Annenberg Institute
BROWN UNIVERSITY

EdWorkingPaper No. 26-1551

One Click Away: AI Tutoring with Khanmigo in a Two-Year School Experiment

Philip Oreopoulos
University of Toronto

Nina Low
Charles River Associates

Generative AI has been promoted as the technology that could transform education by providing every student a personal tutor. We provide some of the first large-scale experimental evidence, from a two-year cluster randomized trial in 18 Tennessee middle schools in which randomly assigned students used Khan Academy with its AI tutor, Khanmigo, configured to coach rather than give answers, during existing daily remedial mathematics sessions. Assignment raises math achievement by 1.3 national percentile ranks per term, or about 0.06 to 0.08 standard deviations over a school year; the implied effect of a full year of active participation reaches 0.14 standard deviations. These gains resemble those from Khan Academy practice without AI assistance. One explanation is that students used the tutor infrequently and, when they did, rarely engaged it in substantive mathematical dialogue: 96 percent of students tried Khanmigo at least once, but the median student messaged it on only a third of the days they practiced, and in only 17 percent of the exercise sessions in which they made a mistake. Messages that students did send were mostly bare answers or clicks on suggested prompts. The binding constraint appears to be engagement: realizing the promise of AI tutoring will require getting students to use it, not just giving them access.

VERSION: August 2026

Suggested citation: Oreopoulos, Philip, and Nina Low. (2026). One Click Away: AI Tutoring with Khanmigo in a Two-Year School Experiment. (EdWorkingPaper: 26-1551). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/kner-hv33

One Click Away: AI Tutoring with Khanmigo in a Two-Year School Experiment

Philip Oreopoulos* Nina Low†

Abstract

Generative AI has been promoted as the technology that could transform education by providing every student a personal tutor. We provide some of the first large-scale experimental evidence, from a two-year cluster randomized trial in 18 Tennessee middle schools in which randomly assigned students used Khan Academy with its AI tutor, Khanmigo, configured to coach rather than give answers, during existing daily remedial mathematics sessions. Assignment raises math achievement by 1.3 national percentile ranks per term, or about 0.06 to 0.08 standard deviations over a school year; the implied effect of a full year of active participation reaches 0.14 standard deviations. These gains resemble those from Khan Academy practice without AI assistance. One explanation is that students used the tutor infrequently and, when they did, rarely engaged it in substantive mathematical dialogue: 96 percent of students tried Khanmigo at least once, but the median student messaged it on only a third of the days they practiced, and in only 17 percent of the exercise sessions in which they made a mistake. Messages that students did send were mostly bare answers or clicks on suggested prompts. The binding constraint appears to be engagement: realizing the promise of AI tutoring will require getting students to use it, not just giving them access.

Keywords: Artificial intelligence; AI tutoring; Khanmigo; computer-assisted learning; educational technology; field experiment; math achievement; student engagement; response to intervention

JEL Codes: I21, I24, J24, O33

*Department of Economics, University of Toronto, and NBER. Email: philip.oreopoulos@utoronto.ca.
†Charles River Associates. Email: nlow@crai.com.

Acknowledgments: We are very grateful to Hamilton County Schools, participating teachers, school leaders, and students for making this study possible. We thank John Rice, Kassidy Cunningham, Victoria Hales, and Courtney Lucas, our partners at Hamilton County Schools, for their collaboration and support throughout this project. We thank Keli Hill and colleagues at Khan Academy for guidance on the platform and its features, for facilitating the district partnership that enabled account rostering and dashboard-based monitoring, and for coordinating the data-sharing agreement under which platform data were provided. We thank Oliver Keyes-Krysakowski, Ruochong Dong, and Hiba Jamshedi for excellent research assistance. This research was supported by the Smith Richardson Foundation, the Social Sciences and Humanities Research Council of Canada (SSHRC), the Natural Sciences and Engineering Research Council of Canada (NSERC), and the Abdul Latif Jameel Poverty Action Lab (J-PAL). The study was registered with the AEA RCT Registry under trial number AEARCTR-0013519 and received research ethics approval from the University of Toronto Research Ethics Board. The views expressed herein are those of the authors and do not necessarily reflect the views of the funders or Charles River Associates.

The most effective way to teach a struggling student is also the most expensive. A tutor who can diagnose a student's specific error, explain the step they missed, and adapt the next problem to their misunderstanding produces some of the largest gains in education, a third of a standard deviation or more from the best one-on-one programs (Nickow et al., 2024; Guryan et al., 2023). That effectiveness comes from a skilled adult's attention delivered one student at a time, which is exactly why it does not scale well: high-dosage tutoring often costs several thousand dollars per student per year, and efforts to expand it confront persistent staffing and fidelity problems (Kraft and Falken, 2021; Bhatt et al., 2024). The binding constraint has been human labor rather than curriculum or technology. School systems serving students below grade level therefore fall back on group-based remediation through Response to Intervention and similar tiered models, which reach many students cheaply but cannot tailor feedback to each one in real time. The individualized feedback that works best has been the hardest to deliver at scale.

Large language models change the cost of producing that feedback. Earlier computer-assisted learning raised achievement by customizing pacing and problem sequences while delivering fixed, pre-authored hints (Escueta et al., 2020). A large language model can instead generate adaptive feedback in natural-language dialogue—helping a student get started on an unfamiliar problem, walking through a worked example, or interpreting a specific error and surfacing the misconception behind it—at a marginal cost per interaction close to zero. In principle this reproduces part of what makes human tutoring effective, at a cost structure that scales. Whether it does so in practice is far less clear: the experimental evidence on LLM tutoring remains small, short-horizon, and concentrated outside ordinary public school instruction because few AI tutoring systems have been deployed widely enough, or long enough, to study.

This paper reports what we believe is the largest and longest experimental evaluation of an AI tutoring program in public schools to date. Over the 2024–25 and 2025–26 school years, we conducted a cluster randomized trial across 18 middle schools in Hamilton County, Tennessee. Within each school, we randomly assigned one or two grades (out of three) to use Khan Academy together with Khanmigo—one of the first AI tutors deployed at scale in U.S. schools, designed to coach rather than tell—during the district's existing daily remedial mathematics sessions, twenty-five- to forty-minute small-group blocks that below-grade-level students already attend. Khanmigo withholds answers, asks guiding questions, and sits one click away as students work through each problem. Students in control grades continued their usual remedial instruction, a mix of teacher-led practice and other math software without AI tutoring. Because access is embedded in mandatory session time rather than offered for optional use, the estimate is not attenuated by the selective take-up that plagues opt-in

evaluations of educational technology.

Three findings emerge. First, the program generated substantial practice and correspondingly higher achievement. Treated students averaged about 30 additional minutes of weekly practice on Khan Academy—and among students actively enrolled in the remedial program, over 45 minutes per week in the second year, as implementation matured. Assignment increased mathematics scores on the district's benchmark assessment by 1.3 national percentile ranks per term, or roughly 0.06 standard deviations per school year and about 0.08 in the second year. These intent-to-treat estimates average over students who exit the remedial program as their achievement improves; the implied effect of a full year of participation in the second year is 0.14 standard deviations. Effects track realized practice: gains are concentrated in the second year, when practice was highest.

Second, these gains are similar to those predicted by the same platform without AI assistance. Quasi-experimental estimates of Khan Academy practice without Khanmigo imply returns per hour of practice close to what we estimate here (Eames et al., 2026), and our effects sit within the range that computer-assisted learning field experiments have produced for two decades (Escueta et al., 2020). Nothing in the achievement data requires the AI tutor to explain it.

Third, we observe how students actually used the AI tutor, and the answer is: not much. Message-level records matched to exercise logs are available for treated students in Year 2—the year of strongest implementation and largest effects—and they show that access was nearly universal but engagement was thin. Ninety-six percent of students messaged Khanmigo at least once. Yet the median student sent it no messages on two-thirds of the days they practiced, and messaged it in only 17 percent of the exercise sessions in which they made a mistake. The messages students did send were mostly one- or two-word answers to Khanmigo's own questions or clicks on suggested prompts; 14.5 percent contained a mathematical question or a step of reasoning.

These findings point to engagement, rather than access, as the margin that limits AI tutoring in this setting. The deployment removed the barriers that typically lower take-up of educational technology—the software was free, embedded in mandatory session time, and supported by a teacher, and the AI tutor itself prompted students with suggested questions—yet while these conditions generated substantial practice, they did not generate much tutoring dialogue. Seeking help with one's own confusion remained a choice, and most students declined it most of the time, a pattern consistent with evidence of behavioral barriers such as present bias, reliance on routine, and help avoidance (Lavecchia et al., 2016). Because assignment moves every component of the program together, we cannot rule out some contribution from the tutor where it was engaged, and whether richer features—like being more

proactive, button driven, with voice, vision, or memory of the student—would elicit more engagement. What the results do establish is that realized AI dosage was low, and that the program's effects likely operated through structured practice, at a cost per unit of learning far below human tutoring: at roughly $15 per student per year, the program delivers a meaningful share of the individualized feedback margin that has historically required expensive human attention, and unlike tutoring it scales without a corresponding increase in skilled labor. Projections of what AI tutoring can achieve in schools should rest on how students actually use it rather than on what the technology can do in principle.

The paper proceeds as follows. Section 1 places the study in the literature on tutoring, computer-assisted learning, and the emerging evidence on AI tutors. Section 2 describes the setting, experiment, and data. Section 3 presents take-up: platform practice and Khanmigo use. Section 4 presents effects on achievement. Section 5 interprets the findings and their cost-effectiveness, and Section 6 concludes.

1 Background and Related Literature

Individualized tutoring is among the most effective interventions in education. Meta-analytic evidence from randomized evaluations puts the average effect of tutoring programs at roughly 0.3 standard deviations, among the largest of any schooling intervention (Nickow et al., 2024; Dietrichson et al., 2017), and intensive daily tutoring has produced gains of similar magnitude for struggling adolescents (Guryan et al., 2023). The challenge is cost: high-dosage programs often run several thousand dollars per student per year, and efforts to scale them across districts confront persistent staffing, scheduling, and fidelity problems (Kraft and Falken, 2021; Bhatt et al., 2024). For students below grade level, school systems therefore rely on cheaper group-based remediation, which reaches many students but cannot respond to each one individually.

Computer-assisted learning (CAL) may help to close part of that gap. Experimental evaluations of math practice software generally find positive effects in the range of 0.05 to 0.20 standard deviations (Escueta et al., 2020; Barrow et al., 2009; Roschelle et al., 2016; Pane et al., 2014), with larger effects where adaptive software substitutes for especially weak instruction, as in Muralidharan et al. (2019). Intelligent tutoring systems, the pre-LLM generation of software offering structured hints and mastery-based progression, show comparable averages in meta-analysis (Kulik and Fletcher, 2016). These technologies personalize the sequencing and pacing of practice, but their feedback is pre-authored: the software can tell a student that an answer is wrong and display a canned hint, but it cannot converse with the student about why.

A recurring lesson from this literature is that realized use, not software capability, is the binding margin. When platforms are merely made available, usage is low and concentrated among students who need help least: the median U.S. Khan Academy user logs only a few hours of practice per year (Eames et al., 2026), opt-in educational resources are claimed disproportionately by advantaged families (Robinson et al., 2025), and the strongest platform effect sizes often come from the small, self-selected minority of high users, a pattern Holt (2024) labels the "5 percent problem." Usage responds to structure. Effects concentrate where schools dedicate session time and teachers actively integrate the software (Ferman et al., 2019); training and supporting teachers to structure CAL practice raises both usage and achievement (Oreopoulos et al., 2024); hybrid models that alternate human tutoring with platform practice show promise of preserving most of tutoring's effect at lower cost (Bhatt et al., 2024). Per-hour returns to platform practice, moreover, appear lower for weaker students (Eames et al., 2026), precisely the population remediation targets.

