Must Remediation Be Deep? Experimental and Administrative Evidence from India*

Emily Cupito Alex Eble Guthrie Gray-Lobe Saloni Gupta Michael Kremer
Sabareesh Ramachandran Wendy Wong

Preliminary: please do not cite or circulate without permission.

Latest draft

Last updated: September 26, 2026

Abstract

In many developing countries, school curricula outpace children's learning, and children in higher grades often learn very little. Targeting teaching to children's actual levels raises learning, but nearly all such programs are NGO-delivered and require remediation several grades back, something unpalatable to schools and parents and thus hard to maintain at scale. Using multiple experiments and administrative data from an at-scale policy rollout in Andhra Pradesh, we show government can deliver large learning gains through targeted instruction with lighter remediation, and these results replicate. Administrative data from three contexts shows learning hinges on fidelity.

JEL Codes: I21, I25, O15, C93. *Keywords:* Personalized Adaptive Learning, EdTech, Scale, Government, India, Math Education.

*Cupito: Development Innovation Lab, University of Chicago. Eble: Teachers College, Columbia University; eble@tc.columbia.edu. Gray-Lobe: Development Innovation Lab, University of Chicago; graylobe@uchicago.edu. Gupta: Brown University; saloni_gupta@brown.edu. Kremer: Development Innovation Lab, University of Chicago; kremermr@uchicago.edu. Ramachandran: Indian School of Business, Hyderabad; sabareesh_ramachandran@isb.edu. Wong: CRI Foundation. We acknowledge funding from the *Learning for All Initiative* at J-PAL and the What Works Hub for Global Education. We received helpful comments from Andy de Barros, Ale Ganimian, Morgan Hardy, Karthik Muralidharan, and many seminar participants. We are grateful to our implementation partners at the Government of Andhra Pradesh, the ConveGenius team, and the Central Square Foundation for their collaboration, support, and partnership throughout the study. We thank Aishwarya Rebelly, Diego Mendoza M., Michelle Cherian, Nandita Gupta, Niharika Bhagavatula, Prarthna Iyer, Sharad Hotha, and Yiyuan He for excellent research assistance. This trial is pre-registered at the AEA RCT Registry (AEARCTR-0014271; https://doi.org/10.1257/rct.14271-2.0). Ethics approval from the University of Chicago IRB (protocol IRB23-0807, approved June 26, 2023).

I Introduction

A consensus among international organizations, philanthropies, and researchers has emerged that when students lack foundational skills, instruction must remediate those skills to improve outcomes. This view reflects a large body of rigorous research showing that adjusting instruction to suit an individual student's level can produce large learning gains. Experimental support for this view dates to some of the earliest randomized evaluations in development economics (e.g. Banerjee et al., 2007; **?**; Muralidharan et al., 2019) and was highlighted in the materials accompanying the 2019 Nobel Prize in economics, making this research a prominent example of randomized evaluations shaping policy.

However, remediation is not always viable and has potential downsides, which are often noted by parents, teachers, and policymakers. Education systems reward schools, teachers, and students themselves according to students' mastery of grade-level skills, rather than their ability to acquire the skills the students were supposed to have picked up in earlier grades. This is particularly true in the later grades of middle school and high school, when teachers spend much of their time preparing students for school-leaving exams that determine entrance into higher schooling. This tension has led to resistance towards targeted learning programs and, in some cases, even efforts by teachers and parents to undermine them (Banerjee et al., 2016, 2017). Furthermore, providing differentiated instruction tailored to each student's individual level is typically infeasible in LMIC classrooms. Many approaches to leveling instruction do so at the group level, potentially having distributional consequences for top-performing students. In part due to these constraints, the consensus around the importance of foundational skills has not yet materialized into widespread adoption of remedial education.

Can students learn grade level without revisiting foundational material on which they are struggling? If the answer to this question is negative, then much of the time and resources spent on education globally is wasted.

This study uses multiple experiments and administrative data to evaluate a digital math tutor implemented at scale through a government program in Andhra Pradesh, India. Developed by ConveGenius, an Indian education technology firm, in collaboration with the state government, the tablet-based program was delivered by government teachers in government lower secondary schools during normal school hours, replacing two of the eight weekly math classes. The software assesses students' skills and adapts instruction to their learning levels, but deliberately limits remediation to make the program more appealing to teachers, schools, and

families. It assigns 66.5 percent of content at students' enrolled grade level and remediates no more than two grades below that level. By comparison, over 90 percent of the content in the personalized adaptive learning program evaluated by Muralidharan et al. (2019) was below students' enrolled grade level.

This design may seem problematic if foundational learning gaps must be remediated for students to learn. According to the ASER assessment, only 40 percent of eighth-grade students could solve simple division problems, a skill they should have mastered five grades earlier. Concerns that the software provided insufficient remediation to address these gaps motivated an additional experiment testing the effects of allowing deeper remediation. We estimate that introducing the standard, less remedial version of the software raises test scores by 0.39 standard deviations over two years, a gain equivalent to almost two additional years of schooling. Gains are broad-based, indicating that educational resources can generate substantial learning gains without addressing the full extent of students' foundational learning gaps.

Limiting remediation is not without costs. First, the experiment varying remediation depth—the number of grades below a student's enrolled grade that the software can draw content from—indicates that allowing deeper remediation can raise test scores by up to 0.06 standard deviations. This suggests scope to improve the program through additional remediation, although the estimated gains are modest relative to the impact of the standard version. Second, the program's intention-to-treat effects are larger at the top of the achievement distribution. These differences may reflect heterogeneity in usage, given substantial within-classroom variation in time spent using the tablets. They are also consistent with the possibility that higher-quality instructional resources offering limited remediation disproportionately benefit higher-performing students, increasing learning inequality even as they generate broad-based gains.

We further examine the relationship between the amount of usage of the program and learning gains. We find that using the software for longer leads to better learning outcomes. We exploit variation in the student-to-tablet ratio across schools to generate an instrumental variables (IV) estimate of the per-hour learning gain from using the program. We find that students gain $+0.011$ SD per hour of use ($p < 0.001$). This estimate hews closely to the observational correlation of $+0.009$ SD per hour estimated in data from the wider government rollout. We show that this causal relationship is robust to other measures of quantity and quality of usage.

We examine the cost-effectiveness of the intervention through the lens of learning-adjusted years of schooling (Angrist et al., 2020). We calculate that it cost $38.66 per student over two years. This means that the intervention yielded 1.27 learning-adjusted years of schooling (LAYS) per $100. This falls in the lower

half of the range of estimates from a recent meta-analysis of comparable adaptive-learning interventions (0.24–2.66 LAYS per $100). We emphasize that this estimate pertains to the configuration of the intervention evalued in the trial: 35.3 hours of use per student over two years, delivered under active state monitoring. Later in the paper, we explain how the cost per unit of learning depends on the quality of the dose, which, in turn, depends on fidelity of implementation.

We further examine the logistics that enabled the intervention and discusses the external validity of our results past the experimental window and beyond the experimental sample. We first describe the quality of implementation and the learning trajectories of students in three separate groups of schools: in the experimental sample; in the broader set of government schools involved in the original rollout; and in a separate set of residential schools, which the government began implementing towards the end of the second year. In the project schools in this third year, multiple dimensions of the state's delivery infrastructure contracted or completely stopped.

The intervention was primarily led by the teachers and the state officials. All officials had access to a dashboard that showed the total duration that PAL was used for in each week by the average student in each school. This allowed the officials to identify schools which were underperforming on usage or performance metrics, as well as to solve any technical or logistical issues they may have been facing as well as nudge them to take the students to the lab and get the students to use the software.

Using observational data, we show that this monitoring appears to be material to the success of the program. We find that in the periods of time when the schools were being monitored the usage was higher and falls subsequently. In this third year, Monitoring by the central authority fell to roughly a quarter of what it was during the study; the field-support workforce was entirely absent for a substantial period of time because their contract was allowed to lapse; and the replacement of tablets, which degrade rapidly, stalled. As a result of these failures, usage in the trial schools collapsed, and the treatment-control gap in test scores narrowed from $+0.31$ SD to $+0.1$ SD.

In the 444 government schools involved in the broader program rollout but not in the study, the lab used the tablets to teach three subjects, not just math. These schools also had older tablets on which the software was much slower. We find the usage per student in these schools was about half that in the trial schools during the first two years but collapsed to less than an hour in the entire third year. Most tablets in these schools were broken by then. We estimate that students in these schools gain 0.025sd of learning per hour of usage. In the targeted residential schools, by contrast, the program was introduced in the third year with a new set

of tablets. The delivery, monitoring, and usage was sustained throughout the third year, students averaged 13.6 total cumulative hours, and their observational learning trajectories suggest learning gains in line with our experimental results. Pairing our usage data to test scores, we estimate that students in these schools gain $+0.015$ SD of learning per hour of usage. Indeed, across all three settings, observational estimates of the correlation between hours of usage and learning gains hew closely to the trial's causal IV estimate; what varies, rather, is the dose each delivery configuration produces.

Finally, we report an evaluation that randomizes the frequency with which monitors focused on technology troubleshooting and proper usage visit schools. The government engaged one external Field Management Staff (FMS) for every twelve schools to provide on-ground technical support. While they played a critical role in getting the program set up and in helping teachers initially and their presence was likely essential, eventually their role in sustaining usage was limited. This experiment halved the frequency at which the external staff visit schools. It finds no impact of monitoring intensity on a school-level measure of weekly software usage per student.

This paper contributes to two active bodies of work in applied microeconomics: that studying how the state delivers services in low- and middle-income countries, that studying the delivery of education services via in-school adaptive learning programs. There are two sets of questions in the state delivery literature on which we make progress. The first relates to technology; prior work on technology-enabled state delivery documents how technological infrastructure can extend the state's administrative reach, via a diverse range of technological solutions and monitoring programs (Muralidharan et al., 2016, 2021; Okunogbe and Pouliquen, 2022; Beg, 2022). Our paper extends this to study a comprehensive technological platform that captures the delivery and uptake of classroom lessons in government schools. The platform was built in part by the government of Andhra Pradesh, and we study its impacts, challenges, and adaptation as it was taken to scale.

We advance understanding of the economics of adaptive learning. This research finds that implementation structure—typically a hired lab-in-charge per school—is often what determines whether these platforms can produce learning at scale. Muralidharan and Singh (2025) adapt the Mindspark platform for in-school delivery in Rajasthan public schools through a hired lab-in-charge (LIC) per school(Muralidharan et al., 2019). They find a 0.22 SD math gain over 18 months. An evaluation of a period in which LIC support was reduced led to a 50 percent decrease in usage of the platform. Oreopoulos et al. (2026) add a part-time LIC to Uttar Pradesh boarding schools that already had Khan Academy access and raise math achievement by 0.44 SD in 31 weeks, isolating implementation structure as the binding margin. Relatedly, in a study of

contract teachers,Bold et al. (2016) show that an intervention adding contract teachers to schools raised test scores in Kenya when implemented by an NGO but not when implemented through the government's regular bureaucracy. The same literature increasingly recognizes that depth of targeting is one component of these programs, not their singular engine. de Barros and Ganimian (2024) hold the Mindspark platform constant and vary only whether students receive level-targeted or grade-level content: students gain only 0.05 SD on average, with the effect concentrated among bottom-quartile students (+0.22 SD). Ardington et al. (2026) synthesize forty-six estimates across teaching-at-the-right-level evaluations and attribute cross-study variation to who delivers, how much time is allocated, and the intensity of monitoring—rather than to the depth of level-matching. The analysis in Kaffenberger et al. (2026) decomposes the TaRL approach into nine core components, of which aligning instruction to current learning levels is one.

We extend this literature on three dimensions: who delivers, how deep the remediation goes, and how the program performs as it scaled through public school systems. On delivery, the implementation chain—scheduling, monitoring, and content integration—runs through the state's regular school hierarchy, supported by a far thinner contracted technical workforce (roughly one field staffer per 12 schools) than the per-school LIC configurations of the comparison studies. On content depth, our setting contrasts sharply with the Rajasthan Mindspark configuration, where over 97 percent of math items presented to grade 8 students were below grade level (Muralidharan and Singh, 2025); in our trial, roughly 82 percent of the analogous items sit at the enrolled grade—and our within-school experiment directly tests, and largely rejects, the hypothesis that deeper remediation would add substantial learning. On performance at scale, we show that the per-hour return replicates wherever delivery is sustained, while the learning produced at scale depends on the delivery infrastructure that turns program access into engaged hours per student.

The remainder of the paper proceeds as follows. Section II describes the Andhra Pradesh school system, the program's rollout, and the intervention itself. Section III describes the experimental design, data, and empirical specifications. Section IV reports the trial's effects on learning, their heterogeneity, the per-hour IV return, and cost-effectiveness. Section V reports the within-school experiment on remediation depth. Section VI follows the program through its third year across the trial schools, the program rollout, and the targeted residential schools. Section VII concludes.

II Background, setting, and intervention

Our study takes place in the Indian state of Andhra Pradesh. The state has roughly 54 million people, making it India's tenth most populous state. Its educational performance places it in the middle of the distribution of Indian states: below high performers such as Kerala, but ahead of the lowest performers, such as Uttar Pradesh and Bihar.

We work in government schools in the state. Private provision of education in Andhra Pradesh is widespread across the state, and the offerings range from low-cost private schools in rural areas and lower-income urban areas to higher-cost, elite schools, most often found in urban areas (Singh, 2015). As a result, the students who remain in public schools are often negatively selected on income and family educational background, among other traits (Muralidharan and Sundararaman, 2015).

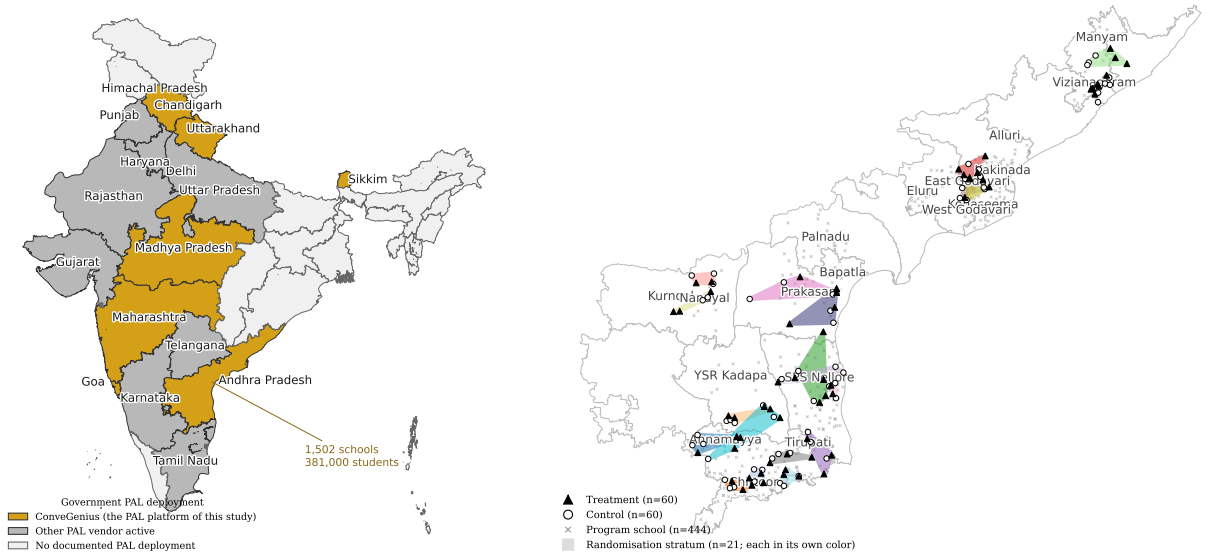
II.A. Program rollout and study sample

In the past decade, personalized adaptive learning (PAL) has gained prominence as a potential solution to the "learning crisis" in India's schools, in which learning levels fall below expected skill mastery early in most children's academic careers and this wedge grows as children progress through the system. We study one such PAL program, developed in partnership between the Government of Andhra Pradesh and the education technology firm ConveGenius. ConveGenius has a large national presence in India: in the left panel of Figure 1, we indicate in gold the states using a version of the ConveGenius PAL product. The figure shows that PAL use is even more common across the country; in gray, we indicate the states using other PAL programs in their school systems.

The program we study is tablet-based, math-focused, and targeted at middle school students. It launched in 2019 and suspended during the COVID-19 pandemic; when it restarted in 2022, it operated in 464 government middle schools across 17 of the state's 26 districts.[1]

As part of the program's expansion, we partnered with the state to conduct a randomized evaluation in a prospectively selected set of 120 schools, stratified across 8 districts and 21 mandal-cluster groups. Panel B of Figure 1 plots these 120 study schools alongside the broader, contemporaneous 464-school rollout; the 21 randomization strata appear as colored polygons in the figure. The intervention ran in the treatment schools for two academic years, AY 2023–24 and AY 2024–25. In AY 2023-24, the roll out began in the month of

[1]To date, the program has been implemented in roughly 1,200 schools statewide.



(A) PAL DEPLOYMENT ACROSS INDIAN STATES (B) RCT DESIGN WITHIN ANDHRA PRADESH

FIGURE 1. RCT SCHOOLS WITHIN THE ANDHRA PRADESH PAL ROLLOUT

Notes: Panel (a): Indian states color-coded by PAL deployment status. Gold = states where ConveGenius operates (the PAL platform of this study). Dark gray = states with another documented PAL vendor. Light gray = no documented government PAL deployment. *Panel (b):* Andhra Pradesh with all district boundaries. The 120 RCT schools (60 treatment = filled triangles; 60 control = open circles) sit within the 444-school government PAL rollout (gray × markers) covering 17 of AP's 26 districts. The 21 randomization strata, formed by grouping adjacent mandals within each of the 8 RCT districts, are shown as translucent polygons in distinct colors. *Source (Panel a):* Authors' compilation, April 2026, from PAL Works (palworks.in), ConveGenius, Educational Initiatives (Mindspark/Ei Shiksha), and ∼30 other public sources.

September and continued until the summer break in April. In AY 2024-25, the program resumed in July and continued until the endline data collection in February.

II.B. The PAL software being evaluated

The program's content is organized into chapters mirroring the state's mathematics textbooks. As in other PAL-style programs, students work individually on tablets during scheduled class time. Lessons comprise instructional videos, practice problems, and end-of-lesson assessments. The software operates primarily offline, with teachers syncing usage data to the server approximately once per week.

The application's key design feature is a more restricted remediation strategy than most other widely used PAL programs in the country. For example, the Mindspark software studied in prior work and currently used in several Indian states delivers remedial content from as many as five grades below the student's enrolled grade, and this remedial content typically constitutes over 90 percent of what students encounter. The ConveGenius software limits remediation on both margins: content reaches at most two grade levels below the student's enrolled grade, and remedial content makes up less than 30 percent of what students see.

The PAL program provides lessons in bundles called chapters, and each chapter progresses through three phases. First, the program diagnoses the student's learning level vis-a-vis the grade-level content in the chapter. Second, it delivers any remedial content indicated by the diagnosis. Note that this per-chapter routing is student-specific: different diagnostic outcomes deliver different bundles of activities to different students. Third, it delivers grade-level lessons. We discuss the process and logic behind it in greater depth in Appendix A.

Because students have autonomy over how they proceed through the software, not all students complete all of the content in a given lesson (Figure A1, Panel B). Across $N = 6{,}052$ treatment students, the mean fraction of prescribed learning objectives a student opens—which we term *compliance*—is 0.61 (median 0.65).[2] Compliance varies meaningfully within each baseline-ability quintile but is essentially flat across quintiles: students in the bottom quintile open 60 percent of their prescribed activities on average, students in the top quintile 62 percent. Time on the platform and compliance are correlated ($r = 0.58$) but distinct: hours measure raw exposure during PAL periods, while compliance measures the fraction of the prescribed sequence the student actually completes.

II.C. School-level delivery

Each treatment school dedicated a classroom to serve as a "lab," which housed the tablets and in which the intervention was administered. Each lab was equipped with 30 learning stations; each station comprised a tablet loaded with the PAL software, a charging point, and earphones. The number of stations did not vary with the size of the school or cohort.

Students in our study, as in much of Andhra Pradesh, were expected to receive eight 40-minute periods of math instruction per week from their schools. Treatment schools were instructed to convert two of these periods to PAL sessions, with the aim of reaching roughly 80 minutes of software use per week. During these sessions, the math teacher took the students from eligible grades to the lab and supervised their independent use of the tablets. Math teachers in treatment schools reported spending less time on remedial instruction and practice problems during the remaining six regular math periods, instead focusing on teaching key concepts in the classroom and directing students to use PAL for practice on each chapter.

[2] Compliance is defined only for students with at least one chapter started on the platform: a student who logs in but never opens a chapter has no prescription assigned and therefore no denominator. About 690 treatment students with no module-level data are dropped here but appear in the Table 2 hours sample (where they are coded as zero hours and zero activities). The remaining gap between $N = 6{,}052$ here and the full treatment sample is driven by missing baseline-ability data, which compliance-by-quintile statistics require.

In sections with more than 30 students, tablets had to be shared: teachers were asked to split the section into batches that alternated between PAL and regular instruction, though in practice we also observed students being paired on a single tablet or taking turns. Across the 60 treatment schools, 44.7 percent of the 300 sections exceeded the 30-tablet count or were "oversubscribed"; the median oversubscribed section held 37 students (a 1.23:1 student-to-tablet ratio), and 59.6 percent of treatment students were in oversubscribed sections (Appendix Figure 1).

II.D. State implementation infrastructure

An important feature of this program was the state infrastructure designed to monitor implementation and ensure the fidelity and quality of the program. At the top of the monitoring chain was the office of the State Project Director (SPD) with a designated State PAL Coordinator who exclusively handled the day-to-day operations of this program. Each district had a District Nodal Officer (DNO) for PAL who relayed state instructions and reprimands to the headmasters and schools. In the school, the headmaster and teachers were responsible for running the lab sessions.

The State PAL Coordinator and the DNOs had access to the weekly PAL usage per student in each school in near-real time.[3] A WhatsApp broadcast group from the State PAL Coordinator to all DNOs shared weekly school- and district-level usage rankings, issued nudges to increase usage, and coordinated operational pushes such as data syncing, orientation sessions, and tablet maintenance. The State Project Director and the PAL State Coordinator led district-by-district sessions with headmasters, math teachers, and DNOs in the program's first weeks. They issued in-person summons to low-performing DNOs. All 28 RCT treatment headmasters in our Year-3 survey report receiving these state communications periodically, and 71 percent report having seen the school-level usage report card.

A key difference between this PAL program and others is the ratio of local supervisory staff to PAL labs. Most successful PAL programs have a staff cadre responsible for on-the-ground management; in two recent PAL studies, one such staff member was allocated for each lab (Muralidharan and Singh, 2025; Oreopoulos et al., 2026). In our program, there was one such staff member (here called a Field Management Staff, or FMS) was assigned to a cluster of approximately 12 schools. These staff's main remit was to visit each school regularly and ensure smooth operations, including: fixing technical issues with tablets and software,

[3]The dashboard updates every time the data was synced from the tablets to the server in each school. The teachers or FMS were instructed to sync the tablets every week.

syncing data from the offline platform; monitoring usage and (when necessary) flagging low usage to the headmaster. Crucially, the FMS are not a tier in the government chain of command; rather, they comprise a parallel technical-support workforce that interfaces between the school and the software vendor. The purpose of the FMS is to take responsibility for ensuring that tablets work and that data reach the dashboard, allowing other school staff to complete their normally scheduled tasks.

III Experimental Design and Study Sample for the Cluster-RCT

Our cluster-randomized controlled trial was conducted in 120 government schools in Andhra Pradesh. The government identified eight districts, within which eligible schools were grouped into 21 geographic strata based on mandal clusters. To be eligible, a school had to meet four criteria: (i) it was a co-educational government school serving grades 6–10; (ii) it had fewer than 60 students per section in all grades, ensuring a maximum student-to-tablet ratio of 2:1; (iii) it had fewer than 20 total sections across grades 6–9, ensuring that all sections could access the lab at least twice a week; and (iv) its pupil-teacher ratio was not in the top decile of schools in the state.[4]

Our study sample comprises the 18,039 students enrolled in grades 6–9 across the 120 study schools in AY 2024–25.These students' exposure to the program varied by cohort: students in grades 7–9 at endline received two full years of the program, while grade 6 students (a new cohort entering in AY 2024–25) received one year. Table 1 presents summary statistics on school (Panel A) and student (Panel B) characteristics by treatment arm, along with tests of balance. Treatment and control schools are balanced on both school-level characteristics (15 traits spanning enrollment, staffing, and ICT infrastructure; joint $F = 1.14$, $p = 0.336$) and student-level characteristics (nine traits, including grade composition, gender, age, parental education, and socioeconomic status/SES;joint $F = 1.07$, $p = 0.387$).[5] We present a CONSORT-style diagram of the flow of participants through the study in Appendix Figure 2.

We enrolled participants using administrative data from government assessments administered in AY 2022–23, prior to the start of the intervention. Endline assessments took place in February and March 2025. Of the 18,039 enrolled students, 14,215 (78.8 percent) completed the endline assessment—6,742 in treatment

[4]practically, this means that we excluded all schools with a student-to-teacher ratio of greater than 70:1; in practice, this dropped less than 10% of schools.

[5]In the student-level pre-intervention government math score, treatment students score 2.6 points lower than control students. Because there is much greater missingness of this variable among control students than treatment students, we argue that this is not a reliable test for balance. We discuss this further in Section III.A and Appendix C).

schools and 7,473 in control schools. Attrition was 3.9 percentage points higher in treatment schools (23.7 versus 19.8 percent). The majority of students who did not take the endline assessment remained enrolled in school: 81 percent appeared in government assessment records for 2024–25, and 70 percent had non-zero math scores. We present formal attrition analysis and Lee bounds in Appendix D.

III.A. Data

Our data come from three main sources: independent learning assessments at endline (the primary outcome) conducted by a private survey firm; administrative data from the state government; and student usage and learning data collected through the software by ConveGenius.

We conducted math assessments at endline to capture both grade-level and earlier-grade learning. The instrument comprises 72 multiple-choice items spanning five broad math topics—algebra, geometry, measurement, statistics, and number systems—with questions drawn from the Andhra Pradesh state curriculum, spanning grade 2 through grade 9. Each student answered items from grade 2 up to their enrolled grade, yielding assessments between 50 and 72 items long depending on grade level. Using item response theory (IRT), we transform each student's performance into a single ability parameter, which we standardize to the control group mean and standard deviation. Following other work using this method to generate comparable scores across children and grades, we call this our IRT score (Muralidharan et al., 2019; Muralidharan and Singh, 2025). We describe the development and piloting of the instrument in Appendix E.

Baseline math ability, used in robustness checks and in heterogeneity analyses, comes from pre-intervention government assessments—specifically, the government's "Summative Assessment 1" exams. The source is cohort-specific: grade 7–9 endline cohorts use the AY 2022–23 exam (the year immediately before PAL began), while the grade 6 endline cohort, who entered the study schools in AY 2024–25, use the AY 2023–24 exam from their prior (grade-5) school. Overall coverage is 61 percent, but missingness is sharply asymmetric across arms.[6] For analyses that condition on baseline, we include a missing-baseline flag and interpret the resulting estimates with caution. Appendix F validates these government scores against our independent endline IRT. This analysis documents why we use them only for within-school rankings and not as a cross-school ability measure.

We administered the endline assessment on tablets using SurveyCTO, conducted by enumerators inde-

[6]This asymmetry arose because the government extracted baseline data for treatment and control schools at different times using different systems.

pendent of the schools and the intervention. This choice—having our endline assessment on tablets rather than on paper exams—served several goals. First, responses were mandatory, and each question had an explicit "don't know" option; this avoided the problem of missing values common on written exams. Second, randomized presentation of both topic order and question order across three test forms reduced opportunities for copying. Third, the platform recorded per-item response times, enabling us to detect guessing and rushing. Fourth, same-day data syncing allowed us to monitor school-level completion rates in real time and schedule efficient return visits to high-absence schools.

A natural concern with a tablet-administered endline is that treatment students' greater familiarity with tablets could inflate their measured performance. We conducted two data collection exercises to assess this risk. First, we ran a pre-pilot of the endline exam measuring response times and accuracy on the device across treatment and control groups Second, We conduct and report a sub-study in Appendix G that directly tests whether treatment affected students' ability to answer questions on a tablet relative to answering the same questions on paper. It finds no significant interaction between test modality and treatment status, ruling out tablet familiarity as a driver of the headline effect.

III.B. Empirical Specification

Our main estimating equation is an ordinary least squares regression implementing an intent-to-treat analysis:

$$Y_{isg} = \alpha + \beta\, T_s + \mathbf{X}_i'\gamma + \delta_g + \mu_m + \varepsilon_{isg} \tag{1}$$

On the left side, Y_{isg} is the endline math score for student i in school s and grade g: the continuous IRT score described above, standardized to the control group mean and standard deviation. On the right side, T_s is a treatment indicator; $\mathbf{X}_i$ is a vector of student-level controls comprising an SES index,[7] indicators for each level of father's and mother's education, age, gender, and a missing-value indicator for each control with missing data; δ_g are grade fixed effects; and μ_m are randomization-stratum (mandal-cluster) fixed effects. Standard errors are clustered at the school level.

Our main coefficient of interest is β, the group-average ITT effect of being offered the program. Baseline math ability is not included in $\mathbf{X}_i$ for the headline specification because its missingness is sharply imbalanced across arms; a missing-value indicator would itself be treatment-correlated and could absorb experimentally

[7] We take the first principal component of a PCA of indicators for the possession of nine household assets collected at endline: refrigerator, laptop, the father owning a mobile phone, the mother owning a mobile phone, air conditioner, washing machine, car, scooter, and bicycle.

TABLE 1—SAMPLE CHARACTERISTICS AND BALANCE

	AP (1)	Control (2)	Treatment (3)	Difference (4)	(SE) (5)	p-value (6)	N (7)
Panel A: School Characteristics [AP: $N = 5,298$ schools; Treatment: $N = 60$ schools; Control: $N = 60$ schools]							
Enrollment (grades 6–9)	217.1	160.0	155.1	-4.333	(17.848)	0.808	120
Total enrollment	309.1	234.6	223.2	-11.020	(22.825)	0.629	120
Urban	0.129	0.100	0.117	0.027	(0.055)	0.627	120
Has primary section	0.493	0.683	0.583	-0.111	(0.087)	0.203	120
Math teachers	2.1	2.0	1.9	-0.098	(0.153)	0.525	120
Physical ed. teachers	1.7	1.5	1.3	-0.205*	(0.109)	0.060	120
Total teachers	11.8	10.8	10.5	-0.286	(0.706)	0.685	120
Classrooms	10.8	9.4	8.9	-0.512	(0.629)	0.416	120
Functional computers	1.9	1.3	1.6	0.212	(0.610)	0.728	120
Projector	0.614	0.600	0.550	-0.051	(0.092)	0.584	120
Smart TV	0.681	0.767	0.650	-0.108	(0.076)	0.158	120
Digital board	0.286	0.267	0.283	0.020	(0.070)	0.773	120
Uses ICT tools	0.794	0.817	0.833	0.020	(0.073)	0.782	120
Digital content available	0.695	0.717	0.733	0.037	(0.079)	0.638	120
Generator/UPS	0.133	0.133	0.117	-0.017	(0.060)	0.780	120
Joint F-statistic		1.14		[$p = 0.336$]			120
Panel B: Student Characteristics [AP: $N = 1,150,232$ students; Treatment: $N = 6,742$ students; Control: $N = 7,473$ students]							
Grade 6	0.223	0.237	0.228	-0.011	(0.008)	0.168	14,215
Grade 7	0.241	0.235	0.237	0.006	(0.007)	0.414	14,215
Grade 8	0.255	0.248	0.262	0.013	(0.008)	0.109	14,215
Grade 9	0.281	0.280	0.272	-0.007	(0.009)	0.407	14,215
Female		0.492	0.479	-0.020	(0.019)	0.301	14,215
Age		12.30	12.35	0.039	(0.037)	0.293	14,124
Father education		2.14	2.07	-0.057	(0.059)	0.336	12,744
Mother education		2.10	2.07	-0.037	(0.045)	0.412	13,135
SES index (PCA)		-0.02	0.02	-0.030	(0.040)	0.454	14,215
Joint F-statistic		1.07		[$p = 0.387$]			12,267
Baseline math score (govt SA1)		35.41	32.40	-2.482*	(1.300)	0.056	10,161

Notes: Column (1) shows means for all 5,298 government schools with grades 6–10 in Andhra Pradesh (1,150,232 students in grades 6–9). School characteristics from government administrative records (AY 2023–24); ICT variables from UDISE+ 2024–25. Columns (2)–(3) show unconditional group means for the 120 randomised study schools. Column (4) and standard errors are from OLS of each variable on a treatment indicator and 21 randomization strata (mandal cluster) fixed effects. Panel A reports school-level characteristics (including enrollment in grades 6–9, the study grades) with HC1 robust standard errors ($N = 120$ schools). Panel B reports student-level characteristics for the grades 6–9 study sample, with standard errors clustered at the school level. Baseline math score (government SA1, 2022–23) is excluded from the joint F-test due to differential administrative data coverage (72%; see text). * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

induced variation. We report a robustness specification adding baseline in Appendix Table D1.

A separate parameter of interest is the per-hour learning gain from usage of the program. We estimate this by exploiting variation in per-student dosage across schools arising from tablet congestion. Because

all treatment schools received 30 tablets regardless of enrollment, students in larger sections face a higher student-to-tablet ratio: sections with more than 30 students must rotate access, so as section size grows beyond 30, each student gets less time on the program in a given week. We instrument for usage hours using the interaction of treatment status and section enrollment:

$$\text{First stage:} \quad H_{isg} = \pi_0 + \pi_1 (T_s \times S_{sg}) + \pi_2 T_s + \pi_3 S_{sg} + \mathbf{X}_i' \pi_4 + \delta_g + \mu_m + \nu_{isg} \tag{2}$$

$$\text{Second stage:} \quad Y_{isg} = \alpha + \lambda \hat{H}_{isg} + \phi T_s + \psi S_{sg} + \mathbf{X}_i' \gamma + \delta_g + \mu_m + \varepsilon_{isg} \tag{3}$$

Here H_{isg} is total PAL usage hours and S_{sg} is section enrollment (the number of students in the student's class section). The exclusion restriction requires that the difference in treatment effects between larger and smaller sections is driven entirely by the difference in PAL usage between them—equivalently, that if usage were equalized across section sizes, treatment effects would be the same in larger and smaller sections.

We conduct a series of prespecified heterogeneity analyses. These proceed in two manners. One estimates equation (1) separately by subgroup. The other includes interaction terms with treatment status.

IV Estimates of the Impact of PAL on Learning from a Cluster RCT

In this section, we present four sets of analyses of the data from the cluster RCT. The first describes the uptake of the program and how delivery varied across time. The second presents the overall impacts on learning. The third presents analysis of the relationship between the quantity of usage and learning outcomes using both descriptive and instrumental variables approaches. The fourth presents an analysis of the relationship between the quality of usage and learning. For purposes of length, we present our main heterogeneity analysis in an appendix

IV.A. Delivery and Uptake of the Intervention

Our first results describe the delivery and uptake of the intervention. Table 2 summarizes PAL usage across both academic years. Each week, students were scheduled to sit two of their eight regularly schedule math classes in the PAL lab. On the school calendar this amounted to 80 minutes of per-week instructional time for PAL, but because of between-class transitions and setup time, this translated into roughly 30 per-class minutes (and 60 per-week minutes) of actual PAL-delivered instruction. The program was not used in weeks

TABLE 2—PAL USAGE STATISTICS (TREATMENT GROUP)

	Mean	Median	SD	P25	P75	*N*
Panel A: Overall Usage						
Total usage (hours, 2 years)	34.2	30.2	23.3	17.4	47.3	6,742
Year 1 usage (hours)	14.7	12.6	15.5	0.0	23.8	6,742
Year 2 usage (hours)	19.5	17.8	12.5	10.7	26.2	6,742
Video time (hours)	18.9	16.5	14.2	9.6	25.2	6,742
Question time (hours)	15.3	13.5	10.5	7.3	21.3	6,742
Months active (Y2)	6.6	7.0	1.9	6.0	8.0	6,742
Chapters attempted	23.0	24	12.1	14	31	6,742
Chapters completed	15.4	14	11.2	6	23	6,742
Compliance (overall)	0.60	0.66	0.21	0.49	0.76	6,742
Compliance — remedial (BG)	0.74	0.78	0.20	0.66	0.86	6,742
Compliance — at-grade (AG)	0.56	0.62	0.22	0.44	0.73	6,742

Panel B: Usage by Grade

	Y1 hours	Y2 hours	Total	Months (Y2)	Compliance	*N*
Grade 6	—	20.4	20.4	6.0	0.53	1,540
Grade 7	18.9	20.5	39.4	6.8	0.58	1,599
Grade 8	19.7	20.0	39.7	6.9	0.63	1,769
Grade 9	18.4	17.5	35.9	6.8	0.65	1,834

Notes: Treatment group only ($N = 6{,}742$ students in 60 schools). Panel A reports individual-level usage aggregated across both academic years (AY 2023–24 and 2024–25). Video time is passive instruction; question time is active practice. Chapters are curriculum units within the PAL platform. *Compliance* is the student-level mean across chapters of unique learning objectives (LOs) completed divided by the LOs prescribed by PAL's diagnostic for the student's (chapter × cascade signature × prior-incomplete history) cell, capped at 1; the BG (below-grade) and AG (at-grade) rows apply the same ratio separately to the remedial and at-grade portions of each prescription. Treatment students with no module-level data are coded as zero across hours, chapters, and compliance to keep the Panel A sample at $N = 6{,}742$. Panel B shows usage by endline grade. Grade 6 students entered the study in Year 2 only. Usage data from ConveGenius platform logs, capped at the endline assessment start date (February 17, 2025). Each school was allocated one 80-minute PAL period per week; accounting for school closures and setup time, two-year students achieved 89% of the calendar-adjusted effective target (Appendix Table 1). Topic-level usage breakdown is reported in the appendix.

where there was no school because of holidays, nor during school weeks when were scheduled assessments or pre-exam revision periods. As a result, there were 22 weeks of scheduled PAL instruction in Year 1 and 21 in Year 2.

Against this calendar-adjusted target of 43 hours over two years, students in grades 7–9 averaged 38.2 total hours of instruction, or 89 percent of the target. The total hours of PAL instruction were higher in Year 2 (92 percent of the target) than Year 1 (86 percent), consistent with a period of ramp-up during the program's first year.[8]

We next present a few more descriptive analyses of who among our student population uses the PAL

[8]We provide a breakdown of this in Appendix Table 1. In Appendix Figure 3 we trace the monthly usage profile across both years alongside the content grade levels students worked on.

program, how they use it, and what they think of it. We show correlates of the intensity of PAL usage in Appendix Table 2. We find that usage is predicted by baseline math scores and age, though not by family SES. Our analysis of how students use the tablets loads on the types of content they spend PAL time, and is presented in Appendix Table B1. We find that they spend slightly more time watching videos (54 percent) than answering questions (46 percent). We present analysis from a survey of student perceptions of the program in Appendix Table B2. We find that eighty percent of surveyed treatment students report enjoying PAL "a lot," though 37 percent report language comprehension and video loading as implementation challenges. Appendix B characterizes the platform's instructional pipeline using module-level data, showing that student experience of PAL is largely uniform across the baseline-ability distribution.

