Can Adaptive Learning Platforms Improve Student Outcomes? Experimental Evidence from the Dominican Republic

Carolina Lopez
Development Research Group, World Bank

Astrid Camille Pineda
World Bank *

July 15, 2026

Click here for the latest version

Abstract

Most students in developing countries receive grade-level instruction despite lacking prerequisite skills, a mismatch that widens learning gaps as students progress. This paper reports evidence from a randomized controlled trial across 38 classrooms and 1,333 ninth-grade students in the Dominican Republic in which classrooms were randomly assigned to one of three arms: one in which computer-adaptive learning (CAL) software replaced two of seven weekly mathematics hours, another where CAL was combined with small-group tutoring, or a business-as-usual control. CAL improved test scores by 0.29–0.31 standard deviations over six months, a magnitude on the larger end of the range of the effects documented for education technology interventions in developing countries. The addition of small-group tutoring yielded no detectable improvement over CAL alone; if anything, classrooms assigned both interventions showed smaller gains than those receiving CAL alone, though this difference is not statistically significant. Evidence on mechanisms indicates that tutoring classrooms exhibited reduced engagement with the CAL platform. These results suggest that CAL software can be effectively integrated into classroom instruction to improve learning at scale but that multiple interventions in combination may not produce additive benefits and may instead generate unintended behavioral responses that offset their effectiveness.

JEL Classification: C93, I21, I28, J24, O15

Keywords: Computer-adaptive learning, education technology, randomized controlled trial, mathematics education, developing countries, Dominican Republic

*We thank the Ministry of Education of the Dominican Republic, especially Ancell Scheker Mendoza, Vice Minister of Technical and Pedagogical Affairs, and the participating schools for facilitating this research. We are especially grateful to Sarah Cohodes, Alex Eble, Alejandro Ganimian, Juan Barón, Matteo Magnaricotte, Peter Bergman, Donald Green, and Chris Neilson for valuable comments and support. Aurora Salvador provided excellent research assistance. The study was funded by the Tinker Foundation and is registered as AEARCTR-0012383 in the AER registry. It was approved under protocol number 24-444 by the Teachers College IRB. Pineda was supported by the Institute of Education Sciences, US Department of Education, through grant R305B200017 to Teachers College Columbia University. The findings, interpretations, and conclusions expressed in this paper are entirely those of the authors. They do not necessarily represent the views of the World Bank and its affiliated organizations, or those of the Executive Directors of the World Bank or the governments they represent. All errors are our own. Lopez: World Bank, carolina_lopez@worldbank.org, Pineda: World Bank, apineda@worldbank.org

1 Introduction

Most students in developing countries receive grade-level instruction despite lacking prerequisite skills, and this mismatch widens learning gaps over time (Muralidharan et al., 2019). The problem is widespread: Less than 20 percent of students in low- and middle-income countries meet minimum proficiency standards in mathematics (World Bank, 2017), and the typical student in Latin America performs approximately five years behind their OECD counterparts (IDB, 2023). When students who lack foundational knowledge are taught content beyond their current level, they fall further behind instead of catching up, contributing to both poor achievement and high dropout rates.

Computer-adaptive learning (CAL) platforms offer a potentially scalable solution to this instructional mismatch. By continuously assessing student knowledge and adjusting content difficulty accordingly, CAL can deliver personalized instruction at a fraction of the cost of one-on-one tutoring. Recent evidence from India demonstrates that CAL can generate substantial learning gains when substituted for regular classroom instruction, with effects ranging from 0.20–0.22 standard deviations (SD) at scale (**?**) to 0.45 SD under intensive implementation support (Oreopoulos et al., 2026). However, as policymakers seek to maximize learning recovery, a critical question remains unexplored: Can combinations of CAL with other effective interventions—such as small-group tutoring—generate complementary benefits, or might they produce unintended interference?

This paper addresses this question through a randomized controlled trial (RCT) integrating CAL into regular mathematics instruction for 1,333 ninth-grade students across 38 classrooms in the Dominican Republic. We employed ALEKS, a CAL platform developed by McGraw Hill and adapted to the Dominican Republic's math curriculum.[1] All students in treated classrooms received licenses to access the software during two of their seven weekly math instruction hours, such that CAL replaced rather than supplemented traditional instruction. Moreover, to test for complementarities between technology-based and traditional instructional approaches, we implemented a second randomization within the treated classrooms, providing up to three weekly hours of small-group tutoring to academically at-risk students. The tutoring sessions were scheduled during existing extracurricular periods to prevent confounding from additional instructional time.

We find that the use of the adaptive software during math classes improved student outcomes by 0.29–0.31 SD by endline (six months after the start of the intervention). These positive effects are robust across model specifications and consistent across the mathematical domains students studied, including algebra and geometry, and difficulty levels, including content below and at grade level. This estimate falls close to the upper end of the magnitude range of documented effects of CAL interventions. The addition of small-group tutoring did not enhance these gains; if anything, classrooms assigned both interventions showed smaller effects than classrooms treated with CAL alone, though the difference is not statistically significant.

[1]The intervention was implemented by the Ministry of Education from 2018 to 2023, with support from the World Bank.

Why did the two treatment arms have different impacts? Implementation fidelity does not explain the difference: Both the CAL and tutoring interventions showed similar compliance rates (approximately 65–68 percent of the intended dosage) and barriers, predominantly school closures and infrastructure failures. Instead, the evidence points to behavioral mechanisms. Students in tutoring classrooms—including nontutees—were less likely to use the adaptive platform and scored lower on a measure of internal locus of control (i.e., belief in their ability to influence their own outcomes) than the CAL-only group. In addition, the tutoring may not have been well targeted, as the students selected by teachers to receive tutoring were not among the lowest performers at baseline, as was intended by the intervention.

This paper makes three contributions to the literature on educational technology in developing countries. First, we provide experimental evidence on CAL as a *substitute* for, rather than supplement to, traditional instruction. Most studies have evaluated CAL in after-school settings or in combination with tutoring (Muralidharan et al., 2019; Guryan et al., 2023), leaving open questions about its effectiveness when it is integrated into regular school schedules or implemented alone (Muralidharan et al., 2019). Recent evidence from India demonstrates that intensive implementation support can generate CAL effects of 0.45 SD when dedicated personnel ensure high-fidelity platform usage (Oreopoulos et al., 2026), but such models require dedicated staff and face scalability constraints; the impact of CAL alone relative to that of additional instructional time remained difficult to disentangle in prior designs (Glewwe and Muralidharan, 2016; Muralidharan and Singh, 2025). By evaluating a program replacing two of students' seven weekly math instruction hours with CAL sessions, we add to a small but growing body of evidence on its efficacy when CAL is implemented within the standard school curriculum, rather than as an after-school add-on (Muralidharan and Singh, 2025).

Second, we provide a rare experimental test of complementarities between technology-based and traditional instruction. Tutoring is among the most effective interventions for improving learning outcomes, with a recent meta-analysis reporting an average effect size of 0.37 SD Nickow et al. (2020), but its scalability is constrained by high costs and the difficulty of recruiting qualified tutors (Kraft and Falken, 2020; Nickow et al., 2020; Kraft et al., 2024). Small-group tutoring and CAL have each been proposed as more scalable alternatives to one-on-one models (Augustine et al., 2016; Guryan et al., 2023), and a natural policy response is to combine them, on the assumption that the two together would yield benefits beyond those of either intervention alone—an assumption rarely tested directly. Our two-stage randomization design isolates the CAL$\times$Tutoring interaction effect, providing causal evidence on whether adaptive technology and human tutoring are complementary or competing inputs into student learning.

Third, we document behavioral mechanisms through which intervention complementarities may fail. Contrary to expectations, CAL's combination with in-person tutoring did not reinforce its effects; instead, the gains from the combined treatment were indistinguishable from zero relative to the effects on the control group. This pattern is associated with reduced platform engagement and a lower degree of internal locus of control among students in the tutoring classrooms, including those not directly tutored, which suggests that the introduction of human support may erode the self-directed engagement that drives CAL's effectiveness. These findings underscore how important it is for the learning effects of bundled interventions to be tested empirically rather than assumed to be additive.

The deepening of the global learning crisis following COVID-19 has highlighted the need for more resilient, scalable approaches to instruction. Our findings show that adaptive learning technology can deliver substantial learning gains even amid the infrastructure constraints typical of developing-country settings. At the same time, the reduced effectiveness we observe for CAL plus tutoring—which is linked to decreased platform engagement and shifts in student agency—shows that layering of interventions does not guarantee additive effects and that successful implementation requires an understanding of how components interact, not simply how many are deployed.

This paper proceeds as follows. In Section 2, we describe the context, intervention, and implementation. In Section 3, we describe our data. In Sections 4 and 5, we present the empirical framework and results. Lastly, in Section 6, we discuss mechanisms, and in Section 7, we conclude.

2 Setting and Experimental Design

2.1 Background

The Dominican Republic exemplifies the mathematics learning crisis facing much of Latin America and the Caribbean (LAC). In 2022, the country ranked third to last globally on the OECD's Programme for International Student Assessment (PISA), with 92 percent of its students classified as low-performing in mathematics (OECD, 2023a,b). The consequences of these learning deficits become particularly acute during the transition to secondary education. The Dominican education system progresses from primary (grades 1–6) through lower secondary (grades 7–9) to upper secondary (grades 10–12), with students typically entering lower secondary around age 12.[2] Research indicates that up to 40 percent of students drop out between 8th and 12th grade (IDEICE, 2017).

Ninth grade represents a critical intervention point in this trajectory. Students at this stage have likely accumulated substantial learning deficits that threaten their ability to complete secondary education, yet they remain within the compulsory education system, where remedial interventions can still be effectively implemented. Our study evaluates an intervention designed to address these challenges by providing targeted remedial mathematics training to ninth-grade students. To the extent that poor academic performance drives dropout decisions, improvements in learning outcomes may also promote educational retention; however, if students face binding nonacademic

[2]The school year runs from late August to late June, organized over two academic semesters. The 2023–24 school year officially started on August 28, 2023, and ended on June 21, 2024.

constraints such as economic pressures or family obligations, learning gains alone may prove insufficient to affect persistence. This study contributes to the broader efforts to disentangle the roles of academic performance and nonacademic constraints in driving dropout in the Dominican Republic and the wider LAC region.

2.2 Intervention

The study and intervention were implemented in partnership with the Ministry of Education of the Dominican Republic (MINERD) and Abdul Latif Jameel Poverty Action Lab (J-PAL).[3] The intervention was implemented between October 2023 and May 2024 in a sample of 38 ninth-grade classrooms in 9 schools.[4] The program had two core elements: a CAL intervention and a tutoring component. The first component introduced the use of a CAL platform (ALEKS) as part of the school's regular schedule, with CAL substituting two out of students' seven weekly hours scheduled for math instruction. Half of the schools' classrooms were randomly selected (with stratification by classroom size) to receive CAL licenses for all students and their math teachers. The second component provided small-group tutoring during extracurricular hours to approximately 20 percent of students in classrooms randomly selected from the first treatment group. Each component is summarized in more detail below.

ALEKS platform. The program provided access to ALEKS, a math-focused CAL software. ALEKS uses predictive algorithms to deliver personalized learning experiences through adaptive assessments and assignments. These tools evaluate students' knowledge by identifying the math topics they have mastered, those they have not, and those that they are ready to learn. This approach is grounded in knowledge space theory (KST), which posits that a subject's knowledge can be mapped as a set of solvable problems (Doignon and Falmagne, 1985; Falmagne et al., 1990). Accordingly, ALEKS assesses students' mastery by their ability to solve representative problems within each topic.

The software encompasses over 600 mathematics topics for ninth graders. Given the infeasibility of covering all topics within an academic year, the curriculum was narrowed to grade-appropriate content and essential prerequisites, which yielded 337 topics organized into nine modules corresponding to content areas scheduled for instruction during the academic year.[5] Beyond practice exercises, ALEKS provides functionality for teachers to assign and evaluate homework and tests, along with a dashboard for tracking class and individual student progress.

All teachers assigned to a CAL classroom received an initial 2-hour training on the use of the learning technology. This training was provided by McGraw Hill, the developer of the learning

[3]The study is registered as AEARCTR-0012383 in the AEA registry and was approved under protocol number 24-444 by the Teachers College IRB.

[4]Figure 1 shows the full timeline of the study and intervention.

[5]This selection was performed by two local experts on math pedagogy on the basis of a review of the curriculum. Most of the prerequisite topics were identified by the learning software.

platform. In addition, during the first 3 months of the program, teachers met one-on-one with a CAL aide for approximately one hour every two weeks. The aide's role was to support them in accessing and monitoring students' assessments, tracking their progress, and identifying topics that students had not yet mastered for lesson planning. Importantly, one of the primary aspects of this extended training was to help teachers utilize the software's specific tools to ease their workload, including features for assignment and automatic grading of homework, thus enabling the teachers to save time that could be devoted to other teaching activities.

Tutoring Intervention. The second component was designed to reveal any complementarities between CAL and traditional instruction through the provision of up to three weekly hours of tutoring to academically at-risk students. We targeted approximately 20 percent of the students in each classroom selected for tutoring for this intervention. Sessions were managed by one or two tutors, depending on the number of tutees (typically 5–8) per classroom. To select tutoring recipients, mathematics teachers initially identified and ranked up to 10 students on perceived academic need. To minimize selection bias, we subsequently provided the teachers with baseline mathematics test results and gave them the opportunity to revise their selections, which approximately 70 percent of the teachers opted to do. We used the final teacher selections to determine tutoring assignments. Despite this safeguard, the students ultimately selected for tutoring were not, on average, among the lowest baseline performers in their classrooms—a targeting gap we revisit in Section 6, where we consider its implications for the program's effectiveness.