Large language models relax the feedback constraint that limited earlier CAL. A generative tutor can interpret a student's specific error, answer a question posed in the student's own words, and walk through a worked example in dialogue, at a marginal cost per interaction near zero. The prospect attracted immediate attention, accompanied by predictions of transformative effects on education (Khan, 2023). Khan Academy, granted early access to GPT-4, launched Khanmigo in 2023 as among the first LLM tutors deployed in U.S. schools. Reflecting concerns that unconstrained chatbots would do students' work for them, Khanmigo was designed to withhold answers and guide students with questions. Subsequent evidence is consistent with the reasoning behind this design: Bastani et al. (2025) find that high school students given unrestricted GPT-4 access solved more practice problems but performed *worse* on subsequent unassisted exams, while a version restricted to tutoring-style guidance avoided the harm.

The experimental evidence on LLM tutors is recent, small, and largely classroom-based. Kestin et al. (2025) randomize the order in which Harvard undergraduates experience two physics lessons, one taught in a well-designed active-learning classroom and one through an AI tutor purpose-built for the course, and find the AI condition roughly doubles learning gains on an immediate posttest. Henkel et al. (2024) evaluate Rori, a WhatsApp-based conversational math tutor, across eleven Ghanaian schools and find growth-score gains of about 0.36 standard deviations from two weekly in-school sessions. LearnLM Team, Google DeepMind and Eedi (2025) integrate a pedagogically fine-tuned model into regular classroom mathematics practice in five UK secondary schools: when a diagnostic question reveals a misconception, the online platform opens a text chat with a tutor presented to the student as human, whose AI-drafted messages an expert tutor vets before sending, and randomizes

each session between AI-drafted and human tutoring; students in the AI sessions perform at least as well as those tutored by humans alone. Wang et al. (2025) find that giving human tutors real-time LLM suggestions raises students' topic mastery by four percentage points, with the largest gains for the least experienced tutors. These studies establish that generative tutoring can produce learning, but they are short-horizon, their comparisons are identified conditional on a tutoring exchange already underway, and in each the dialogue is text-based, human-mediated, or initiated for the student rather than by the student.

What these studies leave open is what happens when a conversational tutor is simply made available to ordinary students, within routine instruction, over an extended period—the margin on which adoption at scale currently rests, and the one this study measures. The closest evidence on that margin comes from an adaptive literacy platform in U.S. elementary schools: students given dedicated time engaged with it for only minutes per week, and assigning human tutors to encourage engagement raised usage without raising achievement (Robinson et al., 2026). In a companion one-class experiment in the same district as this study, we randomize access to an AI tutor and whether students were required to use it after mistakes. We find that AI assistance improves next-question accuracy, and find positive but small and imprecise effects on a short test one week later (Oreopoulos et al., 2026).

Khanmigo's own trajectory illustrates the gap between capability and engagement. Three years after launch, Khan Academy's leadership acknowledged that the tutor had been, for most students, "a non-event": most students made little use of it, the anticipated transformation in personalized instruction had not materialized, and students often struggled to formulate the questions a conversational tutor requires (Barnum, 2026). In 2026 the organization redesigned its student experience to activate Khanmigo automatically during practice, explaining that students had not been seeking out the tutor's help on their own (Barnum, 2026). These practitioner assessments echo the take-up findings from the CAL literature, but they rest on internal observation rather than on randomized deployment with directly measured usage and achievement outcomes.

This study contributes evidence of that kind. The trial spans two school years and fifty-three randomized grade-level clusters across eighteen schools; it is embedded in routine public school remediation rather than added to it, so that delivery inside a mandatory daily session addresses the access margin that depresses voluntary take-up; and message-level records allow us to observe not only whether the AI tutor was available but how, and how much, students used it.

2 Setting, Experiment, and Data

2.1 The RTI Program and the Intervention

Hamilton County Schools, in Tennessee, operate a tiered Response to Intervention (RTI) program in mathematics for grades 6–8. RTI is the standard framework through which U.S. public schools deliver supplemental instruction to struggling students (Fuchs and Fuchs, 2006). The district identifies students for RTI through trimester universal screening on the Measures of Academic Progress (MAP) assessment, supplemented by teacher referral; screening thresholds vary by school, and schools typically prioritize reading remediation when a student scores low in both subjects. Identified students attend an intervention block of twenty-five to forty minutes, scheduled daily, in groups of roughly four to eight, led by an intervention specialist or classroom teacher. Scheduled and delivered time differ in practice: blocks are shortened, repurposed, or cancelled for assemblies, standardized testing windows, field trips, staff absences, and the ordinary interruptions that consume instructional time in U.S. schools (Kraft and Monti-Nussbaum, 2021), and students are sometimes pulled for competing services. Realized program dosage is therefore below the actual scheduled ceiling.

The intervention replaced the mathematics content of these sessions with practice on Khan Academy accompanied by Khanmigo, Khan Academy's generative AI tutor. Practice is individualized through Khan Academy's integration with the MAP assessment itself: each student's most recent RIT scores are imported automatically and place the student into personalized learning pathways in the four MAP instructional areas of middle school mathematics—operations and algebraic thinking, the real and complex number systems, geometry, and statistics and probability—so that every student practices exercises and topic videos pitched at their current level in each area, with teachers able to review and adjust placements.

Khanmigo is available throughout via a text-to-text chat window, and students can also reach it through suggested prompts that appear alongside exercises. The tutor is configured for guided help: it does not supply answers, instead responding with hints, questions, and step-by-step prompts tied to the student's current problem. Teachers can monitor progress through the platform dashboard and manage session logistics.

During Year 1 and most of Year 2, Khanmigo appeared to students as a chat icon on the exercise page that students clicked to open, alongside suggested help prompts available throughout (for example, a prominent "Tutor Me" control at the lower right of the screen, beside the buttons for advancing to the next question). On January 29, 2026—between the Winter 2026 and Spring 2026 delivery windows—Khan Academy switched the tutor to open

by default shortly after page load.[1] A broader redesign that activates Khanmigo throughout practice, adopted in response to low student-initiated use, was rolled out to district partners after the close of our delivery window (Barnum, 2026).

Throughout the paper we refer to the combined program—Khan Academy with Khanmigo, delivered within the existing RTI block—as KWiK, the label used in our exhibits.

Control grades continued business-as-usual RTI mathematics. This counterfactual is heterogeneous and is not technology-free: control teachers drew on district-approved resources spanning small-group instruction, teacher-prepared worksheets and workbook practice, and computer-assisted learning platforms without conversational AI—most commonly Waggle, the adaptive practice program of the district's curriculum provider, along with platforms such as IXL, Zearn, and DeltaMath. Of the eighteen participating schools, four reported using no CAL in control sessions, four used CAL exclusively, and ten used combined approaches. The experiment therefore identifies the effect of the KWiK program against the realistic mix of remedial practice districts currently use, rather than against a never-software benchmark. Two implications follow. First, because control practice varies across teachers, the estimate averages over the realized distribution of business-as-usual remediation. Second, because many control students also received adaptive practice software, the treatment-control contrast should not be read as isolating the AI tutor: the program differed from control-arm CAL in several respects at once, including the structure of the practice itself—mastery-based learning paths placed directly from each student's MAP scores, implemented uniformly across treated sessions—alongside the availability of Khanmigo. We return to what can and cannot be attributed to the AI component in Sections 3 and 5.

2.2 Randomization, Exclusions, and Identification

Randomization was conducted by the research team before Year 1 and held fixed for both study years. Within each school, we first randomly assigned the number of treated grades (one or two of the school's three), then randomly selected the specific grade or grades from among the school's RTI-active grades. No further stratification was applied. The design yields 53 grade-within-school clusters, 28 treated and 25 control, across 18 schools.

Because the draw by chance placed treated cells disproportionately in higher grades, treated students are a slight majority of the Year-1 sample; the share moves to one-half in Year 2 as the heavily treated grade-8 cohort graduates and new cohorts enter more evenly. Cell assignments are fixed throughout, so arm shares vary only through these compositional

[1] Message-level records show no detectable change around this switch in the share of exercise sessions with any student message, the share of mistake sessions with a message, or messages per engaged session; Section 5 draws on this when interpreting how interface design relates to engagement.

flows.

Twenty schools were originally in scope; the three departures from the original roster were documented in the study's pre-registration.[2] One is excluded because school staff reassigned students across grades in violation of the randomized assignment. A second school is excluded because staffing challenges left the school without a functioning RTI program; the pre-registration anticipated excluding this school in Year 2 only, but the same staffing problems applied throughout, and we exclude it from both years. A third participated in Year 1 but was closing; in Year 2 the school received program access for all students and exits the experimental sample. Because that school ran RTI mathematics in two of its three grades, its departure removes two clusters, which is why per-term cluster counts fall from 53 to 51 in Year-2 terms. Section 4 shows the main estimate is robust to reconstructing East Lake's original assignment and including it (Table A10).

Identification comes from the within-school, across-grade contrast. Treated and control grades share the same school environment, leadership, and non-mathematics staffing. RTI mathematics sessions are grade-specific, and in the large majority of schools no teacher led RTI mathematics in both a treated and a control grade. We examine cross-grade spillovers descriptively in Section 4 and find no detectable signal, though the design has limited power on this margin.

2.3 Sample Construction and Data

The estimating sample follows a once-in, always-in rule: a student enters at their first scheduled RTI mathematics term and remains in the sample in every subsequent term with the required test scores, whether or not they remain in active RTI placement. The rule ensures that the sample is never conditioned on a post-treatment variable. The share of the sample no longer scheduled in RTI grows from zero in Winter 2025 to 19, 25, 31, and 40 percent across the four subsequent terms. Exit concentrates at the summer boundary, when schools re-roster from scratch against the spring assessment: among Year-1 students continuing into Fall 2025, 46 percent return above the placement threshold and enter the sample as spillover, roughly double the exit rate at within-year transitions.

The alternative—keeping only students still enrolled in RTI each term—would select the sample on the outcome. RTI exit is achievement-based: students leave upon reaching proficiency. A positive program effect therefore pushes treated students over the exit threshold at a higher rate, and the students it pushes out are precisely those who responded to it. The treated students remaining in RTI are increasingly those the program did not move,

[2] AEA RCT Registry, `https://doi.org/10.1257/rct.13519-2.0`.

while control students face the same threshold without the push, leaving their composition unfiltered. Whatever limits a student's response to the program—disengagement, weaker preparation, irregular attendance—plausibly also predicts slower subsequent growth, so the restricted comparison would stack the treated arm with its weakest growers, biasing estimates downward. Conditioning on baseline scores cannot repair this, because the filter selects on responsiveness within score levels, not only on the level itself; and the problem compounds mechanically, deepening each term in proportion to the program's own success. The balance diagnostics below display the observable trace of this selection: exit rates are similar across arms, yet imposing the restriction opens a significant baseline gap, with remaining treated students scoring below remaining controls.