IV.B. Impacts on Learning

In this section, we present intent-to-treat estimates of the project's impact on student learning from the cluster RCT. In Table 3, we report our main regression estimates. The ITT effect, pooled across grades, is a 0.39 SD gain in learning relative to the control group ($p < 0.001$). Disaggregating these estimates by grade, we find that the effects of the program range from 0.33 SD in grade 9 to 0.51 SD in grade 7.The null hypothesis that effects are equal across grades is marginally rejected ($p = 0.091$), with the pattern suggesting a declining gradient from younger to older cohorts. We investigate this gradient—and its explanation through tablet congestion—in later sections.

In Figure 2, we plot kernel density estimates of endline math scores by treatment status and grade. The treatment distribution shifts visibly rightward in all four grades, with the largest shift at grade 7 and the smallest at grade 9. The shift is not confined to one tail: treatment widens the distribution as well as shifting its mean, indicating that the program generates variance in gains across students.

These results are robust to a battery of alternative specifications and inference procedures. In Appendix Table D1, we show that the treatment effect estimate is stable within a range of 0.01 SD across three distinct control specifications. The significance of our ITT estimate is also robust to the use of randomization inference (RI $p < 0.001$; see Appendix Figure D1). We also present estimates of the sensitivity of our results to differential attrition via Lee (2009) bounds in Appendix Table D2. These are also large and positive (0.27–0.48 SD). For space reasons, we present our analysis and its results in Appendix C.

TABLE 3—INTENT-TO-TREAT ESTIMATES OF THE EFFECT OF PAL ON MATH LEARNING

	Pooled (1)	Grade 6 (2)	Grade 7 (3)	Grade 8 (4)	Grade 9 (5)
Treatment	0.390***	0.393***	0.509***	0.346***	0.327***
	(0.054)	(0.070)	(0.069)	(0.068)	(0.062)
Controls	Yes	Yes	Yes	Yes	Yes
Strata FE	Yes	Yes	Yes	Yes	Yes
Grade FE	Yes				
Control mean	0.000	-0.338	-0.157	0.116	0.316
Observations	14,215	3,313	3,354	3,623	3,925
R^2	0.137	0.153	0.125	0.097	0.086
p-value (equality across grades)			0.091		

Notes: Dependent variable: IRT math ability parameter standardized to control group mean and SD. Each column reports OLS estimates of the treatment effect with 21 randomization strata (mandal cluster) fixed effects. Column (1) includes grade fixed effects (grade 6 = base). Controls: SES index, father education, mother education, age, gender, and missing flags for each imputed variable. Standard errors clustered at the school level (120 clusters) in parentheses. The p-value tests the null hypothesis that treatment effects are equal across grades (from a pooled interaction model). * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Effects on Government Test Scores

We can also examine whether treatment effects appear on government-administered math assessments. Some of these are set and graded by individual schools, and others are graded externally. We present results for each in turn.

In Appendix Table 3, we present estimates for tests that are graded by teachers from within the school in which the test was administered. For these, effects are positive and significant for several grade-year combinations: grade 6 in Year 2 (0.34 and 0.27 SD on the mid-year and end-of-year summative assessments), grade 8 in Year 1 mid-year (0.24 SD), and grade 8 in Year 2 mid-year (0.36 SD). Other cells are positive but not statistically significant. Effects tend to be stronger on the mid-year than the end-of-year assessments, potentially reflecting the alignment of program content with mid-year topics.

In Appendix Table F1, we present results using scores on from the grade 10 board exam, the one government outcome graded externally to the schools students attend. We report the estimate because the external grading makes it the cleanest cross-school government measure, while its specification sensitivity cautions against treating the negative direction as a settled finding.[9] Our point estimate for this exam is

[9]Because of the timeline of our data collection, we have this only for the grade 9 cohort at the time they have one year of PAL exposure.

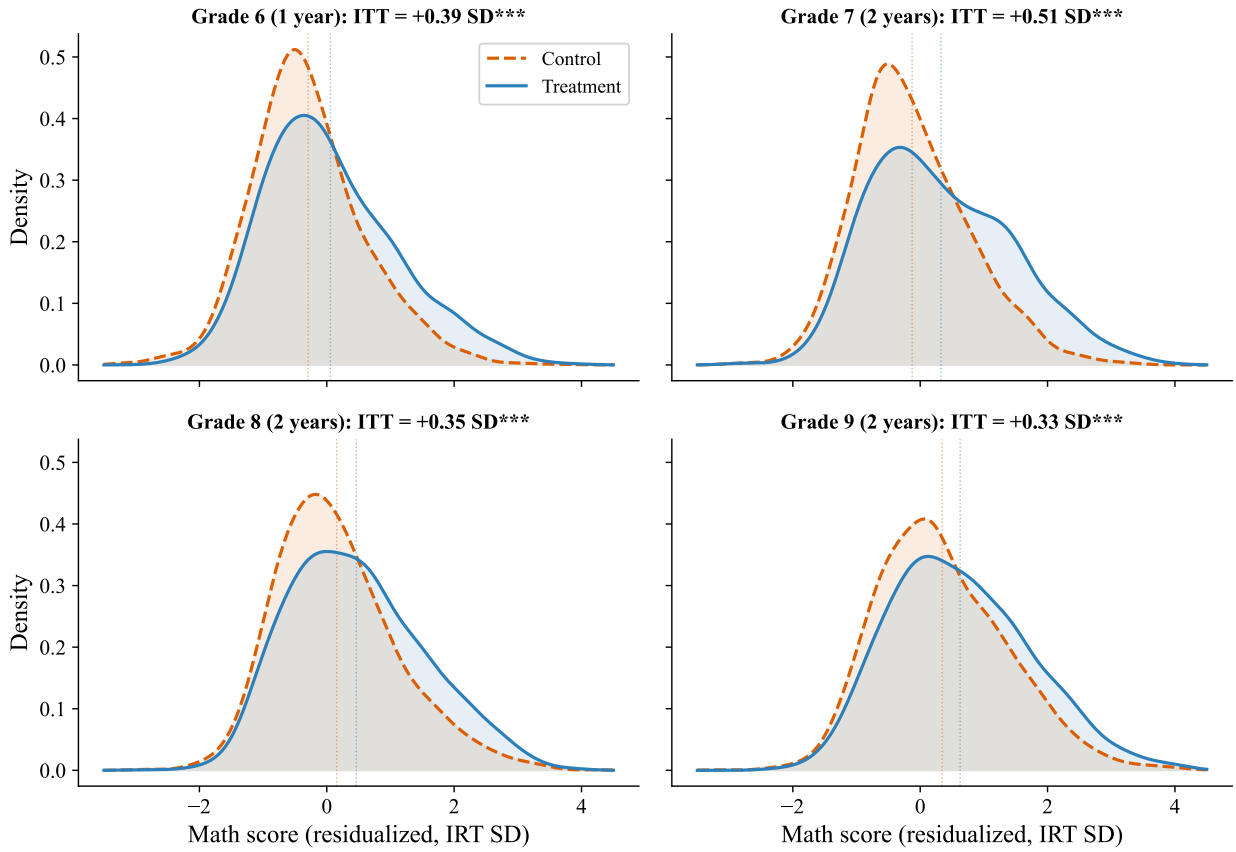


FIGURE 2. SCORE DISTRIBUTIONS BY GRADE: TREATMENT VS. CONTROL

Notes: Kernel density estimates of endline math scores by treatment status and grade. Scores are residualized for student covariates (SES index, father/mother education dummies, age, female, with missing-flag indicators) and mandal-cluster strata fixed effects, then re-centered at the grade-level mean to remain on the standardized IRT scale. Dotted vertical lines mark group means. Panel titles report the regression-adjusted ITT (matching Table 3) from OLS of *mathscore_z* on treatment, the same covariates, and strata fixed effects, with school-clustered standard errors. By Frisch–Waugh–Lovell, the T–C mean shift in each panel equals the panel-title ITT up to the residual correlation between treatment and the partialled-out covariates, which is approximately zero under random assignment. Grade 6 students had 1 year of PAL exposure; grades 7–9 had 2 years. Bandwidth = 0.25.

negative (-0.14 SD, $p = 0.04$), but is significant only with strata fixed effects (no student-level controls are available for this cohort).

One possibility often raised by concerned teachers during the rollout of PAL programs is that they could displace time spent on exam preparation, and lead to such a negative effect. The patterns of usage and performance in our data, shown in Appendix Table F2, reveals a different explanation. The negative impact of being in a program school on grade 10 board exam performance is concentrated among low-usage students. The performance of high-usage students on this exam, on the other hand, is indistinguishable from controls. What this suggests is that, perhaps consistent with other studies of deep remediation and other studies of PAL programs, if a student has classes that are replaced with no instruction, then their learning will decline.

Government school assessments have well-documented limitations as measures of absolute learning. The convergent validity, measured at the student level, between Year 2 within-school scores and our independently administered endline is modest ($r = 0.35$). Measured at the school level, it is negative ($r = -0.14$; Appendix Figure F1), consistent with school-specific grading standards introducing noise. The positive effects on these assessments should therefore be read as broadly consistent with the main IRT results, rather than as precise estimates of magnitude.

IV.C. The Relationship Between Quantity of Usage and Learning

In this section, we present estimates to quantify the relationship between the quantity of program uptake (as measured by hours of usage) and the amount of learning that study children exhibit over the period of the cluster RCT. We begin with descriptive results and finish the section with causal estimates from an instrumental variable strategy that exploits differences in the ratio of students to tablets in a given school.

In column 3 of Table 4, our estimates suggest that one additional hour of PAL usage is associated with 0.0093 sd higher test score at endline. However, the duration of PAL usage is potentially endogenous to the student's ability (higher ability students may be using PAL for a longer duration) and these estimates may be biased for the true causal effect. To address this endogeneity, we use variation in access to tablets as an instrument.

All treatment schools received exactly 30 tablets, regardless of enrollment. Appendix Figure 1 shows the full distribution of section sizes relative to this cap: 44.7 percent of treatment sections have more than 30 students and the constraint binds. Even in sections with fewer than 30 students, when a tablet breaks or when a tablet is not charged there may be more students than the number of tablets. In classes with more students than tablets, the teachers were instructed to create two batches and teach one batch while the other is in the PAL lab. However, in practice, the teachers would either ask students to take turns in using the tablets. As a result, students in the bigger sections used PAL for shorter durations.

However, it is possible that section size is directly related to learning outcomes. We instead use section size interacted with treatment as an instrument for duration of PAL usage and impute usage of 0 for the students in the control schools. The exclusion restriction assumed is that if the schools were given tablets in proportion to the section size, the treatment effect would be the same across classrooms of different sizes.

This assumption may be violated if: (a) the bigger sections are noisier making it harder for students to hear the tablets and hence gaining less, or (b) the bigger sections have more heterogeneity in the classroom

TABLE 4—DOSE-RESPONSE: IV ESTIMATES USING TABLET CONGESTION

	First Stage Usage (hrs) (1)	Reduced Form Math score (2)	OLS Math score (3)	2SLS Math score (4)
Treatment × Section enrollment	-0.370***	-0.0103***		
	(0.110)	(0.0036)		
PAL usage (hours)			0.0093***	0.0110***
			(0.0018)	(0.0014)
Treatment	48.3***	0.728***		
	(4.2)	(0.126)		
Controls	Yes	Yes	Yes	Yes
Strata FE	Yes	Yes	Yes	Yes
Grade FE	Yes	Yes	Yes	Yes
Control mean		0.001		0.001
First-stage F	11.2			11.2
Observations	14,123	14,123	6,665	14,123

Notes: Dependent variable in columns (2)–(4): IRT math score standardized to control mean and SD. Column (1): first stage regresses PAL usage hours on treatment × section enrollment and controls. Column (2): reduced form regresses math score on the same instrument. Column (3): OLS of math score on usage hours (treatment group only; biased by selection into usage). Column (4): 2SLS using treatment × section enrollment as instrument for usage hours, with the linear projection anchored at the control mean at zero usage (structural equation drops the treatment dummy so the slope attributes the full T–C gap to usage hours). The instrument exploits the fact that all treatment schools received 30 tablets regardless of enrollment; larger sections share more students per tablet, reducing per-student PAL time. Section enrollment does not affect control student scores ($\beta = 0.001$, $p = 0.75$). All regressions include controls, 21 strata FE, and grade FE. Standard errors clustered at the school level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

making the relative gain from personalization higher. The students were given earphones to use with the tablets, which reduces the possibility of (a). In the control classrooms we find that measures of within section heterogeneity like interquartile range and mean absolute deviation are constant across section sizes.

We see in Column 1 of Table 4 and Figure 4 that the duration PAL was used for is negatively correlated with section size. The F-statistic for the first stage of the two-stage least squares regression ($F = 11.2$) exceeds the conventional Staiger–Stock rule-of-thumb of 10. Column 4 in Table 4 reports the estimates. The 2SLS estimate of the causal return to program usage is 0.011 SD per hour ($p < 0.001$). The magnitude of the estimate is quite close to the conditional correlation between hours of usage and learning estimated from the treatment group via OLS estimate (0.0093 SD per hour).

In Appendix H we show the dose response relationship holds with other measures of usage like lesson completion rate which also capture the quality of usage.

IV.D. Cost Effectiveness

The program is funded through state and central (Samagra Shiksha) government education budgets and implemented by the state's regular school administration. We calculate the government's cost of delivery using administrative expenditure data from the 524-school combined sample where state-deployed PAL is in use (444 program schools, 60 RCT treatment schools, and the 20 schools of the experiment in Section V). On an annualized basis, the program costs $19.33 per student-year (INR 1,662): $13.81 for hardware, $3.67 for monitoring led by state officials and vendor's field management staff, $1.71 for software licensing, and $0.14 for teacher and official training. The hardware figure amortizes the cost of tablets and charging devices over a 3-year tablet depreciation schedule, together with annual replacement of headphones, and SD cards. Over the two-year study period, the per-student cost was therefore $38.66 (INR 3,325).[10]

Hardware is the single largest cost component, accounting for approximately 71 percent of annualized expenditure, and this cost is effectively recurring because of the rapid depreciation of tablet computers at this price point. A January 2026 inventory across the 524 schools with state-deployed tablets found that only 57 percent of tablets remained functional, with 29 percent of schools reporting zero working tablets. Furthermore, the need for tablet replacement is driven both by physical decay and by the evolution of software requirements. For example, since inception, the program has transitioned from using 8-inch to 10-inch devices, which come at a higher unit cost. Software licensing accounts for only 9 percent of costs, though current pricing reflects a volume discount that may vary. Monitoring and implementation support account for the remaining 19 percent.

To place these costs in context, we express learning gains in two complementary metrics. First, using our own data: in control schools, the cross-grade learning gradient is 0.22 SD per grade, so the treatment effect of 0.39 SD is equivalent to approximately 1.8 additional years of learning in the Andhra Pradesh government school system. Second, to enable comparison with the broader literature, we convert to Learning-Adjusted Years of Schooling (LAYS), which expresses gains relative to a high-performance global benchmark of 0.8 SD per year (Angrist et al., 2020):

$$\text{LAYS} = \frac{\hat{\beta}}{\delta_h} = \frac{0.39}{0.80} = 0.49$$

[10]All costs in 2024–25 prices, converted at $1 = INR 86 (Indian Rupees). Per-student figures assume 204 students per school, the average across the 80 research schools. Monitoring covers district officers, Field Management Staff (paid to ConveGenius), state oversight, and Samagra Shiksha; it excludes research-partner advising time, which the government does not fund. Software licensing and vendor service rates were set at program inception in 2019 and have remained nominally constant in state contracts since.

The intervention produced learning equivalent to just under half a year of highly effective schooling. Given a total program cost of $38.66 per student over two years, this translates to:

$$\text{LAYS per \$100} = \frac{0.49}{38.66} \times 100 = 1.27$$

This cost-effectiveness of 1.27 LAYS per $100 falls slightly below the midpoint of the range reported for adaptive and computer-assisted learning programs (0.24–2.66 LAYS per $100) (Angrist et al., 2020). Using previously available estimates, the Global Education Evidence Advisory Panel classified adaptive learning software as "Promising but Limited Evidence," noting that no study had demonstrated cost-effectiveness at scale in the public sector (Global Education Evidence Advisory Panel, 2023). Our estimates suggest that government-delivered adaptive learning can be cost-effective and highly impactful. While they are not among the least expensive interventions per unit of learning, there are two features of the intervention that indicate the potential of the project. One, the magnitude of the ITT estimate is very high, and demonstrates large gains learning gains at scale, as opposed to small gains or gains that are inflated via ToT estimation. The second is that much of the cost is driven by hardware expenditures, which is a general school infrastructure constraint (rather than a feature specific to this intervention's design) and one that is likely to decline with greater usage of tablets in schools.

V Remediation

The program's two-grade remediation cap is partly a pedagogical choice and partly an accommodation to pressures to ensure students are learning grade-level content: serving multi-grade-below content to enrolled-grade students makes visible to teachers, parents, and administrators that a substantial share of students are operating several grades below curricular expectation. This raises a natural question: does restricting the depth of remediation leave some unrealized learning on the table? In this section, we report a pre-registered within-school randomized experiment that answers this question directly. We find that deeper remediation produces a small additional gain—one consistent with deep content being roughly as productive per minute as the rest of the program, but delivered in such small additional quantities that the aggregate effect is modest. The shallow-remediation design, in other words, sacrifices little measured learning.

V.A. Design and the Experimental Contrast

We conducted the experiment in 20 schools drawn from the high-usage tail of the state's program, distributed across three districts (Bapatla, Prakasam, and Vizianagaram). Within each school, students in grades 6–9 were individually randomized to one of two configurations of the PAL algorithm: *Status Quo*, the standard configuration, which caps remediation at two grades below the student's enrolled grade; or *deep remediation*, which extends remediation up to five grades below and provides a small amount of additional instructional time to make sure that the internal curricular logic of the ConveGenius PAL intervention is preserved. Lab sessions are scheduled by the classroom teacher at the school-grade level, and teachers were not informed of arm assignment, so scheduled exposure to the program is identical across arms by design. The experiment therefore isolates the marginal learning return to access to deeper content, holding scheduled exposure and curriculum delivery fixed.

In Appendix J, we show evidence from administrative data and tablet usage data that the fidelity of the interventions was maintained during the study. This shows balance in terms of number of sessions administered and in quantity of usage with lessons at different levels of remediation. As mentioned above, the deep remediation treatment also entailed extra time. In the appendix, we document this to be approximately 200 additional minutes, or 3.3 hours, as compared to the status quo PAL program. This extra time focused exclusively on deep-remediation content.

V.B. Data and Specification

We restrict our analysis to students who completed the endline assessment in February–March 2025 and whose treatment exposure can be determined from the platform's module-level records. This yields 3,755 students. We define the treatment indicator as the assignment to treatment and report intent to treat estimates in the analysis below. However, there was some non-compliance to the treatment assignment. In the appendix, we show the treatment effects using the actual version of the software received as well.

The primary outcome is math ability from the same 72-item instrument used in the trial, scored in a joint IRT model estimated on the union of this sample and the trial endline sample ($\approx$ 18,000 students), so ability is measured on a common scale across the two studies. We standardize to the Status Quo group's mean and standard deviation. Our estimating equation regresses this standardized score on the deep remediation indicator, the pre-specified covariate vector, and school-by-grade-by-section fixed effects, which absorb the

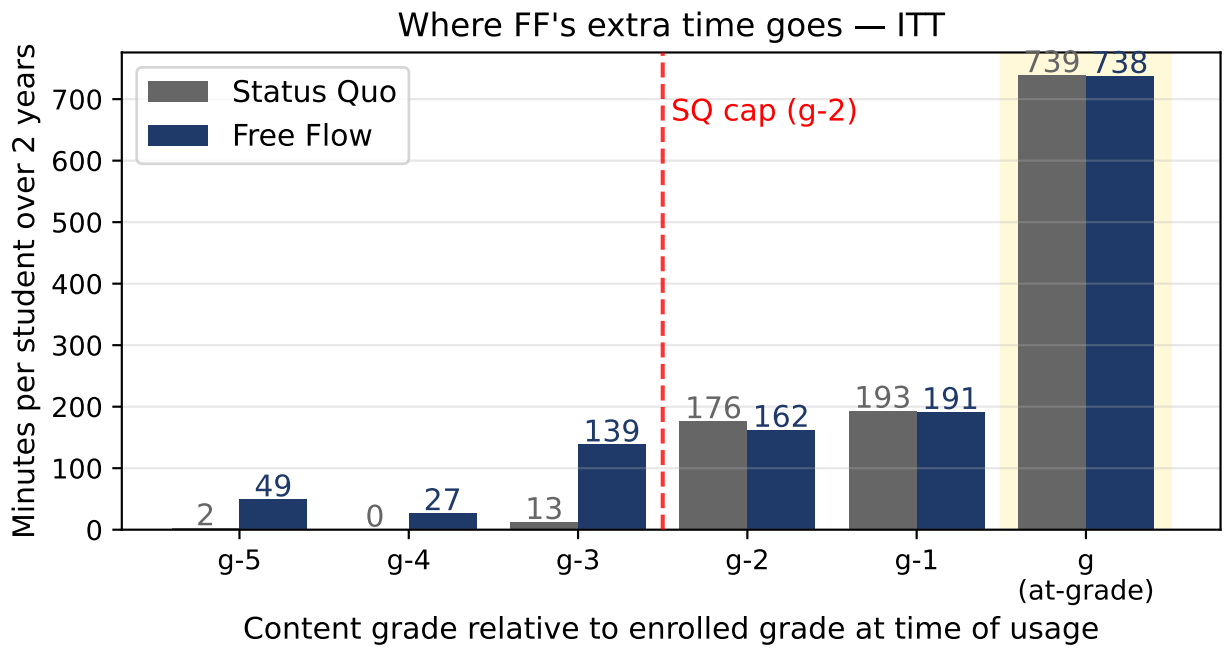


FIGURE 3. DIFFERENCES IN USAGE BETWEEN DEEP REMEDIATION AND STATUS QUO USAGE ARE DRIVEN BY DEEP REMEDIATION CONTENT

Notes: Mean minutes per student of PAL engagement, two-year cumulative (AY 2023–24 + AY 2024–25, December 2023 and January 2024 dropped from both arms; see Appendix J), by content-grade relative to enrolled grade. At $g-3$ through $g-5$ — below the 2-grade Status Quo cap — deep remediation students register an additional $+197$ minutes per student that Status Quo students cannot access. At at-grade (g), deep remediation has a small advantage ($+22$ minutes); at shallow remediation ($g-1$ and $g-2$) the arms are essentially balanced ($+6$ at $g-1$, -5 at $g-2$). The below-cap differential accounts for $\approx 90\%$ of the total Free Flow–Status Quo engagement gap of $+218$ minutes. The figure characterises the intensive margin of the treatment: at how deep a remediation level the algorithm exposes each arm's students to material.

randomization strata; standard errors are clustered at the school level ($S = 20$).

V.C. Results

Table 5 reports the main estimates. The pooled deep remediation effect is small: $+0.046$ SD ($p = 0.16$). Point estimates are positive for each grade with a statistically significant $+0.12$ SD ($p = 0.04$) gain for grade 6 students.-We find no evidence that the effect varies with students' pre-intervention math ability: deep remediation effects are positive across levels of baseline ability, and the joint test of equal effects across terciles of baseline ability fails to reject the null ($F = 0.12$, $p = 0.88$; Appendix Table J2). Despite the prior that low-baseline students have the most room to gain from foundational content, deeper remediation delivers a small, broadly uniform gain across the ability distribution.

The magnitude of these estimates is best read against the dose. The Deep Remediation intervention's distinctive treatment is approximately 3.3 hours of additional deep-remediation content over two years, so

TABLE 5—THE EFFECT OF DEEP REMEDIATION ON STANDARDISED MATH ABILITY (ITT)

	All grades	Grade 6	Grade 7	Grade 8	Grade 9
	(1)	(2)	(3)	(4)	(5)
Free Flow	+0.046	+0.121**	+0.009	+0.034	+0.033
	(0.032)	(0.059)	(0.085)	(0.050)	(0.052)
Status Quo mean	+0.037	-0.318	+0.114	+0.033	+0.242
Observations	3,263	687	786	883	907
R^2	0.228	0.255	0.249	0.210	0.173

Notes: Intent-to-treat estimates of deep remediation on standardised math ability (θ^{std}), pooled across grades (column 1) and separately by enrolled grade (columns 2–5). Specification: OLS of θ^{std} on a deep remediation indicator, a vector of pre-specified covariates, and school-by-grade-by-section fixed effects; standard errors clustered at the school level ($S = 20$). Covariates: age, female, father's and mother's education (dummies for each level), SES index, with missing-indicator imputation. θ^{std} is standardised to the Status Quo control group's mean and standard deviation within the remediation sample; the "Status Quo mean" row reports that group's mean within each column's sample and therefore deviates from zero in the by-grade columns. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

the aggregate effect implies a per-hour return of roughly 0.014 SD—the same order of magnitude as the per-hour return to overall program exposure estimated by the IV strategy in Section IV.C. Deep remediation, in other words, appears roughly as productive per minute as the program's at-grade and shallow-remediation content; the small aggregate effect reflects the small additional dose, not a low marginal return. A direct test of per-minute returns to deep content is unavailable in this design: within the Deep Remediation intervention, the students who log the most minutes on deep remediation content in particular are precisely those the algorithm diagnoses as furthest behind. Note that we are unable to use the instrument for usage (number of students enrolled in a section) because deep remediation assignment is itself the source of the identifying variation.

In Figure 4, we decompose performance on the endline test into performance on questions in one of three bins—questions at the grade level of the student; questions on topics at the level of shallow remediation (e.g., one to two grades below the student's grade); and questions on topics at the level of deep remediation (e.g., three or more grades below the student's current level). Looking at the estimates pooled across all students in all grades, we find that the deep remediation effect is largest for the shallow-remediation items (+0.072 SD, $p = 0.001$); somewhat smaller in magnitude but still statistically significant for the deep-remediation items (+0.051 SD, $p = 0.018$); and positive but not distinguishable from zero for items at grade level (+0.016 SD, $p = 0.56$). The pattern is sharpest at grade 6, where deep remediation students gain +0.164 SD ($p = 0.006$) on shallow-remediation items: improvement one to two content grades *above* where the distinctive deep-content dose was rendered. This pattern is consistent with deep content building prerequisite skills—basic

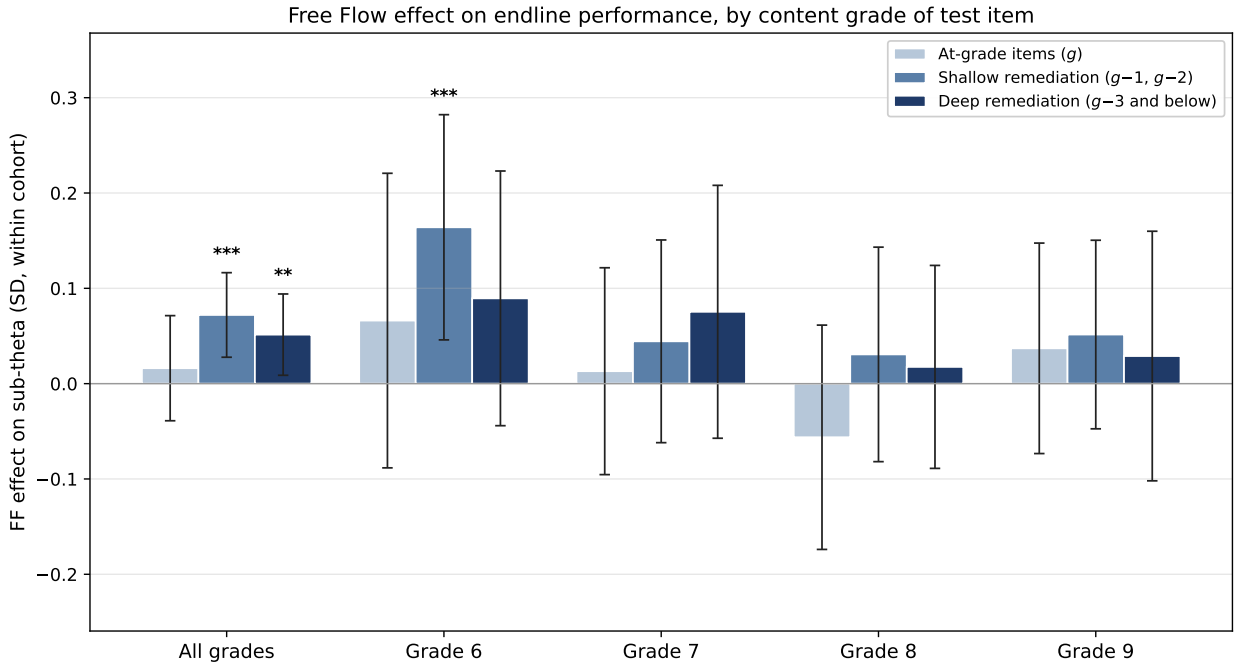


FIGURE 4. WHERE THE DEEP REMEDIATION GAIN LIVES: EFFECT BY ENROLLED GRADE AND ITEM CONTENT DEPTH

Notes: This figure shows the effect of deep remediation on endline math ability, in SD units of the within-cohort distribution, separately for three item buckets defined by content grade relative to the student's enrolled grade: at-grade (g, light), shallow-remediation ($g-1$ and $g-2$, mid-tone), and deep-remediation ($g-3$ and below, dark). Error bars show 95% confidence intervals from OLS with the Section V.B covariate vector and school-by-grade-by-section fixed effects, standard errors clustered at the school level. Stars above bars report significance levels (* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$). The Grade 6 cohort, which receives nearly twice the deep-remediation dose of Grades 7–9, shows a $+0.16$ σ gain on shallow-remediation items (Grade-4 and Grade-5 content), a smaller and statistically indeterminate gain on deep-remediation items (Grade-2 and Grade-3 content), and no detectable gain on at-grade items. The pooled shallow-item effect is highly significant ($+0.07$ σ, $p = 0.001$); other grade cohorts contribute consistent positive but smaller point estimates.

arithmetic, place value, whole-number operations—that transfer to grade-4 and grade-5 material on which students sit closer to a competency threshold. Whether the gain would propagate further, to at-grade items given longer exposure, is a question this experiment cannot answer.

Two limitations bound these inferences. Statistical power is concentrated in the pooled tests; grade- and tercile-level estimates are individually noisy. And the twenty schools are drawn from the high-usage tail of the program: in schools with weaker implementation, both the students who respond and the deep-remediation dose that Deep Remediation induces may differ. Within these bounds, the experiment's answer to the section's opening question is direct: extending remediation from two grades to five adds little measured learning at the dose the algorithm delivers, and the program's shallow design carries a small learning cost.

VI Logistics, replication, and voltage drops

In this section we first describe how the state and the external staff implemented the intervention and describe how scalable each component was. Beyond the technology and software, the implementation of the intervention included two main components: the technical support provided by the Field Management Staff (FMS) and the monitoring led by the state. Both these factors influenced the duration of PAL usage per student. We show that while the FMS played an important role in the implementation, the frequency of FMS visits eventually has limited impact on the usage of PAL. The implementation is primarily led by teachers and successful implementation is associated with the state-level authorities closely monitoring the implementation and sending nudges to teachers.

The results we have presented so far are internal-validity claims: they hold for the schools we randomized, during the two years the state actively supported the experiment. We then examine the implementation of the program in three other delivery configurations: the original trial schools in the third year after the state's attention moved elsewhere; the 444 schools where the program was rolled out in 2019 and continued being implemented since then; and roughly 700 Kasturba Gandhi Balika Vidyalaya (KGBV) residential girls' schools—henceforth, targeted residential schools—that the state added to the rollout in January 2025. The policy question we attempt to answer in this section is this: what can the parameters estimated from the cluster-RCT, particularly per-hour return, teach us about the impacts of the program at scale and past the experimental window?

VI.A. State-led monitoring and usage

The implementation was monitored directly by the office of the State Project Director. The State PAL coordinator sent 48 and 50 WhatsApp nudges in the first two academic years. The nudges included information on usage data and flagging schools with low usage, instructions on syncing data with the server, and information on review meetings. During the review meetings, the State Project Director recognized the teachers in schools with higher PAL usage.

Figure 5 also overlays the state's monitoring activity—review meetings (red) and WhatsApp usage nudges to district officers (orange)—on the usage trajectories. Some meetings precede visible rises in weekly usage. Notably, in the treatment schools in the trial, usage was initially very low in 2023 but picked up considerably after the nudges from the state. Towards the end of the trial period in Jan–Feb 2025 the usage is much lower

than the same months in the previous year. This coincides with a time period when the State PAL coordinator was assigned other responsibilities and the number of WhatsApp nudges considerably declined. We describe the implementation in greater detail in Appendix B.

VI.B. Replication in other configurations

Figure 5 organizes the section. It plots weekly math PAL minutes per enrolled student across three academic years for the three samples, and each of its panels tells a different story. The trial schools (top panel) peak at roughly 70 minutes per student per week in parts of Year 1, sustain at 30–50 minutes per student per week through Year 2, and collapse to under 10 minutes per student per week in Year 3, when the government's support pulled back. In the government's broader program rollout, shown in the middle panel, we see usage at roughly one-third the level of that in the cluster-RCT during the first two years, collapsing to essentially zero in Year 3. In the targeted residential schools, shown in the bottom panel, we see usage jump to roughly 50 minutes per student per week a few months after its January 2025 launch, and these schools sustain usage at 20–40 minutes per student per year through Year 3, despite the decline in the other two groups. We tell each story in turn, then show that a single dose-response curve organizes all three.

To measure learning under each configuration, we assessed three samples at a Year-3 endline in February–March 2026 on a parallel math assessment, with Year-3 ability placed on the Year-2 IRT scale through Mean–Sigma linking on 45 common anchor items: a random 80-school subset of the parent trial (40 treatment, 40 control); a 40-school subsample of the 444-school program rollout; and 40 of the targeted residential schools. The first sample re-tests the original experiment one year past the experimental window; the second and third reach out-of-experiment populations operating under different delivery configurations.

Figure 5 is descriptive and we cannot infer the causal impact of the state monitoring. But it previews the section's organizing fact: usage moves with the intensity of the state's implementation effort and, ultimately, its focus.

VI.C. The PAL Program in the Cluster-RCT Schools After the End of the Study

In the year after the cluster-RCT concluded, several facets of the implementation and monitoring of the PAL program collapsed simultaneously in these schools.[11] The number of nudges sent to teachers collapsed from 46 and 48 in the previous two academic years to 4 in the first half of the academic year. The FMS

[11]The broader delivery infrastructure is described in Appendix II.D.

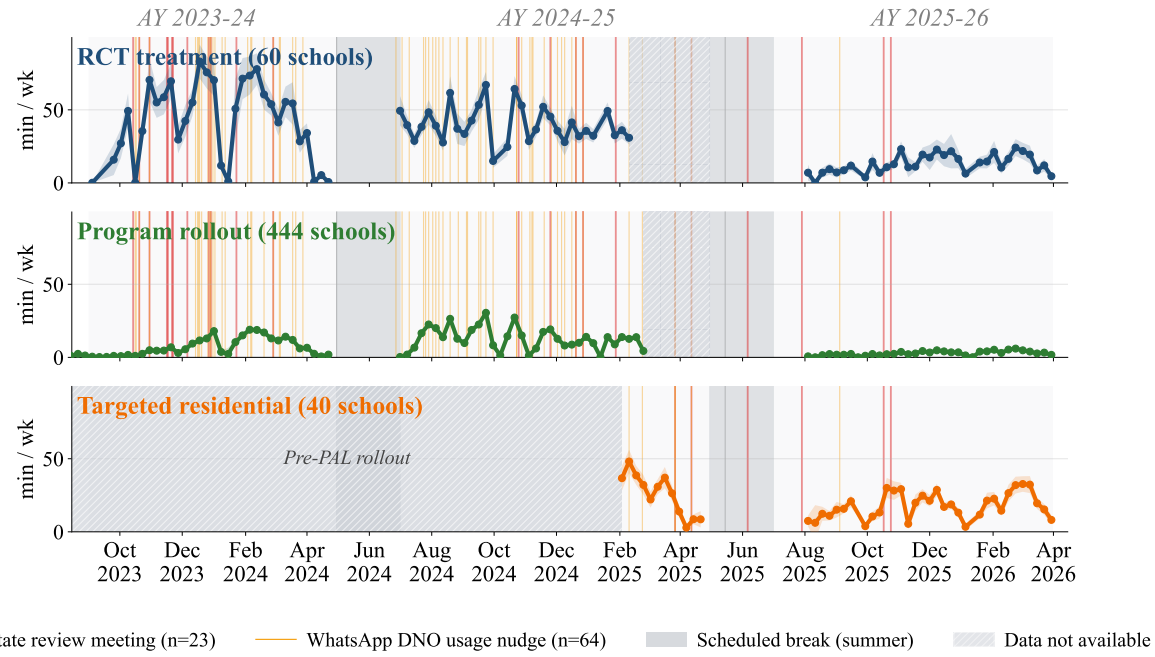


FIGURE 5. THREE-YEAR MATH PAL USAGE: RCT TREATMENT, PROGRAM ROLLOUT, AND TARGETED RESIDENTIAL SCHOOLS

Notes: Weekly mean math PAL minutes per enrolled student across three academic years, by delivery configuration. RCT treatment (60 schools, blue) operated AY 2023–24 and 2024–25 as an intensive math-only intervention; AY 2025–26 reflects post-experimental delivery. Program rollout (444 schools, green; 391 in Y1 due to a 2023–24 ConveGenius export gap) shares tablet-time across math, Telugu, and English. Targeted residential (KGBV) schools (40 schools, orange) joined the AP rollout in January 2025; pre-2025 traces do not exist. Per-school enrollment is taken from each year's official register; AY 2025–26 uses the AY 2024–25 register as a placeholder while the Y3 register is being finalized. Vertical markers: red for state-coordinator review meetings ($n = 23$); orange for WhatsApp DNO usage nudges ($n = 64$). Summer breaks are shaded in solid gray; periods where data are not available (e.g., targeted residential schools before their January 2025 launch) are shaded with a hatched gray, as indicated in the legend.

hired for the previous two years were let go because the state had not approved the PAL budget for the year. The State PAL coordinator was also tasked with implementing the intervention in 700 additional residential schools in that year with no additional staff for monitoring. Further, many schools reported having many tablets broken by the start of the third academic year. As a result, the usage in the third year collapsed to

The impacts of this contraction on project implementation varied substantially across schools. Six of the 60 treatment schools delivered effectively zero PAL in Year 3, and an additional 28 delivered fewer than five hours per student. Notably, the zero-delivery schools were low-fidelity implementers during the RCT as well—their Year-1 and Year-2 doses were already roughly half those of the high-implementing schools. Once the external pressure from government attention receded, however, this dropped to zero.