Tutors were selected through a competitive process that included interviews. All tutors were either recent graduates or in the final stages of earning a bachelor's degree in mathematics with a focus on secondary education and had experience using information technology for teaching. The tutoring sessions were integrated into students' schedules during elective workshop periods, covering approximately three academic hours per week, typically divided into one to three weekly sessions.[6] Elective workshops, typically on subjects outside the core curriculum such as computing or art, are offered for all Dominican students attending extended-day schools.[7]

Tutors received support to ensure quality and consistency from a supervisor with training and experience in mathematics education for high school students. The supervisor reviewed individual session plans before each session, confirmed that topics and instructional methods had been prepared, and provided feedback to tutors as needed. Tutors also communicated closely with the mathematics teacher to align their sessions with the classroom curriculum and were instructed to incorporate the CAL platform into tutoring sessions.[8] During the sessions, tutors reviewed

[6] Out of the 9 tutoring classrooms, 2 sections had one weekly meeting of 3 hours, 5 sections had two weekly meetings of 2 hours and then 1 hour, and 2 sections had three weekly meetings of 1 hour each.

[7] During the 2021-2022 school year, approximately 67 percent of public high schools in the Dominican Republic offered an extended school day format (MINERD, 2023), where the school day starts at 8 am and ends at 4 pm.

[8] For example, tutors were encouraged to use the platform to identify topics on which students lagged behind and to review these during the tutoring sessions or use the platform's accompanying exercises as practice. If a given session had a higher tutee/tutor ratio because of a tutor's absence, the tutor was encouraged to alternate the use of CAL with small-group tutoring.

topics covered in class during that week or revisited prior material, including prerequisite topics that students found challenging; for instance, if a class was working on linear equations, a tutor might review the rules of signs before addressing the current topic.

The effectiveness of the tutoring sessions was monitored through unannounced visits, intended to identify potential issues or areas for improvement, followed by one-on-one feedback meetings between the supervisor and the observed tutor. Tutors also monitored attendance in weekly surveys reporting the number of tutoring hours conducted for each class and student.

2.3 Sample and Randomization

The intervention was administered in 9 public high schools (38 ninth-grade classrooms) in the Dominican Republic. The sample size was constrained by schools' ability to meet the required technical and operational conditions. Schools were selected on four criteria: (1) having an active computer lab, (2) having regular access to electricity and internet, (3) having an extended-day format, and (4) having at least three ninth-grade classrooms. We further limited the sample to schools that had participated in Programate, a joint collaboration between MINERD and the World Bank that provided access to the ALEKS software to approximately 60 schools between 2018 and 2023. These criteria yielded an initial list of 34 schools. Between April and August 2023, the research team, with support from MINERD, contacted these schools by phone to verify eligibility and assess interest. Though all the contacted schools expressed strong interest, only 14 confirmed reliable internet access and a functional computer lab.

Between September 4 and 15, a member of the research team conducted school visits to validate the information collected during the spring and summer terms. Only 9 of the 14 visited schools had functioning internet and computers. All the participating schools were located in Santo Domingo and Santiago—the two largest cities in the country and together home to approximately 45 percent of the national student population (Ministerio de Educación de la República Dominicana (MINERD), 2023). Thus, although the sample is geographically concentrated, it captures a substantial share of the national ninth-grade population.

Treatment was assigned through a two-stage randomization process at the classroom level. Classrooms were first ranked by size and paired with the closest matching classroom in terms of size; within each pair, one classroom was randomly assigned to the treatment group and the other to the control group. This pairing and randomization process was then repeated within the treatment group to assign half of its classrooms to receive the additional tutoring component: New pairs were formed on the basis of classroom size within the CAL group, and the tutoring intervention was randomly assigned within each pair. The study ultimately comprised 38 classrooms, with 19 assigned to receive the CAL software and the remaining 19 serving as controls.[9] Of the 19

[9] Students in control classrooms received temporary licenses to access the software approximately two months after endline data collection.

CAL classrooms, 9 were further selected to receive group tutoring. The randomization design is summarized in Figure 2, and power calculations are included in Appendix A.1.

2.4 Implementation

The CAL intervention started on October 23 and concluded on May 3.[10] During the first week of implementation, all treated students took a mandatory initial knowledge check, the first activity within ALEKS. From this initial assessment, the platform predicts the topics each student has mastered, has not yet mastered, and is ready to learn and generates an individualized learning path accordingly.

The first tutoring session took place on October 23 and the last on April 26. The sessions took place one to three days per week during extracurricular hours, which, in all but one school, corresponded to the hours allocated for optional workshops.[11] Workshops and tutoring sessions typically took place between 2:00 pm and 3:30 pm. In total, 61 students were selected to participate in tutoring.

3 Data

Our primary measure of program effectiveness is students' academic performance. We also examine a set of secondary outcomes—socioemotional skills, psychological well-being, and student-reported teacher behavior—to explore potential channels through which the interventions may have operated. These measures were collected by means of a combination of survey data, administrative records, data from the adaptive software, and brief math assessments, each described below.

The research team administered student surveys and math assessments at baseline (September 2023), midline (December 2023), and endline (May 2024). Administrative records were collected and digitized after the end of the school year, between June and September 2024.

Academic performance. Our main outcome of interest is students' academic performance, which we use to examine whether the CAL platform or tutoring component improved student learning in math. The assessments were designed with items of varying difficulty, adapted from MINERD's national assessments. The endline tests were designed by a pedagogical expert and followed the content and cognitive areas used on the 2019 Trends in International Mathematics and Science Study (TIMMS) eighth-grade math assessments, covering seven content domains and three cognitive areas (Mullis and Martin, 2017).[12] We examine student performance across these

[10]This is equivalent to roughly 21 weeks, excluding winter, Easter week, and two weeks when the use of the platform was interrupted for training and data collection.

[11]One of the nine schools did not offer optional workshops, so students stayed after school hours to receive the tutoring sessions.

[12]The eighth-grade TIMMS math assessments evaluated the following content domains: (1) numbers, (2) algebra, (3) geometry, and (4) data and probability. The cognitive domains evaluated were: (1) knowing, (2) applying, and (3) reasoning. TIMMS is administered only in the fourth and eighth grades.

content areas and cognitive domains at both midline and endline and classify items by difficulty level to assess differential impacts by grade-level content.

Secondary outcomes. Beyond academic performance, we collected several secondary measures to explore the channels through which the interventions may have affected students, though, as discussed in Section 6, we observe limited or no statistically significant treatment effects on most of these outcomes and interpret our estimates as exploratory.

Administrative records. Between June and September 2024, we additionally collected and digitized schools' administrative records. These records include students' end-of-year grades in math and Spanish, attendance, dropout status, and grade promotion outcomes. End-of-year Spanish grades are included to test for adverse spillover effects from time reallocated away from other subjects.

Socioemotional skills. As students increasingly rely on digital devices for learning, there is a risk that socioemotional or noncognitive skills may be negatively impacted. In baseline and endline surveys we measured the following three skills: perseverance, locus of control, and grit. Perseverance was measured through a real effort task adapted from Alan et al. (2019), in which students answered a logic question and then chose between attempting another question of the same difficulty, tackling a harder question, or giving up; their choice serves as our measure of perseverance. Internal locus of control—students' belief that they have control over the outcome of events in their lives—is measured with an adaptation of Rotter's locus of control scale (Rotter, 1966). Grit is measured with a 4-item scale adapted from Carlana and La Ferrara (2021). The grit and locus of control indices are normalized to range between 0 and 1.

Student well-being. Because each student works independently on ALEKS, this learning mode may offer fewer opportunities for peer interaction or support than regular instruction, potentially increasing feelings of isolation in the classroom. This concern is compounded by a broader literature linking increased screen time to lower psychological well-being among teenagers (Stiglic and Viner, 2019; Twenge and Campbell, 2018). To assess whether these dynamics affected our sample, we measure student well-being using items from the Children's Depression Screener (ChilD-S) developed by Frühe et al. (2012).

Increased use of CAL could plausibly affect math anxiety in either direction: An improvement in students' math skills or command of foundational learning could reduce anxiety, but greater exposure to diagnostic feedback highlighting unmastered topics could increase it. Following Araya et al. (2019), we construct a math anxiety index from students' Likert-scale responses to two statements: "I get nervous before math tests" and "I am worried about having bad grades in math."

Teacher behavior. The CAL platform may free up teacher time by automating tasks such as developing and grading homework and tests, potentially shifting how teachers allocate time and effort across classroom activities. To explore this, we ask students about their perceptions of teacher behavior (e.g., whether the teacher provided additional support outside class or clarified a lesson that was not understood) and time management (e.g., whether the teacher started class late or came to school). We use these responses to construct a time mismanagement index and a teacher engagement index.[13] The indices are standardized to have mean zero and standard deviation of 1 in the control group. The questions used were adapted from questionnaires used by De Hoyos et al. (2021) and Ferguson and Danielson (2015).

CAL data. Used for both monitoring and assessment purposes, the CAL software records students' interactions with the platform, including the number of sessions held, date and duration of each session, and number of newly mastered topics per student. We use this information to assess compliance with the intervention and measure progress within the treatment group. These data are available for the treatment group only.

3.1 Balance Checks and Attrition

Table 1 presents balance checks using baseline survey data. There are no statistically significant differences between the treatment and control groups across observable characteristics, which supports the integrity of the randomization. The sample is approximately 56 percent girls and an average age of nearly 15—somewhat above the typical enrollment age of 13–14 for ninth graders, which reflects the high rates of grade repetition in the Dominican Republic (approximately 19 percent of the students in our sample had repeated at least one grade). Baseline academic performance is relatively low: Students answered an average of 6 out of 12 questions correctly. The sample also skews toward students of lower socioeconomic status (SES), with fewer than 40 percent of students having a parent who attended college.

Attrition was low throughout the study. As shown in Columns 1 and 5 of Table 2, 92 percent of the baseline sample completed the surveys and assessments at midline and 90 percent at endline. Completion rates were similar across treatment and control groups, with differences of 0.5 percentage points at midline and 2.9 percentage points at endline. Although the latter change is statistically significant at the 5 percent level, it is small in magnitude and unlikely to affect our estimates meaningfully.

More importantly, the students who completed the surveys were broadly similar to those who did not, as shown in Columns 2–4 and 6–8 of Table 2. A few minor differences in observable characteristics do emerge: Older students were slightly less likely to respond, with each additional year of age reducing response probability by 3.7 percentage points; girls were about 4 percentage points likelier to complete surveys than boys; and students with higher baseline test scores were marginally likelier to remain in the sample (0.9 percentage points per standard deviation). None

[13] The engagement index is built from student responses to the following items, each beginning with "Your math teacher...": (1) "keeps students engaged," (2) "asks questions to check for understanding," (3) "knows when everyone understands," (4) "makes every student work hard," (5) "assigns homework that helps you learn," (6) "reviews past topics," (7) "marks your work to help understanding," and (8) "wants you to explain your answers."

of these differences are large enough to raise substantial concerns about sample composition. As a robustness check, in Appendix Table A1, we present Lee (2009) bounds for the main performance outcomes, which confirm that our findings are not driven by differential attrition.[14]

4 Empirical Framework

We estimate the impact of the program on student outcomes using the following regression:

$$Y_{ci} = \beta_0 + \beta_1 \text{CAL}_c + \beta_2 \text{Tutoring}_c + \beta_3 X_{ci} + \epsilon_{ci}$$

where Y_{ci} is the outcome of student i in classroom c, CAL_c is a binary variable equal to 1 if classroom c was assigned to the CAL treatment and 0 otherwise, and $Tutoring_c$ is a binary variable equal to 1 if the classroom was additionally assigned to receive small-group tutoring and 0 otherwise. X_{ci} is a vector of baseline characteristics for student i in classroom c. Under this specification, β_1 captures the average treatment effect (ATE) of the CAL software on the outcome relative to the control group's, and β_2 captures the additional impact ("value added") of the tutoring component on this outcome relative to its change in classrooms that received CAL alone. In the results section, we report each coefficient separately and their sum,$\beta_1 + \beta_2$, which represents the total effect for classrooms assigned both the CAL and tutoring interventions, relative to the control group outcome.

Throughout this paper, we use randomization-based inference as our preferred estimation strategy, deriving statistical inferences directly on the basis of the randomization procedure rather than assumptions about the sample distribution (Athey and Imbens, 2017; Green et al., 2011; Young, 2019). This approach is particularly well suited to our setting for two reasons. First, it naturally accommodates the pairwise randomization structure used in this study, so that the inference respects the design under which treatment was assigned. Second, it yields nearly exact p-values without requiring large-sample approximations, which is especially valuable when the number of clusters is small, as is the case here.

