including interaction terms between treatment assignment and the specified characteristic. Treatment Effect shows the main treatment effect for the reference group with school-clustered standard errors in parentheses. Interaction terms show the additional treatment effect for the specified subgroup relative to the reference group. Total treatment effect for any subgroup equals Treatment Effect plus relevant interaction coefficient. Reference groups are: All-Boys schools (Panel A), Below Median HDI districts (Panel B), UP Board curriculum schools (Panel C), Grade 6 students (Panel D), and High Tercile baseline performers (Panel E). HDI refers to Human Development Index at the district level with above/below median split. Students classified into terciles based on baseline test scores: Low Tercile (0-33rd percentile), Medium Tercile (34th-67th percentile), High Tercile (68th-100th percentile). Panel E not available for Endline Only Sample (column (1)) due to incomplete baseline data. All regressions include district group fixed effects and school-level clustering. District group fixed effects control for geographic clustering in treatment assignment. Baseline controls include standardized baseline test scores and apply only to column (3). Control group means and standard deviations in square brackets reflect standardized endline scores. Endline Only Sample: 7,888 students from 80 schools. Baseline + Endline Sample: 5,535 students from 74 schools. Significance levels: *** $p \leq 0.01$, ** $p \leq 0.05$, * $p \leq 0.10$.

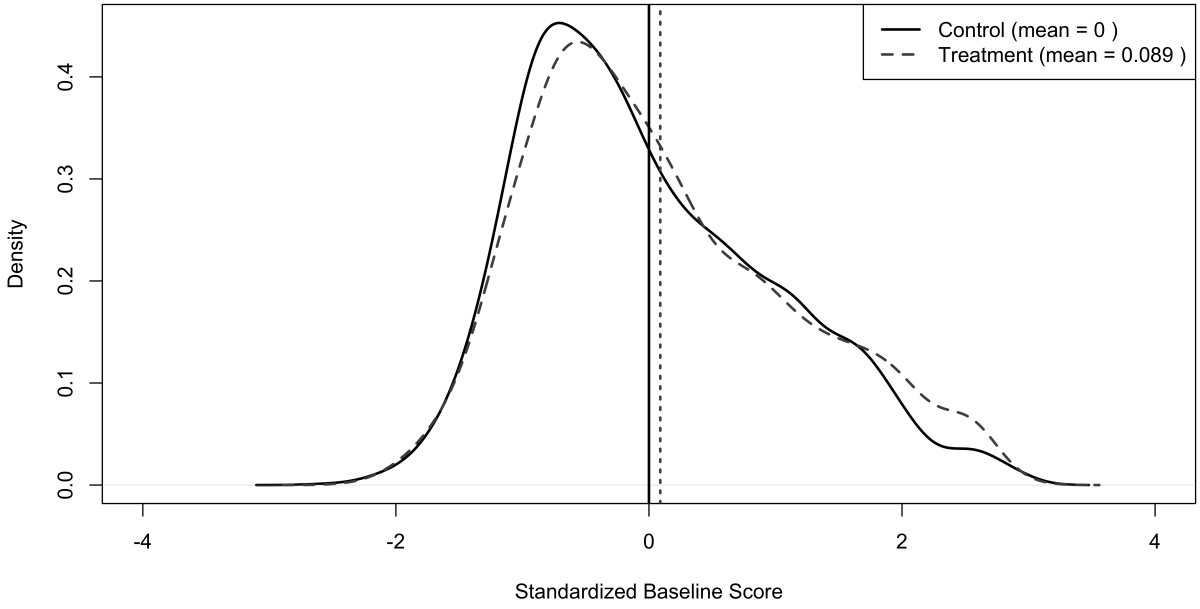
APPENDIX

Table A1: Full Sample School-Level Balance Tests

	Full Sample (N=83)	
Variable	**Control Mean [SD]**	**T-C Diff. (SE)**
District HDI Score	0.598	0.005
	[0.052]	(0.009)
All-Girls School	0.309	-0.010
		(0.105)
CBSE Curriculum Adoption	0.436	0.069
		(0.116)
School Size	149.491	12.373
	[46.459]	(9.454)
Missing Baseline Survey	0.319	0.027
	[0.151]	(0.033)
Missing Endline Survey	0.496	-0.079*
	[0.205]	(0.047)
Regression specifications		
District Group Fixed Effects		✓

Notes: School-level balance tests for the full randomized sample of 83 schools (28 treatment, 55 control), including three control schools that served as board examination centres and could not administer endline tests. Each cell shows control group means with standard deviations in square brackets for continuous variables and treatment-control differences with standard errors in parentheses. District group fixed effects control for geographic clustering in treatment assignment. District HDI Score is Human Development Index (0-1 scale). All-Girls School is proportion of all-girls schools versus all-boys schools (binary variable). CBSE Curriculum Adoption is proportion using CBSE (Central Board of Secondary Education) versus UP (Uttar Pradesh) Board curriculum (binary variable). School size is number of students per school. Missing Baseline Survey is the fraction of registered students without baseline test scores. Missing Endline Survey is the fraction of registered students without endline test scores. The single marginally significant difference in endline missingness reflects the three control schools without endline data; these schools were board centres that conducted internal examinations earlier in the year and could not be tested during the scheduled endline assessment. Removal of these three schools does not alter the pattern of successful randomization on pre-treatment characteristics (see Table 2). Significance levels: *** $p \leq 0.01$, ** $p \leq 0.05$, * $p \leq 0.10$.

Figure A1: Baseline Test Scores - Endline Only Sample



Notes: Kernel density plots comparing standardized baseline test score distributions between treatment and control groups. Solid line shows control group distribution, dashed line shows treatment group distribution. Vertical lines indicate group means (control mean = 0, treatment mean = 0.089). Endline Only Sample includes 5,980 students with available baseline data.