Evidence from surveys of headmasters, teachers, and students in the second and third year of the study sheds light on the mechanisms behind this drop. This survey finds that 52 percent of treatment schools experienced a tablet shortage. Across the 164 PAL-school teachers surveyed, 29 percent report having either

TABLE 6—YEAR-3 ATTENUATION IN THE RCT PANEL

	(1) Year 2 θ_z	(2) Year 3 θ_z
Treatment	0.309***	0.101
	(0.072)	(0.065)
Control mean	-0.005	0.381
Treatment mean	0.329	0.534
Observations	3,092	3,092
Controls	Yes	Yes
Strata FE	Yes	Yes

Notes: The sample is the balanced panel of 3,092 RCT students observed at both the Year-2 and Year-3 endlines (grades 7 and 9). Each column regresses the standardized IRT math score (θ_z, standardized to the control-group mean and SD) on a treatment indicator, using the same specification as the main outcomes: student covariates, strata fixed effects, and standard errors clustered at the school level. Column (1) is the Year-2 score; column (2) is the linked Year-3 score. *Control mean* and *Treatment mean* are unadjusted group means of the outcome. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

reduced the number of PAL sessions held during the year or cancelled them altogether, citing pressure from multiple sources: competing school activities (75 percent); the need to complete the material covered in the curriculum for that year (40 percent); and technical issues with tablets, electricity, or internet (36 percent). Students report the same contraction: while at the Year-2 endline survey 91 percent of treatment students reported having two or more PAL labs per week and only 0.6 percent reported none; at the end of the third year, only 53 percent reported that they had at least one PAL session each week. A further 19 percent reported that the PAL lab ran only before exams or not at all.

Likely as a result of this drop in fidelity, we find that at the end of the third year, the learning gains that students in the treatment group experienced, relative to the control group, shrank. We present these results in Table 6. The analysis restricts to students assessed in both the years. We find that at the end of Year 3, the Year-2 treatment effect estimate (+0.31 SD) declined to +0.1 SD ($p = 0.12$). While the collapse in usage lead to no additional learning gain in the PAL schools, the treatment effects also attenuate in the year because the control students catch up at a faster rate.

VI.D. The Program Rollout

The 444-school program rollout ran alongside the trial for all three years and delivers a structurally smaller dose. Program students receive roughly one-fifth the per-student math PAL of trial students—6.9 versus 35.3 cumulative hours over AY 2023–25 (Table 7)—and the gap is structural rather than operational. Program schools have more than twice the grade 6–9 enrollment of trial schools (326 versus 155) under the

TABLE 7—PAL USAGE BY COHORT: YEARS 1–3

	RCT Treatment (60 schools)	Program (all)	Program (>0 tabs)	Targeted residential (40 schools)
Schools	60	444	375	40
Students (active on PAL)	6,532	134,090	110,458	—
Mean enrollment (grades 6–9)	155	326	319	159
Working tablets (reported, mean)	27.1	19.0	22.5	—
Peak monthly active students per grade (mean)	26.5	66.5	65.1	—
Student-tablet ratio per grade-cohort (mean)	1.2	4.4	4.4	—
Math usage Y1+Y2, hours (mean)	35.3	6.9	7.1	—
Math usage Y1+Y2, hours (median)	31.1	5.5	5.7	—
Months active Y1+Y2 (mean)	6.9	5.2	5.3	—
Math usage Y3, hours (mean)	5.4	0.64	—	13.6
% math-active in Y3	60%	17%	—	97%
Y3 θ_z (mean, linked)	+0.32	+0.21	—	+0.54

Notes: RCT treatment schools used PAL for math only. Program schools used PAL for math, Telugu, and English (3 subjects). Targeted residential schools joined the AP rollout in January 2025; Years 1 and 2 rows are not applicable. Years 1 and 2 usage statistics are for math only in both groups, summed across AY 2023–24 and AY 2024–25. Year-3 rows are computed on Year 3 endline-panel students (n: RCT treatment 2,967; Program 5,957; Targeted residential 5,452); Year 3 θ_z is standardized to the RCT-control mean and standard deviation. "Active" students are those with any recorded PAL usage across the relevant period (Years 1 and 2 for the upper block, Year 3 for the lower block). "Working tablets (reported)" is from government/CG infrastructure records. "Peak monthly active students per grade" is computed per school-grade pair: the maximum, across months of AY 2024–25, of the count of students in that grade with any usage that month; averaged across school-grade pairs. Grade-level disaggregation avoids pooling four cohorts who use the same tablets at different periods. "Student-tablet ratio per grade-cohort" is the mean of grade-cohort enrollment / school working tablets, computed across (school × grade) cells. A PAL period serves one grade-cohort sharing the school's tablet allocation; the ratio exceeds 1 when the cohort is larger than the working tablet count. For program schools we use student-grade counts as the cohort proxy (section-level enrollment data are not available in program records). Program schools were part of the government's PAL rollout (17 of 26 AP districts) beginning in 2019. RCT schools began implementation in AY 2023–24. All schools—both program and RCT treatment—have a 30-tablet cap regardless of enrollment.

same 30-tablet allocation, producing a per-cohort student-to-tablet ratio of 4.4:1 against the trial's 1.2:1. And program schools split the lab across three subjects, allocating math two of four weekly PAL periods, where the trial configuration was math-only. Using administrative data at the session-level, we see that the gap in dosage (i.e., having fewer PAL classes over the year) operates through session frequency, not the number of students per classroom. While the number of students per PAL class is similar in program schools and the cluster RCT schools (median 20 versus 17 students), students in program schools attend one fewer session per month than those in the RCT schools (2.7 versus 3.7; see Appendix Figure 5). Put differently, the bottleneck is the school deciding to have a given math class be a PAL session, not a constraint in terms of number of devices per student.

Another key failure point was the deterioration and ultimate failure of many tablets in the third year. In the third year of the study, the number of working tablets per school declined by approximately 19 percent, and

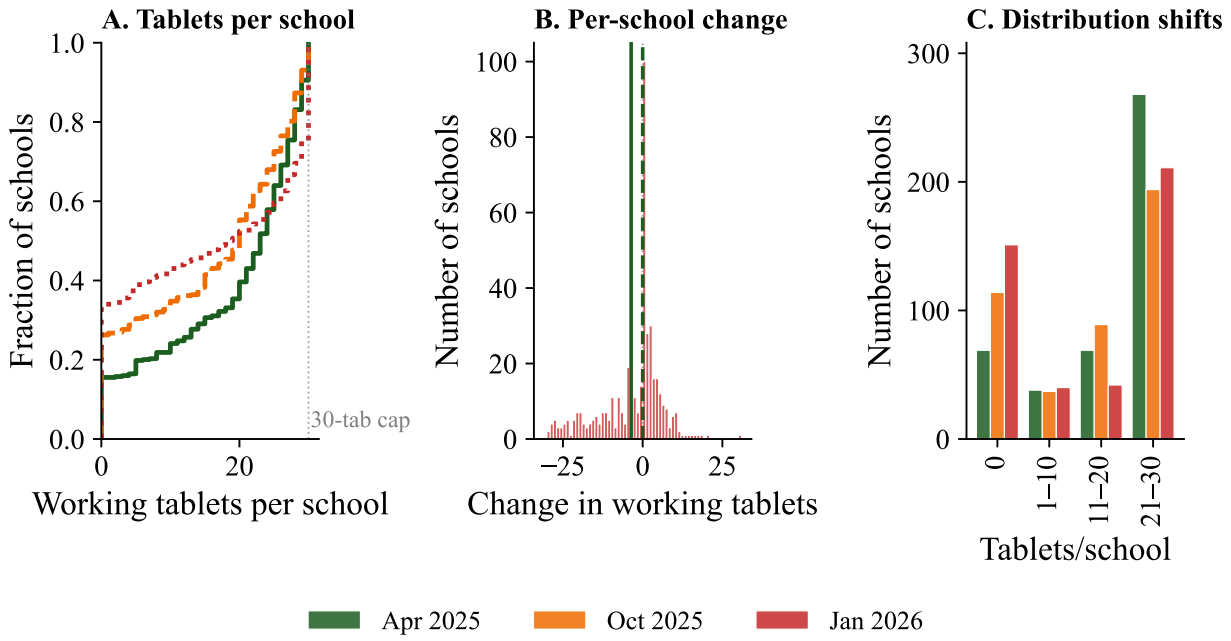


FIGURE 6. WORKING TABLET INVENTORY ACROSS 444 PROGRAM SCHOOLS, 2025–2026

Notes: Three inventory snapshots: April 2025 (end of AY 2024–25; $n = 444$, median 23), October 2025 (start of AY 2025–26; $n = 434$, median 20), and January 2026 (mid AY 2025–26; $n = 444$, median 18). *Panel A*: empirical CDF of working tablets per school at each snapshot. *Panel B*: per-school change between April 2025 and January 2026 (mean −3.6, median 0); 174 of 434 schools (40%) lost tablets, 142 (33%) were stable, 118 (27%) gained. *Panel C*: distribution of schools across 0/1–10/11–20/21–30 working-tablet buckets. Mean attrition over the nine-month window is −3.6 tablets per school (−19%).

this masked substantial heterogeneity. By the end of the year, 34 percent of program schools had zero working tablets (Figure 6). Because the original 2019–2022 deployment has not been systematically replaced, the dose at scale reflects accumulating hardware attrition layered on top of the structural three-subject allocation.

Within program schools, regressing Year-3 ability on cumulative PAL hours with school fixed effects yields a slope of +0.025 SD per hour—slightly steeper than the dose response slope observed in the trial schools.

VI.E. The Targeted Residential Schools

The third configuration is the one the state added last and ran best. The targeted residential schools are government boarding schools for girls from educationally disadvantaged groups. In January of 2025, the state announced it would rollout the PAL program in 700 of them.

We collected a learning assessment in 40 of these schools in the end of the third year of our study. We show these results in the bottom panel of Figure 5. In terms of implementation and uptake, these schools rapidly ramped to high levels of usage, reaching roughly 50 minutes per student per week within months of

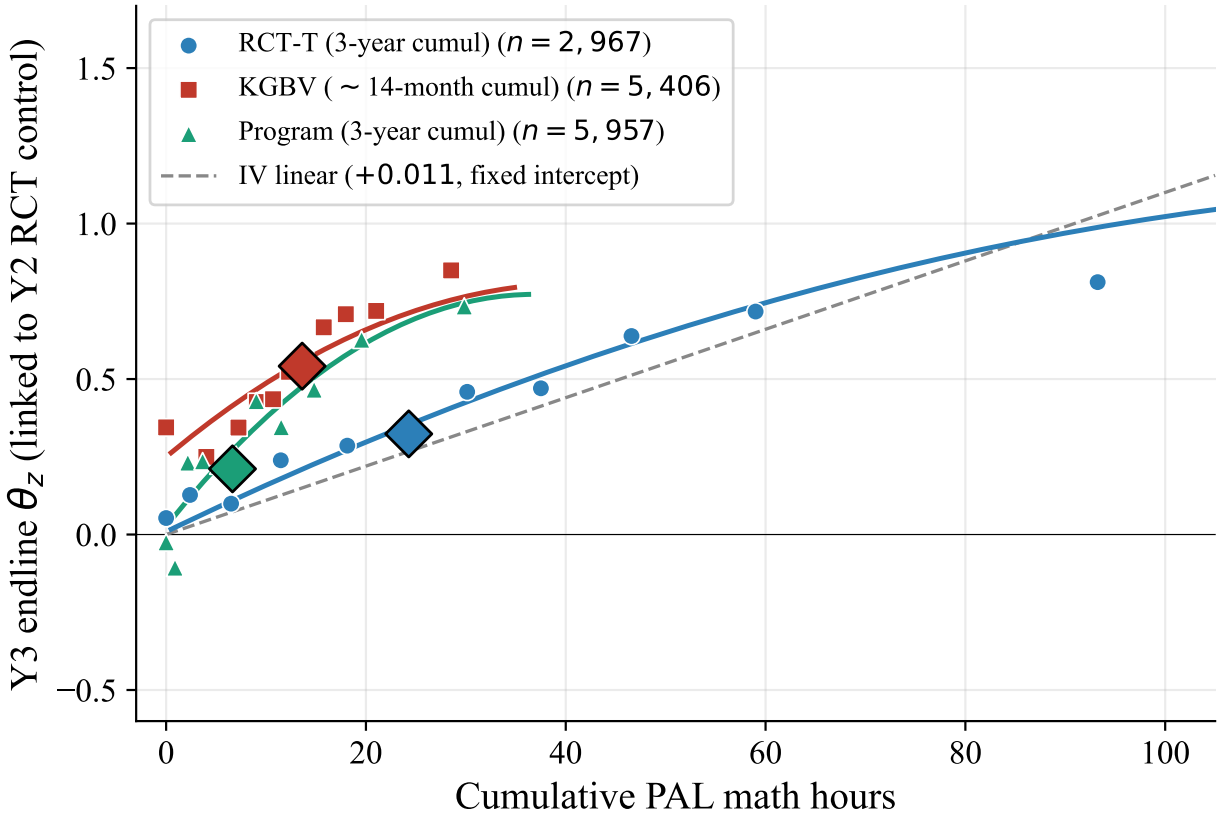


FIGURE 7. CROSS-SAMPLE WITHIN-SCHOOL DOSE-RESPONSE ON CUMULATIVE PAL HOURS

Notes: Within-school binscatter and quadratic fit (zero-bin plus 10 active deciles per sample) of Y3 endline IRT θ (linked to the Y2 RCT-control scale via Mean–Sigma on 45 anchor items) on cumulative PAL math hours, for three delivery configurations: RCT-T treatment arm (blue circles, 3-year cumul AY 2023–26, $n = 2{,}967$), targeted residential (KGBV) girls' schools (red squares, ∼14-month cumul since the January 2025 launch, $n = 5{,}406$), and program-rollout subsample (green triangles, 3-year cumul AY 2023–26 with Y1 student-level hours recovered via SHA-256 hash linkage to the cleartext Y3 student_code, $n = 5{,}957$). Sample-mean diamonds mark each sample's mean hours and mean Y3 θ. Quadratic fits clipped to each sample's observed support (approximately 0–100 hrs RCT-T, 0–35 hrs residential, 0–12 hrs program). Gray dashed reference is the fixed-intercept IV slope of +0.011 identified causally on the RCT 2-year sample. Each sample's per-hour slope orders by where its mean sits on the curve — consistent with a single concave dose-response whose averaged slope depends on the sample's exposure region. Per-hour coefficients with significance tests are reported in Appendix Table 5.

launch. Furthermore, they sustained this level of usage over the next 12 months, staying at roughly 20–40 minutes through Year 3—even as delivery in the the 60 cluster-RCT schools and the 444 program schools collapsed. By the Year-3 endline, residential school students had received a total 13.6 hours of math PAL instruction, with 97 percent of students active on the platform.

Within residential schools, the dose-response slope on cumulative hours is +0.015 SD per hour, very close to the slope observed within the trial schools.

Figure 7 assembles the three stories into a single picture. Within each configuration, regressing Year-3 ability on cumulative hours with school fixed effects recovers a similar dose-response curve with slopes:

+0.025 SD per hour in program schools (mean 6.6 hours of usage), +0.015 in residential schools (mean 13.6 hours of usage), and +0.010 in the trial's three-year panel (mean 24.3 hours of usage)—the last matching the causal IV slope of +0.011 identified on the trial's two-year sample. The comparisons outside the trial remain descriptive, but the pattern is consistent with what a single concave dose-response curve would produce when sampled at three different exposure levels.

Our estimates of the per-hour return to in-school adaptive learning replicates across delivery configurations, school types, and years. Wherever students accumulate hours using the ConveGenius PAL program, learning follows at roughly the rate the trial identified causally. What does not replicate is the dose: the trial configuration produced 35 hours per student under active state monitoring, and its gains halved when that monitoring receded. Contemporaneous rollout of the program in 444 government schools not under the scrutiny of a trial produced almost no hours over two years. A third, later rollout—done in 700 residential schools without the scrutiny of a trial, but with renewed government enthusiasm in the spirit of "ribbon-cutting bias" (Glaeser and Poterba, 2021)—produced a substantial number of hours spent using PAL and yielded learning outcomes to match. Translating the per-hour return into learning at state scale therefore depends on the integrity of the delivery infrastructure—monitoring, field support, working hardware, and tablet-time allocation—rather than on any change to the intervention's design.

VI.F. Randomized FMS visits

Our final analysis in this section reports an experiment conducted in a subset of 123 schools, including both some of the cluster-RCT schools and others in the broader rollout but not the experiment.[12] This experiment took place in the third year of our broader study, i.e., the year after the completion of the cluster-RCT. In this experiment, we randomized the frequency with which the FMS were scheduled to visit schools. Among the roughly 12 schools assigned to each FMS, we created strata of 4-7 schools based on the number of available working tablets in the school. Within each strata half of schools were assigned to low frequency visits and those in the other half were assigned to high frequency visits. The FMS were instructed to visit the schools in the high frequency arm roughly twice as often as they visited the schools assigned to the low frequency arm. They were given a suggested order in which to visit the school but that was not strictly enforced because they needed to visit schools if a teacher asked for help.

[12]We did not include the treated residential schools in this study because the FMS assigned to them were hired and administered separately from those in the other two groups of schools.

We see in Figure 8 the time trends in the FMS visits and usage in the schools assigned to high and low frequency visits. The pre-period (shaded), before any school received a study visit, shows no FMS activity in either arm and balance in the usage metrics. In the top panel we see the mean FMS visits per school per week being higher in the schools assigned to the high frequency visit since the start of the experiment in Nov 2025. In the middle panel we see that the intensive margin of minutes of usage per week per student does not trend differently in the high and low frequency schools. Similarly, in the bottom panel, the trend in the extensive margin of share of students with active PAL usage in a week is similar across both sets of schools.

In column 1 of Table 8 we see that the schools in the low frequency arm received 0.083 visits per week and the schools in the high frequency arm received 0.137 (p-value = 0.025) additional visits per week. The treatment effect on the average duration of usage per student is -0.084 minutes per week per student (mean of 2.479 minutes per week per student in low frequency schools, p-value = 0.89). The effect on the share of students with a non-zero usage is -0.001 (mean of 0.068 in low frequency schools, p-value=0.935). These results are subject to multiple caveats: the usage was very low overall in this period—the usage was on average about 2 minutes per student per week; the usage was partly low because many tablets were broken in many schools and the usage is driven by the few schools with still functional tablets; the FMS hired were in the first few months of their job and may have still been learning on the job; the schools in the experiment had been implementing the PAL labs for over two years by this time and the teachers were aware of the process.

We do not interpret these results as the FMS not having a role in the implementation of the program. On the contrary, they played a critical role in ensuring that the software was correctly loaded on all tablets and provided immediate on-ground technical support to the teachers. However, unlike in the implementation of Mindspark in Rajasthan that Muralidharan and Singh (2025) evaluate or the implementation of Khan Academy that Oreopoulos et al. (2026) study, the intervention in this experiment was led by the state with the external staff having only a skeletal presence and providing technical support. This makes the implementation model very scalable. Once the teachers are trained in the implementation, even a small cadre of staff might suffice to ensure implementation fidelity.

VII Conclusion

In this paper, we've shown that government can implement targeted learning at scale without deeper mediation and achieve very large learning gains in middle school. Using data from a cluster-randomized

TABLE 8—PHASE-1 REGRESSION RESULTS: FIRST STAGE, USAGE, AND SHARE USING PAL OVER 12 WEEKS

	(1) First stage	(2) Usage	(3) Share
Dep. variable	visits/sch/wk	usage/stu	share using PAL
HF	0.137***	-0.084	-0.001
	(0.025)	(0.612)	(0.014)
	[0.000]	[0.890]	[0.935]
Low-frequency mean	0.083	2.479	0.068
Strata FE	Yes	Yes	Yes
Std. errors	HC1	HC1	HC1
Observations	123	115	123
Schools	123	115	123

Each column is a separate school-level regression on the Phase-1 sample over the Nov 3 – Jan 26 active window (about 12 weeks); all include randomisation-strata (FMS $\times$ tablet-count block) fixed effects. HF = high-frequency arm indicator. Column (1): mean weekly FMS visits per school. Column (2): mean weekly PAL usage per enrolled student. Column (3): mean weekly share of enrolled students using PAL. *Enrolment* is grades 6–9 per school from JVK_Data.csv (AY 2023–24), the house per-enrolled-student denominator (cf. impact cleaning/06, surveys/18); usage per enrolled student = school total minutes / enrolment (non-users count as zero). Schools with no ConveGenius records anywhere are dropped from column (2) as missing (not zero), which is why it has fewer schools. *Low-frequency mean* is the control-group mean of the dependent variable. Singleton strata dropped; HC1 standard errors in parentheses, p-values in brackets. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

control trial, a within-school across-student randomized trial, and administrative data on rollout across three different contexts, we show the conditions necessary for the success of a Personalized Adaptive Learning program in the spirit of programs that are currently in place in a plurality of Indian states.

References

Angrist, Noam, David K. Evans, Deon Filmer, Rachel Glennerster, F. Halsey Rogers, and Shwetlena Sabarwal, "How to Improve Education Outcomes Most Efficiently? A Comparison of 150 Interventions Using the New Learning-Adjusted Years of Schooling (LAYS) Metric," Technical Report Policy Research Working Paper 9450, World Bank, Washington, DC 2020.

Ardington, Cally, Alicia Menendez, and Jack Thunde, "The Limits of a 'Great Buy': What Rigorous Evidence Reveals about Teaching at the Right Level," Technical Report, African Foundational Learning Data Hub (AFLEARN), DataFirst, University of Cape Town January 2026.

ASER Centre, "Annual Status of Education Report (Rural) 2024," Available online from ASER Centre (Pratham) 2024. Nationwide household survey of rural learning outcomes and enrollment in India.

Banerjee, Abhijit, Rukmini Banerji, James Berry, Esther Duflo, Harini Kannan, Shobhini Mukerji, Marc Shotland, and Michael Walton, "From Proof of Concept to Scalable Policies: Challenges and Solutions, with an Application," *Journal of Economic Perspectives*, 2017, *31* (4), 73–102.

Banerjee, Abhijit V., Rukmini Banerji, Esther Duflo, Rachel Glennerster, and Stuti Khemani, "Mainstreaming an Effective Intervention: Evidence from Randomized Evaluations of "Teaching at the Right Level" in India," NBER Working Paper 22746, National Bureau of Economic Research 2016.

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

Barnett, Inka, Proochista Ariana, Stavros Petrou, Mary E Penny, Le Thuc Duc, S Galab, Tassew Woldehanna, Javier A Escobal, Emma Plugge, and Jo Boyden, "Cohort profile: the Young Lives study," *International journal of epidemiology*, 2013, *42* (3), 701–708.

Beg, Sabrin, "Digitization and development: Property rights security, and land and labor markets," *Journal of the European Economic Association*, 2022, *20* (1), 395–429.

Bold, Tessa, Mwangi Kimenyi, Germano Mwabu, A Ng'ang'a, and J Sandefur, "Experimental evidence on scaling up education reforms in Kenya," *Journal of Public Economics*, 2016, p. 168.

de Barros, Andreas and Alejandro J. Ganimian, "Which Students Benefit from Computer-Based Individualized Instruction? Experimental Evidence from Public Schools in India," *Journal of Research on Educational Effectiveness*, 2024, *17* (2), 318–343.

Glaeser, Edward L. and James M. Poterba, "Economic Analysis and Infrastructure Investment," in Edward L. Glaeser and James M. Poterba, eds., *Economic Analysis and Infrastructure Investment*, University of Chicago Press, 2021.

Global Education Evidence Advisory Panel, "Cost-Effective Approaches to Improve Global Learning: Smart Buys 2023," Technical Report, World Bank and Foreign, Commonwealth & Development Office, Washington, DC 2023.

Kaffenberger, Michelle, Noam Angrist, Yue-Yi Hwa, Heather Leigh Kayton, Matthew Jukes, and Jonathan Stern, "Core components of teaching at the right level: Unpacking the black box of proven programmes into a set of 'core components' by systematically combining multiple sources of rigorous evidence with implementer insights," Working Paper 2026/001, What Works Hub for Global Education 2026.

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

_ **and** _ , "Adapting for Scale: Experimental Evidence on Technology-Aided Instruction in India," NBER Working Paper 34205, National Bureau of Economic Research September, revised November 2025.

_ **and Venkatesh Sundararaman**, "The aggregate effect of school choice: Evidence from a two-stage experiment in India," *The Quarterly Journal of Economics*, 2015, *130* (3), 1011–1066.

_ **, Paul Niehaus, and Sandip Sukhtankar**, "Building state capacity: Evidence from biometric smartcards in India," *American Economic Review*, 2016, *106* (10), 2895–2929.

_ **,** _ **,** _ **, and Jeffrey Weaver**, "Improving last-mile service delivery using phone-based monitoring," *American Economic Journal: Applied Economics*, 2021, *13* (2), 52–82.

Okunogbe, Oyebola and Victor Pouliquen, "Technology, taxation, and corruption: evidence from the introduction of electronic tax filing," *American Economic Journal: Economic Policy*, 2022, *14* (1), 341–372.

Oreopoulos, Philip, Oliver Keyes-Krysakowski, and Deepak Agarwal, "How In-School Supervised Ed-Tech Support Produces Massive Learning Gains: A Khan Academy Field Experiment in India," NBER Working Paper 34683, National Bureau of Economic Research January 2026.

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

_ , "Improving administrative data at scale: Experimental evidence on digital testing in Indian schools," *The Economic Journal*, 2024, *134* (661), 2207–2223.

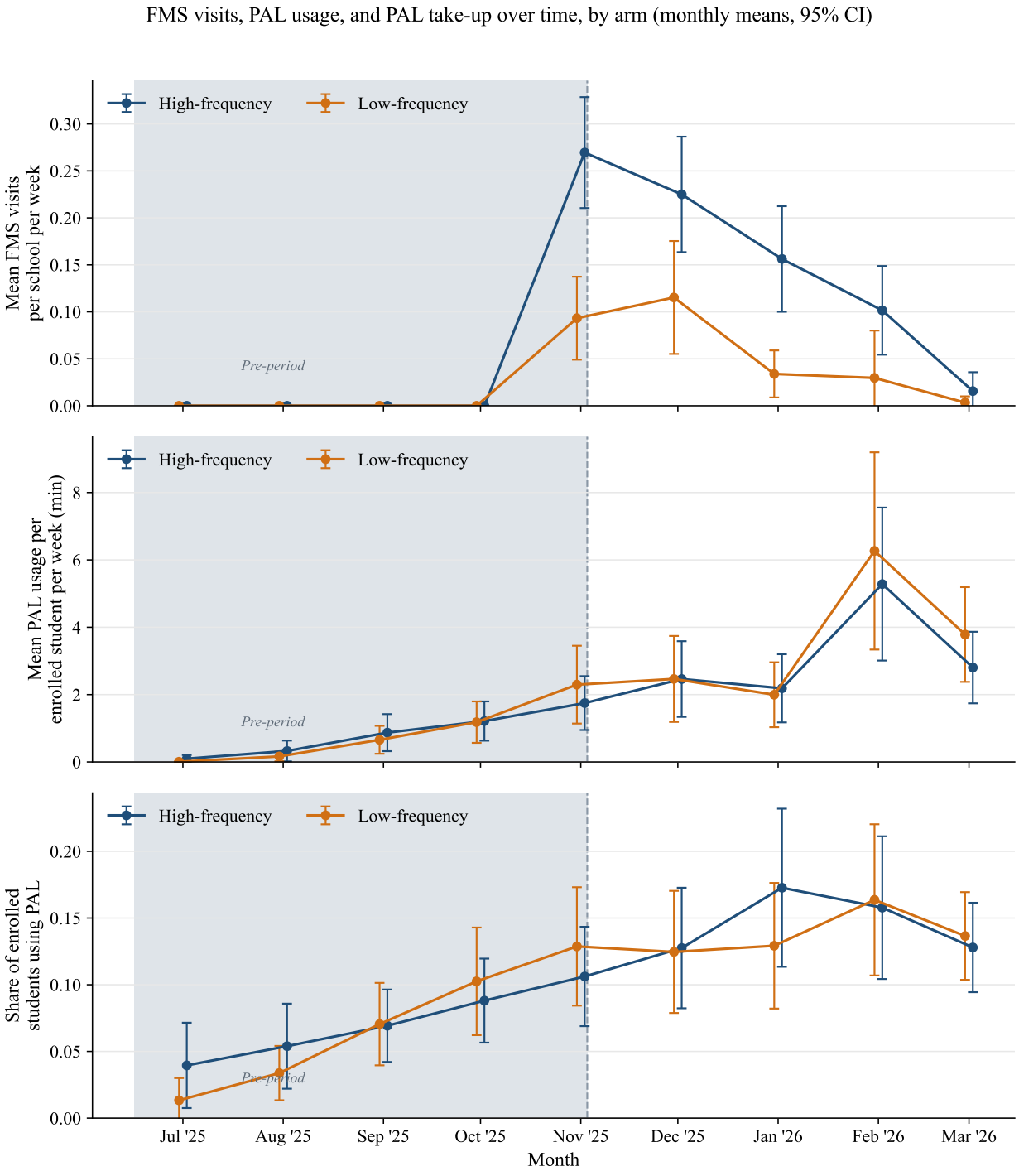


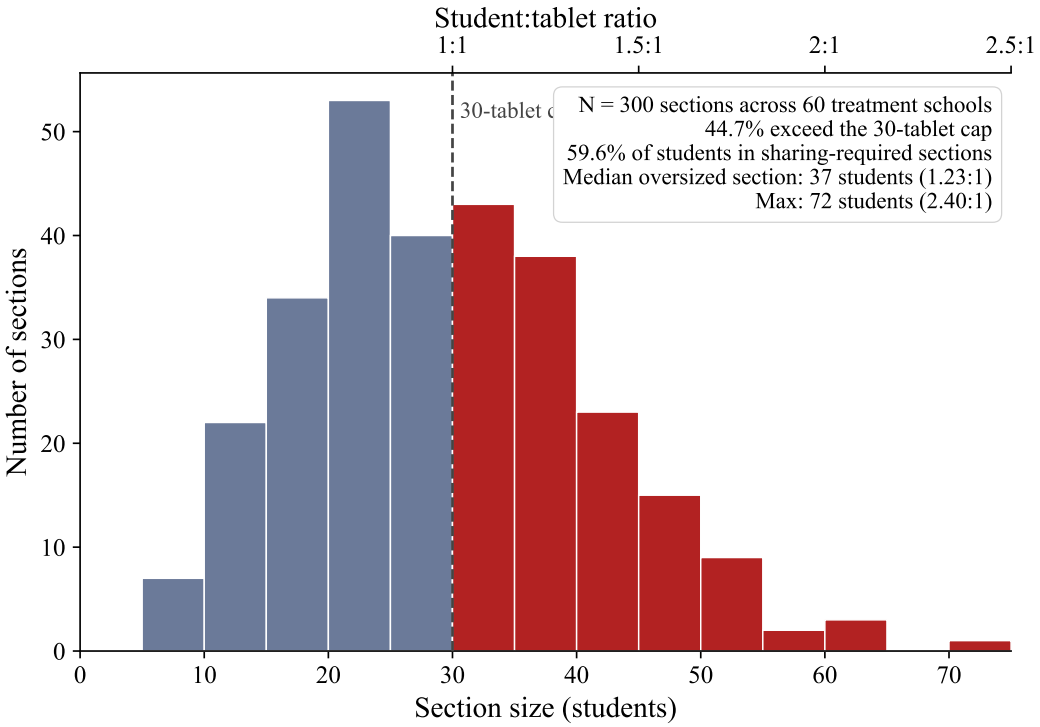
FIGURE 8. FMS VISIT FREQUENCY, PAL USAGE, AND TAKE-UP BY TREATMENT ARM

Notes: Phase-1 schools of the FMS visit-frequency A/B test ($N = 123$: 64 high-frequency, 59 low-frequency), randomized within FMS × tablet-count strata across five districts, with monitoring visits rolled out in Nov, 2025 (after the shaded region). Points are monthly arm means; error bars are 95% confidence intervals across schools. The top panel presents number of FMS visits per school per week, the middle panel presents PAL usage per student per week in minutes, and the bottom panel presents the share of enrolled students with active PAL usage in a week.

Appendices

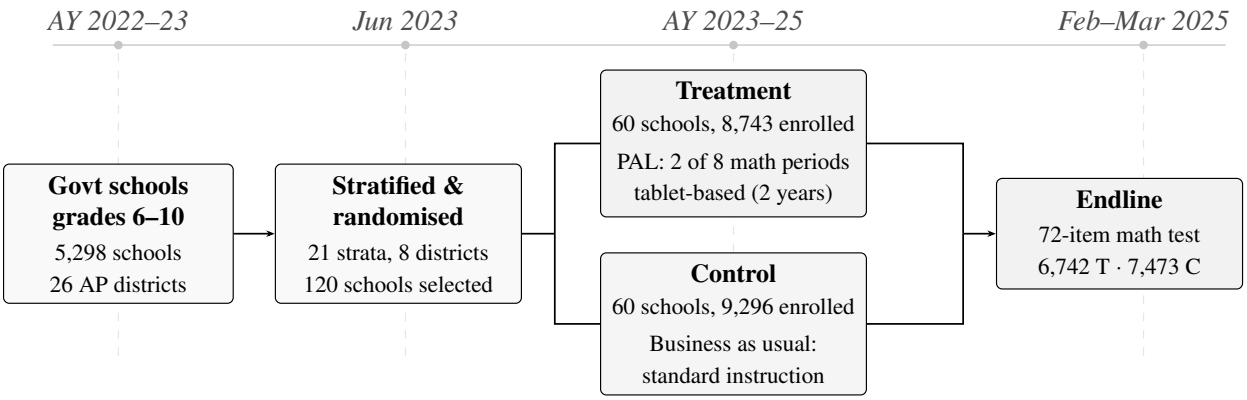
Appendix Figures

This section collects figures cited in the main text that are not placed in the body of the paper. They appear in the order in which they are first cited. Figures that belong to a specific appendix are placed in that appendix and carry its letter.



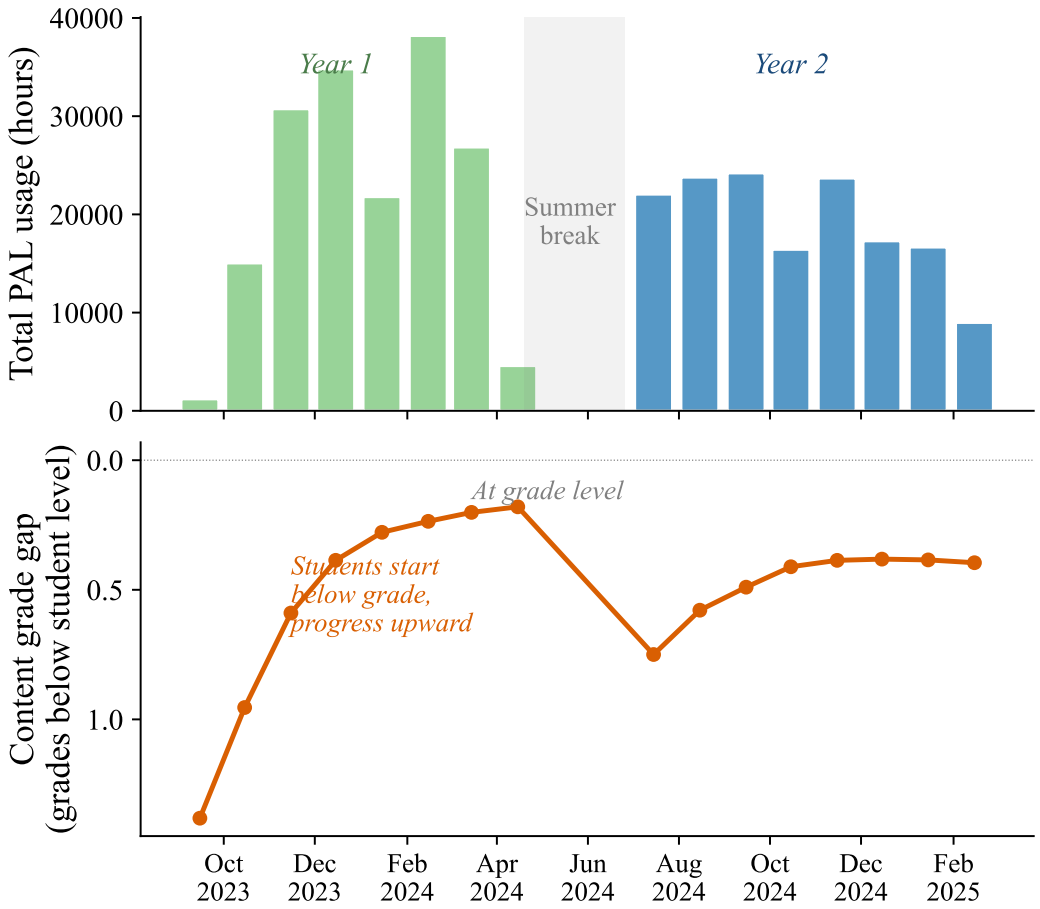
APPENDIX FIGURE 1. CLASS SECTION SIZES RELATIVE TO THE 30-TABLET CAP

Notes: Treatment schools only; $N = 300$ class sections across 60 schools, defined as unique school-grade-section combinations. Section enrollment was missing for 1.1% of treatment students; these were imputed using the mean section enrollment within each student's school-grade cell. Sections ≤ 30 students (blue-gray, 166 of 300) could accommodate every student on a dedicated tablet; sections > 30 (red, 134 of 300) required sharing. The top axis shows the implied student-to-tablet ratio. 59.6% of treatment students were in a sharing-required section.



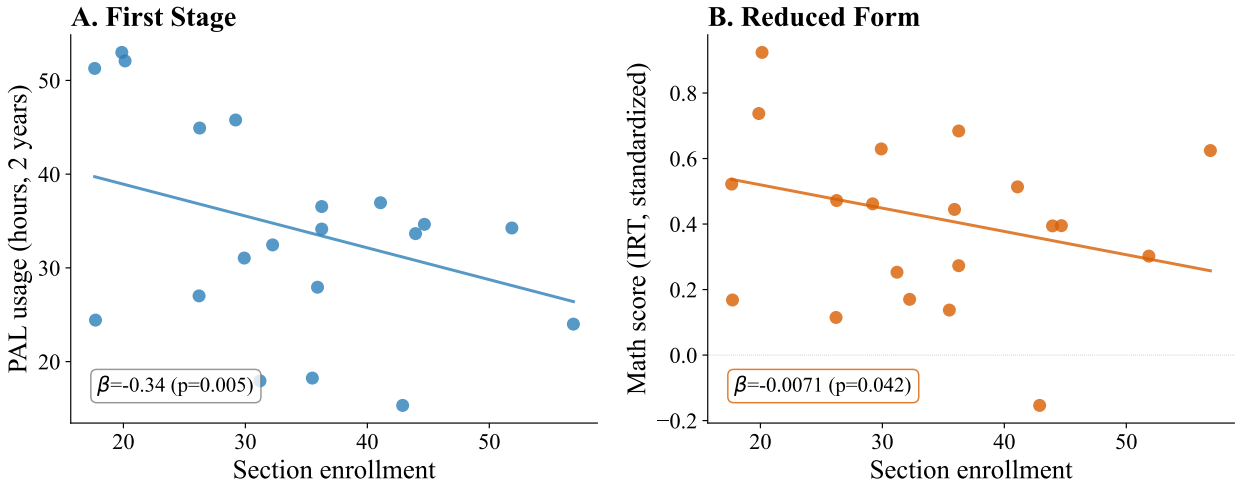
APPENDIX FIGURE 2. CONSORT FLOW DIAGRAM

Notes: Government schools with grades 6–10 in Andhra Pradesh (5,298 schools across 26 districts). Eight districts were selected as study sites and stratified into 21 geographic strata based on mandal clusters within each district. From this pool, 120 schools were randomly selected within strata and assigned to treatment or control. In treatment schools, PAL (Personalized Adaptive Learning) replaced 2 of 8 weekly math periods with tablet-based adaptive learning for 2 academic years (2023–24 and 2024–25). Enrollment counts are from government administrative records (AY 2024–25).