The validity of our findings relies on the stable unit treatment value assumption (Rubin, 1978), which requires that the treatment status of one classroom not affect the potential outcomes of students in another. Several features of the study design support this assumption. To access the ALEKS platform, each student had to log in with a unique username and password, and platform use was closely monitored during designated sessions, which makes it unlikely that students in control classrooms could have accessed or benefited from the software. To assess the possibility of indirect spillovers, through, for example, treated and control students interacting outside class or teachers adjusting their behavior in nontreated classrooms, we conducted discussions with the

[14] We calculated the bounds using a single regression with all three experimental groups ($y = \beta_1 \text{CAL} + \beta_2 \text{CAL} \times \text{Tutoring} + \varepsilon$). All the groups were trimmed to equalize response rates at the minimum observed level (89.8%). The lower bound drops the highest-scoring students from higher-response groups (CAL and CAL+Tutoring); the upper bound drops the lowest-scoring students, ensuring the treatment effect estimates are not biased by differential attrition.

teaching team and focus groups with students. Neither revealed evidence that teachers modified instruction in control classrooms in response to the treatment or that treated students collaborated with control peers on academic tasks. Though this does not definitively rule out interference, we find no evidence that treatment affected the behavior of students or teachers in control classrooms.

5 Results

5.1 CAL and Tutoring Take-Up

The initial ALEKS assessment revealed that, at the time of its completion, the median student had mastered only 2.7 percent (9 out of 337) of the math topics included in the course, confirming the substantial learning deficits documented in the background section. Over the course of the school year, the median treated student spent approximately 14 hours on the platform (approximately 38 minutes per week) and learned an average of 45 new math topics. As shown in Figures 3 and 4, the distributions of both of these variables are left-skewed: Students at the 90th percentile had spent over 38 hours on the platform and learned more than 188 new math topics by the end of the year.

To assess out-of-school usage, we measure the share of total platform time occurring on non-scheduled days. The median student used the platform outside school hours for approximately two hours, representing 15 percent of their total platform time. This modest figure suggests that the effects we document below reflect a reallocation of in-class instructional time rather than an increase in total learning time.

Take-up of the tutoring component was lower than intended. Of the 61 students assigned to receive tutoring, all received at least 2 hours of sessions, with the median student receiving 29 hours over the school year—approximately 1.4 academic hours per week, well below the intended dosage of 3 hours per week. Attendance was a primary driver of this shortfall: On any given week, up to 40 percent of the assigned students failed to show up to at least one scheduled session, according to our monitoring data. We discuss potential explanations for this compliance gap in Section 6.

5.2 Student Performance

Table 3 presents intention-to-treat (ITT) estimates of the learning effects at approximately two months (midline) and six months (endline) after the start of the program. Columns 1 and 2 report midline results, and Columns 3 and 4 report endline results, with and without baseline covariates.

At midline, the CAL-only classrooms showed gains of 0.17–0.21 SD relative to the control group, as shown in Columns 1 and 2. While economically meaningful, these effects are not statistically significant, consistent with the program having had limited time to generate detectable impacts at this stage. These impacts are roughly consistent across specifications (with and without covariates). We similarly find no significant effects of the tutoring component at midline, which is unsurprising given that students had received only 8 tutoring sessions by that point.

By endline, the CAL-only effect had increased to 0.28–0.31 SD, as shown in Columns 3 and 4. These gains are consistent with estimates in the broader literature on CAL in low- and middle-income countries: Araya et al. (2019) report effects of 0.21–0.27 SD from a game-based math platform in Chile after seven months, Muralidharan et al. (2019) find effects of 0.37 SD in India after 4.5 months, and Büchel et al. (2022) document effects of up to 0.17 SD from a CAL software in El Salvador over six months. Our estimates fall within this range and are notable in that they were achieved through direct substitution of regular classroom instruction, without additional instructional time or dedicated implementation staff.

The tutoring component, by contrast, did not enhance these gains. The coefficient on the tutoring add-on (β_2) is negative (-0.19 SD with covariates), though not statistically distinguishable from zero. Nor is the combined effect of CAL plus tutoring ($\beta_1 + \beta_2 = 0.09$ SD) statistically different from the control outcome, which implies that the students in classrooms assigned both interventions performed no better than the control group on average. Though we cannot reject the null hypothesis that the tutoring had zero effect, the direction and magnitude of the point estimates are consistent with the two interventions working *against* rather than reinforcing one another. In line with this interpretation, students in the tutoring classrooms—including those not directly receiving the tutoring—exhibited reduced engagement with the CAL platform (Figure 9) and lower scores on the locus of control index (Appendix Table B8) relative to the outcomes of the CAL-only classrooms, which suggests that the introduction of tutoring may have altered student behavior in ways that eroded the effectiveness of the adaptive technology. We explore these mechanisms in detail in Section 6.

5.3 Heterogeneity in Program Effects

We explore heterogeneity in CAL's effects along three dimensions: content difficulty, mathematical domain, and cognitive skill area.

Following Muralidharan et al. (2019), we first classify the test items into two difficulty categories—below and at grade level—and test whether students fared differentially depending on the complexity of the material. Table 4 shows that the CAL treatment improved performance on content both below and at grade level, with the effects ranging from 0.21 to 0.28 SD, though the estimate for performance on grade-level content loses its statistical significance without covariates. This pattern differs from the finding in Muralidharan et al. (2019) of large impacts on performance on below-grade-level items but no performance impacts on at-grade-level content, as is consistent with the structure of the ALEKS platform, which is designed to cover both difficulty levels.[15] Supporting this interpretation, our midline results—measured when students had spent the majority of their time on modules covering below-grade-level topics—show effects concentrated almost entirely on below-grade-level test items, with at-grade-level effects emerging only by endline after students had

[15] The authors find significant effects for questions in the Hindi subject area both at and below grade level.

progressed through the curriculum.

Table 5 present effects across the four content areas evaluated in the TIMMS eighth-grade mathematics assessment. We find positive effects at endline from the CAL treatment on performance in algebra, numbers, and geometry, though the geometry coefficient is not statistically significant. We find no effect on performance in data and probability, the one TIMMS content area not covered by the platform or included in the the Dominican Republic's ninth-grade curriculum—a pattern that supports the interpretation that the observed gains reflect learning through the platform rather than general test-taking improvements. The results for the tutoring component mirror those in Table 3: The classrooms assigned tutoring show negative but statistically nonsignificant differences from the CAL-only classrooms across all the content areas except data and probability. Table 6 presents results across the three cognitive domains evaluated by TIMMS: knowing, applying, and reasoning. We find large positive effects on all three domains at endline, ranging from 0.16 to 0.31 SD, though the effect on the applying domain does not reach statistical significance.

Lastly, Figure 6 plots the distribution of endline performance for both the treatment and control groups across baseline performance percentiles. Despite the overlap in the confidence intervals at several points, the treatment group consistently outperforms the control group across the full distribution of baseline performance, which suggests that the program's benefits were not concentrated among higher- or lower-performing students.

5.4 Secondary Outcomes

Beyond academic performance, we examine whether the interventions affected students' retention, motivation, socioemotional skills, psychological well-being, beliefs about own performance, and teacher behavior. Most of these secondary outcomes show no statistically significant effects. The notable exceptions are an increase in overconfidence in anticipated test performance and marginally significant reductions in dropout rates and the locus of control index among the CAL+Tutoring students and an increased preference for easier tasks on a real-effort persistence measure among CAL-only students. We summarize these findings briefly here; the full results are reported in Appendix B.

Student Retention and Motivation. Despite the substantial learning gains from the CAL intervention, we find no effects on dropout rates, school absenteeism, intrinsic motivation, or stated preferences for mathematics among CAL-only students. The null effect on dropout is consistent with students facing binding nonacademic constraints—such as economic pressures or family obligations—that learning gains alone cannot address. The CAL+Tutoring group shows a marginally significant reduction in dropout rates at endline. Given the absence of academic performance gains in this arm, this effect is unlikely to have operated through improvements to learning; instead, it may reflect the added benefit of regular human contact and personalized attention from a tutor.

Student Beliefs and Expectations. We examine whether the interventions affected students' calibration of their own academic performance. At endline, students were asked to predict their test score prior to taking the assessment, which allowed us to construct a measure of overconfidence as the difference between predicted and actual performance. As shown in Table 7, students in CAL+Tutoring classrooms exhibited marginally higher overconfidence than the control students (5.9 percentage points, $p < 0.10$ with covariates); the CAL-only students showed no significant difference. Figure 5 illustrates these patterns: Expected performance was similar across all three groups, but actual performance was highest in the CAL-only classrooms, resulting in our measure of overconfidence being greater for the CAL+Tutoring students—whose actual gains were smaller—than for CAL-only students. We discuss the implications of this belief miscalibration for platform engagement in Section 6.4.

Socioemotional Skills. We find no significant effects of either intervention on self-efficacy or grit at endline. Students in CAL-only classrooms were, however, likelier to choose an easier task over a more challenging alternative in the real-effort perseverance exercise, which suggests that sustained exposure to an adaptive platform—which continuously adjusts difficulty to each student's level—may reduce willingness to voluntarily tackle harder problems. On the locus of control measure, the tutoring component produced a marginally significant negative effect at endline ($p<0.10$), which indicates that students in the CAL+Tutoring group were likelier to attribute outcomes to external factors than to their own effort—a finding consistent with the broader pattern of reduced student agency we document in Section 6.

Psychological Well-Being. We find no significant effects on a depression index based on the Children's Depression Screener, nor on the overall math anxiety index at endline. Students in CAL-only classrooms showed a marginally significant increase in nervousness before tests relative to control students; the combined effect for CAL+Tutoring classrooms was indistinguishable from zero. Rather than reflecting distress, this pattern may indicate that students with greater exposure to the platform's diagnostic feedback developed heightened awareness of their own knowledge gaps—consistent with the higher platform engagement we document for CAL-only students than for their CAL+Tutoring peers.

Student Time Use. One potential concern is that CAL might crowd out other productive activities or increase screen time outside school. We find no significant effects on students' time allocation across eight categories: in-person classes, unpaid work at home, unpaid work outside the home, paid work, studying outside school, meeting friends, entertainment, and sports. This suggests that the intervention operated primarily through reallocation of in-class instructional time rather than changes in how students allocated time outside of school.

Teacher Behavior. We find no significant effects on either the teacher time mismanagement index or the teacher engagement index at endline. The absence of detectable changes in teacher behavior suggests that the integration of the CAL platform into regular instruction did not substantially alter how teachers allocated time or engaged with students during nonplatform hours and that the learning gains we document are likelier attributable to the direct substitution of adaptive instruction for traditional class time than to changes in teacher practice.

6 Mechanisms

The impact of the CAL intervention depended critically on students' engagement with the platform and on how the intervention interacted with existing instructional practices. In this section, we first rule out differential implementation as an explanation for the divergent effects between the two treatment arms and then synthesize quantitative evidence on platform usage, student beliefs, and performance outcomes with qualitative findings from focus groups to discuss potential mechanisms.

6.1 Implementation Fidelity

To assess whether differential implementation could explain the divergent effects, we collected weekly monitoring data on session delivery across all 19 CAL classrooms over 21 weeks.[16] Overall, the CAL sessions achieved 67.7 percent of the expected dosage (28.4 of 42 planned sessions). Fidelity did not differ between CAL-only and CAL+Tutoring classrooms: The former achieved 68.8 percent compliance and the latter 66.4 percent. Among the nine classrooms receiving both components, the tutoring sessions achieved 63.5 percent of the expected dosage, a figure statistically indistinguishable from the CAL compliance in the same classrooms.

When sessions were canceled, the reasons were predominantly exogenous. Among the canceled CAL sessions, 74.5 percent were due to school closures, institutional activities, or infrastructure failures; similarly, 75.5 percent of the canceled tutoring sessions were attributable to such factors. The remaining cancellations reflected intervention-specific barriers: laboratory access and ALEKS licensing issues for CAL and tutor availability and space constraints for tutoring. The similarity in compliance rates and cancellation reasons across both components suggests that large differences in implementation quantity are unlikely to explain the attenuated effects in the CAL+Tutoring arm. That said, session-level compliance may not fully capture variation in delivery quality, which is inherently harder to standardize for human-led tutoring than for a software platform; differential implementation quality therefore cannot be entirely ruled out as a contributing factor.

[16] Compliance rates are calculated at the classroom level (sessions held divided by sessions scheduled).

6.2 Engagement Intensity and Dose–Response

A key mechanism through which the CAL intervention appears to have affected learning is sustained engagement with adaptive, self-directed practice. Our platform data reveal substantial heterogeneity in usage: The median treated student used ALEKS for approximately 14 hours over the school year and mastered 45 topics, but students at the 90th percentile spent over 38 hours on the platform and mastered more than 188 topics (see Figures 3 and 4). This variation is economically meaningful. Using random assignment as an instrument for hours of platform use, we estimate that each additional hour of CAL engagement increased endline test scores by approximately 0.026–0.027 SD (see Table 8).

The estimated dose–response relationship exhibits diminishing returns, with gains increasing steeply up to roughly 1500 minutes (25 hours) of use before flattening (see Figure A2). In Table 8, we extrapolate our estimates to 1500 minutes of use—equivalent to roughly 5 months of implementation at 80 percent attendance. The results suggests that fully faithful implementation could generate test score gains of 0.73 SD.[17] Although such an extrapolation should be interpreted cautiously, these estimates suggest that implementation frictions—including connectivity issues, limited device availability, and school disruptions—reduced effective exposure to the platform and may have attenuated its impacts relative to this benchmark.

6.3 Substitution of Effort and Classroom-Level Spillovers

Despite similar access to the platform, students in the CAL+Tutoring classrooms spent less time on ALEKS and mastered fewer topics than students in the CAL-only classrooms. Students in the tutoring arm also exerted less effort while using the platform.[18] Importantly, this pattern is observed not only for students selected for tutoring but also for their nontutored classmates, indicating that the tutoring intervention affected classroom-level learning dynamics rather than only individual tutees.