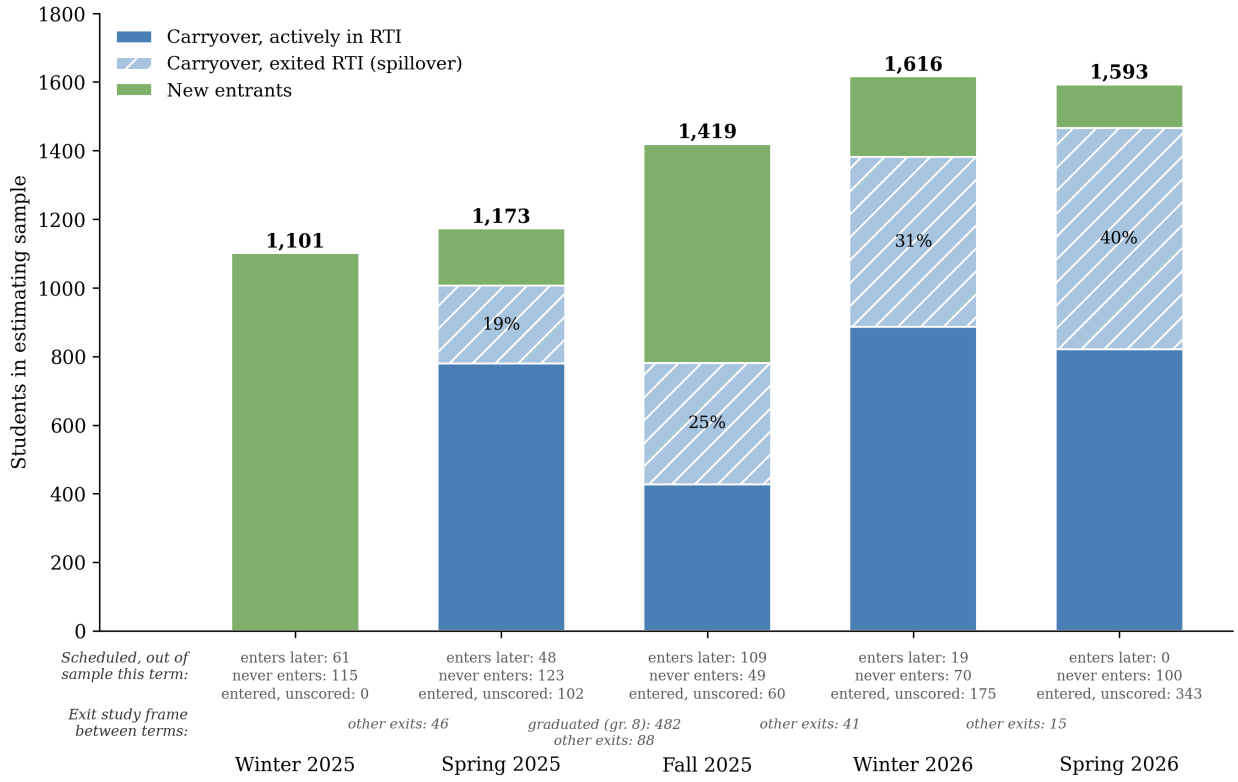
Retention instead carries a transparent cost. Students who have exited receive no contemporaneous treatment, so later-term intent-to-treat estimates are diluted toward zero in proportion to the exited share; the participation-scaled treatment-on-the-treated estimates reported alongside the main results recover the implied effect on participating students. A second, smaller cost is that later-term baselines are themselves post-treatment for previously exposed students; the resulting bias can run in either direction but is bounded by prior-term gains, which are small. Platform records show modest practice even before the first term's baseline, during initial Fall 2024 rostering, so no baseline is perfectly pre-exposure; contamination of that form inflates treated baselines and works against finding effects.

Figure 1 decomposes each term's sample into carryover students actively scheduled in RTI, students retained under the once-in rule after exiting RTI, and new entrants, and accounts for every student-term outside the sample as either transiently unscored or permanently departed.

Figure 1 decomposes each term's estimating sample into carryover students actively scheduled in RTI, students retained under the once-in rule after exiting RTI, and new entrants; beneath each bar it accounts for every scheduled student outside the sample that term, and between bars it records permanent exits from the study frame.

A student-term enters the estimating sample when both a beginning-of-term and an end-of-term MAP score are observed. The primary sample comprises 6,902 student-term observations across the five terms of 2024–25 and 2025–26 (per-term samples of 1,101; 1,173; 1,419; 1,616; and 1,593). Fall 2024 is excluded by design: the district was transitioning to MAP from a different screener that term, and rostering against the new assessment left too few active delivery weeks, a restriction specified in the pre-registration. Within each subsequent term, the schedule divides into a rostering period, a protected minimum-delivery window, and the end-of-term MAP testing window (Table 1); the minimum active delivery weeks vary across terms (4, 7, 4, 8, and 8). Platform records show practice extending

Figure 1: Sample Composition by Term Under the Once-In, Always-In Rule



Notes: Each bar is the term's estimating sample: students with both a beginning- and end-of-term MAP score. Dark blue students were in RTI mathematics in an earlier term and remain in it this term. Light blue students were in RTI in an earlier term but are no longer in it this term; the once-in, always-in rule keeps them in the sample, and their share is printed. Green students are in the sample for the first time. Rows beneath each bar count scheduled students outside the sample that term: "enters later" and "never enters" students have not recorded the consecutive score pair required for entry (the latter never do—236 of 2,708 ever-scheduled students, 8.7 percent, unrelated to assignment, $p = 0.51$); "entered, unscored" students are missing a score this term and return once scores reappear. Unscored counts peak each spring, mainly grade-8 students skipping the year-end test and mid-year leavers; the Spring 2026 count also holds exits that cannot yet be classified as permanent, since that requires observing later terms. Counts between bars are permanent exits at each boundary: 482 grade-8 students who completed Spring 2025 and then aged out of middle school, and 190 other exits over two years. Keeping RTI exiters and unscored students in the study, rather than dropping them, prevents attrition from depending on achievement.

beyond these protected windows into rostering and testing periods, so the minima are floors on exposure rather than its full extent, and Section 4 measures realized exposure directly.

Table 1: Implementation Schedule

Term	Event	Date Range	Min Wks	Max Wks	Note
Year 1 (2024–25)					
0	Fall Testing Period	Aug 7 – Sep 20		6.5	Transition from iReady to MAP; most schools waited to roster using MAP scores
0	**Minimum RTI (Fall 2024)**	Sep 23 – Oct 4	**2**	12.5	
1	Winter 2025 Roster Period	Oct 7 – 31		4	Fall break Oct 11–18
1	**Minimum RTI (Winter 2025)**	Nov 4 – Dec 4	**4**	13.5	Thanksgiving Nov 27–29
1	Winter 2025 Testing Period	Dec 5 – Jan 24		5.5	Winter break Dec 23 – Jan 3
2	Spring 2025 Roster Period	Jan 6 – Feb 14		6	
2	**Minimum RTI (Spring 2025)**	Feb 17 – Apr 11	**7**	19	Spring break Mar 17–21
2	Spring 2025 Testing Period	Apr 7 – May 16		6	
Year 2 (2025–26)					
3	**Minimum RTI (Fall 2025)**	Aug 11 – Sep 5	**4**	6	Rostered using spring 2025 scores
3	Fall 2025 Testing Period	Sep 8 – 19		2	
4	Winter 2026 Roster Period	Sep 22 – Oct 17		4	Fall break Oct 13–17
4	**Minimum RTI (Winter 2026)**	Oct 20 – Dec 12	**8**	15	Thanksgiving Nov 26–28
4	Winter 2026 Testing Period	Jan 5 – 23		3	Winter break Dec 22 – Jan 2
5	Spring 2026 Roster Period	Jan 26 – Feb 6		2	
5	**Minimum RTI (Spring 2026)**	Feb 9 – Apr 10	**8**	15	Spring break Mar 16–20
5	Spring 2026 Testing Period	Apr 13 – May 8		5	

Notes: Operational calendar of the study across the two school years. Each term comprises a roster period, in which schools place students into RTI using the prior MAP screener; a minimum active delivery window (bold), the protected interval in which KWiK or business-as-usual remediation is delivered; and a testing period, in which the MAP assessment serving as the term's endline and the next term's baseline is administered. Min Wks is the guaranteed floor on delivery weeks; Max Wks is the upper bound after any overlap with rostering and testing. School breaks are listed in the final column and are excluded from all per-week dosage measures. Term 0 (Fall 2024) spans the district's transition from the iReady screener to MAP and is excluded from the analysis.

Two features of the resulting panel deserve emphasis. First, students contribute multiple student-term observations, and assignment status can change across years: randomization fixed the treated grades, not treated students, so as students advance a grade between Year 1 and Year 2 they move over the fixed grade-level assignment. A student treated in Year 1 may advance into a treated grade again or into a control grade, and vice versa, generating variation in cumulative exposure—students assigned in both years, in one year only, or in neither—that we exploit in Section 4 when we examine dosage, persistence, and the replication of first-year effects among students newly assigned in Year 2. Second, our term-level estimates measure within-term growth: each specification conditions on the beginning-of-term score, so any gains carried forward from earlier exposure are absorbed into the baseline rather than double-counted, and effects of cumulative exposure are estimated separately.

Platform records provide engagement measures for all treated students across both years:

weekly learning minutes, skills worked and skills brought to proficient mastery, Khanmigo chat counts, and an engagement-consistency measure (the share of active delivery weeks with any use). For Year-2 treated students, we additionally observe message-level Khanmigo records matched to exercise-attempt logs, the basis for the usage analysis in Section 3.

2.4 Outcomes

The outcome is end-of-term mathematics achievement on the NWEA MAP assessment, which the district administers to all students each trimester as part of universal screening, so outcome data collection is identical across arms. MAP reports a RIT scale score, a vertically equated measure comparable across grades, and a national percentile rank. We report effects in standard-deviation units of the RIT score, standardized against the endline distribution of all students in the study schools—including those never placed in RTI—pooled by grade, and show percentile-rank and raw RIT-point estimates alongside. The study and its outcomes were pre-registered in the AEA RCT Registry. As a co-primary outcome we registered the state's annual Tennessee Comprehensive Assessment Program (TCAP) mathematics assessment; TCAP results for both study years will be incorporated as they are finalized.

2.5 Balance and Attrition

Table 2 reports covariate balance by term and pooled, on the estimating sample of student-terms with both scores. Demographic characteristics and baseline achievement are balanced across arms: the pooled treatment-control difference in baseline mathematics is -0.04 population standard deviations (SE 0.04), and differences in gender, race and ethnicity, meal-program eligibility, and special-education designation are small, with marginally significant contrasts concentrated in Year 1. Student-weighted joint tests reflect these ($p = 0.03$ in Winter 2025, $p = 0.05$ pooled), but at the cell level, where randomization operated, joint tests are null in every term ($p \geq 0.42$): the imbalances reflect the composition of a few large cells rather than the assignment mechanism, they run against finding effects—treated students are weaker on ELA—and the estimates are insensitive to demographic controls throughout. Balance also holds when each student is measured once (Table A1).

Treated grades enrolled more RTI students than control grades in the Year-1 terms, a gap of roughly 25 to 30 percent in Winter and Spring 2025. This is largely mechanical: 28 of 53 randomized cells are treated, and the draw by chance placed treated cells disproportionately in higher grades, which enrolled more RTI students in Year 1; the gap closes in Year 2 as the heavily treated grade-8 cohort graduates. Importantly, observed systematic composition is

balanced: if treated-grade staff had rostered additional, or different, students in response to the program, the characteristics and baseline scores of the two arms should diverge. They do not. The alignment of results with this pattern is also reassuring: the program's achievement effects, reported in Section 4, are concentrated in Year 2, when the size gap is absent—the opposite of what one would expect if the imbalance were generating spurious effects.