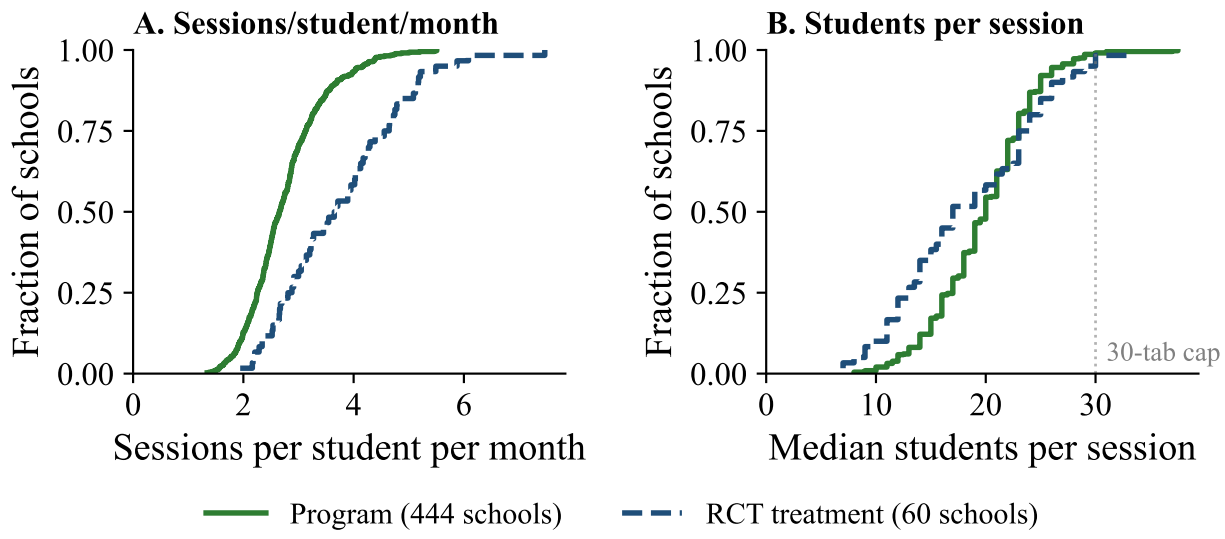
APPENDIX FIGURE 3. MONTHLY PAL USAGE AND CONTENT PROGRESSION

Notes: Top panel: total PAL usage hours per month across all treatment schools (green = AY 2023–24, blue = AY 2024–25). Usage ramps up from September, peaks in December–February, and drops during summer break (May–June). Bottom panel: average content grade gap (student grade minus content grade being studied). The system starts students approximately 1.3 grade levels below their enrolled grade, progressively moves them toward grade-level content, then resets at the start of the new academic year. Data from ConveGenius platform logs for 60 treatment schools.



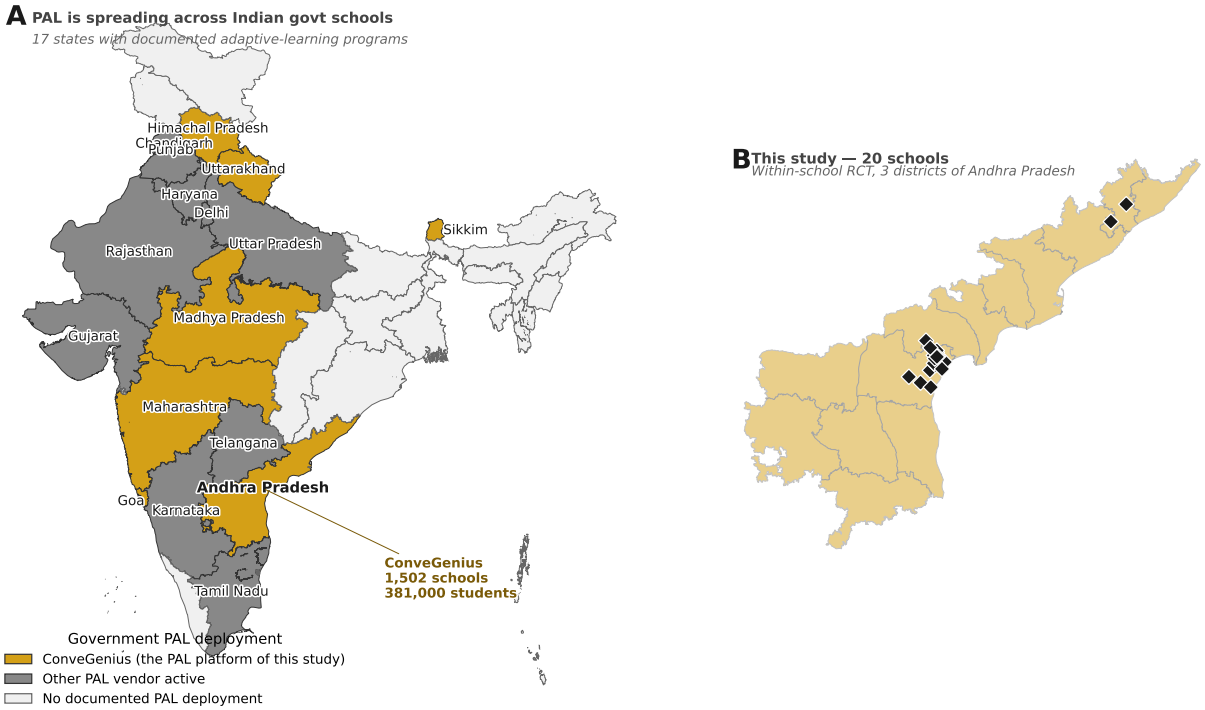
APPENDIX FIGURE 4. IV DIAGNOSTICS: FIRST STAGE AND REDUCED FORM

Notes: Treatment students only. *Panel A*: bin means of PAL usage hours by section enrollment, with a fitted regression line conditional on controls, grade FE, and strata FE (school-clustered SE). Each additional student in a section reduces PAL usage by 0.34 hours ($p = 0.005$). *Panel B*: bin means of endline math scores by section enrollment, same covariate structure. Each additional student reduces scores by 0.007 SD ($p = 0.042$). Bin means are formed within grade.



APPENDIX FIGURE 5. SESSION-LEVEL COMPARISON: PROGRAM VS. RCT TREATMENT SCHOOLS

Notes: ECDFs derived from day-level ConveGenius usage data, both arms in AY 2024–25. A session is defined as a (school × grade × day) combination with at least 5 active math students. *Panel A*: ECDFs of sessions attended per student per month, computed as school-level means across 444 program schools (green solid) and 60 RCT treatment schools (blue dashed). Program median = 2.7, RCT median = 3.7. *Panel B*: ECDFs of median students per session at the school level. Program median = 20, RCT median = 17; session sizes are essentially equivalent across arms. Together the two panels show the dose gap is driven by session frequency rather than by batch size.



Source: Authors' compilation, April 2026, from PAL Works (palworks.in), ConveGenius, Educational Initiatives (Mindspark/Ei Shiksha), and ~30 other public sources.

APPENDIX FIGURE 6. PAL ADOPTION ACROSS INDIAN GOVERNMENT SCHOOLS, WITH THE 20 REMEDIATION-STUDY SCHOOLS

Notes: **Panel A:** Indian states colour-coded by PAL deployment status. Gold = states where ConveGenius (the platform behind this study) operates: Andhra Pradesh, Himachal Pradesh, Madhya Pradesh, Maharashtra, Goa, Uttarakhand, Sikkim. Dark grey = states where another adaptive-learning vendor is documented in government schools (Mindspark/Educational Initiatives, Khan Academy India, iDream Education, Schoolnet/Geneo, Chimple, Eduten). Light grey = no documented government PAL deployment. The annotation for Andhra Pradesh shows the scale of the underlying programme (1,502 schools, 381,000 students; *The Week*, January 2026). **Panel B:** The 20 schools of the within-school remediation RCT (Section V), distributed across 3 districts of Andhra Pradesh (Bapatla, Prakasam, Vizianagaram); randomisation to Deep Remediation / Status Quo arms is at the student level within each school. State-level PAL coverage data compiled April 2026 from public sources. The Avanti Fellows / Pragati / Gurukul programme is excluded as a JEE/NEET test-prep platform with light personalisation, not adaptive learning in the strict PAL sense.

APPENDIX TABLE 1—PAL DOSAGE: TARGET VS. ACTUAL USAGE

	Year 1 (AY 2023–24)	Year 2 (AY 2024–25)	Total (2-year students)
School calendar			
Calendar weeks	31	33	64
Non-instructional weeks	9	12	21
PAL-eligible weeks	22	21	43
Target usage (hours)			
Scheduled (80 min/week)	29.3	28.0	57.3
Effective (60 min/week)	22.0	21.0	43.0
Actual usage (hours, treatment group mean)			
All students	14.7	19.5	34.2
2-year students (G7–9)	19.0	19.3	38.2
% of effective target			
2-year students (G7–9)	86%	92%	89%

Notes: Each school is allocated one 80-minute PAL period per week. The effective target of 60 minutes accounts for student transit to the computer lab and device setup time. Non-instructional weeks include holidays, formative assessment (FA) weeks, summative assessment (SA) weeks, and pre-exam revision weeks, during which PAL sessions are suspended. Year 1 ran late September 2023–April 2024; lab setup was staggered, with 49 schools starting in September and 11 in October. Year 2 ran July 2024–February 2025 (all 60 schools; capped at endline assessment start). Two-year students are grades 7–9 at endline, who had PAL in both academic years.

Appendix Tables

This section collects tables cited in the main text that are not placed in the body of the paper. They appear in the order in which they are first cited. Tables that belong to a specific appendix are placed in that appendix and carry its letter.

APPENDIX TABLE 2—PREDICTORS OF PAL USAGE (TREATMENT GROUP ONLY)

	Bivariate (1)	Joint (2)
Female	1.95***	1.27*
	(0.74)	(0.76)
Age	2.81***	3.03***
	(0.43)	(0.43)
SES index	-0.28	-0.30
	(0.28)	(0.28)
Baseline ability	1.66	1.36
	(1.41)	(1.37)
Enrollment (G6–9)	-0.06***	-0.07***
	(0.01)	(0.01)
Functional tablets	-0.08***	0.00
	(0.02)	(0.02)
R^2		0.273
Observations		6,742

Notes: Dependent variable: total PAL usage hours (2 years). Treatment group only. Column (1): bivariate regressions (one per row). Column (2): joint regression with all predictors. All regressions include strata FE and school-clustered SE. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

APPENDIX TABLE 3—EFFECTS ON GOVERNMENT MATH ASSESSMENT SCORES

	Year 1 (AY 2023–24)		Year 2 (AY 2024–25)	
	SA1	SA2	SA1	SA2
	(1)	(2)	(3)	(4)
Grade 6			0.341***	0.268*
			(0.113)	(0.138)
Grade 7	0.138	0.148	0.181*	-0.013
	(0.100)	(0.119)	(0.098)	(0.130)
Grade 8	0.238***		0.364***	0.092
	(0.082)		(0.115)	(0.153)
Grade 9	0.025	0.131	0.118	0.018
	(0.087)	(0.104)	(0.098)	(0.101)
Grade 10†		-0.108		-0.136**
		(0.095)		(0.067)
Controls	Yes	Yes	Yes	Yes
Strata FE	Yes	Yes	Yes	Yes

Notes: Dependent variable: government math assessment score, standardized within grade to control mean and SD. SA1 and SA2 are mid-year and end-of-year summative assessments administered and graded by schools. Grades 7–9 at endline had PAL for both years; grade 6 had PAL in Year 2 only (new cohort entering AY 2024–25); grade 10 had PAL in Year 1 only (as grade 9 in AY 2023–24, then aged out). †Grade 10 Year 2 column shows the state board exam (externally graded), not SA2. Grade 10 regressions include strata FE only (no student-level controls available). Cells left blank where treatment or control coverage is below 500 students. All other regressions include demographic controls, strata FE, and school-clustered standard errors. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

APPENDIX TABLE 4—IV ESTIMATES BY GRADE AND GENDER

	By Grade					By Gender	
	Grade 6 (1)	Grade 7 (2)	Grade 8 (3)	Grade 9 (4)	Pooled (5)	Boys (6)	Girls (7)
Panel A: First Stage (Dep. var.: PAL usage hours)							
Treatment × Section enrollment	-0.359***	-0.518**	-0.581***	-0.405***	-0.370***	-0.411***	-0.315***
	(0.106)	(0.247)	(0.158)	(0.135)	(0.110)	(0.133)	(0.101)
First-stage F	11.5	4.4	13.5	9.0	11.2	9.5	9.7
Panel B: Reduced Form (Dep. var.: Math score)							
Treatment × Section enrollment	-0.008	-0.006	-0.011**	-0.009**	-0.010***	-0.012***	-0.008*
	(0.006)	(0.007)	(0.006)	(0.005)	(0.004)	(0.004)	(0.004)
Panel C: IV Estimates							
IV (2SLS) estimate	0.0180	0.0125	0.0084	0.0088	0.0110	0.0096	0.0124
OLS estimate	0.016***	0.010***	0.006***	0.011***	0.009***	0.009***	0.009***
	(0.006)	(0.002)	(0.001)	(0.002)	(0.002)	(0.002)	(0.002)
Panel D: Exclusion Restriction (Control group only)							
Section enrollment	-0.000	0.006**	-0.002	-0.000	0.001	0.005*	-0.003
	(0.003)	(0.003)	(0.003)	(0.003)	(0.002)	(0.002)	(0.002)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Strata FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Grade FE	—	—	—	—	Yes	Yes	Yes
Mean usage, T group (hrs)					34.4	33.6	35.4
Observations	3,282	3,328	3,607	3,906	14,123	7,232	6,842

Notes: Each column replicates the IV analysis from Table 4 within a single grade (columns 1–4), pooled (column 5), or by gender (columns 6–7). The instrument is treatment × section enrollment: treatment schools received 30 tablets regardless of enrollment, so larger sections share more students per tablet. Panel A: first stage (section enrollment reduces PAL usage hours). Panel B: reduced form (section enrollment reduces math scores in treatment schools). Panel C: 2SLS and OLS estimates of the return per PAL hour. The 2SLS estimate is stable across grades (0.013–0.023 SD/hour) and similar for boys and girls, indicating the grade gradient in treatment effects (Table 3) reflects differences in usage quantity, not in the per-hour productivity of PAL. The gender gap likewise reflects quantity: girls use PAL ≈2 hours more than boys, but the per-hour return is similar. Panel D: exclusion restriction — section enrollment should not affect control student scores. This holds in all grades except grade 7 ($p = 0.01$). Grade 7 also has a weak first stage ($F = 4.4$); the IV estimate for grade 7 should be interpreted with caution. Grade 6 students had 1 year of PAL; grades 7–9 had 2 years. All regressions include controls and strata FE; pooled and gender columns also include grade FE. Standard errors clustered at the school level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

APPENDIX TABLE 5—CROSS-SAMPLE WITHIN-SCHOOL DOSE-RESPONSE ON CUMULATIVE PAL HOURS

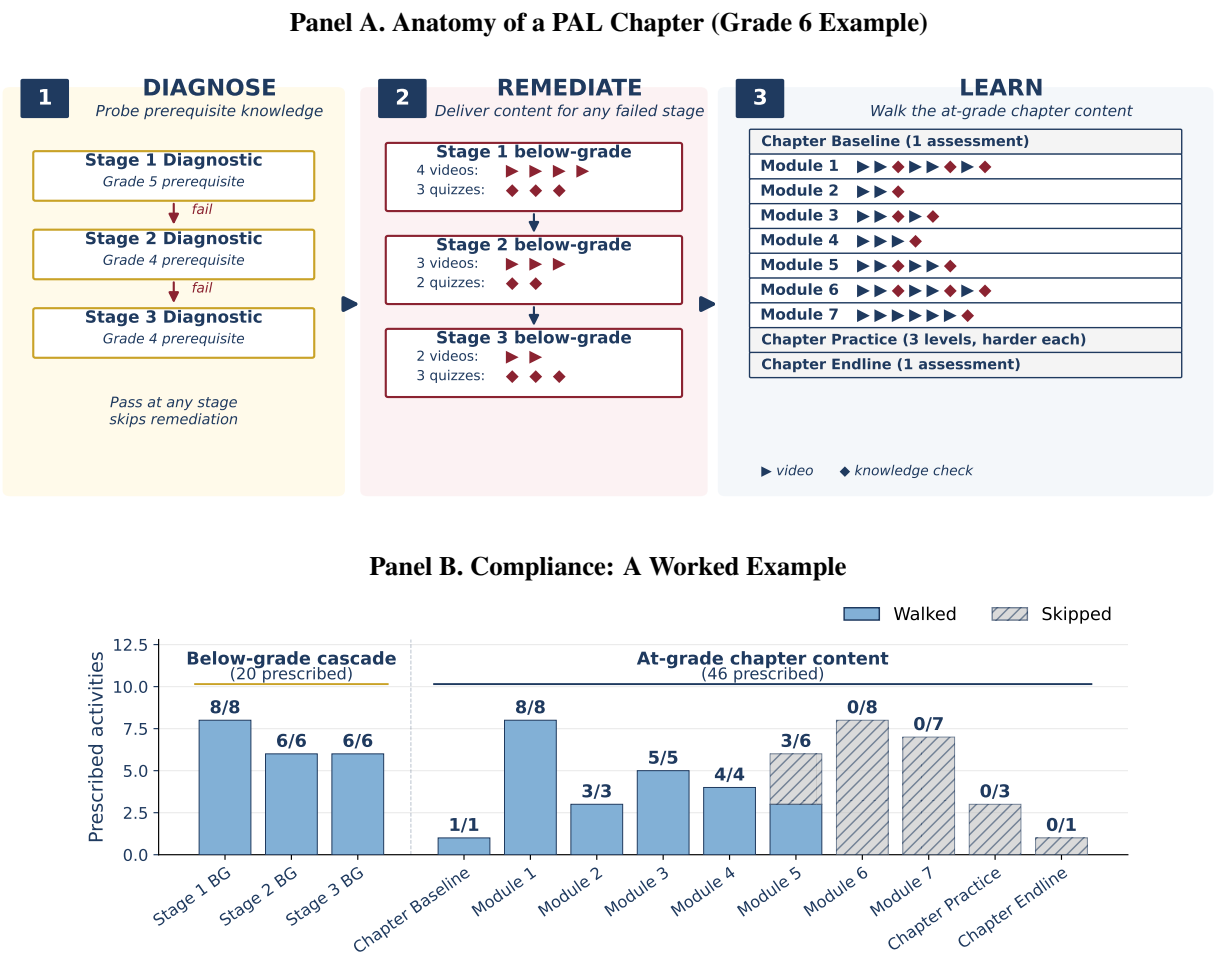
	RCT-T (3-year)		Targeted residential (∼14-month)		Program (3-year)	
	Linear (1)	Quadratic (2)	Linear (3)	Quadratic (4)	Linear (5)	Quadratic (6)
PAL hours	+0.0105***	+0.0155***	+0.0155***	+0.0270***	+0.0249***	+0.0405***
	(0.0012)	(0.0020)	(0.0032)	(0.0055)	(0.0030)	(0.0058)
PAL hours^2		-0.00005***		-0.00033**		-0.00055***
		(0.00001)		(0.00014)		(0.00016)
Students	2,967	2,967	5,406	5,406	5,957	5,957
Schools	40	40	40	40	40	40
Mean hours	24.3	24.3	13.6	13.6	6.6	6.6
Median hours	16.5	16.5	12.8	12.8	3.2	3.2
School FE	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Within-school OLS of Y3 endline IRT θ (linked to the Y2 RCT-control scale via Mean–Sigma on 45 anchor items) on cumulative PAL math hours. Each sample's exposure window matches its actual PAL exposure: RCT-T is 3-year cumul (AY 2023–24 + 2024–25 + 2025–26); Targeted residential is cumul since the residential rollout's January 2025 launch (∼14 months through endline); program is 3-year cumul (AY 2023–26), with Y1 student-level hours recovered via SHA-256 hash linkage to the cleartext Y3 student_code. Per-hour slopes are descriptive within-school and not IV-identified; the RCT-T 3yr linear slope of +0.0105 is close to the fixed-intercept IV slope of +0.011, identified causally on the RCT 2-year sample. Standard errors (in parentheses) clustered at the school level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Appendix A. Detailed description of the PAL program and its internal logic

When a student opens a new chapter, the platform delivers content in three phases (Appendix Figure A1, Panel A): up to three short diagnostic stages testing prerequisite content from earlier grades; below-grade remediation for any diagnostic stage the student fails; and the at-grade chapter content shared by every student. The diagnostic stages—each two to five questions—are completed before any chapter content begins. The first stage tests material one grade level below the student's enrolled grade; depending on the chapter's prerequisite depth, the platform then runs one or two additional stages testing material two grade levels below. The stages a student fails (scores below 80 percent on) determine the remedial content the platform delivers: pass every stage and the student moves directly to the at-grade content; fail the one-grade-below stage and a one-grade-below remedial module is inserted first; fail any two-grade-below stage and two-grade-below remediation is added on top.[13] Across 121,000 student-chapter attempts in our 60 treatment schools, approximately 3 percent of students pass the first stage and skip remedial content; the remaining 97 percent receive one or two grade levels of remediation. This per-chapter routing is student-specific: different diagnostic outcomes deliver different bundles of activities to different students.

[13] In the platform's reference chapter (Basic Geometric Concepts, Grade 6), which has two diagnostic stages, this routing assigns one of three bundles: 54 activities (at-grade only), 61 (adding one-grade-below remediation), or 67 (adding both grade levels of remediation). Each "activity" is either a short instructional video—a *learning objective*, or LO—or a brief multiple-choice quiz testing the videos in a module (a *knowledge check*, or KC).



APPENDIX FIGURE A1. THE PAL CHAPTER AND STUDENT COMPLIANCE

Notes: Panel A: the chapter unfolds in three phases. *Phase 1 (Diagnose):* up to three short diagnostic stages testing prerequisite content from earlier grades, all completed before any chapter content begins. *Phase 2 (Remediate):* below-grade videos and knowledge checks for any failed diagnostic stage. *Phase 3 (Learn):* the at-grade chapter content (a chapter baseline assessment; modules each containing instructional videos and knowledge checks; chapter practice; a chapter endline assessment), identical for every student. *Module* = a topic block within a chapter; *learning objective* (LO) = one short instructional video; *knowledge check* (KC) = a quiz testing the LOs in a module. The figure depicts a Grade 6 chapter with three diagnostic stages and seven at-grade modules; chapters with shallower prerequisites have fewer diagnostic stages. *Panel B:* stylized worked example showing compliance for a single student in a single chapter. Compliance = activities walked ÷ activities prescribed. The illustrative student walks the full below-grade cascade (all three diagnostic-failed stages, with their associated remedial videos and quizzes) and Modules 1–4 of the at-grade content, gets halfway through Module 5, then stops. Activities walked = 44; activities prescribed = 66; compliance ≈ 0.67.

APPENDIX TABLE B1—TIME DIVISION BY MODULE TYPE
(TREATMENT GROUP ONLY)

	Overall	Grade 6	Grade 7	Grade 8	Grade 9
Total hours	34.2	20.4	39.4	39.7	35.9
Video hours	18.9	11.2	21.5	21.6	20.6
Question hours	15.3	9.2	17.9	18.1	15.3
% Video	53.6%	51.7%	52.6%	52.8%	56.9%
N	6,742	1,540	1,599	1,769	1,834

Notes: Treatment group only. Total hours is cumulative PAL usage across 2 academic years. Video hours = passive content viewing. Question hours = active practice (answering questions). Grade 6 had 1 year of PAL; grades 7–9 had 2 years.

Appendix B. Implementation: Tablets, Effort, and the Chapter Pipeline

This appendix documents how PAL is actually delivered—the split between active practice and passive viewing, student-reported implementation challenges, and the standardized six-stage module sequence that proceeds uniformly across ability levels.

B.A. State Monitoring Infrastructure

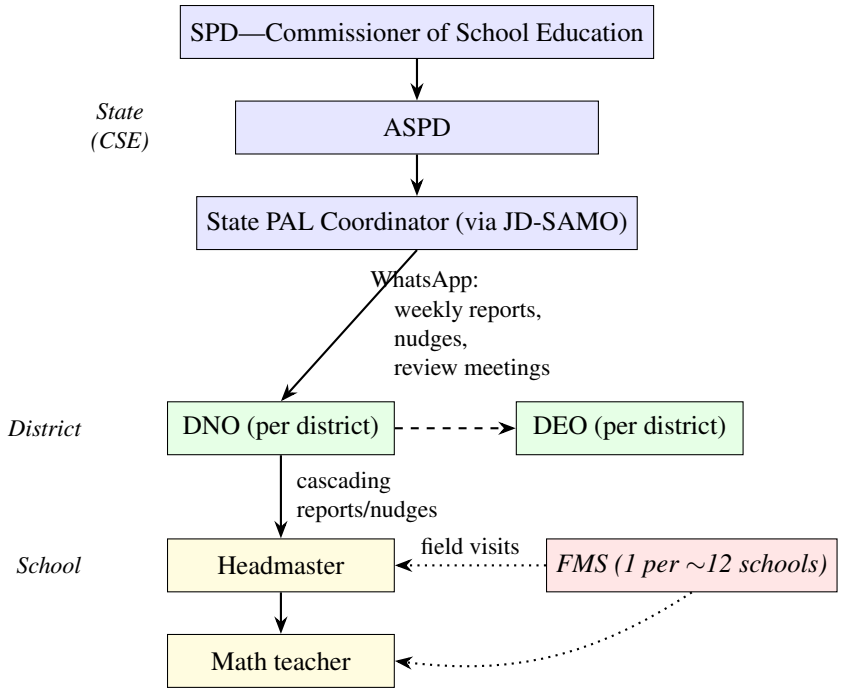
Each treatment school sat at the bottom of a four-tier chain that determined whether the PAL lab actually ran each week. At the top, the State Project Director (SPD) and Additional SPD at the Commissionerate of School Education set policy; under them, a State PAL Coordinator handled day-to-day operational traffic. Each district had a District Nodal Officer (DNO) for PAL, who cascaded state instructions to the schools in their district and coordinated laterally with the District Educational Officer. At the school, the headmaster oversaw delivery and the math teacher ran the lab sessions. We depict this system graphically in Appendix Figure B1.

The chain operated through three concrete channels. First, a WhatsApp broadcast group from the State PAL Coordinator to all DNOs shared weekly school- and district-level usage rankings, issued nudges to increase usage, and coordinated operational pushes such as data syncing, orientation sessions, and tablet maintenance. Second, scheduled review meetings: state-coordinator-led district-by-district sessions with headmasters, math teachers, and DNOs in the program's first weeks; in-person summons of low-performing DNOs to the Commissionerate; and SPD-led review meetings with the broader DNO group. Third, real-time monitoring dashboards reporting per-school usage metrics, with the underlying telemetry coming from the ConveGenius platform itself. All 28 RCT treatment headmasters in our Year-3 survey report receiving these state communications periodically, and 71 percent report having seen the school-level usage report card.

B.B. PAL Instructional Pipeline

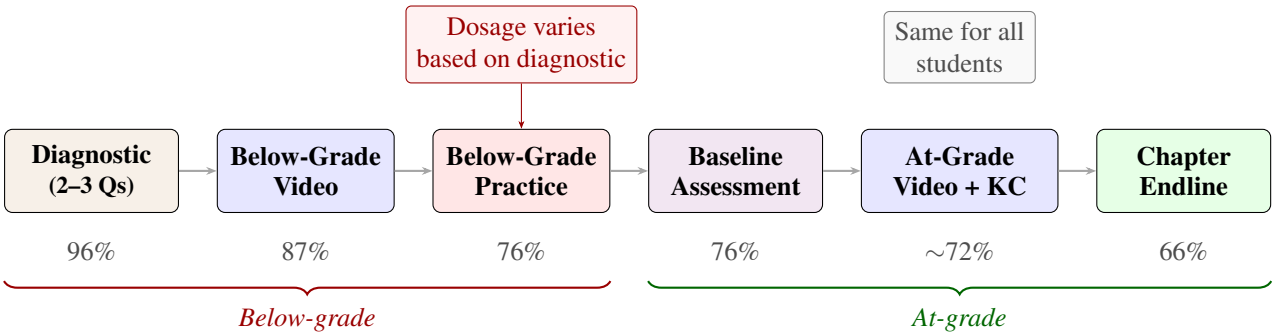
Using module-level platform data (8.78 million learning events across 12,019 students in the 60 treatment schools), we reconstruct the instructional pipeline that students experience within each chapter. Figure B2 shows the standardized six-stage sequence: a short diagnostic on prerequisite content, below-grade remediation (videos and practice), a baseline assessment at grade level, at-grade instruction with formative knowledge checks, and a chapter endline. The chapter sequence within each grade is fixed and identical for all students.

Figure B3 shows the content grade progression over time, tracking five cohorts defined by their enrolled grades across the two study years. All cohorts begin approximately 1.4 grade levels below their enrolled grade and converge toward grade-level material over roughly four months. At the start of Year 2, the system



APPENDIX FIGURE B1. THE SCHOOL–DISTRICT–STATE ESCALATION CHAIN FOR AP PAL

Notes: Solid arrows: direct line of authority. Dashed: lateral district-level coordination. Dotted: technical-support contractor interface. SPD = State Project Director. ASPD = Additional SPD. JD-SAMO = Joint Director, State Academic Monitoring. DNO = District Nodal Officer. DEO = District Educational Officer. HM = Headmaster. FMS = Field Management Staff (contracted through ConveGenius). Funded through state and central (Samagra Shiksha) education budgets and operated by the AP government's regular school administration.



APPENDIX FIGURE B2. PAL CHAPTER PIPELINE

Notes: Standardized chapter pipeline based on 246,829 chapter journeys across 12,019 students. Percentages indicate the fraction of students reaching each stage. The below-grade practice stage (highlighted) is the primary locus of dosage variation; the at-grade stages are uniform across students.

resets continuing cohorts back to below-grade content regardless of Year 1 progress. The Grade 9→10 cohort appears only in Year 1 (students aged out after grade 9); the new Grade 6 cohort appears only in Year 2.

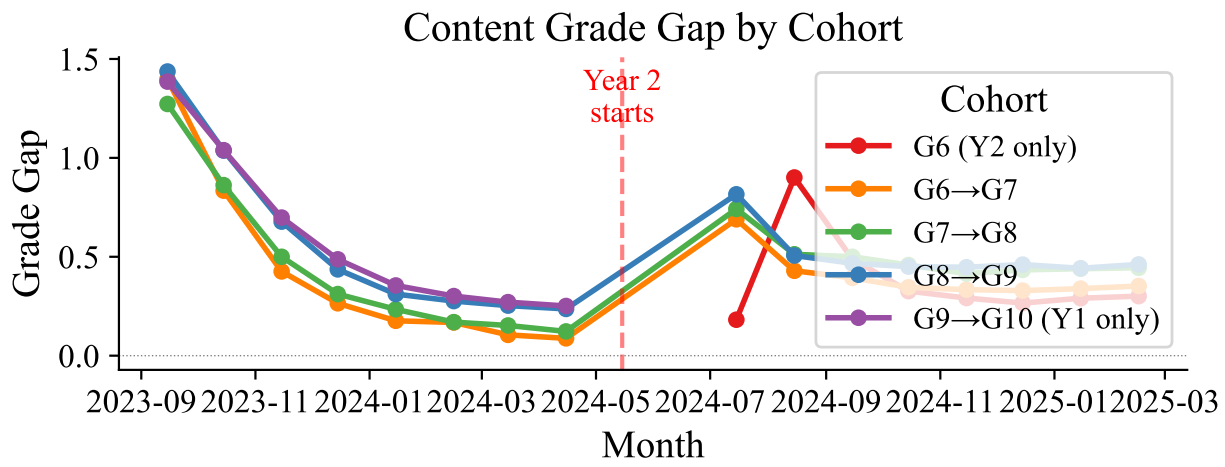
Table B3 shows that the student experience is largely uniform across the baseline ability distribution. Total usage hours, the video–practice split, and the share of below-grade content are all flat across baseline ability quintiles. The Q5–Q1 gap in video share is 3 percentage points; in below-grade content share, 4 percentage points. Figure B4 provides the most direct evidence: the content grade progression for the bottom and top terciles of baseline ability overlaps almost completely in both years. The system moves all students

APPENDIX TABLE B2—PAL EXPERIENCE (TREATMENT GROUP ONLY)

Panel A: Student experience (% reporting)	
Enjoy PAL lab "a lot"	80.3%
PAL helped understand math "a lot"	73.2%
Like watching PAL videos	87.5%
Prefer additional PAL period over math	66.3%
PAL less difficult than regular math	57.5%
Prefer English in PAL lab	52.1%
Always log in with own credentials	74.3%
Always get teacher help to log in	34.4%
PAL classes per week, mean (SD)	2.6 (1.4)
Panel B: Implementation quality (% reporting no problem)	
Video quality is good	88.4%
Knows how to use PAL	87.1%
Enough time allotted	84.1%
Enough tablets available	78.5%
Screen easy to see	77.3%
Lab not too noisy	74.5%
Headphones work	73.8%
No login problems	70.7%
Videos load fine	63.4%
Language understandable	62.8%

Notes: Treatment group only ($N \approx 6{,}740$). Panel A: percentage of students giving the top/positive response to each PAL experience question from the endline student survey (tablet-based, administered by independent enumerators). Panel B: percentage reporting no problem for each implementation dimension (binary: 1=good/no problem, 0=has problem). The most common implementation challenges are language comprehension (37% report problems) and video loading speed (37%).

through the same content at the same speed.



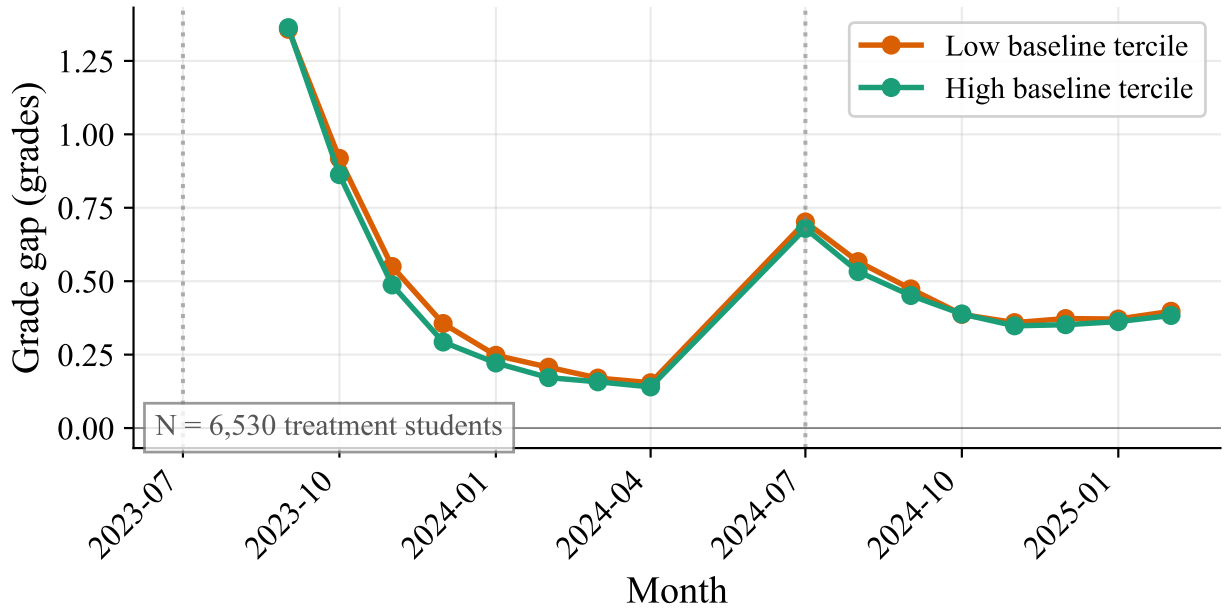
APPENDIX FIGURE B3. CONTENT GRADE GAP BY COHORT

Notes: Grade gap = student's enrolled grade − average content grade of modules used that month. Each line tracks a cohort across its years of PAL exposure: three cohorts (G6→G7, G7→G8, G8→G9) span both academic years; the G9→G10 cohort aged out after Year 1; the Grade 6 cohort entered in Year 2 only. $N = 10{,}920$ students across 60 treatment schools (excluding repeaters and mid-year dropouts).

APPENDIX TABLE B3—PAL USAGE BY BASELINE ABILITY

Baseline quintile	Total usage (hours)	Video (hours)	Practice (hours)	Video share (%)	Below-grade content (%)
Q1	33.9	19.0	14.9	57	33
Q2	33.9	19.1	14.8	56	32
Q3	32.1	18.0	14.1	56	32
Q4	35.3	19.4	15.9	55	30
Q5	35.6	19.0	16.6	54	29
All	34.2	18.9	15.3	55	31
Q5 − Q1	+1.8			-3	-4

Notes: Treatment students with non-missing baseline scores and PAL usage data ($N = 6{,}742$). Total usage is the sum of video and practice hours over two academic years (2023–24 and 2024–25). Below-grade content share is the average fraction of modules within each chapter journey that cover content below the student's enrolled grade.



APPENDIX FIGURE B4. CONTENT GRADE PROGRESSION BY BASELINE ABILITY

Notes: Grade gap = student's enrolled grade − average content grade used that month. Bottom and top terciles defined by within-school × grade percentile rank of government math score. Lines overlap almost completely in both years, indicating identical progression trajectories regardless of actual ability. $N = 12{,}019$ students across 60 treatment schools.

Appendix C. Heterogeneity: Gender, Baseline Ability, and Data Coverage

In this appendix, we present results from the heterogeneity analysis of our ITT estimates in the cluster RCT.

C.A. Heterogeneous Treatment Effects

Table C1 examines treatment-effect heterogeneity along several dimensions.