Figure 9 shows that the average student in a CAL-only classroom spent a total of 1,232 minutes using the platform (56 minutes per week), compared to 1,082 minutes (49 minutes per week) for the average student in a CAL+Tutoring classroom. On the basis of the estimated dose–response relationship, this difference in time usage accounts for approximately 34 percent (0.07 SD) of the observed gap in endline performance between the two treatments arms (see Table 8). Within the CAL+Tutoring classrooms, both tutees and nontutees used the platform less than their CAL-only counterparts, though the gap was larger for tutees, who averaged only 1,000 minutes (or 45 minutes per week).

One interpretation of this result is that tutoring altered how the students allocated learning

[17] This extrapolation requires us to make an assumption about heterogeneity in treatment effects. As shown in Figure 6, we do not find meaningful evidence of effect heterogeneity across baseline performance.

[18] We define effort as the ratio of topics learned to time spent on the platform.

effort. Qualitative evidence suggests that instead of persisting independently, students in the tutoring classrooms often relied on tutors to resolve difficulties. Although the tutoring dosage fell short of its intended levels because of absenteeism and scheduling disruptions, the availability of a tutor may nonetheless have provided an alternative pathway to understanding that substituted for self-directed platform engagement.

6.4 Agency, Beliefs, and Attribution

The differences in engagement across the treatment arms were accompanied by differences in students' beliefs about their own learning. Our survey data indicate that students in the CAL-only classrooms formed more accurate expectations about their academic performance but those in the CAL+Tutoring classrooms exhibited higher overconfidence and a marginally significant reduction in our measure of internal locus of control, which indicates that they were likelier to attribute outcomes to external factors rather than their own effort (see Tables 7 and B8). These patterns suggest that the presence of a tutor may have shifted students' attribution of learning away from individual effort and toward external support.

Platform design features—such as difficulty jumps, loss of progress after errors, and limited explanatory feedback—likely increased the cost of persistence for some students, as highlighted in focus group discussions. In the CAL-only classrooms, students had no alternative source of support and may therefore have been likelier to persevere through those frictions. In the tutoring classrooms, access to human explanations may have reduced tolerance for platform-related challenges, reinforcing both disengagement from the software and a greater tendency to attribute learning outcomes to the tutor rather than to one's own effort. While these mechanisms cannot be conclusively established from the available data, they are consistent with the observed patterns in platform engagement, beliefs, and performance.

6.5 Implications

Overall, the evidence suggests that the absence of additional gains from the tutoring add-on reflects more than implementation shortfalls. Rather, tutoring appears to have interacted with the CAL intervention in ways that weakened the mechanisms through which the technology generates learning gains—by reducing platform engagement, lowering effort on the platform, and shifting students' beliefs about the source of their learning. These behavioral responses were not confined to tutees but extended to their nontutored classmates, which suggests that the presence of tutoring reshaped classroom-level incentives and norms around self-directed learning.

These findings highlight how important it is for behavioral responses to be accounted for when educational interventions are bundled. Adaptive learning technologies may be most effective when they preserve and reinforce student agency and sustained self-directed engagement. When layered with interventions that substitute for these processes, particularly in settings with implementation frictions, their effectiveness may be diminished. More broadly, the results underscore the value of empirically testing intervention bundles instead of assuming their effects are additive.

7 Conclusion

This paper provides experimental evidence on the effectiveness of computer-adaptive learning as a substitute for regular math instruction and on whether combining it with small-group tutoring generates complementary gains. Conducted in the Dominican Republic, one of the world's lowest-performing countries in international assessments, our results speak to both the potential and the limits of technology-based approaches in resource-constrained settings.

We find that replacing two out of seven weekly math hours with an adaptive learning platform improved student performance by 0.28–0.31 SD after 6 months, with the effects proving robust across specifications, mathematical domains, and difficulty levels. These estimates fall within the upper range of the documented effects of computer-adaptive learning in low- and middle-income countries and are notable in that they were achieved through direct substitution of regular classroom instruction, without additional instructional time or dedicated implementation staff.

The addition of small-group tutoring did not enhance these gains. Our point estimates suggest that the classrooms assigned both interventions performed no better than the control group on average and worse than the CAL-only classrooms, though the latter difference is not statistically significant. The evidence points to behavioral mechanisms: Students in the tutoring classrooms—including those not directly tutored—exhibited reduced platform engagement, lower effort on the platform, and a weaker internal locus of control than the CAL-only students. These patterns suggest that the availability of human support altered how students allocated learning effort, substituting tutor-provided explanations for the self-directed persistence that drives the adaptive technology's effectiveness. Reduced platform engagement accounts for approximately 34 percent of the observed performance gap between the two treatment arms, with the belief and attribution shifts we document suggesting additional behavioral channels, though their precise contribution to the remaining gap cannot be quantified from the available data.

Several broader lessons emerge from these findings. First, the effectiveness of adaptive learning technology appears to depend critically on the quality and intensity of student engagement, not merely on access. Implementation frictions that reduce time on the platform—whether from infrastructure constraints or behavioral responses to cointerventions—can substantially attenuate its impacts. Second, bundling of effective interventions does not guarantee that they will yield additive benefits. When two components interact in ways that undermine each other's mechanisms, the resulting bundle may perform worse than either component alone. Third, the classroom-level spillovers we document—where nontutored students in the tutoring classrooms also reduced their platform engagement—suggest that intervention design must account for how components shape broader classroom norms and incentives, not only the behavior of directly targeted students.

As artificial intelligence and adaptive technologies become increasingly integrated into education systems worldwide, questions on how to effectively combine human and technological inputs will grow in importance. This study provides evidence that such integration requires careful attention to how components interact at the behavioral level and that empirical testing of intervention bundles—rather than the assumption of additive effects—is essential for effective policy design.

References

Alan, S., T. Boneva, and S. Ertac (2019). Ever failed, try again, succeed better: Results from a randomized educational intervention on grit. *The Quarterly Journal of Economics 134*(3), 1121–1162.

Araya, R., E. A. Ortiz, N. L. Bottan, and J. P. Cristia (2019). Does gamification in education work?: Experimental evidence from chile.

Athey, S. and G. W. Imbens (2017). The econometrics of randomized experiments. In *Handbook of economic field experiments*, Volume 1, pp. 73–140. Elsevier.

Augustine, C. H., J. S. McCombs, J. F. Pane, H. L. Schwartz, J. Schweig, A. McEachin, and K. Siler-Evans (2016). *Learning from Summer: Effects of Voluntary Summer Learning Programs on Low-Income Urban Youth. RAND Summer Learning Series. Research Report. RR-1557-WF.* ERIC.

Büchel, K., M. Jakob, C. Kühnhanss, D. Steffen, and A. Brunetti (2022). The relative effectiveness of teachers and learning software: Evidence from a field experiment in el salvador. *Journal of labor economics 40*(3), 737–777.

Carlana, M. and E. La Ferrara (2021). Apart but connected: Online tutoring and student outcomes during the covid-19 pandemic.

De Hoyos, R., A. J. Ganimian, and P. A. Holland (2021). Teaching with the test: experimental evidence on diagnostic feedback and capacity building for public schools in argentina. *The World Bank Economic Review 35*(2), 499–520.

Doignon, J. P. and J. C. Falmagne (1985). Spaces for the assessment of knowledge. *International Journal of Man-Machine Studies 23*(2), 175–196.

Dowker, A., A. Sarkar, and C. Y. Looi (2016). Mathematics anxiety: What have we learned in 60 years? *Frontiers in psychology 7*, 508.

Duckworth, A. L., C. Peterson, M. D. Matthews, and D. R. Kelly (2007). Grit: perseverance and passion for long-term goals. *Journal of personality and social psychology 92*(6), 1087.

Evans, D. K. and F. Yuan (2022). How big are effect sizes in international education studies? *Educational Evaluation and Policy Analysis 44*(3), 532–540.

Falmagne, J.-C., M. Koppen, M. Villano, J.-P. Doignon, and L. Johannesen (1990). Introduction to knowledge spaces: How to build, test, and search them. *Psychological Review 97*(2), 201–224.

Ferguson, R. F. and C. Danielson (2015). How framework for teaching and tripod 7cs evidence distinguish key components of effective teaching. *Designing teacher evaluation systems: New guidance from the measures of effective teaching project*, 98–143.

Frühe, B., A.-K. Allgaier, K. Pietsch, M. Baethmann, J. Peters, S. Kellnar, A. Heep, S. Burdach, D. von Schweinitz, and G. Schulte-Körne (2012). Children's depression screener (child-s): development and validation of a depression screening instrument for children in pediatric care. *Child Psychiatry & Human Development 43*, 137–151.

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

Green, D. P., P. M. Aronow, D. E. Bergan, P. Greene, C. Paris, and B. I. Weinberger (2011). Does knowledge of constitutional principles increase support for civil liberties? Results from a randomized field experiment. *The Journal of Politics 73*(2), 463–476.

Guryan, J., J. Ludwig, M. P. Bhatt, P. J. Cook, J. M. Davis, K. Dodge, G. Farkas, R. G. Fryer Jr, S. Mayer, H. Pollack, et al. (2023). Not too late: Improving academic outcomes among adolescents. *American Economic Review 113*(3), 738–765.

Instituto Dominicano de Evaluación e Investigación de la Calidad Educativa (IDEICE) (2017). Title of the document. Accessed: 2023-04-28.

Interamerican Development Bank (2023). Latin America and the Caribbean in PISA 2022: How did the Region Perform? Technical report, Interamerican Development Bank.

Kraft, M. A. and G. T. Falken (2020). A blueprint for scaling tutoring across public schools. edworkingpaper no. 20-335. *Annenberg Institute for School Reform at Brown University*.

Kraft, M. A., B. E. Schueler, and G. Falken (2024, August). What impacts should we expect from tutoring at scale? exploring meta-analytic generalizability.

Lee, D. S. (2009). Training, wages, and sample selection: Estimating sharp bounds on treatment effects. *The Review of Economic Studies 76*(3), 1071–1102.

Ministerio de Educación de la República Dominicana (MINERD) (2023, March). Title of the document. Accessed: 2023-04-28.

Mullis, I. V. and M. O. Martin (2017). *TIMSS 2019 Assessment Frameworks.* ERIC.

Muralidharan, K. and A. Singh (2025). Adapting for scale: Experimental evidence on technology-aided instruction in india. Technical report, National Bureau of Economic Research.

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

Nickow, A., P. Oreopoulos, and V. Quan (2020). The impressive effects of tutoring on prek-12 learning: A systematic review and meta-analysis of the experimental evidence.

OECD (2023a). *PISA 2022 Results (Volume I).*

OECD (2023b). *PISA 2022 Results (Volume II).*

Oreopoulos, P., O. Keyes-Krysakowski, and D. Agarwal (2026, January). How in-school supervised ed-tech support produces massive learning gains: A khan academy field experiment in india. Working Paper 34683, National Bureau of Economic Research.

Rotter, J. B. (1966). Generalized expectancies for internal versus external control of reinforcement. *Psychological monographs: General and applied 80*(1), 1.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of statistics*, 34–58.

Stiglic, N. and R. M. Viner (2019). Effects of screentime on the health and well-being of children and adolescents: a systematic review of reviews. *BMJ open 9*(1), e023191.

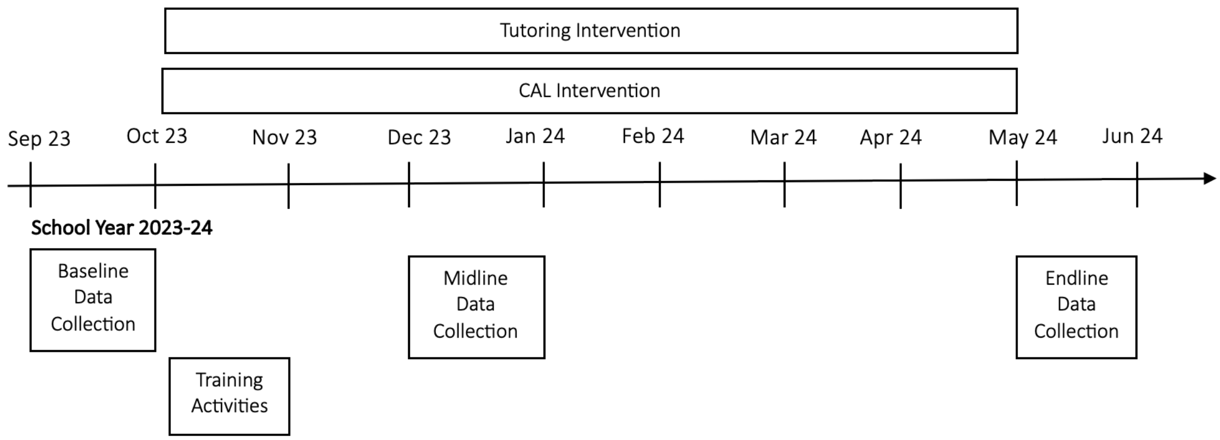
Twenge, J. M. and W. K. Campbell (2018). Associations between screen time and lower psychological well-being among children and adolescents: Evidence from a population-based study. *Preventive medicine reports 12*, 271–283.

World Bank (2017). *World development report 2018: Learning to realize education's promise.* The World Bank.

Young, A. (2019). Channeling fisher: Randomization tests and the statistical insignificance of seemingly significant experimental results. *The quarterly journal of economics 134*(2), 557–598.