Attrition from the estimating sample occurs when a student-term lacks a beginning- or end-of-term score, since testing is administered district-wide to both arms on the same schedule. The count rows of Table 2 show that moving from the full RTI roster to the sample with an endline score, and then to the sample with both scores, reduces treated and control counts nearly proportionally in every term: the treated-control gap does not widen as score requirements tighten, and the composition of the two arms remains balanced across the three sample definitions. Together with the once-in, always-in rule, which prevents achievement-based RTI exit from removing students differentially, we find no indication that missing scores differ by assignment. This holds in the most demanding case: missingness in Spring 2026, the term with the most unscored students, is unrelated to assignment ($p = 0.75$) and to assignment interacted with prior achievement ($p = 0.65$), so differential loss of lower-performing treated students does not explain the term's larger estimate.

Table 2 shows chance imbalances concentrated in Year-1 demographics and the ELA baseline (student-weighted joint tests $p = 0.03$ in Winter 2025 and $p = 0.05$ pooled), while at the cell level, where randomization operated, joint tests are null in every term ($p \geq 0.42$): the student-weighted imbalances reflect the composition of a few large cells rather than the assignment mechanism. The mathematics baseline—the covariate most predictive of the outcome—is balanced in every term, the imbalances that exist run against finding effects, with treated students weaker on ELA, and the term where they concentrate shows the smallest estimate while Spring 2026, the largest, is among the most balanced columns. Estimates are insensitive to adding demographic controls throughout.

Two further diagnostics bear on the once-in, always-in rule itself. First, assignment does not predict exit from scheduled RTI placement (coefficient 0.022, SE 0.047), though the test is imprecise and cannot rule out modest differential exit. Second, treated and control students are balanced on baseline mathematics in the estimating sample (-0.044, $p = 0.22$), but restricting to student-terms in which the student remains actively scheduled in RTI produces a significant imbalance (-0.088, $p = 0.005$), with treated students negatively selected. Exit is achievement-related, so conditioning on continued enrollment removes disproportionately more high-achieving students from the treated arm. Retaining them costs little—the pooled estimate is essentially unchanged, at 0.039, under the restriction—while dropping them forfeits the randomized comparison.

Table 2: Balance Test

	Winter 2025		Spring 2025		Fall 2025		Winter 2026		Spring 2026		Pooled	
	Control	Diff	Control	Diff	Control	Diff	Control	Diff	Control	Diff	Control	Diff
Female	0.517	-0.039	0.531	-0.031	0.529	-0.024	0.529	-0.027	0.535	-0.034	0.529	-0.025
		(0.030)		(0.034)		(0.029)		(0.025)		(0.028)		(0.017)
White	0.228	0.055***	0.233	0.023	0.308	-0.005	0.330	-0.009	0.342	-0.009	0.297	0.008
		(0.018)		(0.018)		(0.023)		(0.020)		(0.019)		(0.014)
Black	0.543	-0.044**	0.527	-0.010	0.465	-0.005	0.440	-0.013	0.430	-0.006	0.472	-0.013
		(0.019)		(0.018)		(0.033)		(0.029)		(0.026)		(0.020)
Hispanic	0.210	-0.013	0.216	-0.016	0.202	0.002	0.202	0.017	0.206	0.007	0.207	0.000
		(0.016)		(0.014)		(0.025)		(0.021)		(0.022)		(0.011)
Other Race	0.020	0.001	0.024	0.003	0.025	0.008	0.027	0.004	0.023	0.007	0.024	0.004
		(0.009)		(0.010)		(0.006)		(0.005)		(0.005)		(0.004)
Free and Reduced-Price Meals	0.705	-0.037*	0.705	-0.017	0.670	0.019	0.655	0.037	0.632	0.042**	0.668	0.017
		(0.022)		(0.020)		(0.028)		(0.022)		(0.018)		(0.014)
Economic Disadvantage	0.467	-0.036*	0.476	-0.030	0.412	0.031	0.412	0.033	0.388	0.037	0.425	0.017
		(0.019)		(0.022)		(0.035)		(0.028)		(0.023)		(0.020)
Dyslexia	0.137	0.019*	0.142	0.019	0.085	0.012	0.089	0.016	0.094	0.008	0.105	0.015
		(0.011)		(0.011)		(0.017)		(0.015)		(0.015)		(0.013)
Special Education Disability	0.174	-0.021*	0.167	0.022*	0.303	-0.010	0.307	0.005	0.322	-0.011	0.267	-0.004
		(0.013)		(0.013)		(0.042)		(0.033)		(0.028)		(0.023)
ELA Baseline (pop. SD)	-0.581	-0.082**	-0.510	-0.107*	-0.540	-0.108	-0.328	-0.039	-0.266	-0.018	-0.426	-0.056
		(0.038)		(0.056)		(0.091)		(0.055)		(0.072)		(0.049)
Math Baseline (pop. SD)	-0.917	-0.017	-0.667	-0.050	-0.705	-0.077	-0.661	-0.038	-0.473	-0.011	-0.665	-0.044
		(0.023)		(0.040)		(0.049)		(0.042)		(0.065)		(0.035)
Joint test, student-weighted (p)	0.03		0.17		0.07		0.16		0.14		0.05	
Joint test, cell-level (p)	0.57		0.59		0.57		0.56		0.42		0.72	
N: full RTI sample (Control / Treated)	556	721	646	800	851	786	956	924	1048	988	4057	4219
N: endline only (Control / Treated)	511	623	582	688	806	761	823	830	850	837	3572	3739
N: baseline & endline (Control / Treated)	505	596	550	623	733	686	805	811	796	797	3389	3513

Notes: Covariate balance by term and pooled, on the analysis sample of student-terms with both a baseline and an endline MAP score under the once-in, always-in rule. Each difference is the coefficient on assignment from a regression of the covariate on treatment with school and grade fixed effects (the pooled column adds term fixed effects), so comparisons are within school, matching the stratified randomization; standard errors, in parentheses, are clustered at the grade-within-school level, the unit of randomization, in every column. The student-weighted joint test is a cluster-robust F-test of all covariates in a regression of assignment on the full set with the same fixed effects; the cell-level test replaces students with equal-weighted school-by-grade cell means, testing the randomization at its unit of assignment. ELA and mathematics baselines are prior-term MAP screeners standardized to the full grade-level population. The final rows report counts under successively stricter score requirements; treated counts exceed control counts in Year 1 because 28 of 53 randomized grade-within-school cells were assigned to treatment, with treated cells by chance concentrated in higher grades, and the gap closes in Year 2 as the heavily treated grade-8 cohort graduates. Attrition across the three definitions is nearly proportional by arm; formal tests of missingness, entry, and exit against assignment appear in the text. $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

2.6 Empirical Strategy

We estimate intent-to-treat effects by ordinary least squares on the stacked student-term panel:

$$Y_{igsty} = \alpha + \beta\, D_{gsy} + \gamma\, Y^{0}_{igsty} + \delta' X_{igsty} + \theta_s + \phi_g + \lambda_{ty} + \varepsilon_{igsty} \tag{1}$$

where Y_{igsty} is end-of-term achievement for student i in grade g, school s, term t, year y; D_{gsy} is grade-level assignment; Y^{0}_{igsty} is the beginning-of-term score; X_{igsty} collects demographic controls in the fully adjusted specification; and θ_s, ϕ_g, and λ_{ty} are school, grade, and term-by-year fixed effects. Conditioning on the lagged score is a precision choice (McKenzie, 2012); the identifying variation is the within-school, across-grade contrast created by randomization. We do not include school-by-grade fixed effects, which would absorb the treatment entirely, but show robustness to alternative structures, including school-by-term effects (Table A6).

Standard errors are clustered at the randomization unit, the grade-within-school cell ($J = 53$). Because the cluster count is modest, we accompany cluster-robust inference for the primary estimate with wild cluster bootstrap p-values (Cameron et al., 2008) and leave-one-school-out estimates. We also report exposure-response relationships that instrument realized platform usage with assignment, and, in the appendix, pre-specified heterogeneity analyses by student characteristics, grade, and baseline achievement, all of which yield interactions that are small and statistically indistinguishable from zero, though with confidence intervals that rule out only large differential effects.

3 Take-up: Practice and Khanmigo Use

3.1 Platform Practice

Assignment produced a large and durable first stage. Treated students averaged 30.2 minutes of weekly Khan Academy practice against a control mean near zero, an intent-to-treat effect of 30.8 minutes per week (Table 3). The practice was productive: treated students practiced 4.3 skills per week and brought 2.9 to proficient mastery, reflecting the platform's placement of exercises at each student's current level. These figures average over all treated student-terms, including students retained after exiting RTI; among students actively scheduled in RTI, weekly practice averages 38 minutes with 3.7 skills to proficiency, rising to 45 minutes and 4.7 skills in Year 2 as implementation matured.[3]

Khanmigo use, by contrast, averaged 1.7 chat sessions per week—2.2 among actively

[3]Table 3 also reports per-hour rates. These average student-level ratios with zeros retained for student-terms logging no learning time, and are therefore lower than the ratio of the weekly means.

scheduled students and 2.7 in Year 2, in roughly fixed proportion to practice time at every scope.[4] The added practice substituted for other remedial activity rather than extending it: RTI attendance is essentially identical across arms (94 percent in both), so assignment changed what happened within the block, not how often students were there.

Table 3: Practice and Khanmigo Intensity

	Control mean	Treated mean	ITT (SE)
Learning time (min/wk)	0.143	30.172	30.764*** (2.662)
Skills worked /wk	0.018	4.256	4.311*** (0.349)
Skills to proficient /wk	0.011	2.906	2.917*** (0.290)
Skills worked /hr	0.035	5.968	5.909*** (0.423)
Skills to proficient /hr	0.020	3.636	3.571*** (0.269)
Khanmigo chats /wk	0.012	1.717	1.752*** (0.156)
RTI attendance rate	0.943	0.942	-0.002 (0.002)

Notes: Intent-to-treat effects of KWiK grade assignment on platform use, pooled across all five terms, on the sample of student-terms with both a baseline and an endline MAP score. ITT coefficients come from regressions of each measure on assignment with school, grade, and term fixed effects; standard errors, in parentheses, are clustered at the grade-within-school level, the unit of randomization. Per-hour measures are set to zero when no learning time is logged, and weekly measures exclude holiday weeks. Weekly measures are computed over all delivery weeks in the term, including weeks in which a student logged no activity and including students retained in the sample after exiting active RTI. The platform was available only in treated grades, so control means are near zero and treated means approximately equal the ITT effects. RTI attendance is observed in both arms. $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Realized dosage should be read against the scheduled ceiling. A daily block of twenty-five to forty minutes implies more than two scheduled hours per week; delivered practice of half an hour reflects the block cancellations, testing windows, and competing session uses described in Section 2, along with ordinary student absence. Thirty minutes per week nonetheless approximates the platform developer's recommended dose, and stands in contrast to voluntary deployments, where median use often is close to zero (Eames et al., 2026; Robinson et al., 2026). Practice was also not confined to the protected windows: calendar-week records show treated students active at similar weekly rates during rostering and testing periods, so the windows understate total exposure, and the exposure-response estimates of Section 4 measure practice over each term's full pre-testing span.

Practice intensity strengthened between years, with weekly minutes rising as delivery periods lengthened (Appendix Table A2). Within Year 2, however, tutor use and practice

[4] A chat is one conversation between a student and Khanmigo, however initiated and however brief, as recorded in the platform's aggregate engagement export; conversations average roughly four student messages. Weekly means include zeros: in a typical week about half of practicing students open no chat, so the average mixes many zero-chat weeks with a right-skewed tail of heavier use. Section 3 turns to the message-level records available for Year-2 treated students, matched to exercise sessions, in which suggested-prompt clicks as well as typed text count as student messages.

moved apart. Part of the movement is compositional—the share of treated students no longer scheduled in RTI grows across the year, pulling down all averages—but the divergence holds among students actively in the program: their weekly practice rose from 42 to 48 minutes between Fall 2025 and Spring 2026 while their weekly chats stayed flat at about 2.6, so conversations per minute of practice declined even as practice deepened. Engagement is also right-skewed across students, with means well above medians for every measure.