Gender. Girls gain significantly more than boys (0.46 vs. 0.33 SD; $p = 0.001$). This interaction survives Bonferroni, Holm, and Benjamini–Hochberg corrections (Appendix Table D3). The advantage is not explained by pre-existing differences in baseline ability or attendance, which are balanced across gender within treatment status. Rather, girls accumulate approximately two more hours of PAL usage than boys, driven by more time on active practice (questions) rather than passive content (videos; Appendix Figure C1). A gender-stratified IV analysis confirms that boys and girls have similar per-hour returns (Appendix Table 4), indicating the gap reflects usage quantity rather than the productivity of each hour. Appendix Table C2 provides the full regression results for this diagnosis.

Baseline math scores. Treatment effects decline with baseline achievement: 0.48 SD for the lowest quintile and 0.31 SD for the highest ($p = 0.08$ for Q1 vs. Q5). The program benefits students across the entire baseline range, but the lowest-baseline students gain the most. A conditional quantile analysis (Appendix Figure L9) shows that within each baseline quintile, effects increase from the 10th to the 75th percentile of the endline distribution before flattening—the program creates within-group variance in outcomes in addition to the across-quintile gradient. The large treatment effect for the "missing baseline" subgroup (0.50 SD) reflects a data-extraction artifact rather than a true subgroup effect: the government provided broader baseline data for treatment schools than for control schools in grades 6–7, creating compositional imbalance in this subgroup (Appendix Tables C3–C4 and Figure C2).

SES and school characteristics. Effects are uniform across SES terciles ($p = 0.42$ for Low vs. High; Appendix Table C5). Students in small sections (at or below 33 students, the approximate tablet supply) gain 0.46 SD compared with 0.34 SD in large sections ($p = 0.05$). Rural and urban schools show similar effects. The heterogeneity by section size foreshadows the congestion mechanism in Section IV.C. Appendix Figure C3 summarizes heterogeneity across all four dimensions visually.

Topic-level effects. Treatment effects are uniform across all five math topics (0.31–0.38 SD; Appendix Table L5 and Figure L5), indicating that the program improves math skills broadly rather than producing topic-specific gains. Effects are slightly larger on harder items (Appendix Figure L6), suggesting the program helps students reach questions they would otherwise get wrong. Treatment students outperform control students at every content grade level, with the largest gains on grade 4–6 content (Appendix Figure L7). The program also increases math enjoyment by 0.19 SD, with positive spillovers to other subjects, and reduces self-reported absences by 0.22 days over two weeks (Appendix Table L4).

C.B. Exploring the reasons behind gender heterogeneity

This appendix resolves why girls gain more than boys (usage quantity, not per-hour productivity), documents the across-quintile gradient in baseline ability, and rules out that the gradient is a data-coverage artefact of asymmetric baseline extracts.

We examine treatment effect heterogeneity along several dimensions. Girls gain more than boys (+0.46 vs +0.33 SD), and this advantage persists across all grades (Figure L1). Effects are uniform across SES terciles

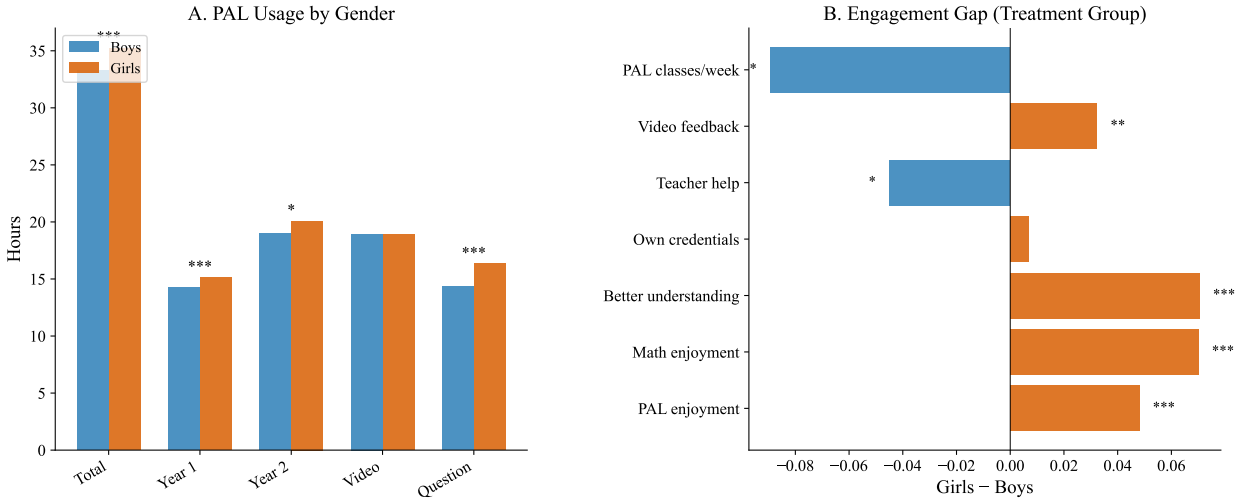
APPENDIX TABLE C1—HETEROGENEOUS TREATMENT EFFECTS

	Treatment	(SE)	Control mean	*N*
Pooled	0.390***	(0.054)	0.000	14,215
Panel A: Gender				
Boys	0.329***	(0.056)	-0.063	7,289
Girls	0.464***	(0.059)	0.068	6,877
p (Boys = Girls)		0.001		
Panel B: Baseline Achievement				
Q1 (lowest)	0.483***	(0.071)	-0.410	2,483
Q2	0.353***	(0.063)	-0.207	1,897
Q3	0.330***	(0.075)	0.057	1,969
Q4	0.339***	(0.079)	0.341	1,936
Q5 (highest)	0.305***	(0.086)	0.929	2,020
Missing baseline	0.498***	(0.077)	-0.155	3,910
p (Q1 = Q5)		0.083		
p (joint)		0.126		
Panel C: SES				
Low SES (T1)	0.350***	(0.059)	-0.057	4,746
Mid SES (T2)	0.422***	(0.066)	0.039	5,182
High SES (T3)	0.381***	(0.060)	0.018	4,287
p (Low = High)		0.418		
Panel D: School Characteristics				
Small section ($\leq$33)	0.463***	(0.059)	-0.014	7,220
Large section ($>$33)	0.336***	(0.075)	0.014	6,995
p (Small = Large)		0.050		
Rural	0.456***	(0.057)	-0.018	11,615
Urban	0.520***	(0.151)	0.095	2,600
p (Rural = Urban)		0.075		

Notes: Dependent variable: IRT math ability parameter standardized to control group mean and SD. Each row reports the treatment coefficient from a separate OLS regression on the indicated subgroup, with 21 strata fixed effects, grade fixed effects, and controls (SES, parental education, age, gender, missing flags). Standard errors clustered at the school level in parentheses. Baseline achievement quintiles are within school $\times$ grade percentile ranks of government math SA1 (AY 2022–23). "Missing baseline" includes students without baseline data (28.8%). Baseline coverage differs by treatment arm for grades 6–7 due to a government data extraction artifact (Appendix Table C3); the treatment effect for this subgroup is not interpretable as a clean subgroup estimate. SES terciles are based on a PCA index of 9 household assets. Section enrollment is the number of students in the student's class section. *p*-values test equality of treatment effects between the indicated groups (from pooled interaction models). * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

($p = 0.62$; Table C5) but decline across baseline ability quintiles, with the largest gains for the lowest-baseline students (Figure C3, Panel C). Girls accumulate more PAL usage hours and report higher engagement than boys (Figure C1); Table C2 provides the full regression results for the gender gap diagnosis.

Baseline ability in our regressions is the student's within-school $\times$ grade percentile rank on the pre-PAL



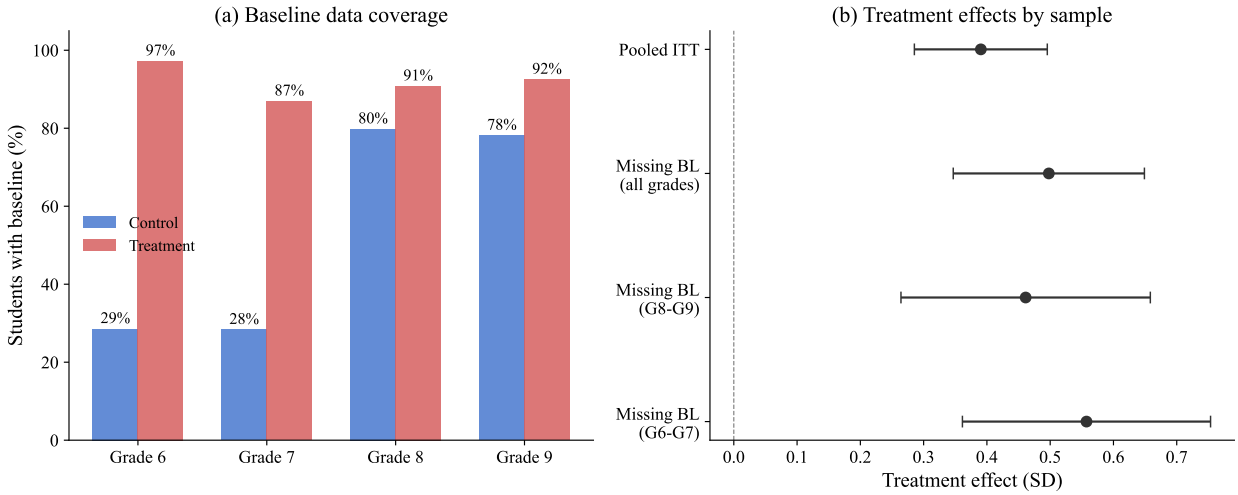
APPENDIX FIGURE C1. GENDER DIFFERENCES IN PAL USAGE AND ENGAGEMENT

Notes: Panel A: Mean PAL usage hours by gender (treatment group only). Girls accumulate more total hours than boys, driven primarily by more time on questions (active practice) rather than videos. *Panel B*: Gender gap in engagement measures (girls − boys, treatment group). Girls report higher PAL enjoyment, math enjoyment, and perceived understanding. Significance stars from regressions with strata FE and school-clustered SE.

government math SA1 assessment, with a cohort-specific source: grade 7–9 at endline use AY 2022–23 SA1, while grade 6 at endline (new cohort, PAL exposure only in 2024–25) use AY 2023–24 SA1—their grade-5 pre-PAL year. Using 2022–23 for grade 6 would measure them as grade 4 students two years pre-endline.

The large treatment effect for the "Missing baseline" subgroup in Table C1 (0.50 SD) reflects overlapping data-coverage issues, not a true subgroup effect, and these issues differ by cohort. For the 2-year cohort (grades 7–9 at endline), missingness reflects a *data-extraction asymmetry*: the AP government delivered AY 2022–23 data for the 524 program + treatment schools in one extract and the 60 control schools in another, with no overlap. In grade 7, this gives 87–89% coverage for treatment vs. 26–29% for control (Table C3, Panel A); in grades 8–9, coverage is balanced at 78–93% in both arms. For the 1-year cohort (grade 6 at endline), missingness is *structural and not closed by choice of baseline year*: grade 6 students are brand-new entrants to the study schools in 2024–25, so their pre-PAL records live in whatever government school they attended as grade 5. Treatment grade 6 students came from the 444 program schools (captured by the 524-school extract, ∼97% coverage), while control grade 6 students came from schools outside the 60-school control extract, yielding only ∼29% coverage. Switching the grade 6 baseline source from 2022–23 to 2023–24 slightly improves total coverage (55% → 61%) but does not close the treatment–control gap, because both years draw from the same asymmetric extracts.

A linear probability model confirms that missingness is predicted entirely by the treatment × grades 6–7 interaction and school category, not by student characteristics (Panel B). Restricting to grades 8–9—where coverage is balanced—the treatment effect among missing-baseline students is 0.46 SD; in grades 6–7, where coverage is asymmetric or structural, it is inflated to 0.56 SD (Table C4). The inflated pooled estimate reflects compositional imbalance in the missing-baseline subgroup, not a real effect on students without baseline.



APPENDIX FIGURE C2. BASELINE DATA COVERAGE AND TREATMENT EFFECT DECOMPOSITION

Notes: Panel (a) shows the percentage of students with a valid pre-PAL government baseline math score (AY 2022–23 SA1 for grades 7–9 at endline; AY 2023–24 SA1 for grade 6) by treatment status and endline grade. The grade 7 imbalance is a data-extraction artifact (treatment and control data delivered in separate extracts without overlap); the grade 6 imbalance is structural (new-cohort students' pre-study schools are not in either extract). Panel (b) shows treatment effects for the pooled sample and for missing-baseline students overall, in grades 8–9 (balanced coverage), and in grades 6–7 (asymmetric coverage). Whiskers show 95% confidence intervals.

APPENDIX TABLE C2—GENDER DIAGNOSIS: WHY DO GIRLS GAIN MORE FROM PAL?

	Boys	Girls	Difference	SE
Panel A: Ruling out alternative explanations				
Baseline ability (observed, within school×grade pctile)				
Control group	0.454	0.532		
Treatment group	0.456	0.508		
Gender TE gap	T×F interaction			
Without baseline control	+0.170***			(0.049)
With baseline control	+0.162***			(0.047)
Attendance (days present, last 2 weeks)				
Control group	10.2	10.4		
Treatment group	10.5	10.6		
Treatment × Female			-0.180**	(0.080)
School composition (T×F×Moderator)				
Enrollment (G6–9)	+0.079*			(0.046)
Urban	+0.015			(0.136)
% female in school	+0.009			(0.063)
Tablets/student	+0.043			(0.047)
	Boys	Girls	Difference	SE
Panel B: Usage and engagement (treatment group)				
PAL usage (hours)				
Total hours (2 years)	33.3	35.2	+2.0***	(0.7)
Year 1 hours	14.3	15.1	+1.2***	(0.5)
Year 2 hours	19.0	20.1	+0.7*	(0.4)
Video hours	18.9	18.9	-0.1	(0.4)
Question hours	14.3	16.3	+2.0***	(0.4)
Engagement (higher = better)				
PAL enjoyment	3.74	3.79	+0.05***	(0.02)
Math enjoyment	3.65	3.71	+0.07***	(0.02)
Better understanding	3.66	3.72	+0.07***	(0.02)
Own credentials	3.63	3.63	+0.01	(0.02)
Teacher help	3.04	2.97	-0.04*	(0.03)
Video feedback	2.85	2.88	+0.03**	(0.01)
PAL classes/week	2.59	2.51	-0.09*	(0.05)
Challenges (total)	7.32	7.91	+0.58***	(0.09)
	Boys TE	Girls TE	Interaction	SE
Panel C: Topic-specific treatment effects (standardized)				
Algebra	0.291	0.425	+0.160***	(0.047)
Number systems	0.309	0.434	+0.159***	(0.047)
Geometry	0.314	0.385	+0.101**	(0.046)
Measurement	0.341	0.438	+0.125**	(0.049)
Statistics	0.263	0.375	+0.145***	(0.050)

Notes: Exploratory analysis investigating mechanisms behind the gender gap in treatment effects (Table C1, Panel A). Panel A tests whether baseline ability, attendance, or school characteristics explain the gap. Panel B compares PAL usage and engagement by gender within the treatment group. Panel C tests whether the gender advantage is concentrated in specific topics. All regressions include strata FE; standard errors clustered at school level. No multiple testing correction. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Panel A: Baseline coverage by treatment status and grade

APPENDIX TABLE C3—BASELINE DATA AVAILABILITY: DIFFERENTIAL COVERAGE BY TREATMENT STATUS

	Control		Treatment	
	Has baseline	%	Has baseline	%
Grade 6	506/1,773	28.5%	1,497/1,540	97.2%
Grade 7	497/1,755	28.3%	1,390/1,599	86.9%
Grade 8	1,479/1,854	79.8%	1,606/1,769	90.8%
Grade 9	1,634/2,091	78.1%	1,696/1,834	92.5%

Panel B: Linear probability model — predictors of missing baseline

	Coefficient	(SE)
Treatment	-0.1243***	(0.0264)
Grades 6–7	0.5462***	(0.0428)
Treatment × Grades 6–7	-0.5054***	(0.0470)
Has primary section (Cat. 6)	-0.1290***	(0.0243)
Female	-0.0095	(0.0100)
SES index	0.0049	(0.0046)
Age	0.0204***	(0.0034)
Observations	14,215	
R^2	0.376	
Strata FE	Yes	

Notes: Panel A shows the number and percentage of students with 2022–23 government baseline math scores (SA1) by treatment status and endline grade. The differential coverage for grades 6–7 reflects a data extraction artifact: the government provided broader baseline data for treatment-eligible schools than for control schools (see text). Panel B reports coefficients from a linear probability model with missing baseline as the dependent variable. Standard errors clustered at the school level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

APPENDIX TABLE C4—TREATMENT EFFECTS BY BASELINE DATA AVAILABILITY AND GRADE

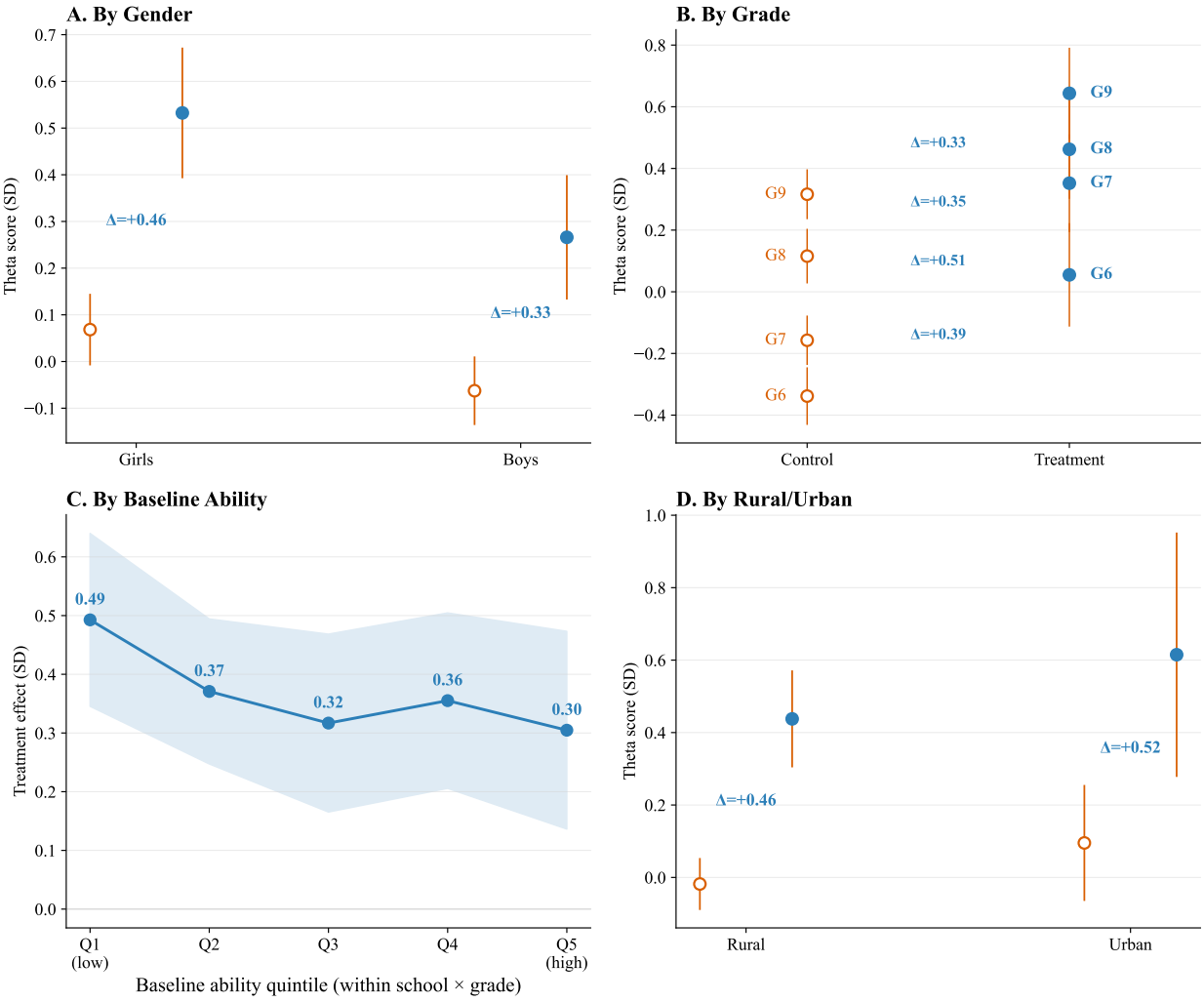
	(1) Full sample	(2) Has baseline	(3) G8–G9 only	(4) G8–G9 has baseline	(5) Missing BL G8–G9	(6) Missing BL G6–G7
Treatment	0.390***	0.361***	0.336***	0.313***	0.461***	0.557***
	(0.054)	(0.059)	(0.059)	(0.062)	(0.101)	(0.100)
Control mean	0.000	0.126	0.222	0.242	0.146	-0.254
N	14,215	10,305	7,548	6,415	1,133	2,777

Notes: Each column reports the ITT treatment effect from OLS of standardized math scores on treatment assignment with controls (SES, parental education, age, gender, missing flags), strata fixed effects, and grade fixed effects appropriate to the sample. Columns (1)–(2) include all grades with full and baseline-restricted samples. Columns (3)–(4) restrict to grades 8–9, where baseline coverage is balanced across arms (∼80–93% in both treatment and control). Columns (5)–(6) isolate students missing baseline data: column (5) in grades 8–9 (balanced missingness), column (6) in grades 6–7 (where control students are disproportionately missing due to the data extraction artifact). Standard errors clustered at the school level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

APPENDIX TABLE C5—HETEROGENEOUS TREATMENT EFFECTS BY SES TERCILE

	Low SES (1)	Middle SES (2)	High SES (3)
Treatment	0.350***	0.422***	0.381***
	(0.059)	(0.066)	(0.060)
Observations	4,746	5,182	4,287
p (equality)		0.619	

Notes: Each column estimates the ITT within a SES tercile. SES index is PCA PC1 of 9 asset indicators. All regressions include controls, strata FE, grade FE, and school-clustered SE. *p* (equality) tests that treatment effects are equal across terciles. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.



APPENDIX FIGURE C3. HETEROGENEOUS TREATMENT EFFECTS BY SUBGROUP

Notes: Open circles show control group means; filled circles show adjusted treatment group means (control mean + regression-adjusted treatment effect); whiskers show 95% school-clustered CIs. Δ annotations show the ITT estimate from regression with controls and strata FE. *Panel A*: Girls start slightly higher and gain more (+0.46 vs +0.33 SD). *Panel B*: Control scores rise with grade (older students know more) while treatment effects decline—consistent with the congestion mechanism. *Panel C*: Treatment effects (line with 95% CI band) decline across baseline ability quintiles (within-school × grade rank), with the largest gains for the lowest-baseline students. *Panel D*: Rural and urban schools show similar effects. Panels A, C, D include grade FE; Panel B estimates are within-grade.

APPENDIX TABLE D1—SENSITIVITY OF ITT TO CONTROL SPECIFICATION

	No controls	Demographics only	Demographics + baseline
	(1)	(2)	(3)
Treatment	0.379***	0.390***	0.393***
	(0.054)	(0.054)	(0.055)
R^2	0.106	0.137	0.238
Observations	14,215	14,215	14,215

Notes: Dependent variable: IRT math score standardized to control. Column (1): treatment + grade FE + strata FE only. Column (2): adds demographic controls (SES, parental education, age, gender). Column (3): adds within-school×grade baseline ability percentile (28.8% missing, imputed at median with missing flag). The treatment effect is stable across specifications. All regressions include strata FE and school-clustered SE. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Appendix D. Robustness of the Main ITT

This appendix confirms that the 0.39 SD headline estimate survives differential attrition, alternative control sets, multiple-testing corrections, non-parametric inference, and the congestion IV's exclusion restriction.

Table D2 presents attrition analysis and Lee (2009) bounds. Differential attrition is modest (3.3 percentage points, treatment higher) but statistically significant. Lee bounds remain large and positive (0.27–0.48 SD), ruling out attrition bias as an explanation for the main treatment effect.

We assess the sensitivity of the main ITT estimate to alternative specifications and inference procedures. The treatment effect is stable across three control specifications (0.38–0.39 SD; Table D1). The gender interaction survives Bonferroni, Holm, and Benjamini-Hochberg corrections (Table D3). Randomization inference, which does not rely on asymptotic assumptions, confirms the main result (RI $p < 0.001$; Figure D1).

Figure D2 visualises the exclusion-restriction placebo for the congestion IV. We plot raw within-grade binned means of endline math scores against section enrollment, separately for treatment and control. Conditioning on grade, controls, and strata FE, section enrollment has a negative effect on T scores ($\beta = -0.007$, $p = 0.07$) but is flat for C ($\beta = +0.001$, $p = 0.75$). The flat control line is direct visual evidence that section size does not affect learning in the absence of PAL, supporting the exclusion restriction.

APPENDIX TABLE D2—ENDLINE ATTRITION ANALYSIS AND LEE BOUNDS

	Pooled	Grade 6	Grade 7	Grade 8	Grade 9
Panel A: Attrition Rates					
Treatment	0.237	0.221	0.210	0.234	0.275
Control	0.198	0.186	0.201	0.185	0.216
Difference	0.033**	0.035	0.009	0.049	0.059
	(0.013)				
N (enrolled)	18,039	4,117	4,186	4,567	5,169

Panel B: Differential Attrition Predictors

	Covariate effect	Treatment × Covariate	*p* (interaction)	*N*
Female	-0.0405***	-0.0279**	0.042	18,039
Age	0.0233***	0.0120*	0.073	18,039
Urban	-0.0081	0.0902**	0.015	18,039
Baseline math (SA1)	-0.0030***	0.0008	0.186	10,638

Panel C: Lee Bounds on Pooled ITT

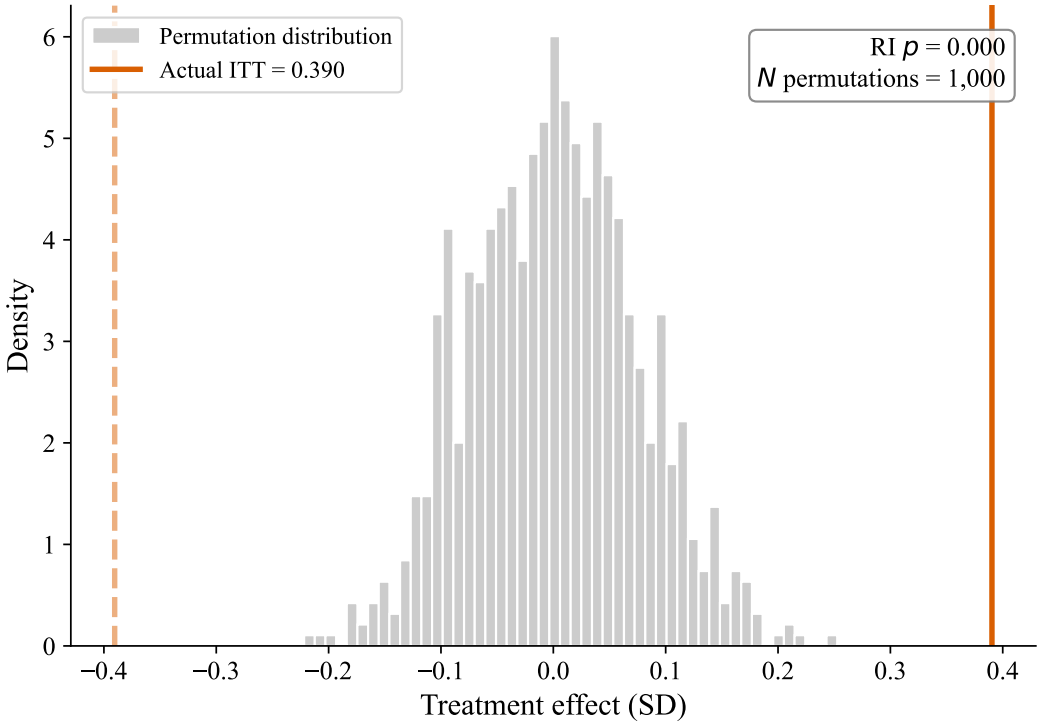
Original ITT	0.390*** (0.054)
Lower bound	0.270*** (0.047)
Upper bound	0.481*** (0.051)
Trim proportion	0.049 (332 students)

Notes: Panel A: attrition rates from the enrollment frame (18,039 students in 120 schools) to endline assessment completion. Difference from OLS of attrition indicator on treatment with strata FE and school-clustered SE. Panel B: each row regresses the attrition indicator on treatment, the baseline covariate, and their interaction (plus strata FE). The interaction tests whether attrition is differentially predicted by the covariate across treatment and control. Panel C: Lee (2009) bounds trim the treatment group outcome distribution to equalize attrition rates. The lower bound trims the top of the treatment distribution; the upper bound trims the bottom. Both bounds being positive rules out attrition bias as an explanation for the main treatment effect. All regressions include strata FE. Panel C includes demographic controls and grade FE. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

APPENDIX TABLE D3—MULTIPLE HYPOTHESIS TESTING CORRECTIONS

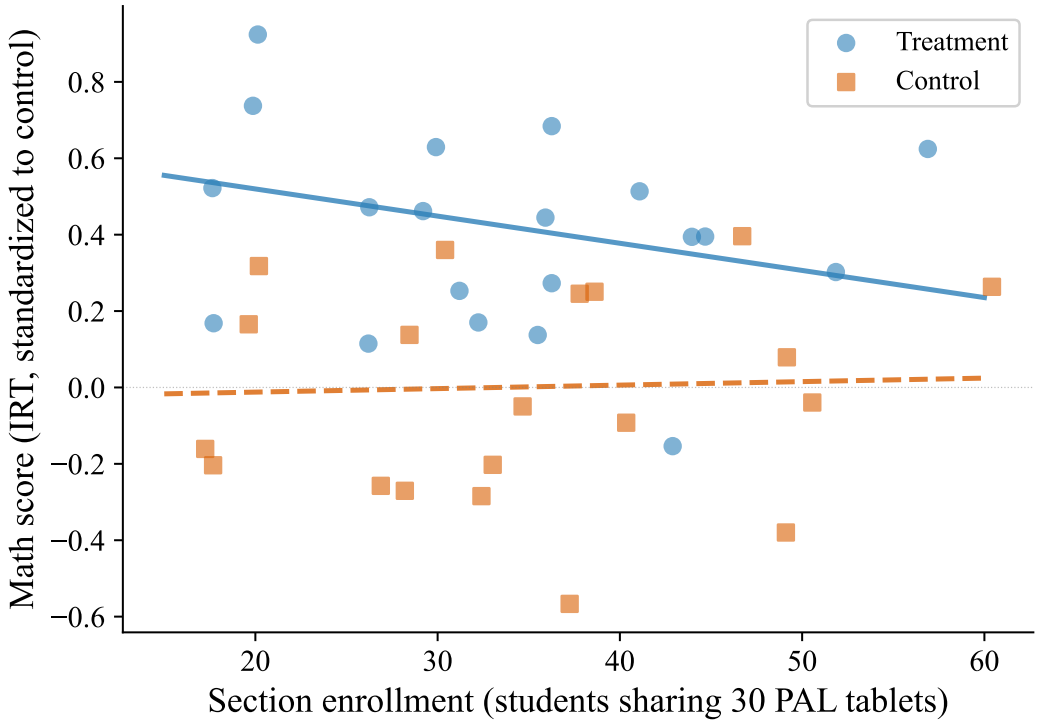
Hypothesis	Unadjusted *p*	Bonferroni	Holm	BH (FDR)
Gender	0.0005***	0.0049	0.0049	0.0049
Grade 7 interaction	0.3680	1.0000	0.6181	0.3680
Grade 8 interaction	0.2733	1.0000	0.6181	0.3075
Grade 9 interaction	0.1236	1.0000	0.6181	0.1949
Baseline Q2 interaction	0.0071***	0.0642	0.0571	0.0321
Baseline Q3 interaction	0.1350	1.0000	0.6181	0.1949
Baseline Q4 interaction	0.0641*	0.5772	0.4489	0.1681
Baseline Q5 interaction	0.1516	1.0000	0.6181	0.1949
Urban	0.0747*	0.6723	0.4489	0.1681

Notes: Unadjusted *p*-values from treatment × subgroup interaction tests. Bonferroni and Holm control family-wise error rate. BH controls false discovery rate. All from regressions with controls, strata FE, grade FE, and school-clustered SE.



APPENDIX FIGURE D1. RANDOMIZATION INFERENCE: PERMUTATION DISTRIBUTION OF TREATMENT EFFECTS

Notes: Distribution of treatment effects from 1,000 random permutations of treatment assignment within strata (at the school level). The solid vertical line shows the actual ITT estimate. The RI *p*-value is the fraction of permuted effects at least as extreme (in absolute value) as the actual effect. This non-parametric test does not rely on asymptotic assumptions about the error distribution.



APPENDIX FIGURE D2. EXCLUSION-RESTRICTION PLACEBO: SECTION ENROLLMENT AND MATH SCORES, TREATMENT VS. CONTROL

Notes: Dots: raw within-grade bin means of endline IRT math scores (5 bins per grade, pooled across grades 6–9). Lines: regression-predicted slopes from full specifications with controls, strata FE, grade FE, and SE clustered at the school level. Treatment slope: $\beta = -0.007$, $p = 0.07$ (larger sections reduce T scores, consistent with congestion). Control slope: $\beta = +0.001$, $p = 0.75$ (flat, consistent with section size not affecting learning in the absence of PAL). The flat control line provides visual evidence for the exclusion-restriction placebo on the congestion IV.

Appendix E. Development of the Test Instrument

E.A. Item Development and Piloting

We developed the Math Master Challenge assessment to measure mathematics achievement across the wide range of ability levels in our sample. Items were drawn from three sources: (i) all questions from every chapter of the Andhra Pradesh state math textbooks for grades 2 through 9; (ii) grade-specific math items from the Young Lives Study (Barnett et al., 2013); and (iii) items from a previous evaluation of adaptive learning technology in India (Muralidharan et al., 2019). From this bank, we selected an initial set of items targeting 10 questions per grade for grades 2–7 and 6 questions per grade for grades 8–9, with two items per topic per grade across five broad mathematical domains: algebra, geometry, measurement, statistics, and number systems.

We piloted the instrument in five schools over the course of one week. Each day, we downloaded the pilot data and estimated a two-parameter logistic (2PL) IRT model, dropping items that exhibited poor discrimination (low slope parameters) or extreme difficulty, with the goal of constructing an instrument whose test information function was sufficiently wide to capture the range of learning levels in our sample.

E.B. Final Instrument

The final assessment comprises 72 unique multiple-choice items spanning the grade 2 through grade 9 curriculum. Each student answered items from grade 2 up to the content level of their enrolled grade, yielding between 50 items (grade 6 students) and 72 items (grade 9 students). The assessment was administered on tablets via SurveyCTO in both English and Telugu, with three randomized question orderings (set assignment uniform at $P = 0.33$ per student) to mitigate order and copying effects. Students completed the assessment in approximately 40–50 minutes on average.

Table E1 shows the distribution of items across content grades and topics. Items from grades 2–7 have 10 questions each, while grades 8–9 have 6 questions each. Number systems has the most items (23), reflecting its central role in the Indian mathematics curriculum from basic arithmetic through rational numbers.

E.C. Psychometric Properties

We estimate a 2PL IRT model, which allows both discrimination (a_j) and difficulty (b_j) parameters to vary across items. Under this model, the probability that student i answers item j correctly is:

$$P(X_{ij} = 1 \mid \theta_i) = \frac{1}{1 + \exp\left(-a_j(\theta_i - b_j)\right)} \tag{4}$$

where θ_i is the latent ability of student i. We standardize the estimated ability parameters to the control group mean and standard deviation, following Muralidharan et al. (2019) and Muralidharan and Singh (2025).

Across all 72 items, discrimination parameters range from 0.08 to 2.16 (mean = 0.88), and difficulty parameters range from -1.54 to 6.57 (mean = 0.84). Lower-grade items are easier and more discriminating (grade 2: mean $a = 1.42$, mean $b = -0.94$), while higher-grade items are harder with lower discrimination (grade 9: mean $a = 0.34$, mean $b = 3.89$), reflecting the intended difficulty progression across the curriculum. Figure E1 displays the joint distribution of discrimination and difficulty parameters, colored by content grade.

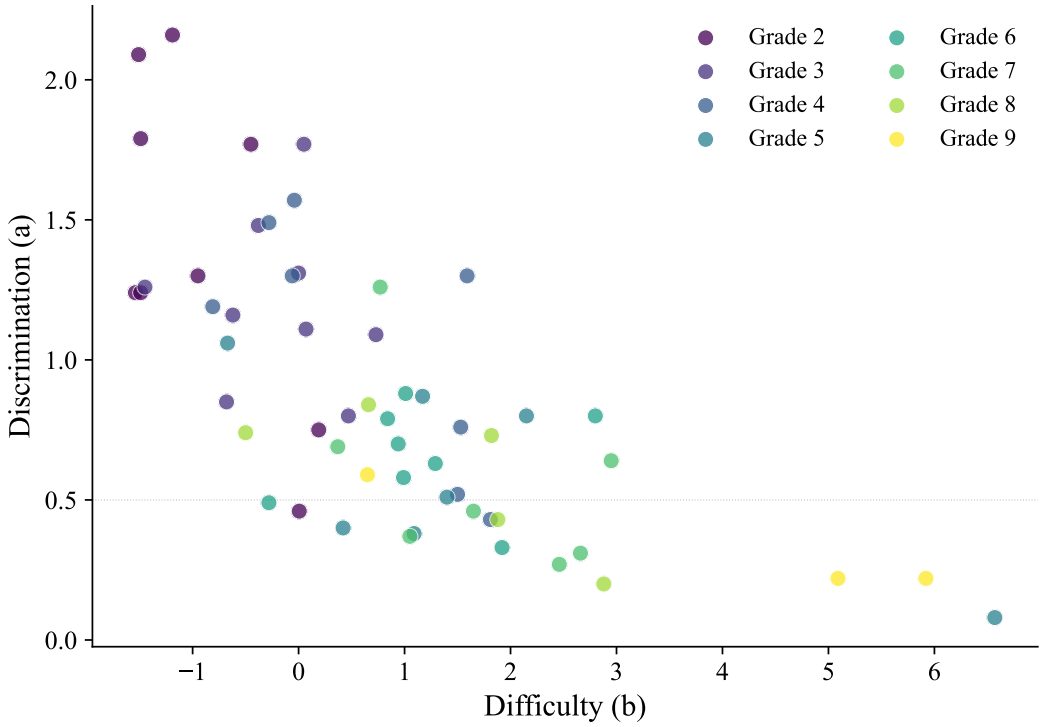
Internal consistency is high across all grade levels: Cronbach's α ranges from 0.90 (grade 6, 50 items) to 0.92 (grade 9, 72 items).