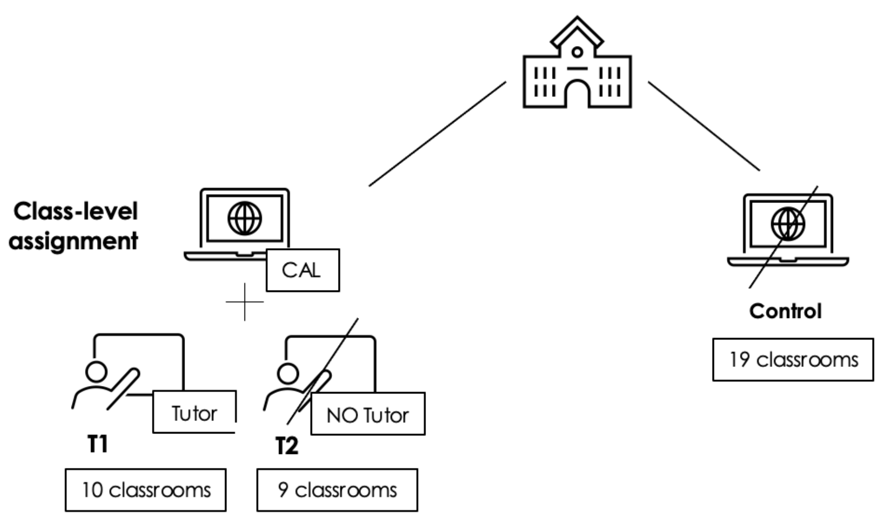
Figures and Tables

Figure 1: Timeline for Intervention and Study



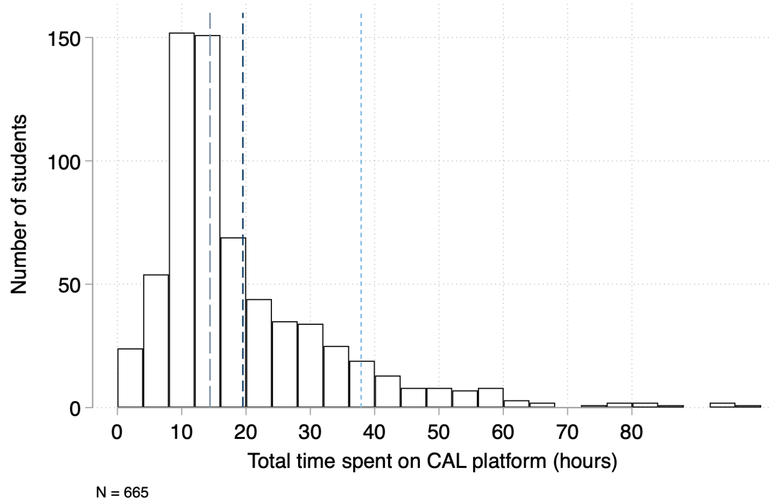
Notes: The computer-adaptive learning (CAL) intervention ran from October 23, 2023, to May 3, 2024 (approximately 21 instructional weeks). The tutoring intervention ran from October 23, 2023, to April 26, 2024. Platform use was interrupted during winter break, Easter week, and two weeks allocated for training and data collection. Baseline surveys were administered in September 2023, midline surveys in December 2023, and endline surveys in May 2024.

Figure 2: Randomization Design



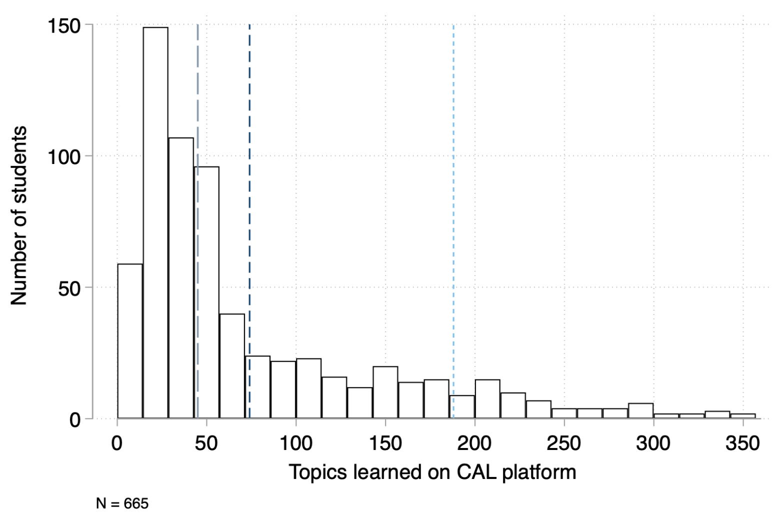
Notes: T2 comprised 9 classrooms assigned to receive access to the computer-adaptive learning (CAL) software. T1 comprised 10 classrooms where in addition to access to the software, the bottom quarter of students, in terms of performance, were to receive tutoring sessions in small groups during extracurricular hours. The control group, containing 19 classrooms, did not receive access to the CAL or tutoring components.

Figure 3: Time Spent on CAL



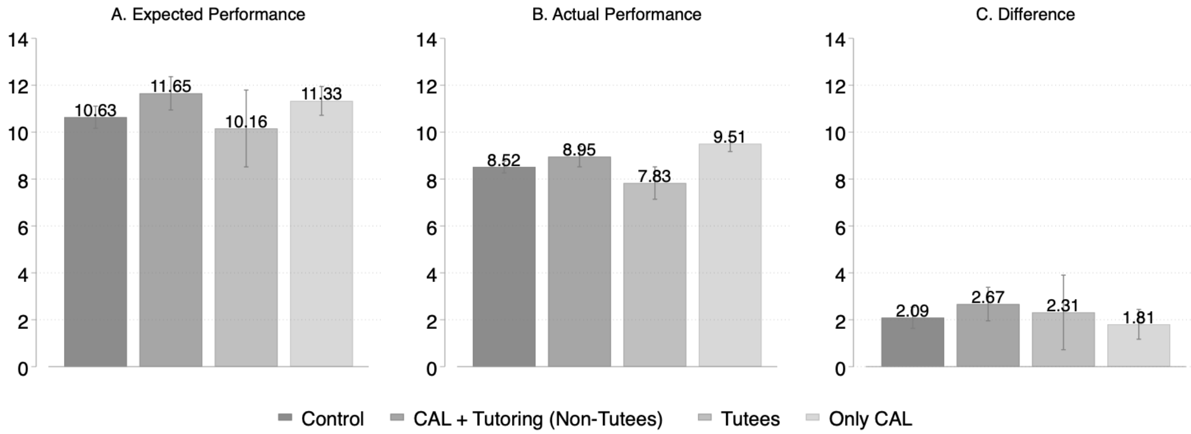
Notes: The figure shows the distribution of total hours spent on the ALEKS computer-adaptive learning (CAL) platform among treated students over the course of the school year. The light gray line indicates the median (approximately 14 hours). The navy line indicates the mean (approximately 20 hours), and the light blue line indicates the 90th percentile (approximately 38 hours). $N = 665$.

Figure 4: Topics Learned on CAL



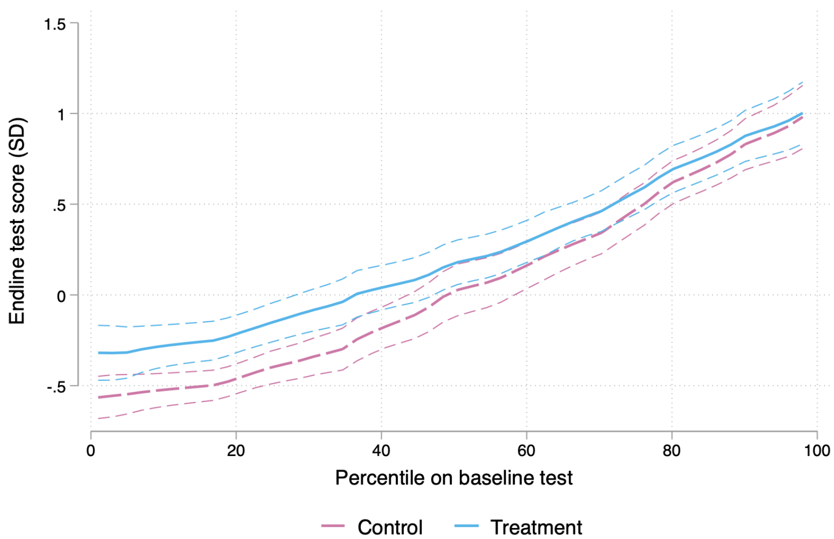
Notes: The figure shows the distribution of total math topics mastered on the ALEKS computer-adaptive learning (CAL) platform among treated students over the course of the school year. The gray line indicates the median (approximately 45 topics), the navy line indicates the median (approximately 74 topics), and the light blue line indicates the 90th percentile (approximately 188 topics). $N = 665$.

Figure 5: Expected Performance on Endline Math Tests



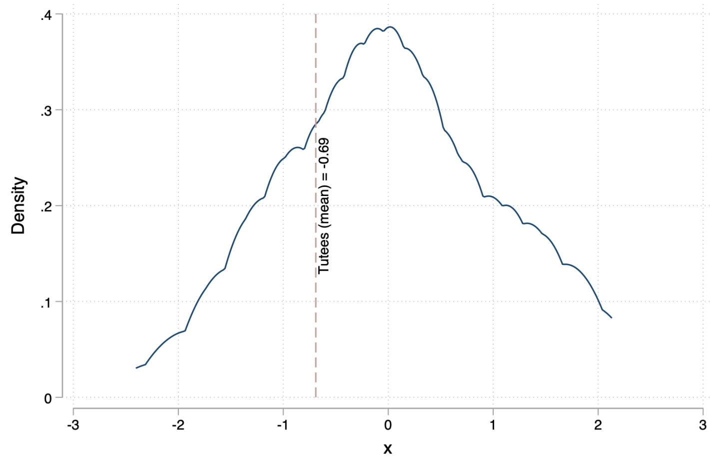
Notes: The figure compares students' expected and actual performance on the endline math assessment across groups. Panel A shows expected performance (number of correct answers predicted by students prior to taking the test), Panel B shows actual performance (number of correct answers), and Panel C shows the difference between the two, where larger values indicate greater overconfidence. The CAL+Tutoring group excludes tutees.

Figure 6: Nonparametric Estimates of Treatment Effects by Baseline Performance



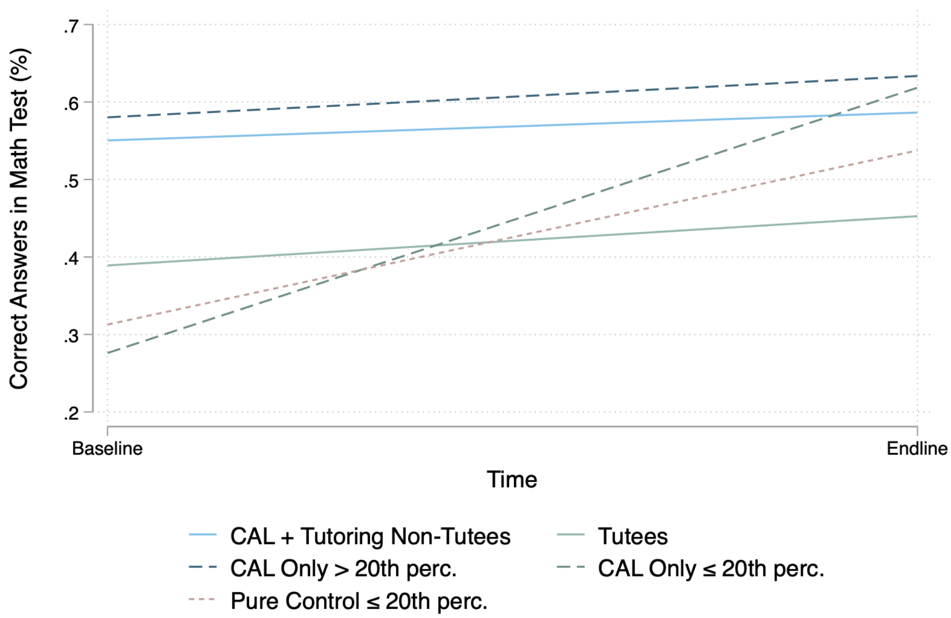
Notes: Following Muralidharan et al. (2019), the figure presents kernel-weighted local mean smoothed plots relating endline test scores to percentiles in baseline tests, separately for the treatment and control groups. The dashed lines show 95 percent confidence intervals.