3.2 Khanmigo Use

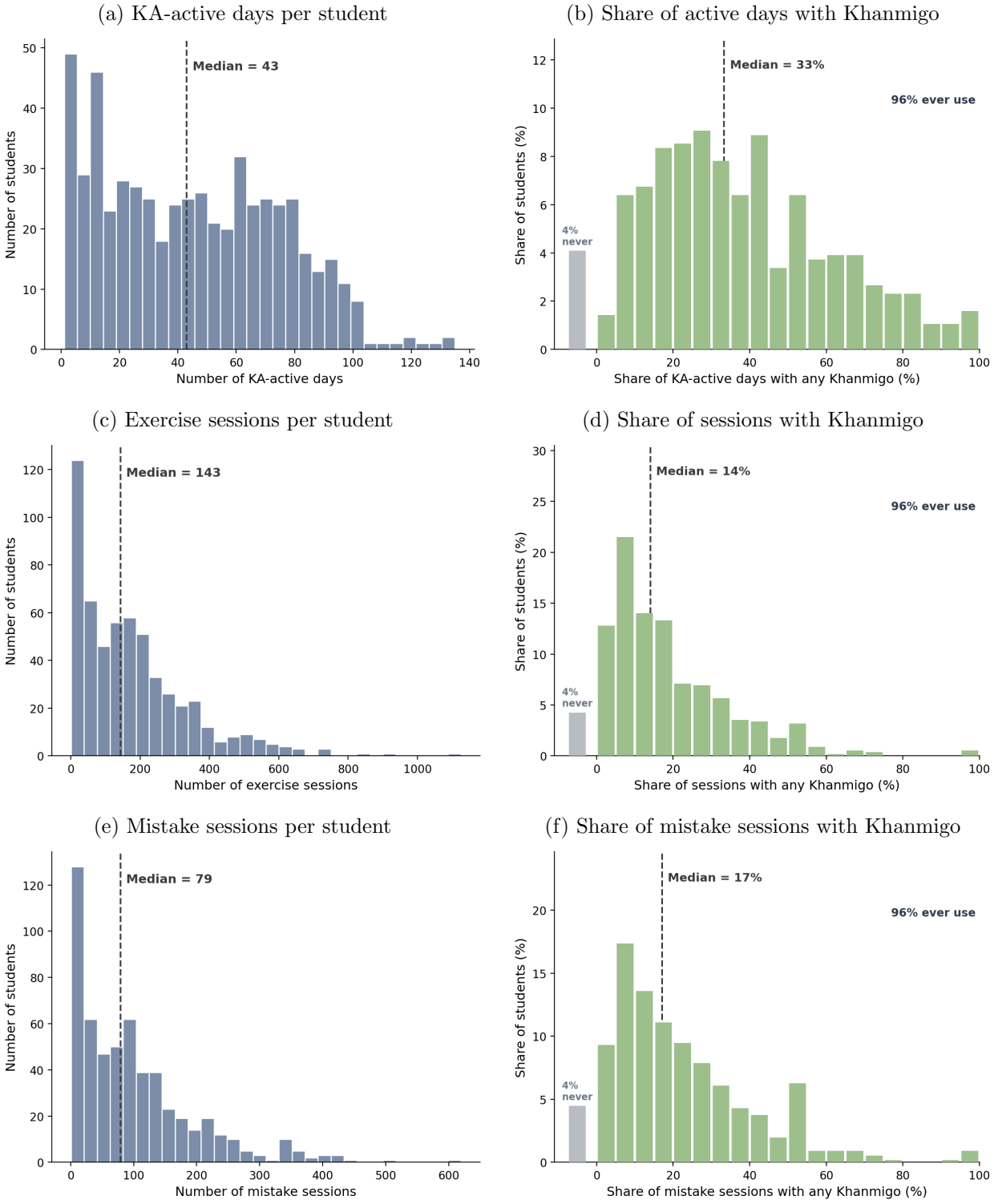
How did students use the AI tutor itself? For Year-2 treated students we match records of student messages to Khanmigo against exercise-attempt logs, allowing us to measure engagement with the tutor directly rather than infer it from aggregate chat counts. The analysis sample comprises the 563 activated Year-2 treated students matched to exercise records, spanning 16 schools, 25,286 platform-active student-days, and 100,017 exercise sessions (Appendix Table A3).[5] The window is the year of strongest implementation and largest achievement effects, so if anything these figures overstate steady-state engagement. The analysis is descriptive, conditional on activation and a successful data match.

Engagement with the tutor is nearly universal at the extensive margin and thin at every intensive margin. Ninety-six percent of students messaged Khanmigo at least once. Yet the median student sent a message on 33 percent of the days they practiced, in 14 percent of exercise sessions, and in 17 percent of mistake sessions—the moments a tutor is nominally for (Figure 2). The typical exercise session, of any kind, contains no message to the tutor: messages per session average 0.9, with a median of zero even in mistake sessions (Appendix Table A4). Engagement is rare rather than shallow when it occurs: messages appear in 9,362 of the 56,362 mistake sessions, and in those sessions conversations run to a median of three student messages and a mean of six and a half.

The form of students' messages qualifies the picture further. Classifying every message sent during math exercise sessions into five categories (Figure 3), the largest share, 39.4 percent, consists of answers to Khanmigo's own questions: numbers or short phrases entered without explanation, which we call bare answers. Another 24.2 percent are clicks on suggested help prompts, and 12.6 percent are low-effort messages such as "idk" or requests for the answer. Messages containing a mathematical question or a step of reasoning account for 14.5 percent, and 9.3 percent are off-task: requests for essays, raps, or jokes, chit-chat, and

[5] An exercise session is the interval from the start of one exercise to the start of the next; a mistake session is an exercise session containing a submitted non-perfect exercise or a start-over, the latter of which in almost all cases follows mistakes within the exercise. Khanmigo use is a student message sent within the session window; suggested-prompt clicks count as messages, and Khanmigo's own responses do not.

Figure 2: Khanmigo Engagement Across Three Margins



Notes: Khanmigo engagement among activated Year-2 treated students matched to exercise records (563 students, 16 schools, August 2025 through May 2026). A KA-active day is a student–calendar day with at least one exercise attempt (25,286 student-days). An exercise session is the interval from the start of one exercise to the start of the next, retained when it has positive learning time and at least one activity marker (100,017 sessions). A mistake session is an exercise session containing a submitted non-perfect exercise or a start-over (56,362 sessions). Khanmigo use is at least one student message, including suggested-prompt clicks, sent within the day or session window. Left panels plot volumes per student; right panels plot each student's share of days or sessions with any Khanmigo message. Grey bars denote the 4 percent of students who never message Khanmigo; dashed vertical lines mark medians across students.

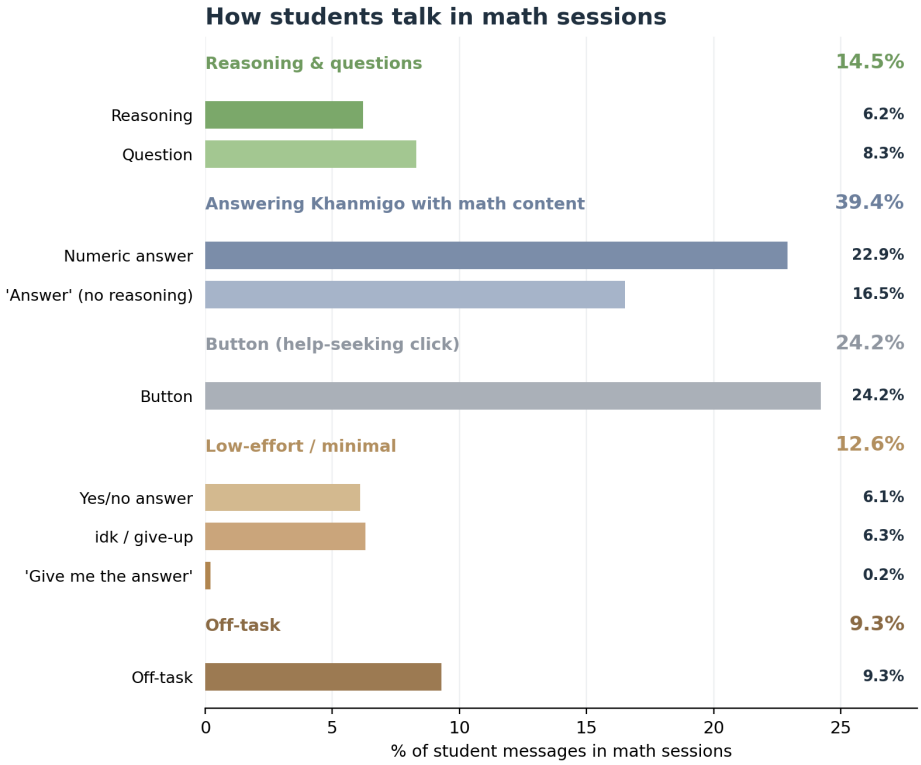
Table 4: Paraphrased Khanmigo Interactions, by Message Category

Category	Example episode
Reasoning & Questions	**S:** is it 60 for mine? **K:** Yes, 60 degrees is correct. How did you figure it out? **S:** cause that one is less than a 90 and all the other ones are bigger then 90 so those couldnt be it so it had to be 60 **S:** what is perpendicular? **K:** Perpendicular means two lines meet to make a square corner, or a 90-degree angle. Can you spot any lines like that in your picture?
Answering Khanmigo with math content	**K:** I see you wrote 48. Can you tell me how you got that? **S:** 6 x 8 **K:** Great start. You found the area of a rectangle. For triangles, do you remember what you do next? **S:** 2 rectangles **K:** What are the side lengths of each rectangle? **S:** 10, 6 **K:** Take another look at the first rectangle. **S:** 2, 10
Button (help-seeking click)	**S:** [Help me solve this. Ask me one question at a time.] **K:** What do you notice about the points on the number line? **S:** [Help me check my answer.] **K:** Great job asking for help. Can you tell me how you picked your points for the rectangle?
Low-effort / minimal	**K:** What do you notice about the hour hands on the clocks? **S:** idk **K:** No worries. Look at the short hand. What number is it pointing to? **S:** whats the answer **K:** What do you notice about the rectangle's rows and columns? **S:** just give me it I already got it wrong so theres no point doing it again
Off-task	**S:** can you write an essay about the history of pizza **K:** For help with writing, you can use Writing Coach. It will guide you through the process step by step. **S:** say a joke **K:** Why did the math book look sad? Because it had too many problems. Want to try a quiz question next?

Notes: **S** denotes a student turn and **K** a Khanmigo turn. Student messages are paraphrased, preserving length, register, and mathematical content; text in square brackets reproduces the platform's suggested prompt chips, which are selected rather than typed by the student. Repeated turns and non-substantive formatting are omitted for space. Answering Khanmigo with math content comprises numeric or short-phrase answers with little or no explanation; low-effort messages include brief acknowledgements, give-up messages, and requests for the answer; off-task messages are unrelated to the math task.

testing behavior. Table 4 demonstrates representative exchanges from each category (slightly altered to maintain anonymity). Two readings of the bare-answer share are possible, and the data do not separate them: entering answers in response to the tutor's questions is minimal engagement, but it is also what compliance with a Socratic tutor looks like, since Khanmigo's design solicits exactly such replies. What the classification establishes is narrower and still informative: student-initiated mathematical articulation—asking a question, explaining a step—is uncommon, roughly one message in seven.