Figure E2 presents the test information function by enrolled grade. The instrument provides the most precise measurement in the ability range $\theta \in [-2, 1]$, with peak information near $\theta \approx -0.7$. The standard error of measurement remains below 0.35 across most of the ability distribution. Higher-grade students receive additional items that extend measurement precision into the upper ability range.

APPENDIX TABLE E1—DISTRIBUTION OF ASSESSMENT ITEMS BY CONTENT GRADE AND TOPIC

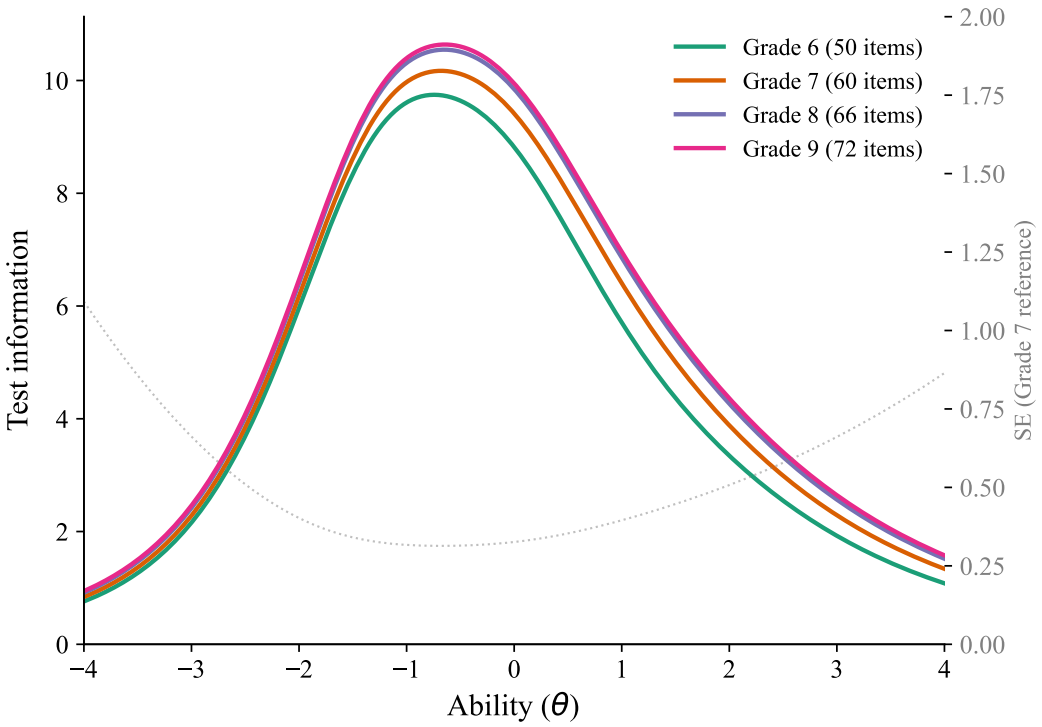
Content grade	Algebra	Geometry	Measurement	Number Systems	Statistics	Total
Grade 2	1	2	1	5	1	10
Grade 3	2	2	2	3	1	10
Grade 4	2	2	2	3	1	10
Grade 5	2	2	2	3	1	10
Grade 6	0	2	2	4	2	10
Grade 7	1	2	2	3	2	10
Grade 8	2	1	1	1	1	6
Grade 9	2	1	1	1	1	6
Total	12	14	13	23	10	72

Notes: Topic classification follows the Indian NCERT curriculum convention, where Number Systems covers arithmetic on specific number types (whole numbers, fractions, decimals, integers) and Algebra begins with variables and unknowns. Students in grade k answer all items from grade 2 through grade k, yielding 50 items (grade 6) to 72 items (grade 9).



APPENDIX FIGURE E1. ITEM DISCRIMINATION AND DIFFICULTY PARAMETERS

Notes: Each point represents one of the 72 items on the Math Master Challenge, colored by the curriculum grade from which the item was drawn. Discrimination (a) reflects how well the item differentiates between students of different ability levels; difficulty (b) indicates the ability level at which a student has a 50% probability of answering correctly. Parameters estimated via a 2PL IRT model. The dashed line at $a = 0.5$ marks a common threshold for acceptable discrimination.



APPENDIX FIGURE E2. TEST INFORMATION FUNCTION BY GRADE

Notes: Test information $I(\theta) = \sum_j a_j^2 P_j(\theta)(1 - P_j(\theta))$ for each enrolled grade's item set. Higher information indicates more precise measurement at that ability level. The dashed gray line shows the corresponding standard error $\mathrm{SE}(\theta) = 1/\sqrt{I(\theta)}$ for grade 7 as a reference. Students in higher grades answer additional items from higher curriculum levels, extending the instrument's measurement range.

Appendix F. Government School Assessments: Validity and the Grade 10 Board

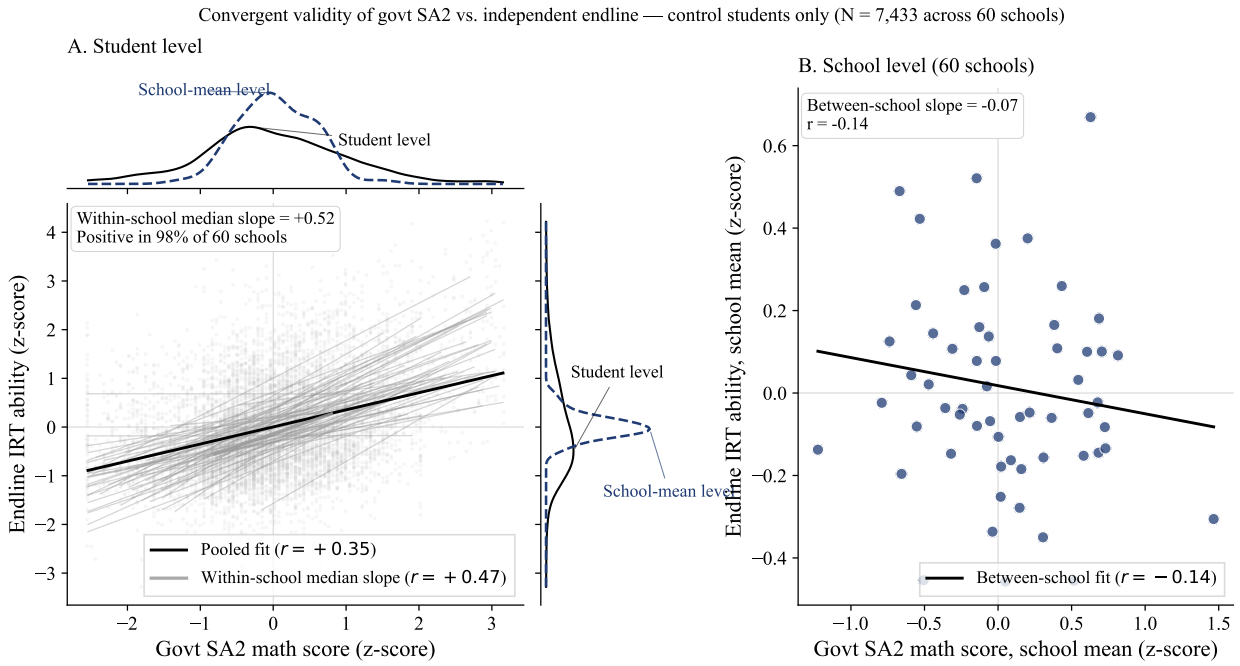
This appendix explains why govt FA/SA scores are useful only as within-school rank controls (not cross-school level measures), reports the negative Grade 10 board exam estimates on the externally graded cohort, and runs a displacement test that rules out PAL-driven exam-prep crowd-out as the mechanism.

Government-administered school assessments (FA and SA exams) are set and graded by individual schools, with teachers proctoring their own students. These assessments have well-documented limitations as measures of absolute learning: teacher-proctored school exams inflate scores by approximately 16–20 percentage points relative to independently monitored tests (Singh, 2024), and grading standards vary across schools, making scores non-comparable.

Figure F1 decomposes the rank and level structure of Year 2 SA2 scores against independently administered endline IRT scores, in the control group only. Two patterns emerge. First, within any given school, govt scores preserve student rank: the median within-school correlation is $+0.47$ and is positive in 98% of 60 control schools (Panel A, gray lines). Teachers rank their own students accurately. Second, across schools the signal disappears: the school-level correlation is -0.14 (Panel B), and the intra-class correlation of govt scores (0.21) is roughly three times that of the endline IRT (0.068). This excess between-school variance is visible in Panel A's marginal kdensities—school-mean govt scores (top, dashed) span nearly as widely as individual students, while school-mean endline scores (right, dashed) collapse to a tight band near zero. Most cross-school variation in govt scores reflects grading heterogeneity rather than ability. These patterns justify the use of govt scores only as within-school ranked baseline controls, and motivate the externally graded Grade 10 board exam as the cleaner cross-school measure.

Table F1 reports Grade 10 board exam effects for the cohort that had one year of PAL exposure in grade 9. Unlike the school-graded FA/SA assessments discussed in the preceding sub-appendix, the board exam is administered and graded externally, removing the grading heterogeneity that dominates cross-school variation in govt scores (Figure F1). Point estimates on the six subjects are all negative (-0.08 to -0.15 SD), but results are sensitive to the inclusion of strata fixed effects: no subject is significant at 5% without strata FE. The uniform direction across non-math subjects is puzzling given PAL is a math-only intervention; one possible explanation is differential exam preparation displacement during grade 9. We report these estimates because the external-grading structure makes them the cleanest cross-school government measure, while the specification sensitivity cautions against treating the negative direction as a settled finding.

To investigate whether the negative board exam effects reflect displacement of exam preparation by PAL usage, Table F2 decomposes the treatment effect by Year 1 PAL usage intensity. If PAL displaced exam preparation, the negative effect should be largest among high-usage students who spent the most time on tablets instead of studying. The evidence points in the opposite direction: the negative effect is concentrated entirely among low-usage students (-0.37 SD composite, $p < 0.001$), while high-usage students are statistically indistinguishable from control ($+0.01$ SD, $p = 0.93$). Within the treatment group, each additional hour of PAL predicts higher board scores across all six subjects ($+0.015$ SD/hour, $p < 0.001$), though this gradient likely reflects selection—higher-ability students engage more with the platform—rather than a causal benefit of usage. Splitting at median school enrollment, the negative effect appears in small schools (-0.14 SD) but not large schools (-0.00 SD), the opposite of the grade 6–9 endline pattern where small schools show the largest positive effects. Taken together, these results indicate that the negative board exam effect is not attributable to PAL usage displacing exam preparation time. The most parsimonious interpretation is that the effect reflects specification sensitivity—it is significant only with strata fixed effects and no student-level controls are available for this cohort—combined with possible school-level confounders that are absorbed differently across strata.



APPENDIX FIGURE F1. CONVERGENT VALIDITY OF GOVERNMENT SCHOOL ASSESSMENTS

Notes: Control students only ($N = 7{,}433$ across 60 schools). Outcome and regressor both z-scored to the control group's mean and SD. *Panel A*: student-level scatter of Year 2 govt SA2 math scores against independently administered endline IRT scores. Gray lines show OLS fits within each of the 60 control schools; the black line is the pooled student-level OLS fit. The gray fit at center of the panel marks the median within-school slope. Marginal kdensities on each axis compare the student-level distribution (solid black) with the school-mean distribution (dashed navy): schools differ moderately in mean govt score (top marginal) but barely in mean endline IRT (right marginal). *Panel B*: school-level means ($N = 60$) with OLS fit. Pooled student $r = +0.35$; within-school median $r = +0.47$ (positive in 98% of schools); between-school $r = -0.14$. The govt score's intra-class correlation (0.21) is $\approx 3\times$ that of the endline IRT (0.068); the excess between-school variance reflects grading heterogeneity rather than ability.

APPENDIX TABLE F1—GRADE 10 BOARD EXAM EFFECTS (1 YEAR PAL EXPOSURE)

	Math (1)	Science (2)	Social Sci. (3)	Lang 1 (4)	Lang 2 (5)	Lang 3 (6)
Panel A: With strata FE						
Treatment	-0.136**	-0.109*	-0.144**	-0.117*	-0.135*	-0.121*
	(0.067)	(0.065)	(0.067)	(0.070)	(0.071)	(0.070)
Panel B: Without strata FE						
Treatment	-0.138	-0.109	-0.152*	-0.139*	-0.121	-0.077
	(0.093)	(0.085)	(0.083)	(0.083)	(0.082)	(0.081)
Control mean (raw)	57.0	65.1	65.4	70.6	56.0	61.7
Observations	4,846	4,846	4,843	4,849	4,849	4,846

Notes: Grade 10 board exam scores for students who had PAL in grade 9 (AY 2023–24, 1 year exposure) and took the state board exam in 2024–25. Each column standardized to control mean and SD. Board exams are externally graded (state level). No student-level controls are available for this cohort (students not in endline sample). Standard errors clustered at the school level. Coverage: 98.4% treatment, 98.9% control. *Panel A* includes 21 strata fixed effects; *Panel B* shows results without strata FE. Point estimates are similar across panels (-0.08 to -0.15 SD), but standard errors are larger without strata FE (0.08–0.09 vs 0.07), and no subject is significant at 5% in Panel B. The negative coefficients across all six subjects suggest a possible effect, but interpretation requires caution: (i) this cohort had only 1 year of PAL exposure, (ii) the board exam was taken *after* students left treatment schools, (iii) the uniform direction across non-math subjects is puzzling given PAL is a math-only intervention, and (iv) results are sensitive to the inclusion of strata FE. One possible explanation is differential exam preparation: PAL may have displaced study time allocated to board exam review during grade 9. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

APPENDIX TABLE F2—GRADE 10 BOARD EXAM: INVESTIGATING THE ROLE OF PAL USAGE

	Math	Science	Social Sci.	Lang 1	Lang 2	Lang 3	Composite
Panel A: Treatment effect by Year 1 PAL usage (vs. control)							
Low usage (4 hrs, N_T=819)	-0.381***	-0.335***	-0.360***	-0.373***	-0.391***	-0.327***	-0.369***
	(0.080)	(0.086)	(0.080)	(0.085)	(0.095)	(0.088)	(0.076)
Mid usage (14 hrs, N_T=818)	-0.077	-0.061	-0.092	-0.069	-0.052	-0.024	-0.061
	(0.080)	(0.070)	(0.073)	(0.087)	(0.084)	(0.082)	(0.068)
High usage (27 hrs, N_T=819)	0.021	0.026	-0.023	0.077	-0.004	-0.059	0.006
	(0.076)	(0.079)	(0.085)	(0.079)	(0.083)	(0.085)	(0.069)
Panel B: Within-treatment usage gradient (OLS)							
PAL hours (Year 1)	0.0139***	0.0145***	0.0156***	0.0180***	0.0190***	0.0098***	0.0154***
	(0.0039)	(0.0035)	(0.0037)	(0.0034)	(0.0036)	(0.0027)	(0.0030)
Panel C: Treatment effect by school size							
Small schools (N_T=1,209, N_C=1,206)	-0.152						-0.138
	(0.099)						(0.086)
Large schools (N_T=1,208, N_C=1,223)	0.018						-0.003
	(0.102)						(0.106)

Notes: This table investigates whether the negative Grade 10 board exam effects (Table F1) are driven by students who used PAL intensively. *Panel A*: treatment students are split into Year 1 usage terciles and each tercile is compared to the full control group (separate regressions). The negative effect is concentrated among low-usage students (-0.37 SD composite, $p < 0.001$); high-usage students are indistinguishable from control ($+0.01$ SD, $p = 0.93$). *Panel B*: within the treatment group, each additional hour of PAL predicts *higher* board scores ($+0.015$ SD/hour, $p < 0.001$), across all six subjects including non-math. This gradient likely reflects selection (higher-ability students engage more) rather than a causal benefit of usage. *Panel C*: splitting at median enrollment, the negative effect is concentrated in small schools, the opposite of the endline pattern. Taken together, these results indicate that the negative board exam effect is not attributable to PAL usage displacing exam preparation. All regressions include strata FE and school-clustered SE. No student-level controls are available for this cohort. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Appendix G. Tablet vs. Paper Assessment-Mode Equivalence

A natural concern with our tablet-administered endline is that PAL exposure may inflate measured ability through tablet familiarity rather than learning, biasing the headline ITT upward. We test this directly using a sub-study embedded in the Year-3 endline (AY 2025–26) that re-tested the *same* Grade 8 RCT students on *the same items* using both modes. Each of 1,953 Grade 8 students in our 80 RCT schools sat both a tablet (SurveyCTO) and a paper (Optical Mark Recognition, OMR) Math Master Challenge in randomized order, with a counter-balanced crossover assigning opposite question sets to the two modes (Tab Set 1 students received OMR Set B and vice versa). Twenty-six "common" items appear on both modes for every student, providing within-student same-item measurement; 40 set-specific items contribute cross-form information through the IRT.

Table G1 reports treatment-effect specifications analogous to the main paper's Table 3, with mode (tablet vs. paper) and its interaction with treatment as the key terms. Panel A uses 2PL IRT θ from a pooled fit across all items administered in the dual-mode design; Panel B restricts the IRT to the items every student answered on both modes (a fully balanced design that does not rely on cross-form linking). Across all six (panel $\times$ specification) cells, the mode $\times$ treatment interaction lies between -0.09 and $+0.003$ SD and is not statistically significant in any specification. The mode main effect is at most -0.09 SD, with $p \geq 0.13$ throughout.

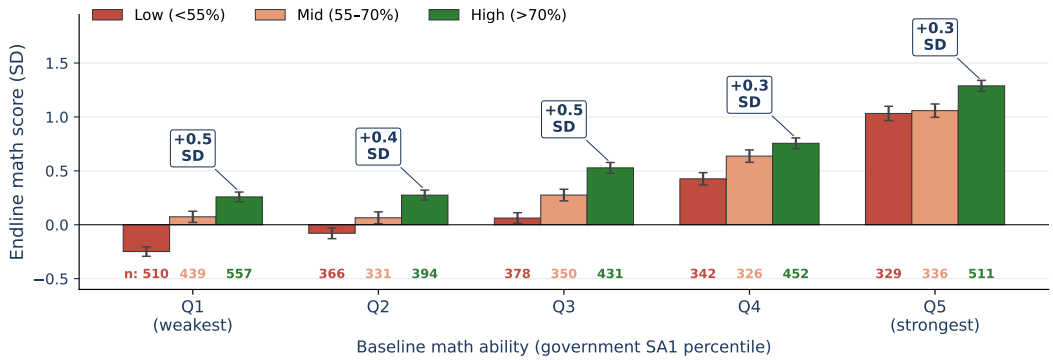
A model-free verification on the common items reaches the same conclusion: mean percent correct is 50.4% on tablet and 50.7% on paper, with a within-student Pearson correlation of 0.83 on per-student common-items scores—indicating coherent measurement of the same latent ability across modes. The DiD interaction on percent correct is -1.32 pp (SE 0.91, $p = 0.15$), consistent in sign and SD-equivalent magnitude with Panel A and Panel B.

We conclude that mode of administration does not differentially advantage treatment students on the tablet endline. The tablet-familiarity channel cannot account for the headline $+0.39$ SD ITT.

APPENDIX TABLE G1—G8 MODE COMPARISON: TABLET VS PAPER ASSESSMENT (Y3 RCT STUDENTS)

	(1) Model 1 (main)	(2) Model 1 (robust)	(3) Model 2 (1st exam)
Panel A. Primary specification: 2PL IRT on 67 items			
1.mode	0.006	0.006	-0.063
	(0.031)	(0.031)	(0.062)
1.mode × 1.treatment	-0.092	-0.094	-0.019
	(0.060)	(0.060)	(0.094)
R^2	0.914	0.913	0.095
Panel B. Robustness: 2PL IRT on 26 common items only			
1.mode	-0.038	-0.038	-0.094
	(0.030)	(0.030)	(0.062)
1.mode × 1.treatment	-0.070	-0.071	0.003
	(0.045)	(0.045)	(0.090)
R^2	0.922	0.922	0.087
N	3,906	3,886	1,953
Fixed effects	Student	Student	Mandal cluster

Notes: Outcome is θ standardized to the control × tablet reference cell (mean 0, SD 1). Panel A uses 2PL IRT θ from a pooled fit across all items administered in the dual-mode design; Panel B uses 2PL IRT θ from a fit restricted to the items every Grade 8 RCT student answered on both modes (a fully balanced spec). Models 1 use student fixed effects; Model 2 (first exam) uses mandal-cluster strata FE plus covariates (gender, age, SES, parental education with missing flags). Standard errors clustered at school. A negative interaction would indicate that treatment students gain a relative tablet advantage. * p<0.10, ** p<0.05, *** p<0.01.



APPENDIX FIGURE H1. ENDLINE MATH SCORE BY BASELINE ABILITY QUINTILE × COMPLIANCE TERCILE

Notes: Mean endline math score (in SD units, standardized to control-group mean and SD) by baseline math ability quintile (x-axis; from government SA1 percentile rank within school × grade) and compliance tercile (color; equal-frequency cuts at compliance ≈ 0.56 and 0.72). Error bars are ±1 standard error. Cell counts shown below each bar (range 343–538 students). Boxed annotations report the bottom-to-top tercile gap within each quintile. Treatment students only ($N = 6{,}052$); compliance defined as in Section II.B.

Appendix H. The Relationship Between Quality of Usage and Learning

We turn next to characterizing how learning within the treatment group varies with *how* students use the tablets, i.e., the relationship between quality of usage and learning. We proxy for quality with the parameter *compliance*, which we define to be the fraction of the prescribed activity sequence the student completes.[14] A higher level of compliance captures the student using the tablets closer to the way intended by the PAL program's designer, ConveGenius. We summarize the distribution of compliance by student, alongside hours used, in Table 2.

In Figure H1, we report endline math scores in the control group estimated separately by baseline-ability (broken into quintiles) and, within baseline ability, compliance (broken into terciles). Within every quintile of baseline ability, we observe a monotonic relationship between compliance and learning. Students in the top-tercile of compliance score 0.3–0.5 SD better on the endline assessment than those in the bottom tercile.[15] Furthermore, this gradient is steepest among students in the lowest quintile of baseline ability; put differently, what we learn from this analysis is that the students who arrive at the program with the largest foundational skill gaps benefit the most. Nonetheless, this gradient remains large and positive across the remaining quintiles, demonstrating that the quality of engagement with the PAL program predicts higher learning at every point in the baseline-ability distribution. Because compliance is not randomly assigned, unobserved motivation correlated with both engagement and learning would bias this gradient upward; we read this results as bounding, rather than identifying, the returns to the quality of usage.

[14] This is also described in Section II.B.

[15] The top tercile of compliance corresponds to the student opening more than 72 percent of prescribed activities; the bottom tercile corresponds to opening less than 56 percent of prescribed activities.

Appendix I. Dose-Response: IV Diagnostics, Shape Robustness, and Subgroup IV

This appendix supports the IV identification behind Table 4 and Figure 4: the first-stage and reduced-form scatters, the shape robustness of the dose-response curve under alternative fixed effects, and the subgroup IV replications by grade and gender.

First Stage and Reduced Form

Figure 4 visualises the first-stage and reduced-form relationships that identify the IV in Table 4. Panel A shows that each additional student in a section reduces PAL usage by 0.34 hours over two years ($p = 0.005$). Panel B shows that each additional student reduces endline math scores by 0.007 SD ($p = 0.04$). The ratio of these slopes delivers the 2SLS estimate reported in Table 4.

Shape of the Dose-Response

Figure 4 in the main text characterizes the within-school shape of the dose-response. Table I1 supports that figure by reporting the pooled quadratic and linear-spline specifications under two alternative sets of fixed effects. The quadratic coefficient on hours^2 is negative across all six specifications; under 21 mandal-cluster strata FE the coefficient is significant at the 5% level ($p = 0.024$), and under 60 school FE (within-school identification only) precision tightens to $p = 0.002$. The linear spline's joint Wald test rejects linearity under school FE ($p = 0.012$). The shape finding—steep learning between 10 and 30 hours, flat above—is robust to including the one-year Grade 6 cohort and to retaining students with 100+ hours of usage.

Cross-Sample Per-Hour Returns

Figure 7 in Section VI overlays within-school dose-response binscatter and quadratic fits across three delivery configurations (RCT-T 3-year cumulative, targeted residential $\sim$14-month cumulative, program 3-year cumulative). Table 5 reports the underlying linear and quadratic per-hour coefficients with school-clustered standard errors. The cross-sample slope ordering 0.025 (program) > 0.016 (targeted residential) > 0.011 (RCT-T) tracks each sample's mean position on the dose-response curve — program students are concentrated near the steep activation region, residential students in the mid-range, and RCT-T students in the flattening high-dose region. The RCT-T 3-year linear slope of $+0.0105$ matches the IV slope of $+0.011$ identified on the RCT 2-year sample (Figure 4). All three samples show concave shape (negative hours-squared, $p \leq 0.02$).

Subgroup Replications

Table 4 replicates the IV analysis from Table 4 separately by grade and by gender. Per-hour estimates are broadly consistent across subgroups (0.008–0.018 SD per hour), though statistical power is limited within any single subgroup. Girls accumulate approximately 2 more hours of PAL usage than boys over two years; per-hour returns are similar.

APPENDIX TABLE I1—SHAPE OF THE POOLED DOSE-RESPONSE

	Strata FE			School FE		
	Linear (1)	Quadratic (2)	Spline (3)	Linear (4)	Quadratic (5)	Spline (6)
Usage hours	0.0091***	0.0128***	-0.0023	0.0068***	0.0101***	-0.0070
	(0.0018)	(0.0029)	(0.0078)	(0.0012)	(0.0018)	(0.0076)
Usage hours^2 ($\times 10^3$)		-0.036**			-0.033***	
		(0.016)			(0.010)	
Slope change at 10 hrs			+0.0235**			+0.0263**
			(0.0115)			(0.0118)
Slope change at 20 hrs			-0.0061			-0.0063
			(0.0111)			(0.0109)
Slope change at 30 hrs			-0.0082			-0.0083
			(0.0074)			(0.0058)
Spline joint Wald p			0.086			0.012
Strata FE (21)	Y	Y	Y			
School FE (60)				Y	Y	Y
Controls + grade FE	Y	Y	Y	Y	Y	Y
Observations	6,742	6,742	6,742	6,742	6,742	6,742

Notes: Treatment-group students only. Dependent variable is endline IRT math score standardised to the control mean and SD. Columns (1)–(3) use the 21 mandal-cluster strata FE; columns (4)–(6) replace strata FE with 60 school FE, absorbing between-school composition. The quadratic coefficient on hours^2 is negative across all six specifications, indicating concavity in the pooled within-treatment dose-response. Precision tightens substantially under school FE. Spline knots at 10, 20, and 30 hours; rows report the change in slope at each knot. Standard errors clustered at the school level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Appendix J. Deep Remediation: Design Details and Extended Results

This appendix supports the within-school remediation-depth experiment reported in Section V. It documents that the experimental contrast is the content-depth contrast the design intended (attendance, curriculum overlap, and the anatomy of the engagement gap); details the content-based treatment definition, sample, and balance; and reports the full heterogeneity and item-level results.

J.A. Intervention Fidelity: Attendance, Curriculum, and the Engagement Gap

Randomization is at the student-within-school level: within each of the 20 study schools, students in grades 6–9 are individually randomized to Deep Remediation (`GOAPFF_1023`, remediation depth unlimited) or Status Quo (`GOAPM_0823`, remediation capped at 2 grade levels below enrolled grade). PAL lab sessions are scheduled by classroom teachers at the school-grade level; teachers are not informed of arm assignment. Baseline scheduled exposure to PAL—lab days, lab start times, lab lengths—should therefore be identical across arms. This sub-appendix documents that scheduled exposure is indeed equal, that the PAL algorithm exposes the same curriculum to both arms, and that the treatment contrast manifests entirely as a difference in *content depth*: access to below-cap remediation material that the Status Quo algorithm does not render.

Physical lab attendance is identical across arms. Figure J1 shows the weekly session pattern across all 20 schools: at every school $\times$ day-of-week cell, the Deep Remediation (navy) and Status Quo (grey) session bars lie on top of one another. Across 4,666 (school $\times$ grade $\times$ day) cells in which at least three students from each arm were active over the two-year intervention (AY 2023–24 and 2024–25 pooled, with December 2023 and January 2024 dropped from both arms because Deep Remediation data for those months is incomplete; see footnote 16), the median difference between arms in session start and end time is essentially zero, and the cells align within ± 10 min on both edges in roughly four out of five cases (Figure J2). Physical compliance is full and equal.[16]

Conditional on identical attendance, we examine whether the PAL algorithm renders the same instructional material to each arm. For each (school $\times$ grade) class, we compute the set of distinct content-chapters each arm's students touched over the two-year intervention (excluding the December 2023 and January 2024 gap window), keyed by (content-grade, chapter number) so that Telugu-medium and English-medium renderings of the same chapter resolve to a single identifier. Table J3 reports the arm-level overlap by content-grade. At at-grade content, the Deep Remediation and Status Quo chapter sets overlap at Jaccard 0.99—the two arms cover essentially identical curricula. Overlap remains high at shallow remediation (Jaccard 0.85), where Status Quo's chapter set is 95% contained within Deep Remediation's; Deep Remediation students touch a small handful of additional chapters per class but never skip content Status Quo visits. At deep remediation, virtually 100% of Status Quo's (small) chapter footprint is a subset of Deep Remediation's, and Deep Remediation students access on average 3.3 additional chapters per class that the Status Quo algorithm does not render. The PAL algorithm is content-conservative: the material exposed to Status Quo is always a subset of what Deep Remediation students see.

Conditional on identical attendance and overlapping curricula, Deep Remediation students log modestly more total engagement than Status Quo students. Across the bilateral two-year intervention window, Deep Remediation students average 20.2 hours of wall-clock PAL engagement per student versus 17.9 for Status Quo (a $+2.3$-hour Deep Remediation advantage); the in-app `total_time_spent_new` timer, which accumulates pause and background time within each open module and therefore exceeds wall-clock by a small multiplicative factor, registers a $+3.7$-hour Deep Remediation advantage (21.2 versus 17.6 hours, a $+21\%$

[16] The ConveGenius re-export delivered on 2026-04-27 cleaned four GOAPM (Status Quo) files but did not include re-exports of the GOAPFF (Deep Remediation) data for December 2023 (entirely missing) or January 2024 (partial, $\approx$25% of expected rows). Including those months in their current state would bias the FF-vs-SQ comparison against Deep Remediation. We therefore drop December 2023 and January 2024 from both arms throughout the fidelity analyses, retaining bilateral parity.

APPENDIX TABLE J1—COVARIATE BALANCE BETWEEN DEEP REMEDIATION AND STATUS QUO ARMS

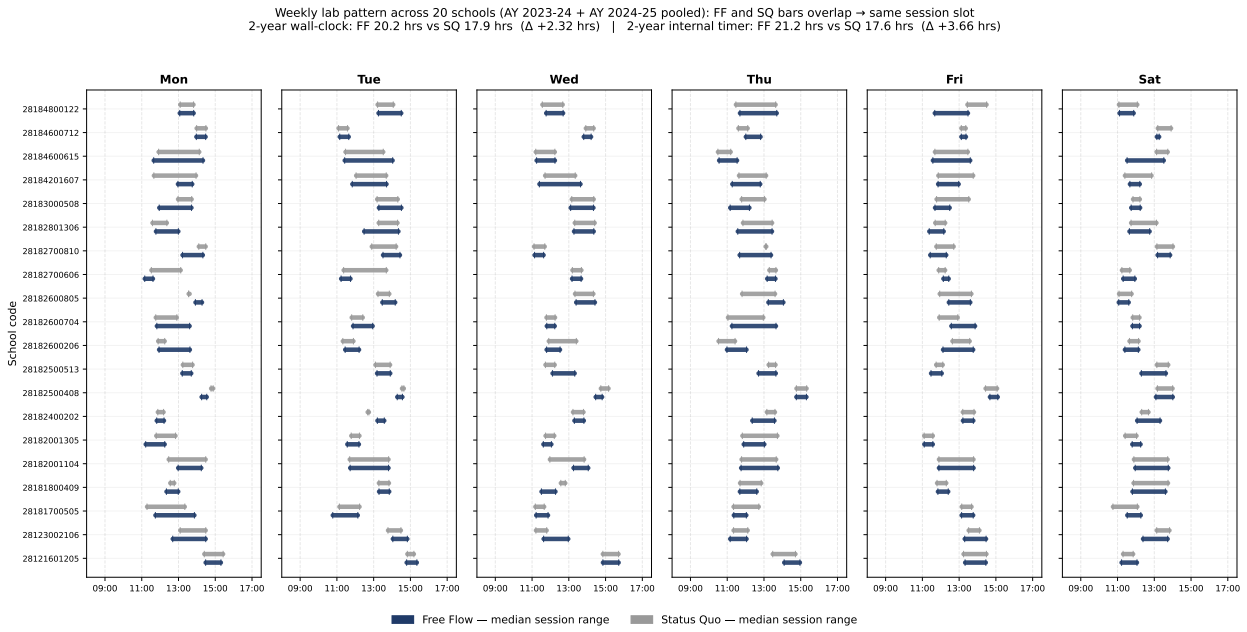
	Status Quo (1)	Deep Remediation (2)	Difference (3)	(SE) (4)	p-valu (5)
Female	0.438	0.491	0.057*	(0.031)	0.067
Age (years)	12.42	12.46	-0.017	(0.035)	0.624
Scheduled tribe	0.365	0.342	0.007	(0.015)	0.650
Scheduled caste	0.156	0.168	0.010	(0.011)	0.345
Backward class–A	0.048	0.039	-0.010	(0.007)	0.157
Backward class–B	0.432	0.451	-0.007	(0.012)	0.545
Father's education (0–5)	2.21	2.20	-0.047	(0.063)	0.456
Mother's education (0–5)	1.95	1.98	0.027	(0.067)	0.688
SES index (PCA)	0.02	-0.02	-0.026	(0.030)	0.394
Joint F-statistic			2.63		[$p = 0.0$
Pre-PAL baseline math (cohort-specific govt SA1)	32.50	31.72	-0.282	(0.548)	0.606

Notes: Means of each covariate by arm (columns 1–2) and the estimated Deep Remediation − Status Quo difference (column 3) from OLS of the covariate on Deep Remediation plus school-by-grade-by-section fixed effects. Standard errors clustered at the school level ($S = 20$). Caste dummies follow the Andhra Pradesh reservation categories; Backward Class B is the omitted reference in the joint F-test to avoid collinearity with the other three caste indicators. Pre-PAL baseline math is constructed cohort-specifically: AY 2023–24 SA1 for the Grade 6 endline cohort (their Grade 5 year, before PAL); AY 2022–23 SA1 for the Grade 7–9 endline cohorts (their Grade 5–7 year, also pre-PAL). Coverage is 93% of the analysis sample. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

APPENDIX TABLE J2—DEEP REMEDIATION EFFECT BY PRE-PAL BASELINE MATH TERCILE

	FF effect (1)	(SE) (2)	p-value (3)
FF effect at each baseline tercile (within school × grade)			
Tercile 1 (lowest pre-PAL baseline)	+0.058	(0.038)	0.129
Tercile 2 (middle)	+0.055	(0.039)	0.152
Tercile 3 (highest)	+0.077	(0.055)	0.163
Differential effect (vs. T1)			
Tercile 2 − Tercile 1	-0.003	(0.058)	0.962
Tercile 3 − Tercile 1	+0.019	(0.067)	0.772
Joint F (interactions = 0)	0.05		0.954
Mean pre-PAL baseline (T1, T2, T3)	16.9 / 31.6 / 48.2		
Observations	3,492		
R^2	0.348		

Notes: Deep Remediation effects on θ^{std} at each within-school-by-grade tercile of the cohort-specific pre-PAL baseline math score (Grade 6: AY 2023–24 SA1; Grades 7–9: AY 2022–23 SA1). Specification: OLS of θ^{std} on Deep Remediation, two interaction terms (Deep Remediation × Tercile 2 and Free Flow × Tercile 3), two tercile main effects, the covariate vector of Appendix J.B, and school-by-grade-by-section fixed effects; standard errors clustered at the school level ($S = 20$). Terciles are constructed within school × grade (not within section) to keep bins large. FF-effect rows for Tercile 2 and Tercile 3 are linear combinations of the Deep Remediation main effect and the corresponding interaction. The joint F-test asks whether the two interactions are jointly zero. 263 students drop from the sample because their pre-PAL baseline test record is missing. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.



APPENDIX FIGURE J1. DEEP REMEDIATION AND STATUS QUO STUDENTS ATTEND THE SAME LAB SESSIONS

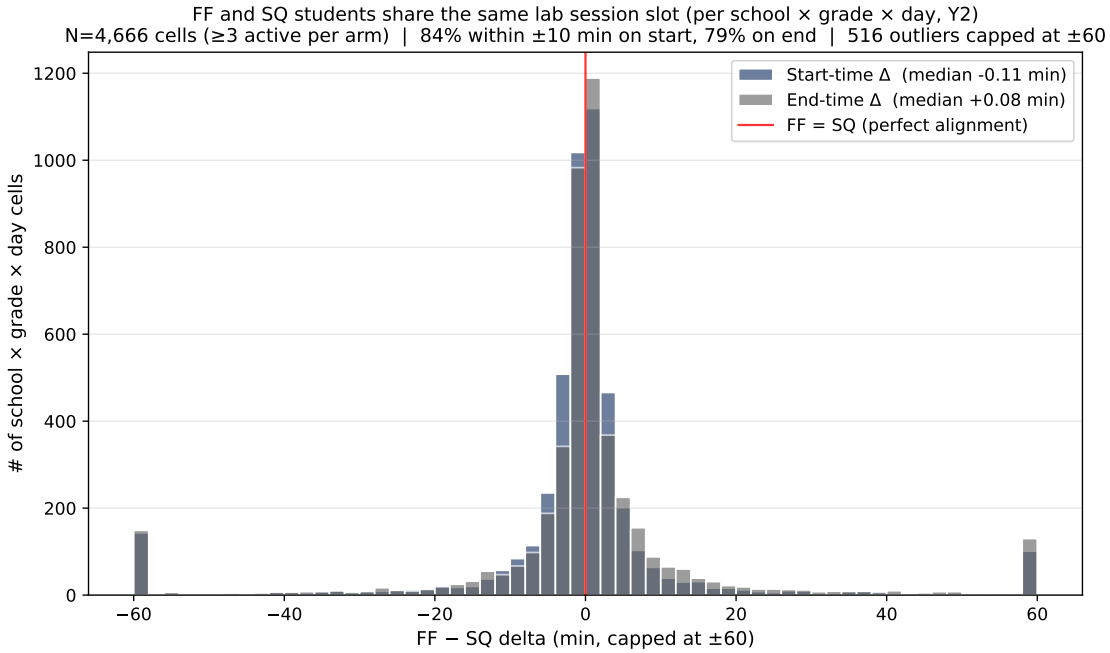
Notes: Weekly PAL lab-session pattern across all 20 remediation-study schools, Monday through Saturday, AY 2023–24 + AY 2024–25 pooled (December 2023 and January 2024 dropped from both arms; see footnote 16). For each (school × day-of-week) cell with at least three qualifying session-days across the two years, horizontal bars show the median session start-to-end time separately for Deep Remediation (navy) and Status Quo (grey) students. Bars lie on top of one another at every cell: arms share the same teacher-scheduled lab slots. Two-year engagement comparison (top of figure): wall-clock PAL engagement is Deep Remediation 20.2 hours versus Status Quo 17.9 hours ($\Delta = +2.3$ hours); in-app timer minutes register 21.2 versus 17.6 ($\Delta = +3.7$ hours), with the wedge between wall-clock and in-app timer reflecting module-timer inflation over wall-clock. Lab-session ranges computed per-student from first to last module submission on a given day; cell medians taken across students within each arm. See Figure J2 for the signed start/end delta across the 4,666 (school × grade × day) cells that qualify.