Figure 7: Selection of Tutees and Distribution of Performance



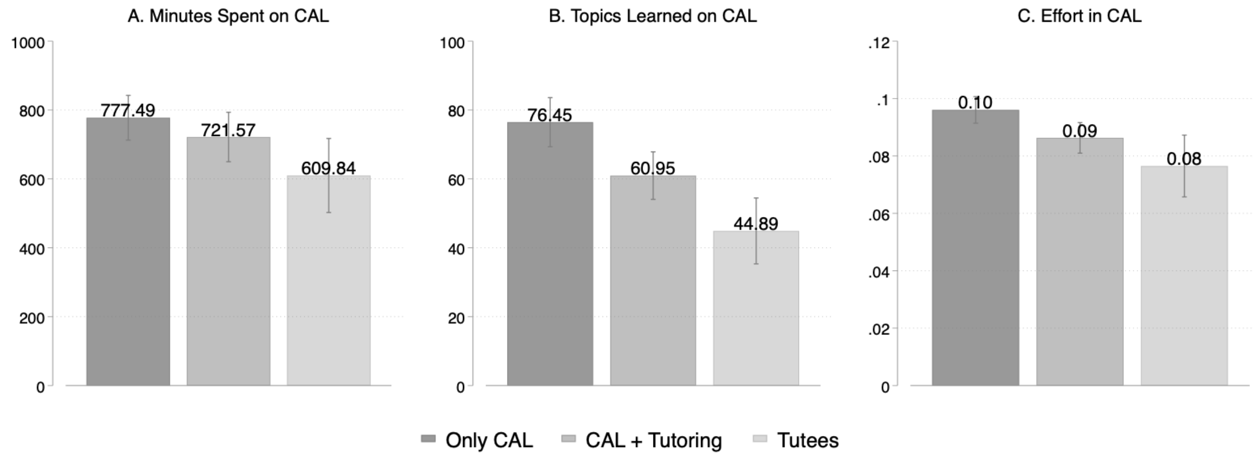
Notes: The figure shows the distribution of standardized baseline math test scores for all students in the CAL+Tutoring classrooms. The dashed vertical line indicates the mean baseline score among the students selected to receive tutoring, showing that selected tutees were not systematically drawn from the lower tail of the baseline performance distribution, contrary to the intervention's targeting intention.

Figure 8: Share of Correct Answers by Performance Group



Notes: The 20th percentile or below threshold of baseline performance is used as an approximate counterfactual for tutees. "CAL+Tutoring Non-Tutees" (blue solid line) shows trends between baseline and endline for students who did not directly receive the group tutoring intervention but were in classrooms where approximately 20 percent of students did. "Tutees" (green solid line) are students who were assigned the tutoring intervention in selected classrooms. "CAL Only $>$ 20th perc." (navy line) is students in CAL-only classrooms who are above the 20th percentile of baseline performance. "CAL Only $\leq$ 20th percentile" (green dashed line) is students in CAL-only classrooms who were below the 20th percentile. "Pure Control $\leq$ 20th percentile" (pink line) is students below the 20th percentile threshold in control classrooms.

Figure 9: Platform Use by Group



Notes: The figure shows platform engagement across treatment groups. Panel A shows total minutes spent on the ALEKS computer-adaptive learning (CAL) platform, Panel B shows the total number of math topics mastered, and Panel C shows effort, defined as the ratio of topics mastered to minutes spent on the platform. The CAL+Tutoring group excludes tutees and is shown separately from the tutee subgroup to distinguish classroom-level spillovers from direct tutoring effects. Error bars indicate 95 percent confidence intervals.

Table 1: Balance Table

Variable	Control N	Control Mean/(SE)	Treatment N	Treatment Mean/(SE)	N	Difference
Female	671	0.553	662	0.565	1333	-0.012
		(0.019)		(0.019)		
Age	617	14.524	627	14.459	1244	0.064
		(0.033)		(0.033)		
Math test scores	622	6.280	629	6.444	1251	-0.164
		(0.109)		(0.110)		
Mother went to college	624	0.295	634	0.317	1258	-0.022
		(0.018)		(0.018)		
Father went to college	624	0.188	634	0.211	1258	-0.024
		(0.016)		(0.016)		
Mother went to high school	624	0.692	634	0.708	1258	-0.016
		(0.018)		(0.018)		
Father went to high school	624	0.566	634	0.557	1258	0.009
		(0.020)		(0.020)		
F-test of joint significance						0.652
F-test, number of observations						1237

Notes: The table reports means and standard errors for observable characteristics across control and treatment groups at baseline. The treatment group pools CAL-only and CAL+Tutoring classrooms. The "Difference" column reports the raw difference in means between treatment and control groups. The F-test of joint significance tests whether all differences are jointly equal to zero; a p-value of 0.652 indicates no evidence of systematic imbalance at baseline. Math test scores are based on a short assessment administered at baseline. Standard errors in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table 2: Attrition from Baseline to Midline and Endline

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Midline Survey				Endline Survey			
	All	All	Treat	Control	All	All	Treat	Control
Treatment	0.005	0.013			0.029*	0.025		
	(0.015)	(0.014)			(0.015)	(0.016)		
Student's age		-0.009	-0.022	0.007		-0.025**	-0.034**	-0.018
		(0.011)	(0.014)	(0.017)		(0.012)	(0.015)	(0.020)
Student is female		-0.028*	-0.020	-0.034		0.008	0.036	-0.013
		(0.015)	(0.021)	(0.023)		(0.017)	(0.022)	(0.026)
Father is unemployed		-0.024	-0.036	-0.002		-0.002	0.036	-0.027
		(0.027)	(0.038)	(0.039)		(0.030)	(0.040)	(0.045)
Mother is unemployed		-0.016	0.012	-0.046*		0.001	0.019	-0.013
		(0.017)	(0.023)	(0.024)		(0.019)	(0.025)	(0.028)
Baseline test score		0.003	0.002	0.005		0.008**	0.011**	0.005
		(0.003)	(0.004)	(0.004)		(0.003)	(0.004)	(0.005)
Repeated a school year		-0.011	0.024	-0.043		-0.005	0.036	-0.038
		(0.024)	(0.033)	(0.036)		(0.027)	(0.035)	(0.042)
Mother went to college		0.020	0.023	0.019		-0.012	-0.018	-0.005
		(0.018)	(0.024)	(0.027)		(0.020)	(0.026)	(0.031)
Mother went to high school		0.031*	0.014	0.044		0.027	0.014	0.039
		(0.019)	(0.026)	(0.028)		(0.021)	(0.028)	(0.032)
Father went to high school		-0.032*	0.013	-0.072***		-0.024	-0.030	-0.018
		(0.017)	(0.024)	(0.025)		(0.019)	(0.026)	(0.029)
Father went to college		-0.006	-0.010	-0.012		-0.010	0.003	-0.017
		(0.020)	(0.026)	(0.032)		(0.022)	(0.028)	(0.036)
Internal locus of control		0.058	0.095	0.021		0.001	0.018	-0.001
		(0.062)	(0.085)	(0.091)		(0.069)	(0.091)	(0.105)
Self-efficacy		-0.031	-0.056	-0.011		-0.056	-0.081	-0.040
		(0.047)	(0.064)	(0.069)		(0.053)	(0.069)	(0.080)
Spanish grade in 2022/23		0.003	0.002	0.006		0.000	0.003	-0.003
		(0.008)	(0.011)	(0.012)		(0.009)	(0.012)	(0.014)
Math grade in 2022/23		-0.010	-0.005	-0.018		0.008	0.013	0.000
		(0.009)	(0.012)	(0.014)		(0.010)	(0.013)	(0.016)
Constant	0.917***	1.075***	1.227***	0.886***	0.900***	1.240***	1.343***	1.184***
	(0.011)	(0.168)	(0.218)	(0.263)	(0.011)	(0.187)	(0.234)	(0.303)
R-squared	0.000	0.025	0.027	0.048	0.003	0.023	0.038	0.022
N	1333	991	492	499	1333	991	492	499

Notes: The table reports ordinary least squares (OLS) estimates of the relationship between baseline characteristics and survey completion at midline (Columns 1–4) and endline (Columns 5–8). Columns 1 and 5 report the difference in completion rates between treatment and control groups. Columns 2 and 6 report coefficients from regressions of survey completion on treatment status and baseline characteristics for the full sample. Columns 3–4 and 7–8 report the same regression results separately for treated and control students. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table 3: Treatment Effects on Math Assessments

	Midline		Endline	
	(1)	(2)	(3)	(4)
CAL Treatment (β_1)	0.2067	0.1670	0.3144*	0.2860**
	(0.1372)	(0.1143)	(0.1668)	(0.1431)
CAL $\times$ Tutoring (β_2)	-0.1133	-0.0770	-0.2486	-0.2045
	(0.1578)	(0.1251)	(0.1708)	(0.1454)
CAL + Tutoring ($\beta_1 + \beta_2$)	0.0934	0.0900	0.0658	0.0815
	(0.1346)	(0.1142)	(0.1486)	(0.1268)
N	1225	1225	1218	1218
Adjusted R^2	0.005	0.277	0.016	0.302
Covariates	No	Yes	No	Yes

Notes: Randomization inference standard errors in parentheses (1000 simulations). Covariates include students' gender, age, baseline math test scores, proxies for confidence and preferences toward math, self-reported grades for the previous year, and parents' employment status. We also control for the baseline level of the outcome variable when data are available. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$ based on randomization inference p-values.

Table 4: Treatment Effects on Test Scores by Content Level

	Below Grade Level		At Grade Level	
	(1)	(2)	(3)	(4)
CAL Treatment (β_1)	0.2846*	0.2648**	0.2400	0.2106*
	(0.1463)	(0.1324)	(0.1445)	(0.1193)
CAL $\times$ Tutoring (β_2)	-0.1747	-0.1405	-0.2623	-0.2204*
	(0.1514)	(0.1397)	(0.1505)	(0.1192)
CAL + Tutoring ($\beta_1 + \beta_2$)	0.1099	0.1243	-0.0223	-0.0098
	(0.1296)	(0.1162)	(0.1302)	(0.1080)
N	1218	1218	1217	1217
Adjusted R^2	0.013	0.218	0.010	0.239
Covariates	No	Yes	No	Yes

Notes: Randomization inference standard errors in parentheses (1000 simulations). Covariates include students' gender, age, baseline math test scores, proxies for confidence and preferences toward math, self-reported grades for the previous year, and parents' employment status. We also control for the baseline level of the outcome variable when data are available. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$ based on randomization inference p-values.

Table 5: Impact on Test Scores by Math Area

	Midline			Endline			
	Algebra	Numbers	Data & Prob	Algebra	Numbers	Data & Prob	Geometry
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
CAL Treatment (β_1)	0.190	0.101	-0.021	0.211*	0.325**	0.027	0.130
	(0.112)	(0.110)	(0.061)	(0.119)	(0.146)	(0.089)	(0.113)
CAL × Tutoring (β_2)	-0.102	0.004	-0.062	-0.220*	-0.115	0.035	-0.192
	(0.116)	(0.126)	(0.078)	(0.119)	(0.152)	(0.098)	(0.126)
CAL + Tutoring ($\beta_1 + \beta_2$)	0.088	0.105	-0.083	-0.010	0.210	0.062	-0.061
	(0.109)	(0.105)	(0.067)	(0.108)	(0.130)	(0.082)	(0.105)
N	1225	1225	1220	1217	1217	1218	1218
Adjusted R^2	0.189	0.187	0.055	0.239	0.180	0.044	0.080

Notes: All specifications include covariates. Randomization inference standard errors in parentheses (1000 simulations). Covariates include students' gender, age, baseline math test scores, proxies for confidence and preferences toward math, self-reported grades for the previous year, and parents' employment status. We also control for the baseline level of the outcome variable when data are available. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$ based on randomization inference p-values.

Table 6: Treatment Effects on Test Scores by Knowledge Domain

	Midline			Endline		
	Knowing	Applying	Reasoning	Knowing	Applying	Reasoning
	(1)	(2)	(3)	(4)	(5)	(6)
CAL Treatment (β_1)	0.201	0.067	0.086	0.320***	0.158	0.208*
	(0.140)	(0.087)	(0.079)	(0.121)	(0.127)	(0.107)
CAL × Tutoring (β_2)	-0.012	-0.062	-0.156*	-0.225*	-0.104	-0.090
	(0.146)	(0.093)	(0.082)	(0.135)	(0.132)	(0.105)
CAL + Tutoring ($\beta_1 + \beta_2$)	0.189	0.004	-0.070	0.095	0.054	0.118
	(0.136)	(0.086)	(0.072)	(0.111)	(0.113)	(0.093)
N	1225	1225	1224	1218	1217	1218
Adjusted R^2	0.178	0.180	0.076	0.186	0.173	0.101

Notes: All specifications include covariates. Randomization inference standard errors in parentheses (1000 simulations). Covariates include students' gender, age, baseline math test scores, proxies for confidence and preferences toward math, self-reported grades for the previous year, and parents' employment status. We also control for the baseline level of the outcome variable when data are available. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$ based on randomization inference p-values.

Table 7: Treatment Effects on Math Overconfidence

	Midline		Endline	
	(1)	(2)	(3)	(4)
CAL Treatment (β_1)	0.0217	0.0359	-0.0032	0.0070
	(0.0446)	(0.0444)	(0.0338)	(0.0353)
CAL $\times$ Tutoring (β_2)	0.0264	0.0275	0.0456	0.0522
	(0.0406)	(0.0424)	(0.0347)	(0.0348)
CAL + Tutoring ($\beta_1 + \beta_2$)	0.0481	0.0634	0.0423	0.0592*
	(0.0427)	(0.0419)	(0.0332)	(0.0347)
N	1205	1205	1188	1188
Control Mean	0.560	0.560	0.580	0.580
Adjusted R^2	-0.000	0.051	-0.000	0.055
Covariates	No	Yes	No	Yes

Notes: The outcome variable takes the value of 1 if the teacher switched classes after program assignment and 0 otherwise. Randomization inference standard errors in parentheses (1000 simulations). Covariates include students' gender, age, baseline math test scores, proxies for confidence and preferences toward math, self-reported grades for the previous year, and parents' employment status. We also control for the baseline level of the outcome variable when data are available. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$ based on randomization inference p-values.