Figure 3: Message-Form Decomposition



Notes: Classification of student messages sent to Khanmigo during math exercise sessions, for the sample of Figure 2. Button clicks and numeric forms are assigned by rule; remaining messages are assigned by a large-language-model classifier. "Reasoning & questions" comprises messages containing a mathematical question or a step of reasoning. "Answering Khanmigo with math content" comprises numeric answers and short answer phrases entered without explanation. "Low-effort / minimal" comprises yes/no replies, give-up messages such as "idk," and requests for the answer. "Off-task" comprises messages unrelated to the math task, including small talk, unrelated requests, testing behavior, and gibberish.

Taken together, the usage data show a tutor that nearly all students touched and a few made a habit of. The median student's practice day, exercise session, and mistake session all passed without a message to the tutor, and most messages that were sent required no mathematical articulation. Section 5 takes up what this implies for interpreting the achievement effects and for the design of AI tutoring systems.

4 Effects on Achievement

4.1 Average Effects

Table 5 reports the pooled intent-to-treat effect of assignment on end-of-term MAP mathematics achievement across all five terms, in the four outcome scalings. Because each specification conditions on the beginning-of-term score, these estimates measure within-term learning gains. In this pre-registered specification, assignment raises achievement by 1.26 national percentile ranks (SE 0.60) per term, equivalent to 0.76 RIT points, 0.050 control-group standard deviations, or 0.040 standard deviations of the full study-school population. Adding demographic controls leaves the estimate essentially unchanged (0.039 population SD), and the specification without the baseline score yields a similar but far noisier estimate. The sign and significance of the effect are stable across every choice of scale. Because assignment attaches to grades while students exit RTI over time, these intent-to-treat estimates average over students no longer receiving treatment; Tables 6 and 7 report the corresponding treatment-on-treated effects on participating students, which exceed the intent-to-treat in proportion to the share of assigned students no longer enrolled in RTI.

4.2 Effects Track Delivery

Table 6 reports term-by-term estimates alongside realized practice. The panels carry over the control- and full-population SD scalings of Table 5 and add, for each, a version standardized by grade-and-term rather than pooled-by-grade SDs (Panels B and D), which puts each term's effect in that term's own units—the appropriate scaling for cross-term comparison; the choice matters little in practice.

The five coefficients across terms are statistically indistinguishable in every scaling ($p = 0.64$ in RIT points, 0.55 in control SD, and 0.58 in population SD). The largest estimate, Spring 2026's 0.074 population SD (0.093 control SD), significant at the five percent level, does occur in a high-practice term, and across years effects and practice rise together; but the design is powered for pooled and annual contrasts rather than term-level ones, so we read the pattern as consistent with a common return to practice applied to different doses, not as term-specific results.

The term-level departures from dose-response run in both directions: Spring 2025's estimate exceeds Winter 2026's on far less practice, and Winter 2026 pairs the study's second-highest realized exposure with a small estimate (0.021, SE 0.032). Departures in both directions are what we might expect from noise around a common effect, and the implied per-hour returns are likewise statistically indistinguishable across terms ($p = 0.29$); the dosage anal-

Table 5: Effect of KWiK Assignment on MAP Mathematics Achievement

	No controls	Baseline	Full controls
Panel A: MAP percentile rank (NPR)			
KWiK Assignment	-0.244	1.264**	1.216**
	(1.325)	(0.597)	(0.568)
Panel B: MAP RIT scale points			
KWiK Assignment	-0.094	0.758*	0.742**
	(0.925)	(0.379)	(0.363)
Panel C: RIT, control SD (pooled by grade)			
KWiK Assignment	0.007	0.050**	0.049**
	(0.057)	(0.023)	(0.022)
Panel D: RIT, full-population SD (pooled by grade)			
KWiK Assignment	0.003	0.040**	0.039**
	(0.046)	(0.019)	(0.018)
Observations	6,902	6,902	6,902
Clusters	53	53	53

Notes: Intent-to-treat effects of KWiK grade assignment on end-of-term MAP mathematics achievement, pooled across all five terms. The unit of observation is the student-term; the sample comprises all student-terms with both a baseline and an endline MAP score under the once-in, always-in rule (6,902 observations, 53 grade-within-school clusters). Each cell reports the coefficient on assignment from a regression with school, grade, and term fixed effects, where term effects are unique across years and absorb term-by-year shocks; standard errors, in parentheses, are clustered at the grade-within-school level, the unit of randomization. Columns add controls in sequence: fixed effects only; the prior-term baseline MAP score; and demographics. $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table 6: Effects by Term

	Winter 2025	Spring 2025	Fall 2025	Winter 2026	Spring 2026
Panel A: control SD, pooled by grade					
KWiK Assignment	0.005	0.049	0.017	0.026	0.093**
	(0.037)	(0.039)	(0.033)	(0.039)	(0.038)
Panel B: control SD, by grade × term					
KWiK Assignment	0.018	0.040	0.019	0.029	0.091**
	(0.047)	(0.038)	(0.040)	(0.042)	(0.035)
Panel C: full-population SD, pooled by grade					
KWiK Assignment	0.005	0.038	0.013	0.021	0.074**
	(0.031)	(0.032)	(0.027)	(0.032)	(0.031)
Panel D: full-population SD, by grade × term					
KWiK Assignment	0.005	0.038	0.016	0.024	0.067**
	(0.032)	(0.030)	(0.031)	(0.034)	(0.028)
Exposure hours per assigned student	3.9	2.7	1.6	4.8	4.8
Exposure hours per participating student	3.9	3.5	2.2	7.1	8.3
Participation rate	1.00	0.78	0.74	0.68	0.58
Implied effect on participants	0.005	0.049	0.018	0.031	0.127
Equality of term effects (p)	0.55–0.64 across panels				
Observations	1,101	1,173	1,419	1,616	1,593
Clusters	53	53	51	51	51

Notes: Intent-to-treat effects of KWiK grade assignment on end-of-term MAP mathematics achievement, estimated in a separate regression per term. Each cell reports the coefficient on assignment from a specification conditioning on the beginning-of-term MAP score with school and grade fixed effects; standard errors, in parentheses, are clustered at the grade-within-school level, the unit of randomization. The sample in each column is that term's student-terms with both a baseline and an endline score under the once-in, always-in rule. The outcome is standardized RIT throughout; panels differ in the standardizing standard deviation: control-group SD pooled within grade (A); control-group SD by grade and term (B); full study-population SD pooled within grade (C); and full-population SD by grade and term (D). Exposure hours are mean platform learning hours per treated student over the term's calendar weeks through the last week before endline testing, as defined in Table 8; hours per participating student divide by the participation rate. A joint test that the five term coefficients are equal does not reject in any panel ($p = 0.55$ to 0.64); term differences, including the small Winter 2026 estimate, are within sampling variation of a common effect. Participation rate is the treated-arm share of student-terms contemporaneously scheduled in RTI, the first stage of a specification instrumenting participation with assignment; the implied effect on participants is the corresponding Wald estimate in the Panel C scaling. It assumes assignment affects achievement only through contemporaneous participation, and is an upper bound on the participant effect if students retain benefits after exiting RTI. Table 8 estimates the corresponding effect per hour with full inference. $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

ysis below estimates that return and benchmarks it against independent estimates from the same platform.

4.3 Effects by Year

Aggregating to the school-year level, Table 7 reports effects on annual achievement growth. The Year-1 effect is 0.020 population SD and indistinguishable from zero; the Year-2 effect is 0.084 (SE 0.041), or 0.116 in control-group SD units, significant at five percent; stacking both years gives 0.062 per school year ($p < 0.10$). The two annual effects are not statistically distinguishable from each other ($p = 0.61$ from a year interaction in the stacked specification), so the concentration of gains in Year 2 describes the point estimates rather than a tested contrast; the improvement nonetheless tracks the rise in realized practice documented in Section 3.

Because annual estimates condition on the fall baseline, gains carried over from Year 1 are netted out by construction; the remaining concern is that prior exposure makes continuing students grow faster. The decomposition at the end of this section indicates the opposite: within Year 2, students assigned for the first time grew more than students in their second year of assignment (0.219 versus 0.093 population SD), though this rests on a smaller, selected panel and we treat it as exploratory. Whether Year-1 gains persisted over the summer cannot be determined: Year-1 assignment leaves no detectable trace in fall Year-2 scores (0.005, SE 0.030), but with a Year-1 effect of only 0.020, the design lacks the power to detect its remnant.

4.4 Exposure-Response

Scaling assignment effects by realized practice time converts the ITT into a per-hour return. Instrumenting annual platform exposure hours with assignment, Table 8 implies 0.0072 population SD per hour of practice in the stacked specification (SE 0.0040), with the Year-2 estimate at 0.0077–0.0079 and significant, and Year 1 imprecise; the stacked estimate sits near the Year-2 value because stacking weights each year by the practice variation it contributes, and most hours occurred in Year 2.

Two benchmarks discipline the magnitude. Eames et al. (2026), using idiosyncratic classroom variation on the same platform and the same MAP outcome, estimate 0.0047 SD per annual hour on average, rising to 0.0086 among classrooms whose practice converts efficiently into mastered skills. Our supervised RTI deployment thus sits at the top of the observational range—despite serving below-grade-level students, for whom the same study finds lower per-hour returns. Structured, monitored delivery appears to buy the efficient end of the distribution of returns to practice.

Table 7: Effects by School Year

	Year 1		Year 2		Stacked	
	Base	Full	Base	Full	Base	Full
Panel A: RTI control SD						
KWiK Assignment	0.016	0.015	0.116**	0.114**	0.066	0.066
	(0.043)	(0.042)	(0.053)	(0.052)	(0.043)	(0.042)
Panel B: Population SD						
KWiK Assignment	0.020	0.018	0.084**	0.082**	0.062*	0.062*
	(0.035)	(0.034)	(0.041)	(0.040)	(0.035)	(0.034)
Participation rate	0.84	0.84	0.61	0.61	0.72	0.72
Implied effect on participants	0.024	0.022	0.142	0.138	0.089	0.089
Observations	1,214	1,214	1,545	1,545	2,759	2,759
Clusters	53	53	51	51	53	53

Notes: Intent-to-treat effects of KWiK grade assignment estimated within each school year. Because students change grades between years while assignment attaches to grades, each year provides a randomized comparison in its own right rather than the same comparison repeated. The two are not statistically independent, however: they share a single randomization draw, roughly two-fifths of Year-2 students also appear in Year 1, and grade progression through the fixed assignment means Year-1 status predicts Year-2 status, so a quarter of the Year-2 control arm was treated in Year 1. Any persistence of first-year effects therefore works against the Year-2 estimate. The unit of observation is the student-year. The baseline is the MAP screener taken immediately before the student's first RTI term of the year and the endline is the year-end spring screener, so each contrast spans only the student's own treatment window; a student first entering in Year 2 contributes only a Year 2 observation. The outcome is end-of-year RIT standardized within year and grade, by the control-group standard deviation (Panel A) and the full study-population standard deviation (Panel B). Base specifications condition on the baseline score; Full specifications add demographics. All columns include school and grade fixed effects, stacked columns add a year fixed effect, and standard errors, in parentheses, are clustered at the grade-within-school level, the unit of randomization. Participation rate is the treated-arm mean share of a student's observed terms spent contemporaneously scheduled in RTI; the implied effect on participants is the Wald estimate scaling the intent-to-treat by that first stage, in Panel B units, and carries the exclusion restriction described in the note to Table 6. $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

The per-hour estimates, the intent-to-treat, and the implied effects on participants reported in Tables 6 and 7 are three expressions of one experimental contrast rather than independent results. Because assignment attaches to grades while students exit RTI over time, the intent-to-treat averages over a growing share of students receiving no contemporaneous treatment—39 percent of the treated arm by Spring 2026. Scaling by the participation rate yields the effect on participating students: 0.127 population SD in Spring 2026 and 0.142 for Year 2, against intent-to-treat estimates of 0.074 and 0.084. Scaling instead by realized hours yields the per-hour returns above. All three share a single reduced form, so the rescalings change interpretation rather than evidence: t-statistics are essentially unchanged across the three, and the larger participant-level magnitudes are correspondingly less precisely estimated.