APPENDIX TABLE J3—DEEP REMEDIATION AND STATUS QUO STUDENTS COVER THE SAME CHAPTERS

Content level	**Jaccard** $\lvert FF \cap SQ\rvert / \lvert FF \cup SQ\rvert$	**FF ⊆ SQ** % of FF ch.	**SQ ⊆ FF** % of SQ ch.	**FF / SQ / shared** mean # chapters
At-grade (g)	0.992	99.6%	99.6%	14.3 / 14.3 / 14.3
Shallow remediation ($g-1$, $g-2$)	0.848	88.4%	95.3%	17.6 / 16.2 / 15.4
Deep remediation ($g-3$, $g-4$, $g-5$)	0.614	61.5%	99.7%	8.0 / 4.7 / 4.7

Notes: Per (school × enrolled-grade) class ($N = 80$ classes with ≥5 common chapters in the at-grade and shallow-remediation rows; $N = 44$ in the deep-remediation row), the set of distinct content-chapters touched by Deep Remediation students and by Status Quo students over the two-year intervention (AY 2023–24 + AY 2024–25), by content-grade relative to enrolled grade. "FF ⊆ SQ" reports the share of Deep Remediation's chapter set also covered by Status Quo; "SQ ⊆ FF" reports the reverse. Chapters are keyed by (content-grade, chapter-number) so that Telugu-medium and English-medium renderings of the same chapter resolve to a single identifier. Shallow-remediation: Status Quo's set is 95% contained in Deep Remediation's; deep-remediation: Status Quo's set is 99.7% contained in Deep Remediation's but Deep Remediation additionally covers 3.3 chapters per class that the Status Quo algorithm does not render.

gap). The in-app gap decomposes multiplicatively into three factors of similar magnitude, shown in Figure J3: Deep Remediation students are active on 9% more days per student (28.2 versus 25.9), complete 3% more modules per active day (25.7 versus 24.9), and each of their modules has a 9% longer in-app timer (1.85 versus 1.70 min). Compounded, $1.09 \times 1.03 \times 1.09 \approx 1.22$, these factors reproduce the observed 21% in-app



APPENDIX FIGURE J2. FF AND SQ SESSION SLOTS ALIGN WITHIN MINUTES AT THE SCHOOL-GRADE-DAY LEVEL

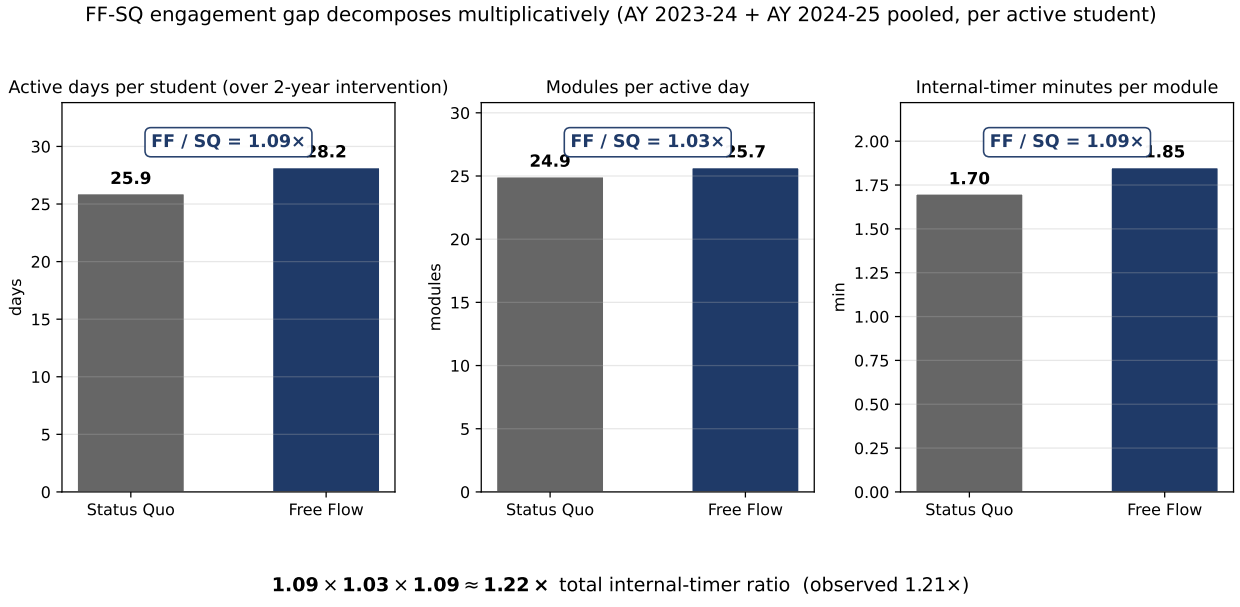
Notes: For each (school × grade × day) cell with at least three active students in each arm (4,666 cells across the two-year intervention), we compute the median first-submission time (navy) and median last-submission time (grey) separately per arm, and take the difference FF − SQ. Both signed-delta distributions are tightly concentrated at zero; the median delta is essentially zero on both start and end times. Roughly four out of five cells align within ±10 min on both start and end. Edge bars at ±60 min represent the ∼5–6% of cells with misaligned sessions (distinct lab periods scheduled on the same day) — these cells are retained in the session-slot alignment claim because the median delta, rather than the mean, is reported.

gap.

The engagement gap also distributes unevenly across module types. Figure J4 reports per-student minutes by module type. The bulk of the Deep Remediation excess lies in Video modules (+238 minutes per student over the two-year intervention, a 42% increment) and in Remedial Practice modules (+22 minutes, 40%). Formative assessments—the second-largest module-type bucket overall—are essentially balanced across arms (+12 minutes on a 271-minute base, 4%). The other four module types (Chapter Practice, Baseline, Diagnostic, Endline) contribute negligibly to the gap. The algorithmic FF–SQ contrast therefore operates at the module-type dimension as well as the content-grade dimension documented in Figure 3 of the main text.

The below-grade dose differential is positive and significant at every grade: Grade 6 +3.3 hours, Grade 7 +3.1 hours, Grade 8 +3.2 hours, Grade 9 +3.6 hours over the bilateral two-year intervention window (all $p < 0.001$). Figure J5 reports the full grade-wise breakdown.

Time inflates with re-attempts: a student who re-watches a video accumulates more minutes without completing more activities. A complementary view comes from compliance—the share of platform-prescribed activities (learning objectives, henceforth LOs) the student completes, capped at one and averaged across chapters. The Deep Remediation–Status Quo time gap of 21% reflects 14% more unique LOs completed and 7.5% more time per LO. Table J4 reports compliance by content depth. Deep Remediation students complete substantially more of their deep-remediation prescription (+37.5 percentage points), and slightly less of their at-grade (−3.2 pp) and shallow-remediation (−3.6 pp) prescriptions; the overall share of prescribed activities completed is marginally lower in Deep Remediation than in Status Quo (40.4% vs 41.6%, $p = 0.04$). The deeper cap therefore reorganises engagement toward below-cap content rather than expanding total work



APPENDIX FIGURE J3. THE DEEP REMEDIATION–STATUS QUO ENGAGEMENT GAP DECOMPOSES MULTIPLICATIVELY

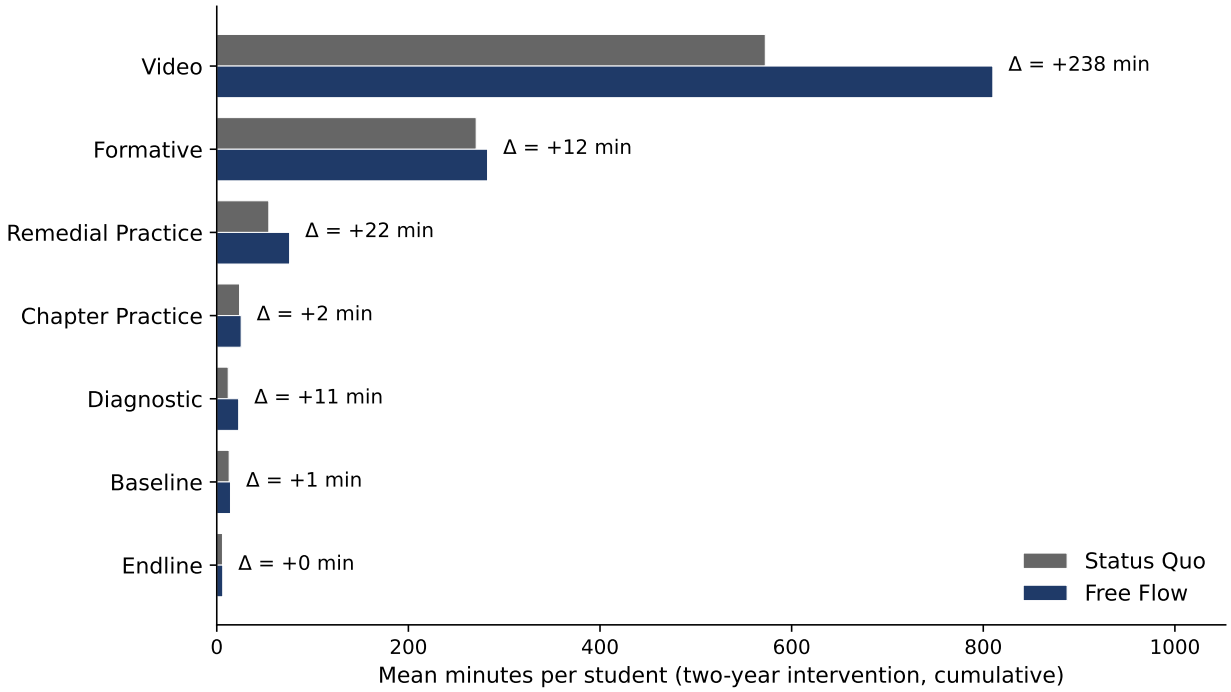
Notes: Per active student, AY 2023–24 + AY 2024–25 pooled (December 2023 and January 2024 dropped from both arms; see footnote 16). The Deep Remediation–Status Quo in-app timer gap of +21% decomposes into three multiplicative factors: Deep Remediation students are active on 9% more days per student (28.2 versus 25.9), complete 3% more modules per active day (25.7 versus 24.9), and each of their modules has a 9% longer in-app timer (1.85 versus 1.70 min). Compounded, $1.09 \times 1.03 \times 1.09 \approx 1.22$. Active days and modules per day are measured from timestamped CG module submissions; in-app timer minutes are the CG platform's internal per-module timer (`total_time_spent_new`), which accumulates pause and background time within a module.

APPENDIX TABLE J4—COMPLIANCE BY CONTENT DEPTH, DEEP REMEDIATION VS STATUS QUO

	Deep Remediation	Status Quo	Δ (DR − SQ)	p-value
Overall compliance	40.4%	41.6%	-1.3 pp**	0.038
At-grade	36.6%	39.9%	-3.2 pp***	< 0.001
Shallow remediation ($g-1$, $g-2$)	47.7%	51.3%	-3.6 pp***	< 0.001
Deep remediation ($g-3$ and below)	37.5%	0.0%	+37.5 pp***	< 0.001
Students	2,017	1,597		

Notes: Compliance is the student-level mean across chapters of platform-prescribed LOs completed divided by LOs prescribed for the (chapter × cascade signature × prior-incomplete history) cell, capped at 1. Sample: students with module-level data ($N = 3{,}614$ of 3,756; the remaining 142 have no logged modules and are dropped, consistent with the engagement-decomposition sample in Figure J3). The deep-remediation row reports the share of prescribed deep-cap activities completed; any positive value for Status Quo students reflects edge cases in the depth-based treatment classification or the small set of students with brief cross-program records, since the structural Status Quo prescription has zero $g-3$-or-below content. Two-sample t-tests assume unequal variances. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

done—and this reorganisation comes with a small reduction in the overall share of the prescription each student walks.



APPENDIX FIGURE J4. THE DEEP REMEDIATION–STATUS QUO ENGAGEMENT GAP CONCENTRATES IN VIDEO AND REMEDIAL PRACTICE MODULES

Notes: Mean minutes per student of PAL engagement, two-year cumulative (AY 2023–24 + AY 2024–25), broken down by CG `module_type`. Module types are sorted by total volume across arms. The Deep Remediation excess of +218 minutes per student concentrates in Video ($\Delta = +238$ min, +42% over Status Quo) and in Remedial Practice ($\Delta = +22$ min, +40%); Formative assessments — the second-largest module-type bucket overall — are essentially balanced across arms ($\Delta = +12$ min on a 271-min base, +4%). Diagnostic, Chapter Practice, Baseline, and Endline modules account for negligibly small per-student volumes and contribute little to the FF–SQ gap.

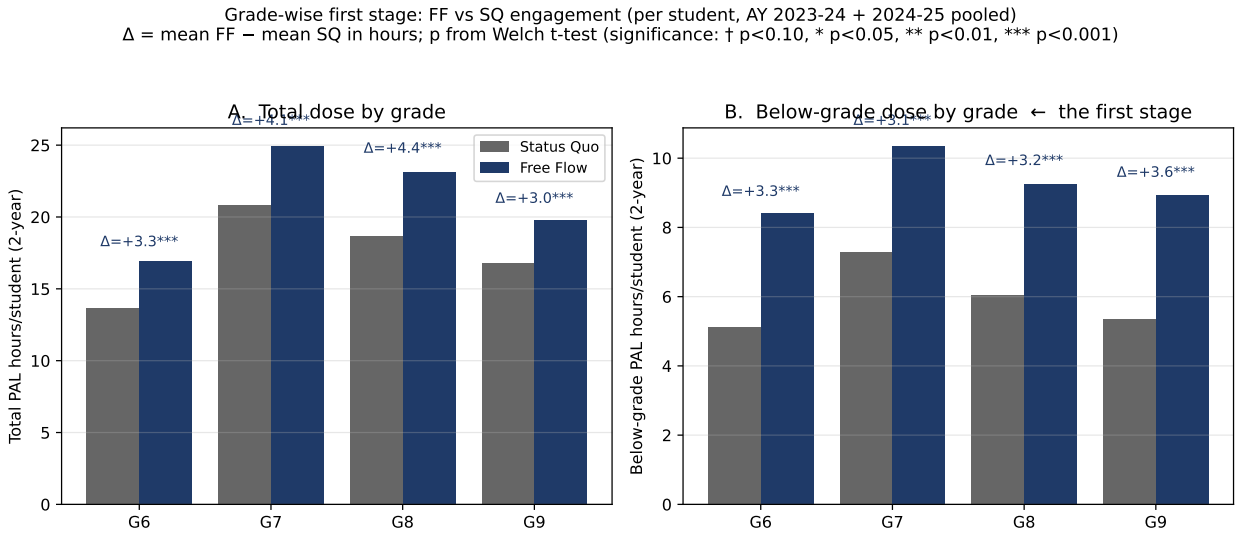
J.B. Treatment Definition, Sample, and Balance

We define the treatment indicator FF_i from *content actually rendered* by the algorithm to each student over the two-year intervention, rather than from the backend randomisation assignment:

$$\text{FF}_i \;=\; 1 \;\text{ iff }\; \text{student } i \text{ logged any module at content-grade } \leq \text{ enrolled-grade} - 3,$$

i.e., received any *deep-remediation* content (three or more grade levels below enrolled grade) at any point in the intervention. The Status Quo PAL algorithm strictly caps at two grade levels below enrolled grade; in our 2.5 million ConveGenius module-level records, no module rendered by the Status Quo algorithm has content-grade below $g-2$. Accordingly, any positive deep-remediation exposure proves that the student was served by the Deep Remediation algorithm during at least some module sessions, while zero deep-remediation exposure indicates the student was served exclusively by the Status Quo algorithm. The treatment indicator inherits the cleanness of this algorithmic dichotomy and is invariant to whatever administrative relabelling occurred between backend assignment and module-level delivery.

This definition departs from a strict intent-to-treat specification based on the original DIL randomisation. We adopt it because backend randomisation and CG-delivered arm disagree for 14% of endline respondents (448 of 3,209 students with both records). Investigation of these mismatches reveals that they reflect within-experiment program switches: for 99% of mismatch students with both years of CG records (307 of 310), CG's algorithm placed them in one arm in 2023–24 and the other arm in 2024–25, so neither year's modal



APPENDIX FIGURE J5. THE DEEP REMEDIATION FIRST STAGE BY ENROLLED GRADE

Notes: Mean PAL hours per student over the two-year intervention (AY 2023–24 + AY 2024–25, December 2023 and January 2024 dropped from both arms; see footnote 16), by enrolled grade × arm. Panel A: total dose. Panel B: below-grade dose — the cap differential. The Deep Remediation algorithm exposes students to remediation up to five grade levels below enrolled grade; the Status Quo algorithm caps at two. Δ = mean FF − mean SQ in hours; significance from two-sided Welch t-test († $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$). All four grade contrasts are positive and significant on both total and below-grade dose: total dose Deep Remediation − Status Quo is +3.3, +4.1, +4.5, and +3.0 hours at Grades 6 through 9 respectively; below-grade dose differential is +3.3, +3.1, +3.2, and +3.6 hours. The below-grade dose differential is roughly the same size as the total dose differential at each grade, indicating that Deep Remediation's extra time is concentrated at content depths the Status Quo algorithm cannot render.

program nor the backend record fully captures their actual exposure. For such cross-over students, the depth-of-content signal is the cleanest single summary of treatment receipt. As a side benefit, the content-based definition recovers 364 endline respondents who would be dropped under strict DIL ITT due to missing or invalid backend records but who do appear in CG module-level data with a determinable arm.

Only one of the 3,756 endline respondents in the master roster is unclassifiable under this definition: a student with neither a backend roster record (in either DIL or CG) nor any PAL usage at all. The remaining 3,755 enter every regression.

Table J5 reports grade-by-arm sample sizes together with demographic descriptives. The sample is 47% female, averages 12.4 years of age, and is drawn from AP's reserved-category population: 35% are Scheduled Tribe, 16% Scheduled Caste, and 48% Backward Class (the bulk of the rest). Father's education averages 2.2 and mother's 2.0 on a 0–5 scale, with the modal family at Grade 8 schooling. The grade composition (Grade 6 to 9) is slightly top-heavy (28% in Grade 9, 22% in Grade 6); we report all analyses both pooled and by-grade.

Our pre-specified covariates are: age, female, four caste dummies, an SES index constructed as the first principal component of nine household assets, and father's and mother's education each on a 0–5 scale. Missing covariate values are imputed with the sample mean (continuous variables) or the omitted category (indicator variables) and accompanied by an indicator for missingness, preserving the full classified sample of 3,755 students. The estimating equation includes school-by-grade-by-section fixed effects, absorbing the original randomisation stratum (school × grade × section). The section variable used for randomisation in 2023 is not preserved in the DIL backend output; we recover it from the AY 2023–24 government enrollment roster (the input file the randomisation script operated on) and, for the Grade 6 cohort that joined in AY 2024–25, from the AY 2024–25 government enrollment file. Together these recover section for 98.4% of randomised students; the residual is folded into a within-school × grade "unknown" cell. Standard errors are

APPENDIX TABLE J5—SAMPLE CHARACTERISTICS BY GRADE AND ARM

	N		Female		Age (yrs)		SES index		Baseline math	
	SQ	FF	SQ	FF	SQ	FF	SQ	FF	SQ	FF
Grade 6	399	434	0.489	0.527	10.76	10.73	0.14	0.02	32.0	32.1
Grade 7	423	460	0.430	0.490	11.92	11.81	-0.05	-0.02	32.2	33.2
Grade 8	448	536	0.431	0.456	12.82	12.83	-0.01	-0.03	28.9	26.6
Grade 9	469	587	0.408	0.498	13.88	13.92	0.01	-0.03	36.7	35.0
All grades	1,739	2,017	0.438	0.491	12.42	12.46	0.02	-0.02	32.5	31.7

Notes: Means of student characteristics for the 3,755-student analysis sample, separately by enrolled grade (rows) and arm (columns). Arm assignment is the content-based treatment indicator defined in Appendix J.B: a student is classified Deep Remediation if and only if they received any deep-remediation content ($\leq g-3$) from the algorithm. *N* reports student count; female, age, and SES index are drawn from the demographic survey, and baseline math is the cohort-specific pre-PAL government Summative Assessment 1 (Grade 6 endline cohort: AY 2023–24; Grades 7–9 endline cohorts: AY 2022–23). Pre-PAL baseline math is missing for approximately 7% of the sample.

clustered at the school level ($S = 20$).

Table J1 reports baseline balance under the content-based classification. Caste shares, age, parental education, and the SES index are tightly balanced across arms, but Deep Remediation students are slightly more likely to be female (+6 percentage points, $p = 0.067$), and the joint F-test on the nine covariates is marginally significant ($F = 2.63$, $p = 0.040$). Pre-PAL math, constructed cohort-specifically (Grade 6 endline $\rightarrow$ AY 2023–24 SA1, taken in their Grade 5 year before PAL; Grades 7–9 endline $\rightarrow$ AY 2022–23 SA1, their Grade 5–7 year, also pre-PAL), is essentially balanced (-0.28 points, $p = 0.61$). The remaining mild covariate-vector imbalance is the price of the as-rendered treatment definition: classification depends on actual algorithm engagement, which correlates with student characteristics. Because pre-PAL baseline math is balanced and the only available pre-PAL test score covers only 93% of the sample, we do not include baseline math as a regression covariate; including it as a control, particularly the contemporaneous AY 2023–24 SA1 used in earlier drafts, would absorb part of the Year-1 PAL effect rather than pre-randomisation heterogeneity for the Grade 7–9 cohorts (whose AY 2023–24 SA1 was administered after they had begun PAL).

J.C. Extended Results: Heterogeneity, Distribution, and Item-Level Decomposition

Heterogeneity by baseline ability. Because Deep Remediation is randomly assigned within school and grade, an interaction with a pre-treatment test-score measure is causally identified by the experimental design. We use the cohort-specific pre-PAL baseline math score described above and partition students into within-school-by-grade terciles; mean baseline scores in the three tercile bins are 17, 30, and 46 respectively, on a 100-point government scale. Table J2 reports Deep Remediation effects at each tercile. Point estimates are positive at all three terciles—+0.049 SD ($p = 0.23$) at the lowest, +0.060 SD ($p = 0.13$) in the middle, and +0.081 SD ($p = 0.11$) at the highest. None of the per-tercile coefficients reaches conventional significance individually; the joint F-test of equal Deep Remediation effects across terciles is far from rejection ($F = 0.12$, $p = 0.88$): with $\sim$1,150 students per tercile, the data are consistent with a uniform Deep Remediation effect across the baseline-ability distribution. We find no evidence that lower-baseline students benefit disproportionately from access to deeper remediation, despite the prior that they have "more room to gain" from foundational content.

Distributional effects. We complement the tercile analysis with three distributional checks on the standardised outcome itself: an F-test for equal SD across arms, its residualised counterpart on within-school-by-grade-by-section deviations, and a quantile regression of the outcome on Deep Remediation (with the

APPENDIX TABLE J6—DISTRIBUTIONAL EFFECTS OF DEEP REMEDIATION ON θ^{std}

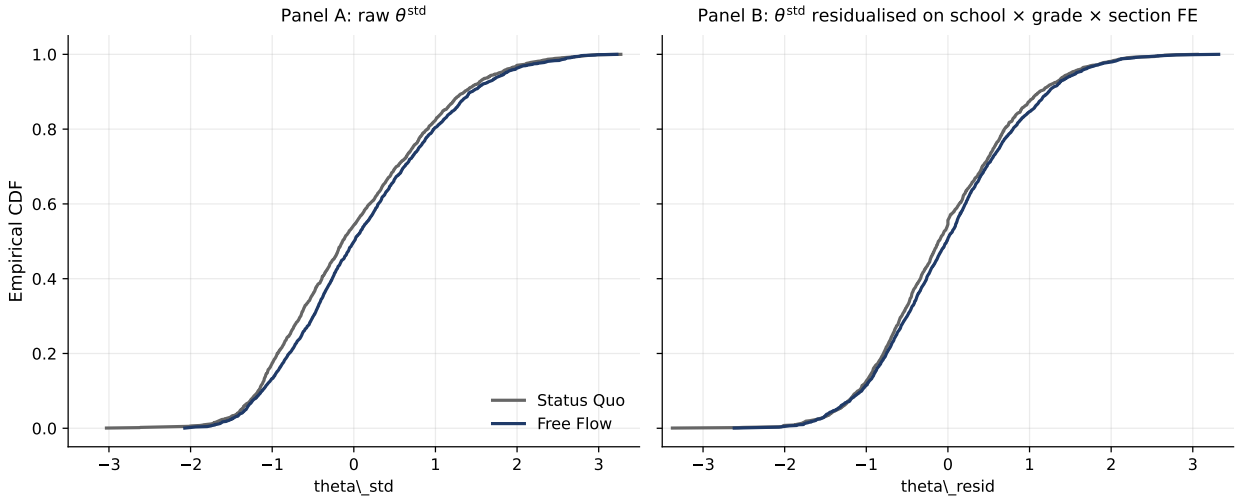
Panel A: F-test on equal SD across arms (variance ratio)					
	SD(SQ)	SD(FF)	F	p	
Raw θ^{std}	1.000	0.990	1.021	0.650	
Residualised on FE	0.888	0.900	1.029	0.543	
Panel B: Quantile regression FF coefficient (cluster-bootstrap SE, $S = 20$ schools)					
	Coef	SE	p	95% CI	N
$q = 0.10$	$+0.028$	(0.040)	0.488	[-0.035, +0.122]	3,755
$q = 0.25$	$+0.083^{***}$	(0.027)	0.002	[+0.022, +0.122]	3,755
$q = 0.50$	$+0.049^{*}$	(0.027)	0.071	[+0.001, +0.101]	3,755
$q = 0.75$	$+0.039$	(0.039)	0.322	[-0.031, +0.120]	3,755
$q = 0.90$	$+0.032$	(0.050)	0.516	[-0.061, +0.139]	3,755

Notes: Panel A reports two-sided variance-ratio (Snedecor) F-tests for equality of SD(θ^{std}) across arms. Row 1 uses raw θ^{std}; row 2 first residualises θ^{std} on the school × grade × section fixed effects and tests the within-stratum residual SD. Panel B reports the FF coefficient from quantile regressions of θ^{std} on Deep Remediation + the covariate vector of Appendix J.B + school × grade × section FE at five quantiles. Standard errors and 95% CIs are pairs cluster-bootstrap (200 replications, schools resampled with replacement, $S = 20$). * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

covariate vector and fixed effects) at $q \in \{0.10, 0.25, 0.50, 0.75, 0.90\}$, with cluster-bootstrap SE (200 reps; schools resampled with replacement, $S = 20$). Table J6 and Figure J6 report the results. The arm SDs are essentially equal—raw 1.000 versus 0.990 ($F = 1.02$, $p = 0.65$); residualised 0.888 versus 0.900 ($F = 1.03$, $p = 0.54$). Quantile regression localises the gain at the lower-middle of the distribution: the FF coefficient is $+0.083$ SD at $q = 0.25$ ($p = 0.002$, 95% CI $[+0.022, +0.122]$) and $+0.049$ SD at the median ($p = 0.07$, $[+0.001, +0.101]$); the $q = 0.75$ point estimate is similar in magnitude ($+0.039$ SD) but not distinguishable from zero ($p = 0.32$), and both tails are indistinguishable from zero ($q = 0.10$: $+0.028$, $p = 0.49$; $q = 0.90$: $+0.033$, $p = 0.52$). The empirical CDFs separate visibly across the lower-to-middle range and converge at the top, matching the quantile pattern. We read this as compatible with the broadly-uniform tercile pattern: the gain is positive across the distribution but slightly larger where students sit somewhat below the within-stratum median rather than at the tails.

Item-level decomposition. For each student we compute three sub-abilities using the same 2PL item parameters from the primary IRT fit, conditioning each estimate on items in only one content bucket relative to the student's enrolled grade: at-grade items (g), shallow-remediation items ($g-1$ and $g-2$), and deep-remediation items ($g-3$ through $g-5$). Each sub-ability is standardised within enrolled-grade cohort, so Deep Remediation coefficients report in SD units of the relevant cohort's distribution of that sub-ability. Table J7 reports the full estimates underlying Figure 4 in the main text. A stacked test of equal Deep Remediation effects across the three buckets returns $F = 1.53$ ($p = 0.242$); the pairwise contrast of shallow versus at-grade effects is $+0.041$ SD ($p = 0.09$), suggestive but not statistically definitive at conventional levels.

Two features of the grade 6 pattern deserve emphasis. First, grade 6 Deep Remediation students engage with roughly 200 minutes of $g-3$-and-below material per year—nearly twice the per-year rate of Grade 7–9 Deep Remediation students—yet the grade 6 gain on the very items that match this distinctive dose ($+0.09$ SD on Grade-2 and Grade-3 items) is not statistically detectable. Second, the *largest* measured gain is on items one to two grades *above* the deepest content they engaged with: the gain on Grade-4 and Grade-5 items ($+0.16$ SD) is roughly twice the gain on the Grade-2 and Grade-3 items where the distinctive content was rendered. Three readings are consistent with this spill-up pattern. The deep content may build prerequisite



APPENDIX FIGURE J6. EMPIRICAL CDFS OF θ^{std} BY ARM

Notes: Panel A: empirical CDFs of raw θ^{std}, separately for Deep Remediation (navy) and Status Quo (grey). Panel B: same after residualising θ^{std} on school × grade × section fixed effects. The Deep Remediation CDF lies visibly to the right of Status Quo through the lower-to-middle range and converges at the upper tail, consistent with a treatment effect concentrated below the median rather than a strict mean-shift. Companion to Table J6.

APPENDIX TABLE J7—DEEP REMEDIATION EFFECT ON ENDLINE PERFORMANCE BY CONTENT GRADE OF TEST ITEM

	All grades (1)	Grade 6 (2)	Grade 7 (3)	Grade 8 (4)	Grade 9 (5)
At-grade items	+0.016	+0.066	+0.013	-0.056	+0.037
	(0.028)	(0.079)	(0.055)	(0.060)	(0.056)
Shallow-remediation items ($g-1$, $g-2$)	+0.072***	+0.164***	+0.044	+0.031	+0.052
	(0.023)	(0.060)	(0.054)	(0.057)	(0.050)
Deep-remediation items ($g-3$ and below)	+0.051**	+0.089	+0.075	+0.018	+0.029
	(0.022)	(0.068)	(0.068)	(0.054)	(0.067)
Joint F (equal FF effect across buckets, pooled)	$F = 1.53$, $p = 0.242$				
Students	3,755	832	883	984	1,056
Items per bucket (G6/G7/G8/G9)	at: 10 / 10 / 6 / 6; shallow: 20 / 20 / 20 / 16; deep: 20 / 30 / 40 / 50				

Notes: Deep Remediation effects on IRT sub-abilities computed from the 72-item Math Master Challenge. For each student we compute three sub-abilities using the globally fitted 2PL item parameters and conditioning each estimate on only items in one content bucket relative to the student's enrolled grade: at-grade (g), shallow-remediation ($g-1$, $g-2$), and deep-remediation ($g-3$ through $g-5$). Each sub-ability is standardised within enrolled-grade cohort. Specification: OLS of the sub-ability on Free Flow plus the covariate vector of Appendix J.B and school-by-grade-by-section fixed effects; standard errors clustered at the school level ($S = 20$). The joint F-test asks whether Deep Remediation effects are equal across the three buckets when all three are stacked into a single student-by-bucket panel and Deep Remediation is interacted with bucket indicators. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

skills—basic arithmetic, place value, whole-number operations—that unlock Grade-4 and Grade-5 content (fraction operations, introductory algebra) on which grade 6 students sit closer to a competency threshold. Mean percent correct on grade 6 students' Grade-2–Grade-3 items (around 0.54–0.58) is already high enough that additional gain has less room to register on the easier items; Grade-4–Grade-5 items at 0.45–0.51 have

APPENDIX TABLE J8—EFFECT OF *assignment* TO DEEP REMEDIATION ON ENDLINE PERFORMANCE

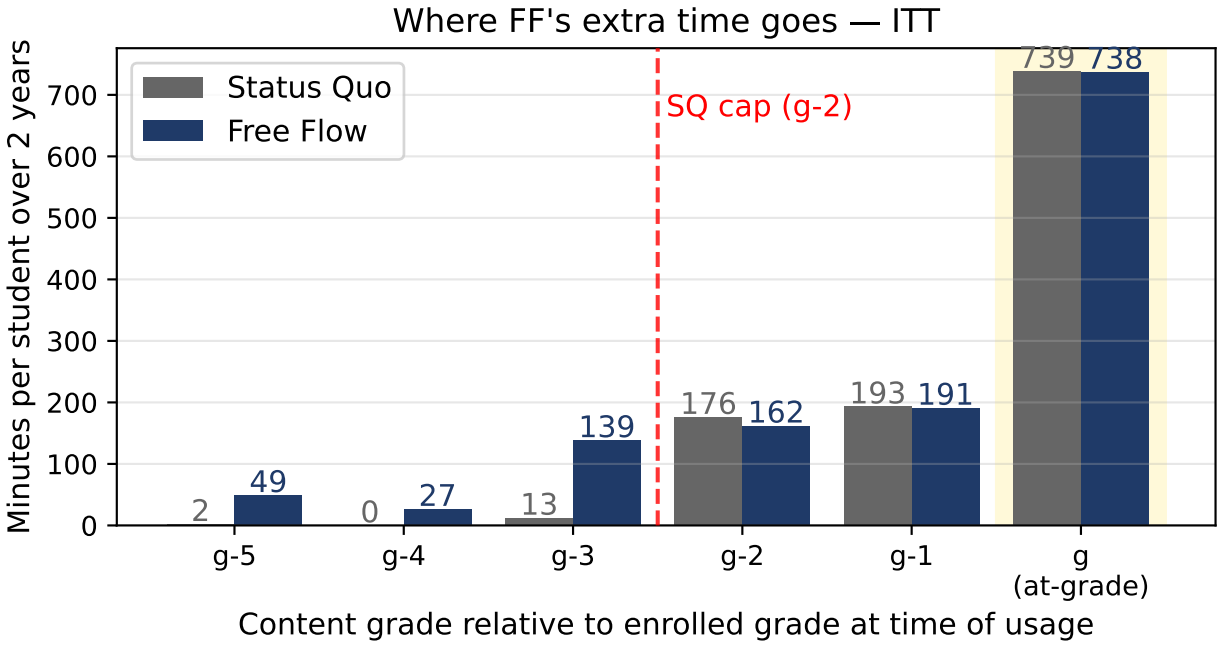
	All grades	Grade 6	Grade 7	Grade 8	Grade 9
	(1)	(2)	(3)	(4)	(5)
Free Flow	+0.046	+0.121**	+0.009	+0.034	+0.033
	(0.032)	(0.059)	(0.085)	(0.050)	(0.052)
Status Quo mean	+0.037	-0.318	+0.114	+0.033	+0.242
Observations	3,263	687	786	883	907
R^2	0.228	0.255	0.249	0.210	0.173

Notes: This is the version of the table with assignment to treatment

more headroom. And the at-grade grade 6 items at 0.42 may sit too far above the foundations the experiment delivered for gains to transfer within the available exposure time. The design does not separate these channels, but the pattern—improvement one to two content-grades above where the distinctive dose landed, with no corresponding improvement at enrolled-grade level—is coherent with a finite per-year transfer distance.

The compliance composition documented in Table J4 maps onto these item-level effects with one revealing exception. Where Deep Remediation students complete more of the deep-remediation prescription (+37.5 percentage points), outcomes follow on deep-remediation items (+0.05 SD). Where Deep Remediation students complete fewer at-grade activities (−3.2 pp), no detectable arm difference appears on at-grade items. The exception is shallow remediation: Deep Remediation students complete fewer prescribed shallow-remediation activities (−3.6 pp) yet score higher on shallow-remediation items (+0.07 SD). This pattern is consistent with backward skill transfer from deep below-grade content to shallow below-grade items: students who complete grade-3 division activities also build the prerequisite skills measured by grade-5 fraction items, even without specifically completing more grade-5 prescribed activities. The deep-cap arm's gain on shallow items therefore reflects skill diffusion rather than direct dose, and the engagement-to-outcome mapping is asymmetric across content depths.

APPENDIX FIGURE J7. IMPACT OF *assignment* TO DEEP REMEDIATION ON CONTENT SERVED



Appendix K. Translating the ITT into the ASER 2024 Classification

This appendix maps the items in our Math Master Challenge to the foundational-skill classification used by the ASER survey, replicates the classification on our control endline, and re-expresses the two-year ITT in the same categorical language used by ASER and by national policy audiences.

ASER Centre (2024) reports for rural Andhra Pradesh grade 8 that 48% of students can perform basic division (down from 52% in 2022), 30.9% can subtract but cannot divide, and 14.2% are stuck at the "recognize 11–99" floor. We replicate the classification on our control endline samples to anchor our control arm's foundational-skill profile against this published reference, and translate the two-year ITT into the same categorical language.

Anchor items. We identify items in our test by *operation content* rather than item-grade label, since calibrated IRT difficulty in our instrument is not monotonic in the question's grade tag. The subtraction anchor (`g2q3`, "Lata has 12 chocolates and gives 4 to her friend") is a word problem without borrowing. The division anchors (`g3q3`, $250 \div 10$; `g3q5`, $65 \div 5$) are clean two-digit-by-one-digit divisions without remainders, materially easier than ASER's three-digit-by-one-digit-with-remainder protocol.

Headline. Two years of PAL raise the share of treated grade 7–8 students with full grade-3 arithmetic mastery from 54.8% in control schools to 61.5% in treatment schools, a 6.7-percentage-point gain that more than reverses the 4-percentage-point decline in basic-division mastery ASER documents for rural AP grade 8 between 2022 and 2024 (52% $\rightarrow$ 48%). The shift breaks down into roughly equal parts: -3.3 pp move off the "cannot subtract" floor, and -3.5 pp clear the subtraction-to-division cliff—both flowing into the "both-operations" bucket.

Why control levels differ from ASER. Our control "both-operations" rate (54.8%) is 6.8 pp higher than ASER's "can divide" rate (48%), consistent with our division anchor being easier than ASER's. Our higher "cannot subtract" rate (28.8% vs 14.2% at ASER's 11–99 floor) is consistent with two contrasts: our 100%-government-school sample lacks the higher-performing rural private-school enrollees ASER's household-based survey includes, and our subtraction item is a word problem requiring reading comprehension while ASER's is bare arithmetic.