Table 8: Dose–Response Relationship for ALEKS Use

	Standardized Math Scores		
	IV estimates (1)	OLS VA (full sample) (2)	OLS VA (treatment group) (3)
CAL hours	0.0269***	0.0188***	0.0254***
	(0.0057)	(0.0031)	(0.0051)
Baseline math score (SD)	0.1433***	0.1459***	0.1129***
	(0.0100)	(0.0099)	(0.0144)
Observations	1218	1218	615
R-squared	0.3229	0.3286	0.3174
Difference-in-Sargan statistic for exogeneity (p-value)	0.0882		
Extrapolated estimates of 2.5-hour treatment	0.0673	0.0469	0.0635
Extrapolated estimates 25 hours (1500 minutes)	0.6731	0.4689	0.6353

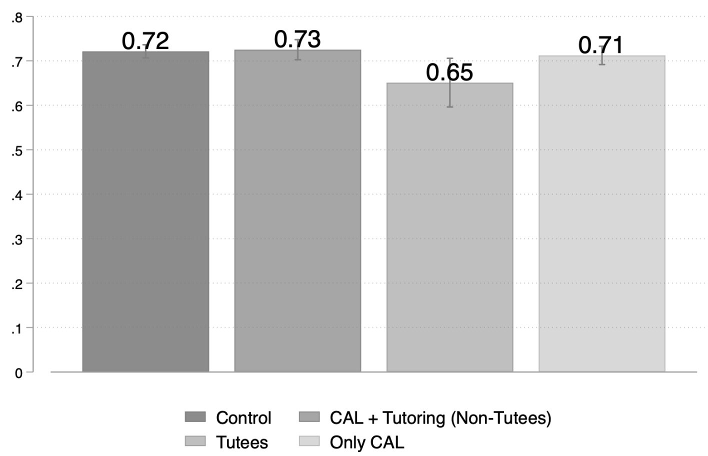
Notes: Robust standard errors in parentheses. * p <0.1, ** p <0.05, *** p <0.01. Treatment group students who were randomly selected for the ALEKS computer-adaptive learning (CAL) platform offer but who did not take up the offer have been marked as having 0 sessions, as have all students in the control group. Covariates include students' gender, age, baseline math test scores, proxies for confidence and preferences toward math, self-reported grades for the previous year, and parents' employment status. Column (1) instruments sessions in ALEKS with the randomized allocation. Column (2) presents estimates from ordinary least squares value added (OLS VA) models in the full sample. Column (3) presents OLS VA estimates using only data on students assigned to treatment. One hour is equivalent to 60 minutes of platform use.

Table A1: Lee Bounds – Single Regression Approach

	Midline Math Test		Endline Math Test	
	Lower Bound	Upper Bound	Lower Bound	Upper Bound
	(1)	(2)	(3)	(4)
CAL Treatment (β_1)	0.161	0.247*	0.235	0.381**
	(0.142)	(0.140)	(0.173)	(0.168)
CAL $\times$ Tutoring (β_2)	-0.185	-0.028	-0.261	-0.239
	(0.157)	(0.167)	(0.171)	(0.179)
CAL + Tutoring ($\beta_1 + \beta_2$)	0.062	0.133	-0.004	0.120
	(0.140)	(0.137)	(0.151)	(0.154)
Observations (before trimming)	1225		1218	
Observations (after trimming)	1206		1199	

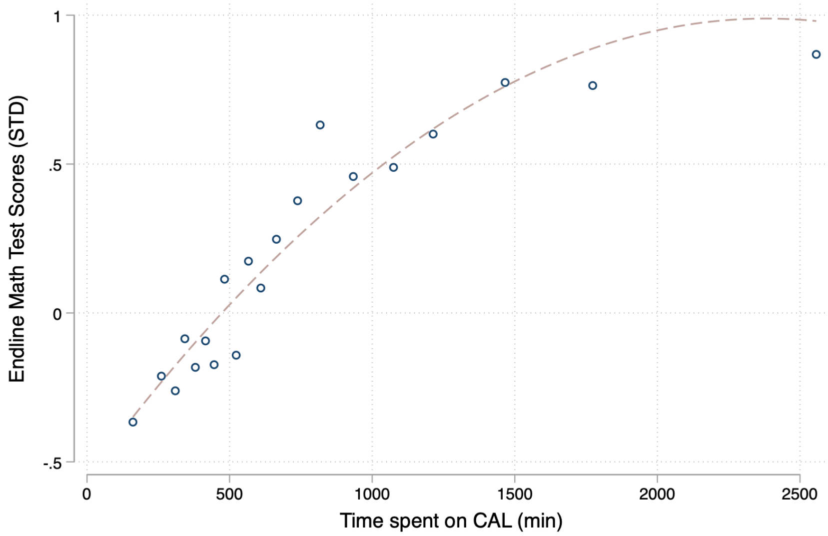
Notes: Lee (2009) bounds calculated with single regression approach for all three groups trimmed to minimum response rate (89.8 percent). Randomization inference standard errors in parentheses (1000 simulations). * p <0.1, ** p <0.05, *** p <0.01 based on randomization inference p-values.

Figure A1: Self-Efficacy Index by Performance Groups



Notes: The figure shows mean self-efficacy index values at endline across treatment groups. The CAL+Tutoring group excludes tutees, who are shown as a separate category. Self-efficacy is measured with a standardized index constructed from student survey responses. Error bars indicate 95 percent confidence intervals.

Figure A2: Returns to Hours of CAL Use



Notes: The figure shows the relationship between total time spent on the ALEKS computer-adaptive learning (CAL) platform (in minutes) and endline math test scores (in standard deviations) among treated students. Each point represents a binned mean. The dashed line shows a fitted curve illustrating the diminishing returns to platform use, with gains increasing steeply up to approximately 1,500 minutes before flattening. Sample is limited to treated students ($N = 664$).

A Additional Analysis

A.1 Power Calculations

To estimate statistical power, we simulate the randomization procedure described in Section 1.3 one thousand times and assume a minimum detectable effect (MDE) of 0.3 SD. Evans and Yuan (2022) find this to be in the 80th percentile of effect sizes of RCTs with learning or access outcomes in low- and middle-income countries. Although this value falls above the median of 0.2 for small RCTs, both tutoring and CAL have been identified as two of the most effective interventions to improve learning outcomes. A recent review of tutoring interventions by Nickow et al. (2020) finds an overall pooled effect size estimate of 0.37 SD. To our knowledge, no meta-studies have looked at effects from computer-*adaptive* learning technologies like the one in this study.

For this exercise, we use 2019 administrative data on ninth-grade students in the same sample of schools selected for this intervention. These data include students' test scores on a national diagnostic assessment, gender, age, and a socioeconomic indicator. We also simulate and include a baseline test score variable. These simulations give us a statistical power of approximately 80 percent, the standard for RCTs in the social sciences. We aim to maximize this number by collecting prognostic variables at baseline, including tests scores on a short math assessment and other student characteristics. In Table A2, we present the results of the simulations using alternative MDEs (0.25, 0.3, 0.35).

MDE	Power
0.25	0.674
0.30	0.799
0.35	0.893

Table A2: Power Simulations

B Additional Results

This appendix presents detailed results for the secondary outcomes summarized in Section 5.4.

B.1 Intrinsic Motivation, Absenteeism, and Dropout

Beyond academic performance, improvements in learning could positively influence students' retention in school by increasing their motivation and interest in the subject and their chances of obtaining a passing grade. In Tables B1, B2, and B3, we present impacts on intrinsic motivation, preferences toward math, absenteeism, and student dropout. Student dropout is a significant concern in our context, especially among the targeted age group. In the Dominican Republic, over 40 percent of students drop out between eighth and twelfth grade (IDEICE, 2017). In the control group, 8 percent of the students dropped out of school during the intervention period (see Table B3). However, we find no evidence that the intervention impacted students' intrinsic motivation, how much they like math (self-reported on a 1–4 scale), school absenteeism, or dropout rates by endline. The null effect on dropout is consistent with the possibility that students face binding nonacademic constraints—such as economic pressures or family obligations—that learning gains alone cannot address. The lack of an effect on intrinsic motivation aligns with the result of Araya et al. (2019), who also observe no effect on this outcome in a study using a game-based learning platform in Chile.

B.2 Math Self-Concept

As students spend more time on the software and master new topics, they receive new information on their abilities and may update their beliefs accordingly. In Table B4, we present effects on students' math self-concept, defined as students' perceptions of their own ability in mathematics compared to others. As in Araya et al. (2019), despite the large increases observed in math performance, we do not find statistically significant differences across experimental groups on a math self-concept index or on the individual measures used to build the index. However, students in the CAL+Tutoring group were 9.3 percentage points less likely to agree with the statement "Math is harder for me than for many of my classmates" than students in the control group or in the CAL-only group. Though these results are not conclusive, they may suggest shifts in students' self-perception and confidence in their math abilities. Despite performing worse on math tests, the students in the tutoring group seem to have felt equally confident about their performance and were likelier to have mismatched expectations about their outcomes in the test than those whose classrooms were not assigned the tutoring component.

B.3 Self-Efficacy, Grit, and Perseverance

By offering personalized learning pathways and continuous feedback on individual progress and milestones, the platform allows students to understand and learn from their mistakes privately, without the stigma of public correction. As students master new topics and receive real-time feedback, this process may enhance their sense of competence and self-efficacy. In this subsection, we examine four related outcomes: self-efficacy, grit, perseverance, and internal locus of control. As shown in Table B5, neither the CAL intervention nor the tutoring intervention had a significant impact on a self-efficacy index.[19] This result suggests that the interventions did not notably alter students' beliefs in their ability to achieve academic success.

Similarly, we do not find impacts on a measure of grit (see Table B6), which assesses students' perseverance and passion for long-term goals (Duckworth et al., 2007). Following Alan et al. (2019), we implemented an alternative measure of perseverance based on a real-effort exercise in which students chose between attempting an easier task, a more difficult task, or giving up. As shown in Table B7, CAL-only students were 9.3 percentage points more likely to choose the easier task and 9.7 percentage points less likely to choose the difficult task relative to the control group, both statistically significant at the 5 percent level. A similar pattern holds for the CAL+Tutoring group, which also shows a reduction in the likelihood of choosing a difficult task relative to the control group (-7.2 percentage points, $p < 0.10$). We do not observe statistically significant differences in the likelihood of giving up after failure for either group. These patterns suggest that exposure to an adaptive platform—which continuously calibrates exercise difficulty to students' current ability—may reduce willingness to voluntarily engage with more challenging material, and that this effect persists even when tutoring is added.

B.4 Student Well-Being

To assess changes in students' well-being, we use two proxies: the Children's Depression Screener (ChilD-S) and a short questionnaire on math anxiety. For the first one, we built an index from students' answers to 8 items about their mental state (e.g., "I feel happy," "I feel lonely," "It is hard for me to concentrate"). For the second one, we used two questions on students' worry about their math grades and their feelings before math tests to build a math anxiety index. Table B9 shows that the intervention had no impact on the first measure of well-being. Though we do not find impacts on a math anxiety index, we find important differences when looking at individual measures of math anxiety. As seen in Table B10, students in the CAL-only group were 6.3 percentage points likelier to report getting nervous before math tests than students in the control and CAL+Tutoring groups. This difference is marginally significant ($p<0.10$). Although math anxiety is typically linked to

[19] This index is built from students' answers to the following items: 1) "I can always solve difficult problems if I try hard enough," 2) "I can solve most problems if I invest the necessary effort," 3) "I can stay calm when I face difficulties because I can trust my coping skills," 4) "When faced with a problem, I can usually find several solutions," and 5) "If I'm in trouble I can find a solution."

lower performance (Dowker et al., 2016), this result may be positive if it signals greater awareness or concern about their math performance in this group.

B.5 Teacher Practices

The CAL platform may free up teachers' time by automating tasks such as developing and grading homework and tests. Additionally, the presence of tutors in the CAL+Tutoring classrooms may have affected teacher behavior through increased accountability or organizational demands. Table B11 presents effects on indices of teacher time mismanagement and classroom engagement, constructed from student reports.

The CAL+Tutoring classrooms showed lower teacher time mismanagement than the control group (0.23 SD without controls, $p < 0.05$), though this result is not robust to the inclusion of covariates. For teacher engagement, the CAL-only classrooms showed marginally lower scores than the control group (0.14 to 0.18 SD), though this effect is not statistically significant. We similarly do not observe significant differences in teacher engagement for the CAL+Tutoring group.

Platform data provide additional context: Though teachers spent an average of 33 minutes per week actively using the platform, they seldom leveraged it to automate routine tasks. Over the course of the academic year, teachers assigned no more than three activities through the CAL platform, suggesting limited adoption of the platform's labor-saving features.

B.6 Student Time Use

One potential concern with technology-based interventions is that they might displace other productive activities or increase screen time outside school. To examine whether the intervention affected how students allocate their time, we collected data on hours spent in eight categories: in-person classes, unpaid work at home, unpaid work outside the home, paid work, studying outside school, meeting friends, entertainment, and sports.

Table B12 presents treatment effects on student time use. We find no statistically significant effects in any of these categories for either the CAL-only or CAL+Tutoring groups. The null effects suggest that the intervention operated primarily through the substitution of regular math instruction during school hours rather than changes in how students allocated time outside of CAL sessions.