Table 8: Effects per Hour of Practice

	Year 1		Year 2		Stacked	
	Base	Full	Base	Full	Base	Full
Panel A: RTI control SD						
Per exposure hour (SD)	0.0028	0.0026	0.0109**	0.0107**	0.0077	0.0077
	(0.0072)	(0.0070)	(0.0052)	(0.0051)	(0.0049)	(0.0047)
Panel B: Population SD						
Per exposure hour (SD)	0.0034	0.0031	0.0079**	0.0077**	0.0072*	0.0072*
	(0.0059)	(0.0057)	(0.0039)	(0.0038)	(0.0040)	(0.0038)
First-stage F	60.7	60.7	129.6	129.6	143.6	143.6
Observations	1,214	1,214	1,545	1,545	2,759	2,759
Clusters	53	53	51	51	53	53

Notes: Two-stage least squares estimates of the effect of an additional hour of platform practice, by school year. The unit of observation is the student-year, with the baseline at the student's first RTI term of the year and the endline at the year-end spring screener. Exposure hours sum each student's logged learning minutes over calendar weeks from each term's start through the last week before endline testing, matched from weekly platform exports and divided by sixty; practice during testing windows is excluded because it cannot affect tests already taken, and including it lowers the coefficients by roughly a quarter. Measured this way, treated students average 5.7 hours in Year 1 and 10.8 in Year 2, exceeding the protected delivery windows underlying the weekly measures of Section 3. Hours are instrumented by grade assignment, so each coefficient is a local average treatment effect. The outcome is end-of-year RIT standardized within year and grade by the control-group standard deviation (Panel A) and the full study-population standard deviation (Panel B). Base specifications condition on the baseline score; Full specifications add demographics. All columns include school and grade fixed effects, stacked columns add a year fixed effect, and standard errors, in parentheses, are clustered at the grade-within-school level, the unit of randomization. First-stage F is the cluster-robust first-stage statistic. $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

4.5 State Test Scores

Tennessee's annual accountability assessment, the Tennessee Comprehensive Assessment Program (TCAP), was registered as a co-primary outcome, and results for both study years will be incorporated as they are finalized. The TCAP analysis serves two purposes beyond replication on a second assessment. First, it measures effects on a test administered once per year, after terms of uneven delivery, rather than at each term's end. Second, it provides the assessment-alignment check noted in Section 2: unlike MAP, TCAP is not linked to the platform's placement system, so it tests whether MAP effects reflect narrowly assessment-aligned practice. Given the within-year MAP pattern—near-zero effects in Year 1, sizable effects in Year 2—the informative TCAP comparison is Year 2.

4.6 Robustness

The main estimate survives several robustness checks; we summarize here and report full tables in the appendix. Restricting to the Term 1 first-entry cohort only removes any influence of treatment-responsive entry: Term-1 entrants, followed across all five terms yield 0.058 population SD per term (SE 0.019), comparable to the pooled 0.040 (Appendix Table A8). Estimates by each entry cohort are positive except for the Fall 2025 entrants, whose effect is essentially zero (-0.004); the largest, 0.320 among the 124 students entering in the final term, rests on 28 clusters and a single term of exposure, and is better read as noise around a small cohort than as a credible effect of that size. Reconstructing East Lake Academy's original assignment with a placement model estimated on its correctly implemented control grades, and returning the school to the sample, moves the pooled estimate from 0.040 to 0.043 population SD (Appendix Table A10)—the exclusion, disclosed in the pre-registration, is not driving results. Alternative fixed-effects structures, including school-by-term effects, leave the estimate between 0.039 and 0.043.

We also examine sensitivity with alternative inference methods. With 53 clusters, conventional cluster-robust standard errors may understate uncertainty, and the pooled estimate's significance holds under the alternatives, with a wild cluster bootstrap at the boundary of the five percent level ($p = 0.051$), school-level clustering at $p = 0.031$, and a leave-one-school-out exercise yielding no sign flips and no single influential school (Appendix Table A5).

Finally, contamination of the control group would bias these comparisons toward zero; a descriptive test regressing control-student outcomes on the treated share of their school's grades yields a negative, insignificant coefficient, providing no evidence that treated-grade proximity raised control outcomes, though the design has limited power on this margin.

Pre-specified heterogeneity analyses by demographics, grade, and baseline achievement

yield interactions that are uniformly small and insignificant, though imprecicely estimated (Appendix Tables A13–A15).

4.7 Exposure Histories

Because randomization fixed grades while students advanced across them, students observed in both years accumulated different assignment histories—assigned in both years, one year, or neither—and a saturated model on this subsample estimates Year-2 outcomes by trajectory (Appendix Table A12). We treat this analysis as exploratory: the both-years panel is small (1,438 students, of whom 1,030 have a Year-2 endline score, across 34 clusters) and selected, since remaining RTI-eligible across years itself depends on achievement. With those caveats, the estimates show two patterns and one non-finding. Students newly assigned in Year 2 gain 0.219 population SD (SE 0.061). The interaction between years is negative (−0.184, SE 0.072), rejecting the additivity restriction that would suggest a single pooled dosage interpretation.

On persistence, the data are uninformative: students assigned only in Year 1 show a Year-2 advantage of 0.058 (SE 0.064), consistent with either persistence or fade-out of a Year-1 effect that was itself small and imprecise. Whether the smaller second-year gain among twice-assigned students reflects selection through RTI exit, diminishing returns to a familiar platform, or noise in a small cell is not distinguishable.

5 Interpretation

5.1 What Carries the Effect

Assignment shifted every program input simultaneously—practice minutes, skills attempted, skills mastered, and Khanmigo use—so the experiment identifies the effect of the combined program rather than of any single component. Three complementary pieces of evidence nevertheless bear on the relative importance of the practice and tutoring margins.

First, within the treated arm, achievement growth is associated with the mastery margin of practice rather than with chat activity. In regressions of end-of-term scores on engagement measures among treated student-terms, each additional skill brought to proficiency per week is associated with 0.015 population SD higher achievement (SE 0.002), an association that is stable across specifications. Weekly Khanmigo chats display a small positive univariate association (0.006, SE 0.003) that declines to approximately zero conditional on skill mastery (−0.001; Table 9). These regressions are descriptive: engagement is not randomly assigned within the treated arm, and students who engage more intensively likely differ in unobserved

respects. We therefore interpret them as a decomposition of the achievement signal rather than as causal estimates.

Table 9: Engagement and Achievement Within the Treated Arm

	(1) Skills only	(2) Chats only	(3) Joint	(4) Joint + Learning Minutes
Skills proficient/week	0.015***		0.015***	0.014***
	(0.002)		(0.002)	(0.002)
Khanmigo chats/week		0.006*	-0.001	-0.002
		(0.003)	(0.003)	(0.004)
Learning time (mins/week)				0.000
				(0.000)
Baseline (pop SD)	0.846***	0.839***	0.846***	0.847***
	(0.032)	(0.032)	(0.032)	(0.031)
Observations	3,513	3,513	3,513	3,513
Clusters	28	28	28	28

Notes: Descriptive associations between platform engagement and end-of-term mathematics achievement, estimated within the treated arm only. The unit of observation is the treated student-term with both a baseline and an endline MAP score (28 grade-within-school clusters). The outcome is end-of-term RIT standardized by the full study-population standard deviation within grade; every column conditions on the baseline score. Columns add engagement regressors in sequence: weekly skills raised to proficiency alone; weekly Khanmigo chats alone; the two jointly; and weekly learning-path minutes added. Engagement is not randomly assigned within the treated arm, so coefficients are conditional associations rather than causal channels. All columns include school, grade, and term fixed effects; standard errors, in parentheses, are clustered at the grade-within-school level, the unit of randomization. $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Second, rescaling the experimental effect by alternative engagement measures yields magnitudes consistent with the practice interpretation. Instrumented by assignment, the effect per weekly skill mastered is 0.025 to 0.028 population SD, close to the within-treated association. The same reduced form expressed per weekly chat implies 0.040 SD per chat, which is difficult to interpret as the causal effect of a single conversation and is more plausibly the practice effect scaled by a small chat margin (Appendix Table A7). With a single instrument, these rescalings cannot separate channels. They are consistent with practice intensity, converted into mastered skills, accounting for the measured effect.

Third, the realized dose of conversational tutoring was small in absolute terms. As Section 3 documents, the median student messaged the tutor in 14 percent of exercise sessions, and roughly one message in seven contained a mathematical question or step of reasoning. Whatever the tutor contributes per engaged interaction, such interactions were infrequent.

Taken together, the evidence supports interpreting our estimates primarily as the effect of structured, individually targeted practice, with the tutor's contribution bounded by the

low intensity of its use. A complementary reading is that the program operates in part as an implementation technology for the remedial block itself. Realized RTI effects are known to hinge on implementation (Balu et al., 2015), and the platform standardizes precisely the elements that vary most across business-as-usual sessions: placement, pacing, and the content of practice, configured automatically from each student's screener scores. On this reading the relevant distinction is not between technology and implementation but between implementation features embedded in the product, which replicate wherever it is adopted, and the support supplied by the research partnership, which may not; the cost-effectiveness discussion below returns to the latter as a bound on external validity.

5.2 Cost-Effectiveness

The program's marginal cost compared to the control group is the license alone—about $15 per student per year, a price that may itself fall as the underlying model costs decline—because delivery occurred within remedial blocks the district already staffed.

High-dosage tutoring, the benchmark intervention for this population, produces larger annual effects of 0.2 to 0.4 SD, but at costs of several thousand dollars per student and with challenges to implement at large scale (Guryan et al., 2023; Kraft and Falken, 2021). The comparison should not be read as a substitution argument for tutors (or teachers): human tutoring delivers more learning per student per year, and nothing in our results suggests that software closes that gap. It is instead a statement about deployment margins: for districts already staffing remedial blocks, directing them toward structured adaptive practice involves little additional expenditure and yields measurable gains. The more consequential question is how much room the same expenditure leaves. Both margins of improvement in this study came free of marginal cost: realized practice roughly doubled between years, possibly through implementation improvements, and the AI tutor—included in the license but contributing no detectable achievement effect as deployed—represents capability districts have already paid for. Whether better implementation can push practice further, and whether redesigned tutoring can convert the AI's latent capability into learning at the same price, are the margins on which this technology's value will be decided.

5.3 Engagement as the Binding Margin

The results in Sections 3 and 4 share a common structure: the program increased use of CAL as intended, but with little interaction with the platform's AI tutor. Delivery within a mandatory, supervised block generated roughly thirty minutes of weekly practice, in a context where voluntary deployments generate almost none, and that practice was associated

with learning at rates comparable to the most effective classrooms observed on the platform (Eames et al., 2026). Conversation with the tutor, by contrast, required student initiative, and student initiative was rare: nearly all students tried Khanmigo, but the median student did not message it on most practice days, in most exercise sessions, or in most sessions containing a mistake.