APPENDIX TABLE K1—ASER-EQUIVALENT CLASSIFICATION, CONTROL VS TREATMENT ENDLINE

	Subtract, not divide	Cannot subtract	Both ($\geq$ G3 arithmetic)	*n*
Y2 grade 7–8 control endline	16.4%	28.8%	54.8%	3,609
Y2 grade 7–8 treatment endline (2yr PAL)	13.0%	25.5%	61.5%	3,368
Treatment − Control	*−3.5 pp*	*−3.3 pp*	*+6.7 pp*	—
ASER 2024 AP grade 8 (ASER Centre, 2024)	30.9%	14.2%	48%	—

Notes: Subtraction anchor: `g2q3` (12−4 word problem). Division anchor: any of `g3q3` (250 ÷ 10) or `g3q5` (65 ÷ 5). ASER "11–99" floor refers to students who can recognize 11–99 but not subtract.

APPENDIX TABLE L1—HETEROGENEOUS TREATMENT EFFECTS BY URBAN/RURAL STATUS

	Rural (1)	Urban (2)
Treatment	0.456***	0.520***
	(0.057)	(0.151)
Observations	11,615	2,600
p (Rural = Urban)	0.075	

Notes: Each column estimates the ITT within rural or urban schools. p (Rural = Urban) tests equality of treatment effects from a pooled interaction model. All regressions include controls, strata FE, grade FE, and school-clustered SE. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

APPENDIX TABLE L2—TREATMENT EFFECT MODERATION BY SCHOOL RESOURCES

Moderator	Treatment × Moderator	SE	p
Enrollment (G6–9)	-0.134	(0.084)	0.109
Math teachers	-0.126	(0.078)	0.107
Total teachers	-0.143*	(0.077)	0.063
Classrooms	-0.043	(0.060)	0.474
Functional tablets	-0.066	(0.076)	0.380
Observations	14,215		

Notes: Each row shows the interaction of treatment with a standardized school-level moderator (mean 0, SD 1) from a separate regression. All regressions include treatment, the moderator, controls, strata FE, grade FE, and school-clustered SE. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Appendix L. Supplementary Results

L.A. Additional Heterogeneity

Tables L1 and L2 examine treatment effect heterogeneity by school urbanicity and infrastructure resources, respectively; effects are similar across rural and urban schools (Table L1) and across schools with above- versus below-median infrastructure resources (Table L2). Treatment-effect heterogeneity by caste cannot be estimated because caste records in government enrollment data are available for treatment students only; Table L3 therefore reports treatment-group mean math scores by caste category for descriptive context.

L.B. Gender Diagnosis

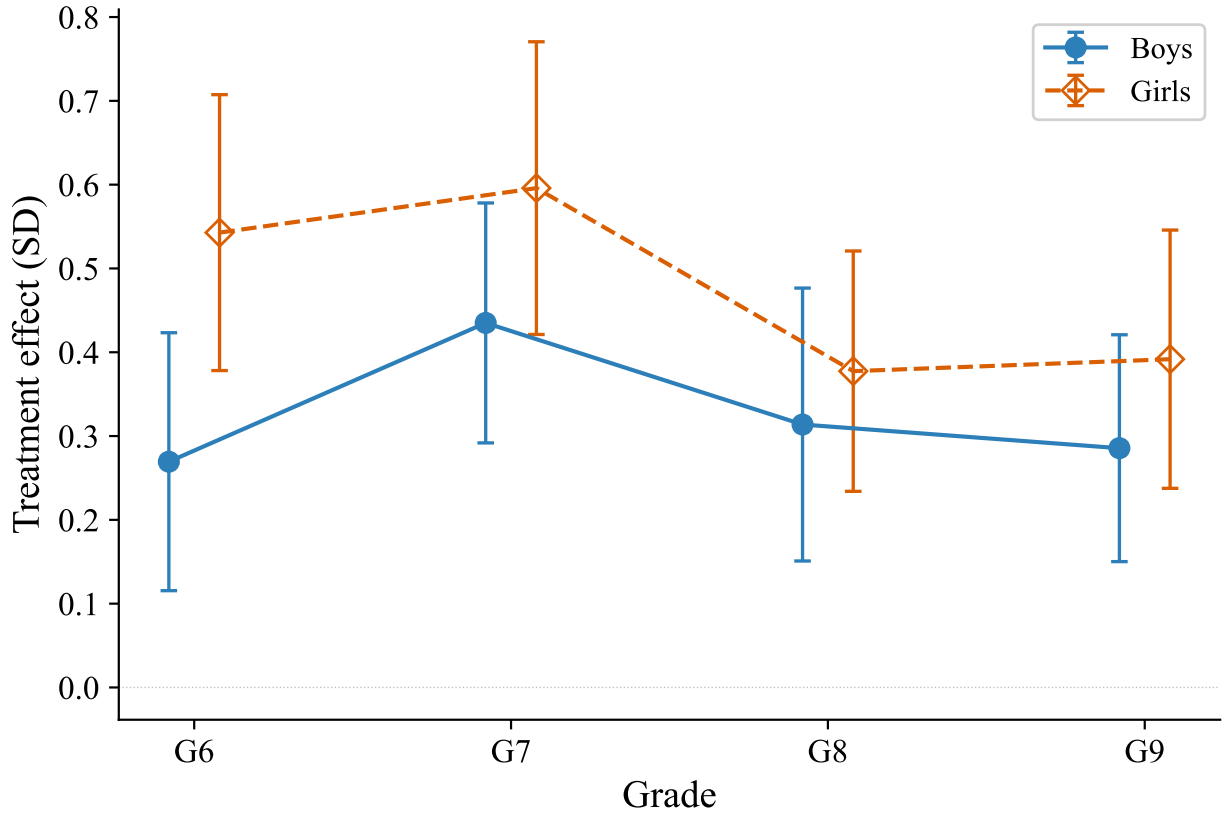
Figure L1 shows that girls gain more than boys in every grade, with the largest gap at grade 6. Figure L2 presents a comprehensive diagnostic of the gender gap in treatment effects, following a three-step narrative.

Step 1: Rule out. We test whether the gender gap can be explained by pre-existing differences. Baseline ability is balanced across gender within treatment status (Panel C). Attendance differences are small and

APPENDIX TABLE L3—MATH SCORES BY CASTE CATEGORY (TREATMENT GROUP ONLY)

	SC	ST	BC-A/Other	BC-B/OBC
Mean IRT score (SD)	0.332	0.769	-0.047	0.430
	(1.066)	(1.211)	(1.059)	(1.150)
N	1,928	808	251	3,669
F-test (equality)		$F = 43.91$, $p = 0.000$		

Notes: Treatment group only. Caste information from government enrollment records, available for treatment students only. Scores are IRT math scores standardized to the overall control group mean and SD. Standard deviations in parentheses. SC = Scheduled Caste, ST = Scheduled Tribe, BC = Backward Class. *F*-test for equality of means across categories.

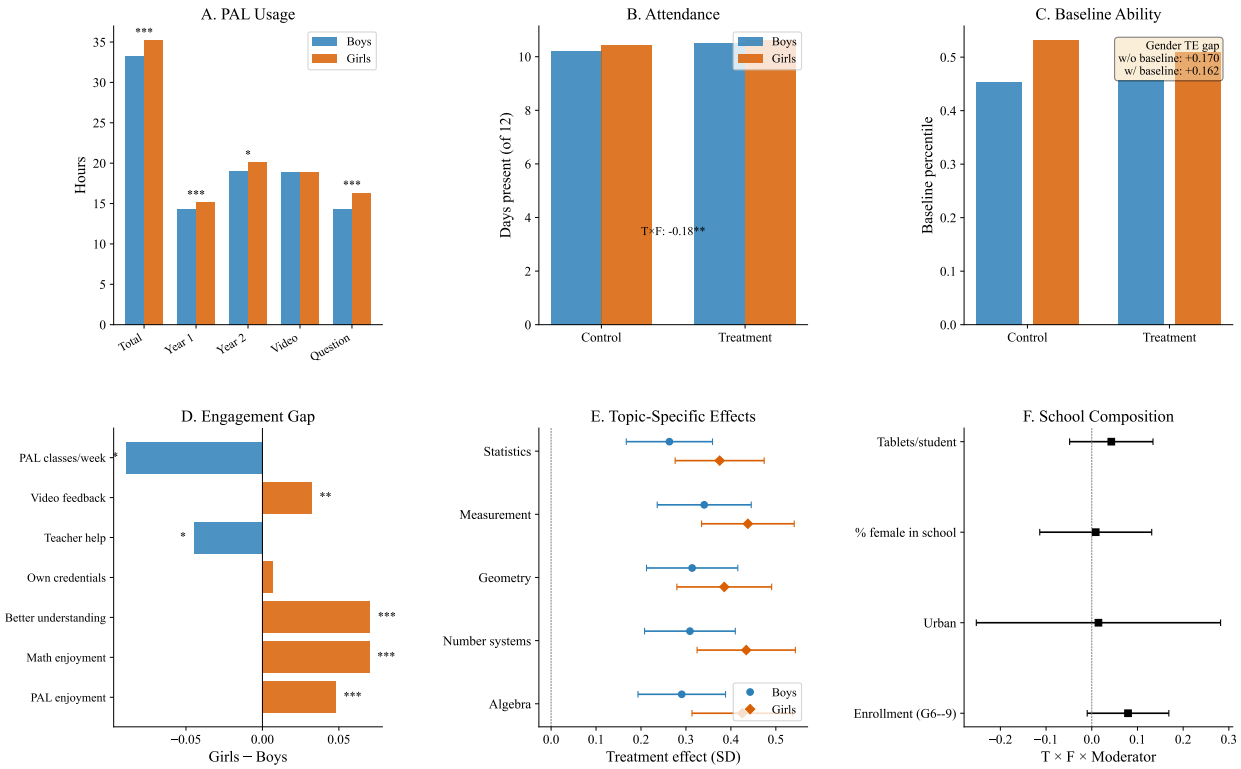


APPENDIX FIGURE L1. TREATMENT EFFECTS BY GENDER AND GRADE

Notes: ITT estimates of PAL treatment effects by gender and grade. Each point shows the treatment effect for a gender × grade subgroup, with 95% confidence intervals. Girls gain more than boys in every grade, with the largest gap at grade 6 (0.55 vs 0.27 SD). All regressions include demographic controls, strata FE, and school-clustered SE.

work against girls, not in their favor (Panel B). School composition—enrollment, urbanicity, percent female, tablets per student—does not moderate the gender gap (Panel F). None of these channels account for the differential treatment effect.

Step 2: Identify. Girls spend approximately 2 more hours on PAL than boys, driven entirely by more time answering questions—active practice—rather than watching videos (Panel A). Girls also report higher



APPENDIX FIGURE L2. GENDER DIAGNOSIS: FULL MECHANISMS PANEL

Notes: Panel A: PAL usage by gender (treatment group). *Panel B*: Attendance by gender and treatment status. *Panel C*: Baseline ability by gender and treatment status. *Panel D*: Engagement gap (girls − boys, treatment group). *Panel E*: Topic-specific treatment effects by gender with 95% CIs. *Panel F*: School composition moderators (T×F×Moderator). All regressions include strata FE and school-clustered SE.

enjoyment of both PAL and math more broadly (Panel D). The gender-stratified IV analysis (Table 4, columns 6–7) confirms that boys and girls have similar per-hour returns (≈0.011 SD/hour), indicating the gap reflects differences in usage quantity rather than per-hour productivity.

Step 3: Assess breadth. The gender advantage is broad-based across all five math topics rather than concentrated in specific areas (Panel E), consistent with a general engagement mechanism rather than topic-specific content effects.

See Table C2 for the full regression results underlying each panel.

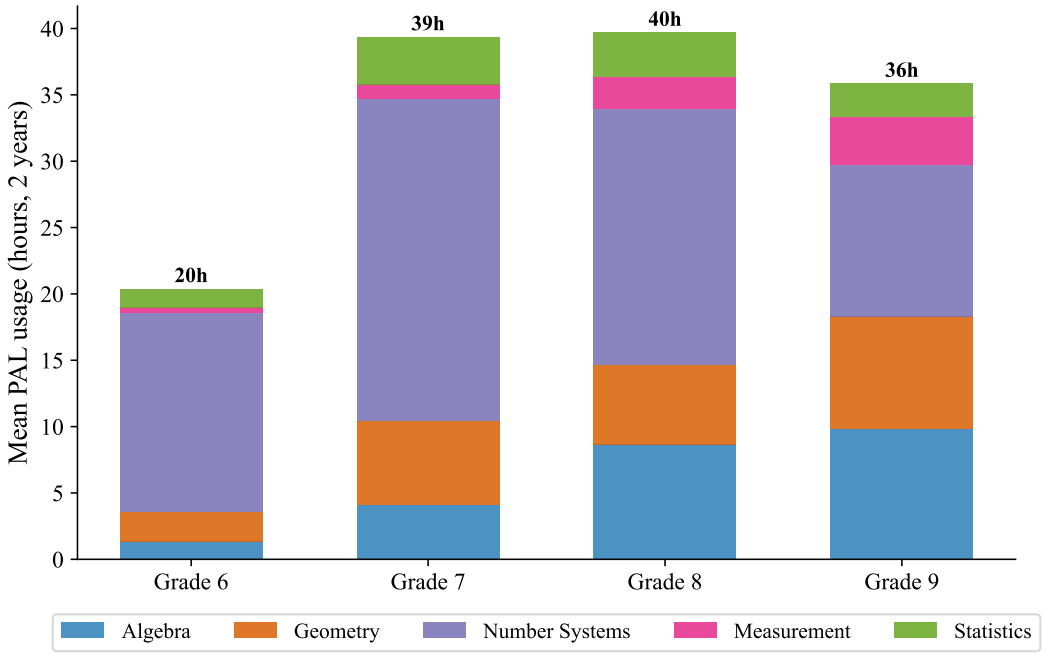
L.C. Descriptive Usage Statistics

Baseline ability and age are the strongest predictors of PAL usage, while SES is not ($p = 0.32$; Table 2). Algebra accounts for the largest share of usage across all grades, consistent with the assessment blueprint weighting (Figure L3).

L.D. Compliance Descriptives

This subsection adds a descriptive companion to our analysis of quality of usage: the relationship between platform hours and compliance.

Hours on the platform and per-student compliance are correlated ($r = 0.57$, $N = 6{,}052$ treatment students with both measures) but not interchangeable (Figure L4). The decile-mean trace rises steeply over the first 30



APPENDIX FIGURE L3. PAL USAGE BY MATH TOPIC AND GRADE

Notes: Mean PAL usage hours (cumulative over 2 academic years) by math topic and student grade. Treatment group only. Grade 6 had 1 year of PAL exposure; grades 7–9 had 2 years. Algebra accounts for the largest share of usage across all grades, consistent with the assessment blueprint weighting.

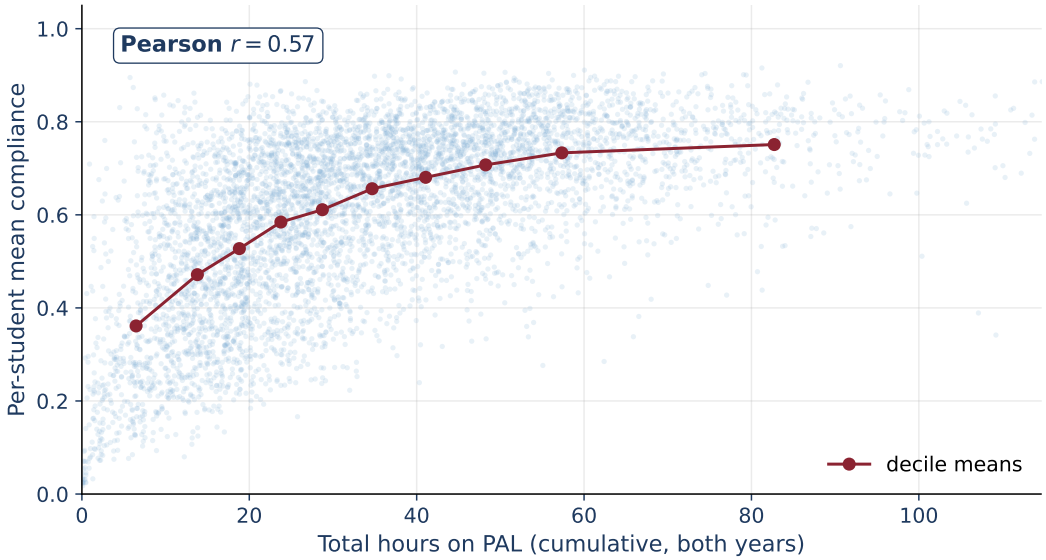
hours of usage and then flattens, while individual-student variation in compliance at any given hours decile remains substantial throughout. The two metrics capture different margins of engagement: hours measure raw exposure during PAL periods, while compliance measures the fraction of the prescribed sequence the student walks. Two students with the same total hours can differ in compliance by 0.4 or more, depending on whether they complete most of what they open or click through and bail.

L.E. Secondary Outcomes and Item-Level Texture

PAL increases math enjoyment by 0.19 SD, with positive spillovers to other subjects (0.10–0.15 SD; Table L4, Panel A). Treatment students also report 0.22 fewer days absent over the last two weeks ($p = 0.003$; Panel B), though strong ceiling effects in enjoyment and the self-reported nature of attendance warrant caution in interpretation.

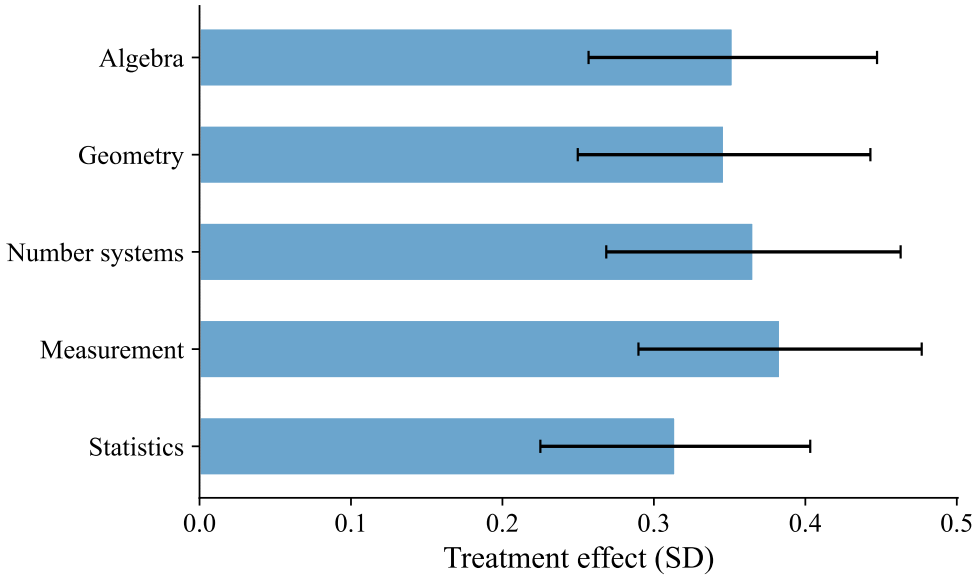
Treatment effects are remarkably uniform across all five math topics (0.31–0.38 SD; Table L5; Figure L5), indicating that PAL improves foundational math skills broadly rather than producing topic-specific gains. This uniformity is consistent with PAL's curriculum-aligned design, which covers all topics proportionally (Figure L3).

We examine treatment effects at the individual item level. Treatment effects are slightly larger on harder items (slope $= 0.10$; Figure L6), suggesting PAL helps students reach questions they would otherwise get wrong rather than reinforcing already-known material. Treatment students outperform control students at every content grade level, with the largest gains on grade 4–6 content (7–10 pp; Figure L7). Both groups show the expected decline in accuracy as content difficulty increases.



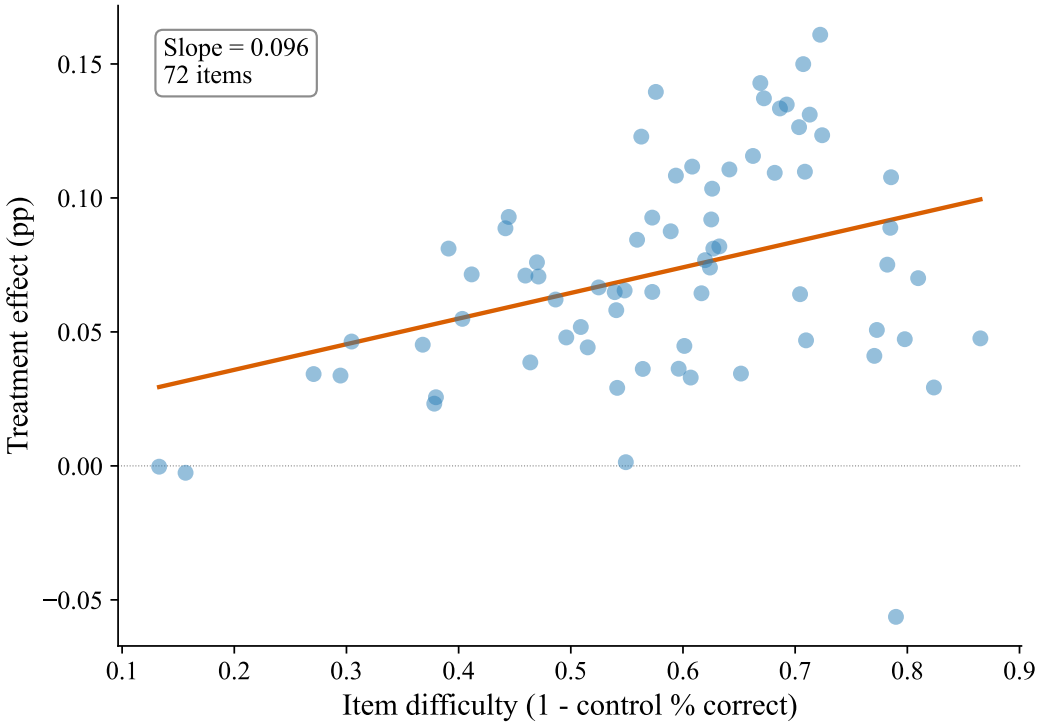
APPENDIX FIGURE L4. TOTAL PAL HOURS AND PER-STUDENT COMPLIANCE

Notes: Per-student total PAL usage (sum of 2023–24 and 2024–25 minutes, converted to hours) on the x-axis against per-student mean compliance on the y-axis, treatment students only ($N = 6{,}052$). Light points are individual students; dark red trace shows decile means of compliance plotted against decile means of hours. Pearson $r = 0.57$. Hours and compliance are correlated but not interchangeable: at any given hours decile, students vary substantially in the fraction of prescribed activities they complete, and the compliance–hours gradient flattens above ~ 30 hours, consistent with engagement having a separate margin from exposure.



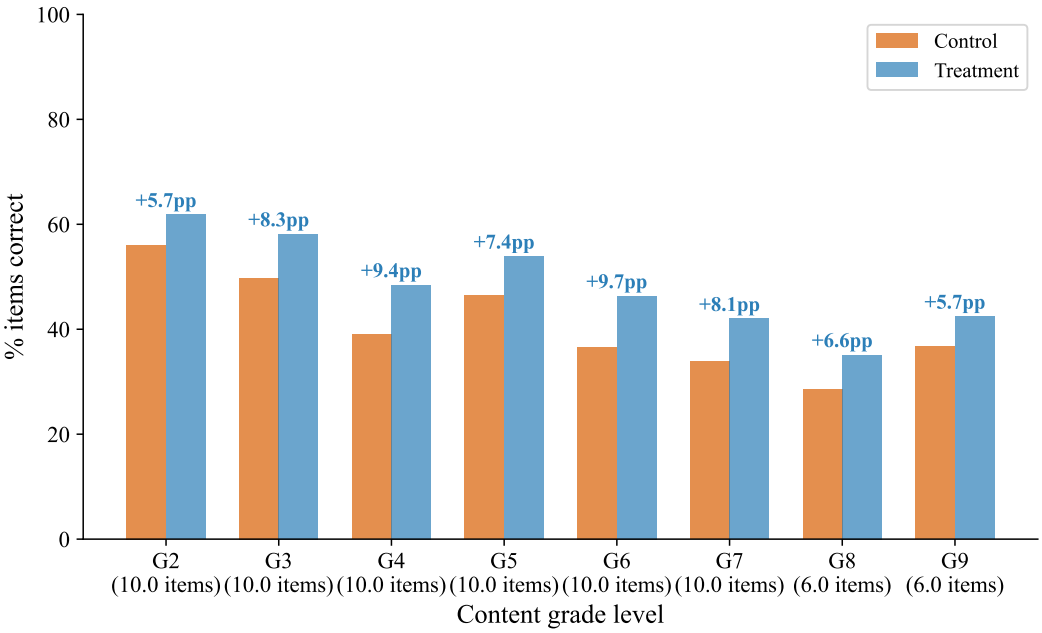
APPENDIX FIGURE L5. TREATMENT EFFECTS BY MATH TOPIC

Notes: Pooled ITT estimates for each of 5 math topics. Bars show point estimates; whiskers show 95% confidence intervals. Effects are remarkably uniform across topics (0.31–0.38 SD), indicating PAL improves foundational math skills broadly rather than topic-specifically. All regressions include controls, strata FE, grade FE, and school-clustered SE.



APPENDIX FIGURE L6. ITEM DIFFICULTY AND TREATMENT EFFECTS

Notes: Each dot represents one of 72 endline assessment items. The *x*-axis shows item difficulty (1 minus control group percent correct); the *y*-axis shows the treatment effect on that item (percentage point difference, T−C). The fitted line shows the relationship between difficulty and treatment effect magnitude. All regressions include strata FE and school-clustered SE.



APPENDIX FIGURE L7. CONTENT MASTERY BY GRADE LEVEL: TREATMENT VS. CONTROL

Notes: Percent of items answered correctly by content grade level (grade 2 = easiest, grade 9 = hardest). Items are organized by the grade-level curriculum they test. Treatment students outperform control students at every content grade level, with the largest gaps on grade 4–6 content (7–10 pp). Both groups show the expected decline in accuracy as content difficulty increases.

APPENDIX TABLE L4—TREATMENT EFFECTS ON SECONDARY OUTCOMES

	Treatment effect	SE	Control mean
Panel A: Subject enjoyment (standardized)			
Math	0.190***	(0.025)	1.44
English	0.131***	(0.026)	1.40
Telugu	0.105***	(0.032)	1.34
Hindi	0.095	(0.065)	1.63
Science	0.144***	(0.029)	1.45
Social science	0.140***	(0.035)	1.53
Panel B: Attendance			
Days present (SD)	0.104***	(0.035)	10.31
Days absent	-0.22***	(0.07)	1.7
Observations		14,215	

Notes: Panel A: enjoyment rating for each subject, reverse-coded and standardized to control mean and SD (original scale: 1=A lot, 2=Somewhat, 3=Not at all, 4=Did not have; positive coefficients = treatment students enjoy the subject more). Strong ceiling effects — nearly all students report high enjoyment regardless of treatment. Panel B: days present out of 12 school days in last 2 weeks, self-reported by students during endline assessment ("In the last 2 weeks, how many days were you absent from school, excluding Sundays and holidays?"). Covers a 2-week window, not the full academic year. Days absent = 12 − days present (raw count). All regressions include controls, strata FE, grade FE, and school-clustered SE. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

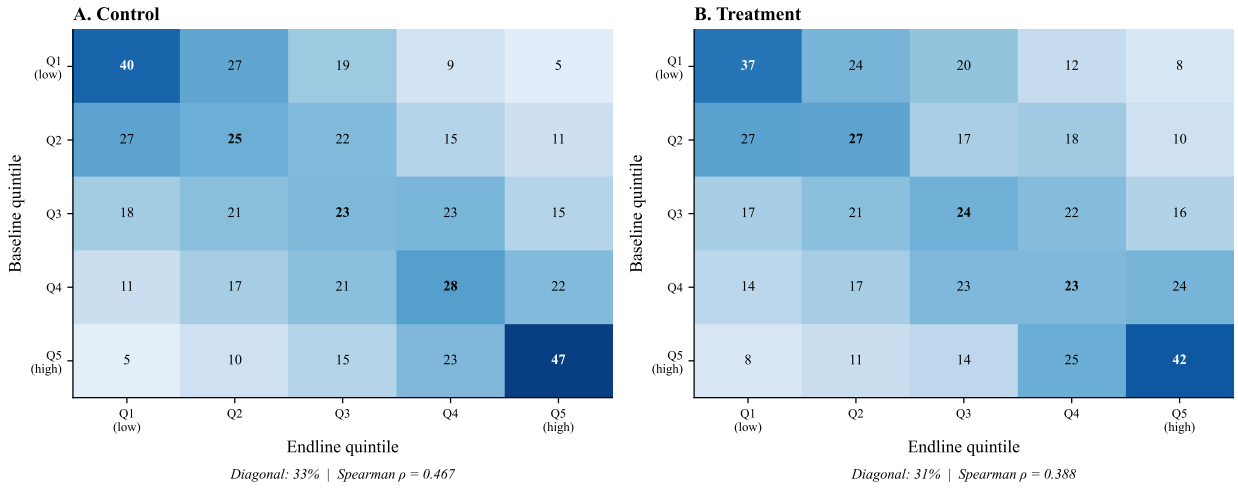
L.F. Rank Invariance and Conditional Quantile Effects

Treatment students are more reshuffled across the ability distribution than control students (Figure L8), violating rank invariance and motivating the conditional quantile analysis. Conditioning on baseline quintile, treatment effects increase from the 10th to 75th percentile of the endline distribution within every quintile, then flatten at the 90th (Figure L9)—indicating that PAL creates within-group variance in addition to the across-quintile gradient documented in Table C1.

APPENDIX TABLE L5—TREATMENT EFFECTS BY MATH TOPIC

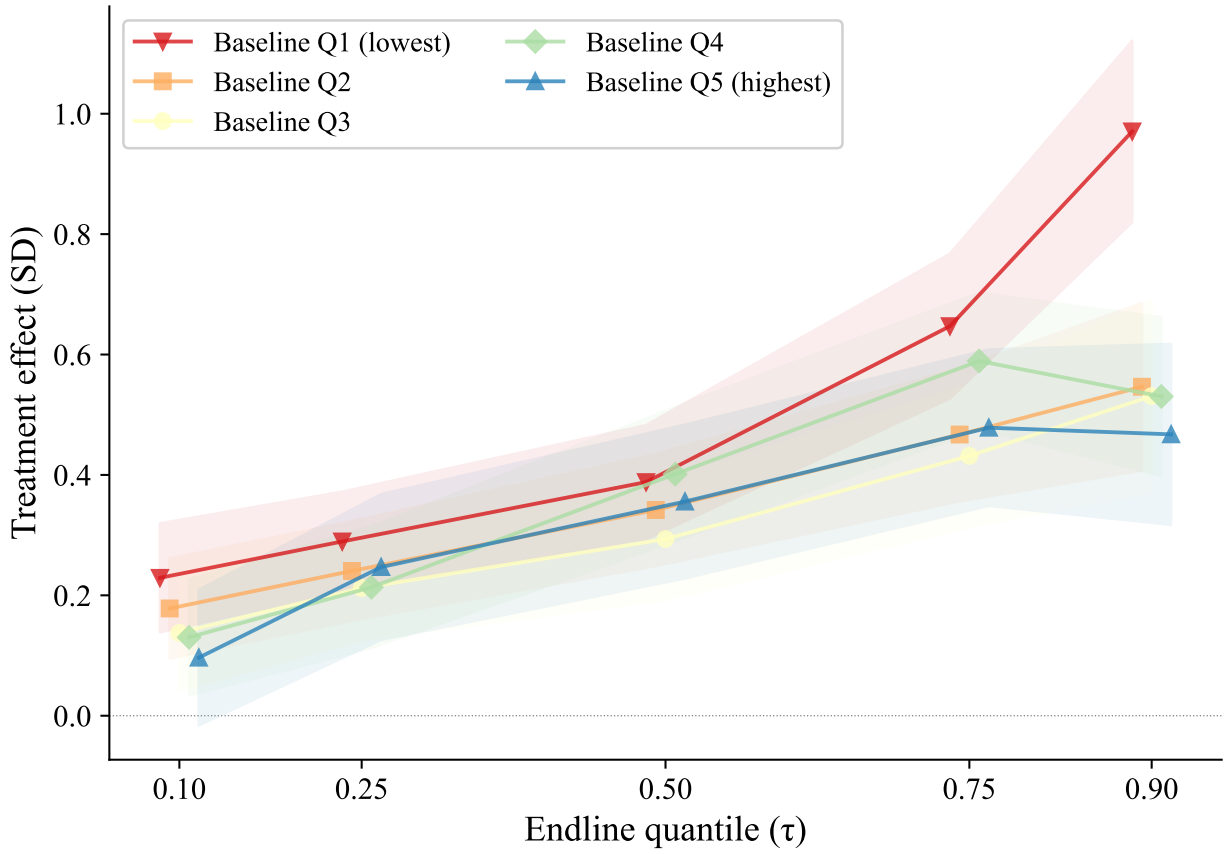
	Pooled (1)	Grade 6 (2)	Grade 7 (3)	Grade 8 (4)	Grade 9 (5)
Algebra	0.352***	0.345***	0.458***	0.327***	0.294***
	(0.049)	(0.067)	(0.070)	(0.059)	(0.054)
Geometry	0.346***	0.353***	0.441***	0.317***	0.286***
	(0.049)	(0.068)	(0.067)	(0.063)	(0.056)
Number systems	0.366***	0.397***	0.471***	0.320***	0.294***
	(0.050)	(0.069)	(0.065)	(0.061)	(0.058)
Measurement	0.383***	0.393***	0.493***	0.325***	0.322***
	(0.048)	(0.075)	(0.063)	(0.061)	(0.052)
Statistics	0.314***	0.417***	0.451***	0.245***	0.188***
	(0.045)	(0.070)	(0.066)	(0.056)	(0.053)
Controls	Yes	Yes	Yes	Yes	Yes
Strata FE	Yes	Yes	Yes	Yes	Yes
Grade FE	Yes				
Observations	14,215	3,313	3,354	3,623	3,925

Notes: Each row shows the ITT estimate for a different math topic. Topic scores are constructed from IRT subscores on items mapped to each topic, standardized to control mean and SD. Topics are not equally weighted in the overall score — algebra has 2–3× more items than statistics. All regressions include controls, strata FE (and grade FE for pooled). School-clustered SE. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.



APPENDIX FIGURE L8. RANK INVARIANCE CHECK: BASELINE TO ENDLINE TRANSITION MATRICES

Notes: Each cell shows the percentage of students in a baseline quintile (row) who end up in a given endline quintile (column). Bold diagonal entries show the percentage staying in the same quintile. If rank invariance held perfectly, all mass would be on the diagonal. Treatment students are more reshuffled (diagonal average 31%) than control (33%), with a lower Spearman rank correlation (0.36 vs 0.47). This violation of rank invariance motivates the conditional quantile analysis in Figure L9. Sample restricted to students with non-missing baseline scores (72% of sample).



APPENDIX FIGURE L9. CONDITIONAL QUANTILE TREATMENT EFFECTS BY BASELINE ABILITY

Notes: Each line shows quantile treatment effects within a baseline ability quintile, addressing the rank-invariance violation in unconditional quantile regressions (Figure L8). Within every baseline quintile, effects increase from the 10th to the 75th percentile of the endline distribution, then flatten or decline at the 90th. Lines for lower-baseline quintiles sit above those for higher-baseline quintiles, consistent with the across-quintile gradient in Table C1. All regressions include controls, strata FE, and grade FE. Shaded bands show 95% confidence intervals.

Appendix M. Tenth boards: IV for usage

This sections presents results from IV estimates of how greater PAL usage affected tenth board exam test scores using the section size as an instrument. This is not as clean an instrument in this case. Students in bigger sections may have experienced just as much disruption as the students in the smaller sections even though they used PAL less. However, in the next two tables we see that the students in the smaller sections who used PAL more had a marginally lower tenth board exam score than those in the bigger sections. As a result, the 2SLS estimates suggest that higher PAL usage is associated with lower board exam scores.

APPENDIX TABLE M1—GRADE 10 BOARD MATH: IV ESTIMATES USING SECTION-SIZE CONGESTION

	First Stage Usage (hrs) (1)	Reduced Form Board math (2)	OLS Board math (3)	2SLS Board math (4)
Treatment × Section enrollment	-0.178***	0.0037		
	(0.056)	(0.0044)		
PAL usage (hours, Year 1)			0.0141***	-0.0082**
			(0.0040)	(0.0041)
Treatment	23.0***	-0.270		
	(2.3)	(0.176)		
Student controls	No	No	No	No
Strata FE	Yes	Yes	Yes	Yes
First-stage F	10.1			10.1
Observations	4,730	4,730	2,365	4,730

Notes: IV estimates of Year-1 PAL usage on Grade 10 board exam math score, using the same instrument as Table 4 (treatment × section enrollment). Grade 10 students used PAL as grade 9 in AY 2023–24, so the instrument is grade-9 *section* enrollment in 2023–24 (from the JVK government enrollment roster), the congestion year for this cohort. The dependent variable in columns (2)–(4) is the board math score standardized to the control mean and SD. Column (4) reports the fixed-intercept 2SLS (structural equation drops the treatment dummy, anchoring the projection at the control mean at zero usage), matching Table 4; the standard LATE (with treatment dummy) is -0.0209 (SE 0.0246). No student-level controls are available for this cohort (students not in the endline sample), so regressions include only strata FE and the section-enrollment level term. Standard errors clustered at the school level. First-stage $F = 10.1$. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

APPENDIX TABLE M2—GRADE 10 BOARD EXAMS: 2SLS EFFECT OF PAL USAGE BY SUBJECT

	Math (1)	Science (2)	Social Sci. (3)	Lang 1 (4)	Lang 2 (5)	Lang 3 (6)	Composite (7)
PAL usage (2SLS, fixed intercept)	-0.0082**	-0.0056	-0.0078*	-0.0065	-0.0076*	-0.0063	-0.0071*
	(0.0041)	(0.0040)	(0.0040)	(0.0043)	(0.0043)	(0.0042)	(0.0036)
PAL usage (2SLS, standard LATE)	-0.0209	0.0156	0.0038	0.0087	0.0041	0.0122	0.0039
	(0.0246)	(0.0237)	(0.0217)	(0.0244)	(0.0261)	(0.0216)	(0.0200)
First-stage F	10.1	10.1	10.1	10.1	10.2	10.2	10.1
Observations	4,730	4,730	4,727	4,733	4,733	4,730	4,736

Notes: Each column reports the 2SLS effect of one hour of Year-1 PAL usage on the standardized Grade 10 board exam score in that subject, instrumenting usage with treatment × grade-9 section enrollment (AY 2023–24). The first row anchors the projection at the control mean at zero usage (fixed intercept); the second row is the standard LATE. All specifications include strata FE and school-clustered standard errors; no student-level controls are available for this cohort. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.