Table B1: Treatment Effects on Student Motivation and Attitudes

	Intrinsic Motivation (1)	Math Enjoyment (2)	Math Confidence (3)	Math Importance (4)
CAL Treatment (β_1)	-0.0218	0.0363	-0.0059	-0.1416**
	(0.0829)	(0.0683)	(0.0874)	(0.0626)
CAL × Tutoring (β_2)	0.1016	0.0769	-0.0137	0.0943
	(0.0925)	(0.0772)	(0.0962)	(0.0848)
CAL + Tutoring ($\beta_1 + \beta_2$)	0.0798	0.1132*	-0.0197	-0.0473
	(0.0802)	(0.0672)	(0.0791)	(0.0600)
N	1194	1191	1179	1181
Control Mean	-0.003	2.731	2.331	2.746
Adjusted R^2	0.101	0.084	0.037	0.065

Notes: Randomization inference standard errors in parentheses (1000 simulations). All regressions include baseline covariates: student gender, age, baseline math score, parental employment status, and student confidence in math. * p <0.1, ** p <0.05, *** p <0.01 based on randomization inference p-values.

Table B2: Treatment Effects on Math Preferences

	Midline		Endline	
	(1)	(2)	(3)	(4)
CAL Treatment (β_1)	-0.0430	0.0109	-0.0163	0.0227
	(0.0868)	(0.0732)	(0.0868)	(0.0831)
CAL × Tutoring (β_2)	-0.0225	-0.0330	0.0243	0.0241
	(0.1079)	(0.0847)	(0.1078)	(0.0970)
CAL + Tutoring ($\beta_1 + \beta_2$)	-0.0655	-0.0221	0.0079	0.0468
	(0.0815)	(0.0696)	(0.0810)	(0.0774)
N	1218	1218	1210	1210
Control Mean	2.787	2.787	2.673	2.673
Adjusted R^2	-0.001	0.322	-0.002	0.235
Covariates	No	Yes	No	Yes

Notes: Randomization inference standard errors in parentheses (1000 simulations). Covariates include students' gender, age, baseline math test scores, proxies for confidence and preferences toward math, self-reported grades for the previous year, and parents' employment status. When available, we also control for the baseline level of the outcome variable. * p <0.1, ** p <0.05, *** p <0.01 based on randomization inference p-values.

Table B3: Treatment Effects on Absence and School Dropout

	Absence		Dropout	
	(1)	(2)	(3)	(4)
CAL Treatment (β_1)	-1.4203	-0.6303	-0.0234	-0.0142
	(2.1908)	(1.7159)	(0.0232)	(0.0186)
CAL × Tutoring (β_2)	1.9477	1.1173	-0.0059	-0.0133
	(2.4394)	(1.8967)	(0.0285)	(0.0238)
CAL + Tutoring ($\beta_1 + \beta_2$)	0.5275	0.4870	-0.0293	-0.0274*
	(2.0593)	(1.6236)	(0.0209)	(0.0165)
N	1333	1333	1333	1333
Control Mean	10.614	10.614	0.077	0.077
Adjusted R^2	0.002	0.138	0.001	0.056
Covariates	No	Yes	No	Yes

Notes: Randomization inference standard errors in parentheses (1000 simulations). Attendance measures are specific to math classes. Covariates include students' gender, age, baseline math test scores, proxies for confidence and preferences toward math, self-reported grades for the previous year, and parents' employment status. We also control for the baseline level of the outcome variable when data are available. * p <0.1, ** p <0.05, *** p <0.01 based on randomization inference p-values.

Table B4: Treatment Effects on Math Self-Concept

	Math Self-Concept Index (1)	I usually do well in math (2)	Math is harder for me than for many of my classmates (3)
CAL Treatment (β_1)	-0.0065	0.0010	0.0302
	(0.0376)	(0.0821)	(0.0685)
CAL × Tutoring (β_2)	0.0127	0.0032	-0.1029
	(0.0408)	(0.0935)	(0.0769)
CAL + Tutoring ($\beta_1 + \beta_2$)	0.0062	0.0042	-0.0727
	(0.0348)	(0.0803)	(0.0624)
N	1184	1182	1175
Control Mean	0.693	2.658	2.523
Adjusted R^2	0.160	0.164	0.063

Notes: Randomization inference standard errors in parentheses (1000 simulations). Covariates include students' gender, age, baseline math test scores, proxies for confidence and preferences toward math, self-reported grades for the previous year, and parents' employment status. We also control for the baseline level of the outcome variable when data are available. * p <0.1, ** p <0.05, *** p <0.01 based on randomization inference p-values.

Table B5: Treatment Effects on Self-Efficacy Index

	Midline		Endline	
	(1)	(2)	(3)	(4)
CAL Treatment (β_1)	-0.0134	-0.0155	-0.0081	-0.0114
	(0.0150)	(0.0124)	(0.0130)	(0.0115)
CAL × Tutoring (β_2)	0.0179	0.0196	-0.0011	0.0045
	(0.0175)	(0.0127)	(0.0174)	(0.0157)
CAL + Tutoring ($\beta_1 + \beta_2$)	0.0046	0.0041	-0.0091	-0.0069
	(0.0134)	(0.0117)	(0.0135)	(0.0133)
N	1217	1217	1194	1194
Control Mean	0.765	0.765	0.721	0.721
Adjusted R^2	0.000	0.248	-0.001	0.129
Covariates	No	Yes	No	Yes

Notes: Randomization inference standard errors in parentheses (1000 simulations). Covariates include students' gender, age, baseline math test scores, proxies for confidence and preferences toward math, self-reported grades for the previous year, and parents' employment status. We also control for the baseline level of the outcome variable when data are available. * p <0.1, ** p <0.05, *** p <0.01 based on randomization inference p-values.

Table B6: Treatment Effects on Grit Index

	Endline	
	(1)	(2)
CAL Treatment (β_1)	0.0104	0.0051
	(0.0134)	(0.0121)
CAL × Tutoring (β_2)	-0.0003	-0.0004
	(0.0152)	(0.0135)
CAL + Tutoring ($\beta_1 + \beta_2$)	0.0101	0.0047
	(0.0127)	(0.0117)
N	1333	1333
Control Mean	0.470	0.470
Adjusted R^2	-0.001	0.033
Covariates	No	Yes

Notes: The outcome is a standardized variable based on students' answers to the following items: 1) "The schoolwork I like the most is what makes me think;" 2) "If I think I will lose in a game, I do not want to continue playing;" 3) "If I set a goal and see that it's harder than I thought, I lose interest;" and 4) "I work hard on tasks." Answer options ranged from 1 ("Doesn't seem like me at all") to 5 ("Very much like me"). Randomization inference standard errors in parentheses (1000 simulations). Covariates include students' gender, age, baseline math test scores, proxies for confidence and preferences toward math, self-reported grades for the previous year, and parents' employment status. We also control for the baseline level of the outcome variable when data are available. * $p<0.1$, ** $p<0.05$, *** $p<0.01$ based on randomization inference p-values.

Table B7: Treatment Effects on Logic Task

	Easy (1)	Difficult (2)	Give Up (3)
CAL Treatment (β_1)	0.0929**	-0.0966**	0.0037
	(0.0459)	(0.0410)	(0.0281)
CAL $\times$ Tutoring (β_2)	-0.0416	0.0250	0.0167
	(0.0551)	(0.0543)	(0.0340)
CAL + Tutoring ($\beta_1 + \beta_2$)	0.0512	-0.0716*	0.0204
	(0.0469)	(0.0409)	(0.0253)
N	1211	1211	1211
Control Mean	0.433	0.380	0.187
Adjusted R^2	0.068	0.072	-0.001

Notes: Randomization inference standard errors in parentheses (1000 simulations). Covariates include students' gender, age, baseline math test scores, proxies for confidence and preferences toward math, self-reported grades for the previous year, and parents' employment status. We also control for the baseline level of the outcome variable when data are available. * p <0.1, ** p <0.05, *** p <0.01 based on randomization inference p-values.

Table B8: Treatment Effects on Locus of Control Index

	Endline	
	(1)	(2)
CAL Treatment (β_1)	0.0105	0.0078
	(0.0087)	(0.0081)
CAL $\times$ Tutoring (β_2)	-0.0246*	-0.0216*
	(0.0111)	(0.0110)
CAL + Tutoring ($\beta_1 + \beta_2$)	-0.0141	-0.0138*
	(0.0084)	(0.0081)
N	1196	1196
Control Mean	0.667	0.667
Adjusted R^2	0.003	0.073
Covariates	No	Yes

Notes: Randomization inference standard errors in parentheses (1000 simulations). Covariates include students' gender, age, baseline math test scores, proxies for confidence and preferences toward math, self-reported grades for the previous year, and parents' employment status. We also control for the baseline level of the outcome variable when data are available. * p <0.1, ** p <0.05, *** p <0.01 based on randomization inference p-values.

Table B9: Treatment Effects on Student Depression

	Depression Index	
	(1)	(2)
CAL Treatment (β_1)	0.0116	0.0090
	(0.0112)	(0.0091)
CAL $\times$ Tutoring (β_2)	-0.0046	-0.0047
	(0.0109)	(0.0089)
CAL + Tutoring ($\beta_1 + \beta_2$)	0.0070	0.0043
	(0.0097)	(0.0079)
N	1205	1205
Control Mean	0.612	0.612
Adjusted R^2	-0.000	0.082
Covariates	No	Yes

Notes: The outcome variable is a standardized index based on students' answers to the following items: 1) "I am happy," 2) "I worry a lot," 3) "I feel sad," 4) "I get upset quickly," 5) "I often think I did something wrong," 6) "It's hard for me to concentrate," 7) "I feel lonely," and 8) "I am not happy." Answers range from 1 (strongly disagree) to 4 (strongly agree). Randomization inference standard errors in parentheses (1000 simulations). Covariates include students' gender, age, baseline math test scores, proxies for confidence and preferences toward math, self-reported grades for the previous year, and parents' employment status. We also control for the baseline level of the outcome variable when data are available. * p <0.1, ** p <0.05, *** p <0.01 based on randomization inference p-values.

Table B10: Treatment Effects on Math Anxiety

	I'm worried of having bad grades in math (1)	I get nervous before math tests (2)	Math Anxiety Index (3)
CAL Treatment (β_1)	0.0333	0.0632*	0.1404
	(0.0326)	(0.0356)	(0.0911)
CAL $\times$ Tutoring (β_2)	-0.0116	-0.0630	-0.1202
	(0.0356)	(0.0390)	(0.0962)
CAL + Tutoring ($\beta_1 + \beta_2$)	0.0218	0.0003	0.0202
	(0.0294)	(0.0358)	(0.0847)
N	1199	1183	1182
Control Mean	0.773	0.713	0.000
Adjusted R^2	0.028	0.036	0.036

Notes: Randomization inference standard errors in parentheses (1000 simulations). Covariates include students' gender, age, baseline math test scores, proxies for confidence and preferences toward math, self-reported grades for the previous year, and parents' employment status. We also control for the baseline level of the outcome variable when data are available. * p <0.1, ** p <0.05, *** p <0.01 based on randomization inference p-values.

Table B11: Treatment Effects on Teacher Behavior Indices

	Time Mismanagement		Teacher Engagement	
	(1)	(2)	(3)	(4)
CAL Treatment (β_1)	-0.0752	-0.0247	-0.1782	-0.1430
	(0.1205)	(0.1155)	(0.1082)	(0.0915)
CAL × Tutoring (β_2)	-0.1532	-0.1546	0.0415	0.0372
	(0.1169)	(0.1100)	(0.1248)	(0.1141)
CAL + Tutoring ($\beta_1 + \beta_2$)	-0.2284**	-0.1793	-0.1367	-0.1058
	(0.1143)	(0.1135)	(0.0963)	(0.0776)
N	1148	1148	1151	1151
Adjusted R^2	0.007	0.117	0.005	0.116
Covariates	No	Yes	No	Yes

Notes: Randomization inference standard errors in parentheses (1000 simulations). Covariates include students' gender, age, baseline math test scores, proxies for confidence and preferences toward math, self-reported grades for the previous year, and parents' employment status. We also control for the baseline level of the outcome variable when data are available. * p <0.1, ** p <0.05, *** p <0.01 based on randomization inference p-values.

Table B12: Treatment Effects on Student Time Use

	In-person classes	Unpaid work (at home)	Unpaid work (out of home)	Paid work (out of school)	Study	Meeting friends	Entertain-ment	Sports
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
CAL Treatment (β_1)	0.0137	-0.2492	-0.0116	0.0393	0.0577	0.0283	0.0444	0.0478
	(0.2324)	(0.2158)	(0.2136)	(0.2400)	(0.1465)	(0.1381)	(0.0911)	(0.0955)
CAL × Tutoring (β_2)	-0.2701	-0.0167	-0.1548	-0.2637	0.0460	-0.1452	-0.0180	0.0013
	(0.2708)	(0.2254)	(0.2199)	(0.2486)	(0.1495)	(0.2028)	(0.1025)	(0.1166)
CAL + Tutoring	-0.2563	-0.2660	-0.1664	-0.2244	0.1037	-0.1170	0.0264	0.0491
	(0.2433)	(0.1878)	(0.2067)	(0.2187)	(0.1365)	(0.1233)	(0.0845)	(0.0938)
Control Mean	6.01	3.80	2.27	2.73	2.03	2.24	3.33	2.08
Adjusted R^2	0.06	0.05	0.11	0.11	0.01	0.04	0.01	0.13
Observations	1182	1168	1156	1152	1177	1169	1169	1166

Notes: * p <0.01, ** p <0.05, *** p <0.1. Covariates include student's gender, age, baseline math test scores, proxies for confidence and preferences toward math, self-reported grades for the previous year, and parents' employment status. Randomization inference standard errors in parentheses (1000 simulations).