The pattern reflects a familiar finding in the economics of education: interventions that rely on student initiative reach fewest of the students who would benefit most (Lavecchia et al., 2016). Struggling students are the least likely to seek help unprompted, and the barriers to doing so bind when students face immediate costs and uncertain long-term benefits, especially when actions being considered are incremental, like considering one mistake on one math problem. The limited dialogue we document is thus less a defect of the AI tutor than a lack of interest in attention and learning.

Subsequent developments are consistent with this interpretation. Khan Academy redesigned Khanmigo in 2026 to activate automatically during practice, citing lower student-initiated use than anticipated (Barnum, 2026), and experimental evidence from other platforms indicates that known human support raises engagement with AI tutoring where availability alone does not (Robinson et al., 2026).

These findings suggest that the returns to further investment in AI tutoring depend on the engagement margin as much as on model capability. In this deployment, the constraint was not what the tutor could do but how often students engaged it, and engagement is responsive to design.—defaults, structure, facilitation, and interface—in ways that capability improvements alone may not address. Whether richer capabilities, such as voice interaction or awareness of the student's work, would themselves raise engagement is an open empirical question. What the present study establishes is the benchmark such designs must exceed: provided free access, mandatory practice time, and an interface that invited use, the median struggling student rarely engaged the tutor. Making AI effective for learning appears to be as much a behavioral challenge as a technological one.

6 Conclusion

Generative AI has been promoted as the technology that might finally deliver individualized tutoring at scale. This paper reports evidence from a two-year cluster randomized trial in which 18 middle schools delivered Khan Academy with its AI tutor, Khanmigo, inside existing daily remedial mathematics sessions. The program raised achievement by roughly 0.06 to 0.08 standard deviations per school year. Effects tracked implementation closely and imply returns per hour of practice in the range of what has been observed for the platform

without AI assistance. Among students who remained in the program, the implied effect of a full year of participation under mature implementation reaches 0.14 standard deviations—approaching half the annual gain of high-dosage tutoring, at a cost orders of magnitude lower.

The distinctive contribution of the study is the observation of how students actually used an AI tutor over an extended deployment. Nearly every student tried Khanmigo. The median student then sent it no messages on two-thirds of the days they practiced, messaged it in 17 percent of the exercise sessions in which they made a mistake, and, when messaging, mostly sent bare answers or clicked suggested prompts; about one message in seven contained an interactive mathematical question or a step of reasoning. The achievement gains the program produced are accordingly attributable primarily to structured, individually targeted practice rather than to conversational tutoring, whose realized dose was small. The design cannot isolate the tutor's marginal contribution, but the usage data bound it: whatever an AI tutor can do in principle, this one was seldom asked.

The intervention we study tried to remove standard access barriers, embedding a free, inviting, guardrailed AI tutor in mandatory supervised practice time to offer assistance when no teacher or human tutor was available. Despite this, student-initiated dialogue remained rare, even after students made mistakes.

It may be that human attention is the key ingredient for realizing the full benefits of personalized learning. More research is clearly needed for exploring AI's full potential. The challenges are just as much behavioral as they are technological. Improving engagement will require a choice architecture that addresses low motivation, effort aversion, lack of confidence, and a fixed mindset. From the technological side, the AI tutor itself is a moving target: systems that speak and listen, generate figures and worked representations on demand, observe how a student is working, or adapt their manner—including humor and encouragement—to the individual student may elicit engagement that a text chat behind a click did not. These capabilities are developing rapidly even if they are not yet mature, and Khan Academy itself continues to redesign Khanmigo toward more active engagement (Barnum, 2026).

For school systems, the practical conclusion is more immediate. Structured adaptive practice delivered within existing remedial time produced real learning gains at trivial marginal cost, and did so under the ordinary frictions of public school implementation. Districts need not resolve the larger question of what AI tutoring may become in order to act on what the bundle already delivers. The larger question remains open. On the evidence here, its answer will depend not only on how capable tutors become, but on whether their capabilities are directed at the margin this study finds binding: getting students to engage.

References

Balu, Rekha, Pei Zhu, Fred Doolittle, Ellen Schiller, Joseph Jenkins, and Russell Gersten, "Evaluation of Response to Intervention Practices for Elementary School Reading," Technical Report NCEE 2016-4000, National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education 2015.

Barnum, Matt, "Why Sal Khan Is Rethinking How AI Will Change Schools," Chalkbeat, April 9, 2026 2026.

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

Bastani, Hamsa, Osbert Bastani, Alp Sungu, Haosen Ge, Özge Kabakcı, and Rei Mariman, "Generative AI Without Guardrails Can Harm Learning: Evidence from High School Mathematics," *Proceedings of the National Academy of Sciences*, 2025, *122* (26), e2422633122.

Bhatt, Monica P., Jonathan Guryan, Salman Khan, Michael LaForest-Tucker, and Bhavya Mishra, "Can Technology Facilitate Scale? Evidence from a Randomized Evaluation of High Dosage Tutoring," NBER Working Paper 32510, National Bureau of Economic Research 2024.

Cameron, A. Colin, Jonah B. Gelbach, and Douglas L. Miller, "Bootstrap-Based Improvements for Inference with Clustered Errors," *The Review of Economics and Statistics*, 2008, *90* (3), 414–427.

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

Eames, Taryn, Emma Brunskill, Bogdan Yamkovenko, Kodi Weatherholtz, and Philip Oreopoulos, "Computer-Assisted Learning in the Real World: How Khan Academy Influences Student Math Learning," *Proceedings of the National Academy of Sciences*, 2026, *123* (10).

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

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

Fuchs, Douglas and Lynn S. Fuchs, "Introduction to Response to Intervention: What, Why, and How Valid Is It?," *Reading Research Quarterly*, 2006, *41* (1), 93–99.

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

Henkel, Owen, Hannah Horne-Robinson, Nessie Kozhakhmetova, and Amanda Lee, "Effective and Scalable Math Support: Experimental Evidence on the Impact of an AI Math Tutor in Ghana," in "Artificial Intelligence in Education (AIED 2024), Communications in Computer and Information Science, vol. 2150" Springer 2024.

Holt, Laurence, "The 5 Percent Problem," *Education Next*, 2024, *24* (2), 26–30.

Kestin, Gregory, Kelly Miller, Anna Klales, Timothy Milbourne, and Gregorio Ponti, "AI Tutoring Outperforms In-Class Active Learning: An RCT Introducing a Novel Research-Based Design in an Authentic Educational Setting," *Scientific Reports*, 2025, *15.*

Khan, Salman, "How AI Could Save (Not Destroy) Education," TED Talk, April 2023 2023.

Kraft, Matthew A. and Grace T. Falken, "A Blueprint for Scaling Tutoring and Mentoring Across Public Schools," *AERA Open*, 2021, *7.*

_ **and Manuel Monti-Nussbaum**, "The Big Problem with Little Interruptions to Classroom Learning," *AERA Open*, 2021, *7.*

Kulik, James A. and J. D. Fletcher, "Effectiveness of Intelligent Tutoring Systems: A Meta-Analytic Review," *Review of Educational Research*, 2016, *86* (1), 42–78.

Lavecchia, Adam M., Heidi Liu, and Philip Oreopoulos, "Behavioral Economics of Education: Progress and Possibilities," in Eric A. Hanushek, Stephen Machin, and Ludger Woessmann, eds., *Handbook of the Economics of Education*, Vol. 5, Elsevier, 2016, pp. 1–74.

LearnLM Team, Google DeepMind and Eedi, "AI Tutoring Can Safely and Effectively Support Students: An Exploratory RCT in UK Classrooms," Technical Report, arXiv preprint arXiv:2512.23633 2025.

McKenzie, David, "Beyond Baseline and Follow-Up: The Case for More T in Experiments," *Journal of Development Economics*, 2012, *99* (2), 210–221.

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

Nickow, Andre, Philip Oreopoulos, and Vincent Quan, "The Promise of Tutoring for PreK–12 Learning: A Systematic Review and Meta-Analysis of the Experimental Evidence," *American Educational Research Journal*, 2024, *61* (1), 74–118.

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

_ , Michael Liut, Alp Sungu, and Nina Low, "Can a Proactive AI Tutor Improve Learning in Real Classroom Math Practice? Experimental Evidence from NUMI," Working Paper, University of Toronto 2026.

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

Robinson, Carly D., Biraj Bisht, and Susanna Loeb, "The Inequity of Opt-In Educational Resources and an Intervention to Increase Equitable Access," *Educational Researcher*, 2025.

_ , David Gormley, Ana Trindade Ribeiro, and Susanna Loeb, "Access Is Not Enough: Human Support Improves Engagement with AI Tutoring," EdWorkingPaper 26-1451, Annenberg Institute at Brown University 2026.

Roschelle, Jeremy, Mingyu Feng, Robert F. Murphy, and Craig A. Mason, "Online Mathematics Homework Increases Student Achievement," *AERA Open*, 2016, *2* (4), 1–12.

Wang, Rose E., Ana T. Ribeiro, Carly D. Robinson, Susanna Loeb, and Dorottya Demszky, "Tutor CoPilot: A Human-AI Approach for Scaling Real-Time Expertise," Working Paper, Stanford University 2025.

A Additional Sample and Design Exhibits

Table A1: Balance Test at First Entry

	Winter 2025		Spring 2025		Fall 2025		Winter 2026		Spring 2026		Pooled	
	Control	Diff	Control	Diff	Control	Diff	Control	Diff	Control	Diff	Control	Diff
Female	0.517	-0.055*	0.640	-0.040	0.572	0.007	0.550	-0.030	0.557	-0.047	0.549	-0.040
		(0.030)		(0.076)		(0.039)		(0.066)		(0.091)		(0.026)
White	0.228	0.031	0.326	0.037	0.287	0.195***	0.440	0.088	0.405	0.044	0.287	0.074
		(0.026)		(0.074)		(0.038)		(0.066)		(0.091)		(0.082)
Black	0.543	-0.007	0.395	0.080	0.483	-0.128***	0.367	-0.123**	0.380	-0.012	0.484	-0.038
		(0.030)		(0.077)		(0.039)		(0.061)		(0.089)		(0.075)
Hispanic	0.210	-0.025	0.256	-0.118*	0.204	-0.073**	0.156	0.023	0.215	-0.032	0.207	-0.040
		(0.024)		(0.061)		(0.029)		(0.049)		(0.073)		(0.033)
Other Race	0.020	0.002	0.023	0.002	0.026	0.005	0.037	0.012	0.000	0.000	0.022	0.004
		(0.009)		(0.024)		(0.013)		(0.027)		(0.000)		(0.007)
Free and Reduced-Price Meals	0.705	-0.009	0.733	-0.020	0.690	-0.059	0.569	0.025	0.633	-0.000	0.684	-0.017
		(0.028)		(0.070)		(0.038)		(0.065)		(0.088)		(0.046)
Economic Disadvantage	0.467	-0.003	0.512	-0.062	0.437	-0.082**	0.385	-0.060	0.342	-0.097	0.445	-0.033
		(0.030)		(0.078)		(0.039)		(0.063)		(0.082)		(0.060)
Dyslexia	0.137	0.051**	0.093	0.032	0.046	0.064***	0.083	-0.018	0.101	-0.060	0.098	0.047
		(0.022)		(0.049)		(0.022)		(0.035)		(0.045)		(0.034)
Special Education Disability	0.174	-0.017	0.105	0.170***	0.270	-0.005	0.294	-0.033	0.392	-0.250***	0.225	-0.022
		(0.023)		(0.060)		(0.035)		(0.059)		(0.075)		(0.024)
ELA Baseline (pop. SD)	-0.581	-0.120**	-0.192	-0.209	-0.439	-0.088	-0.345	0.142*	-0.263	0.323***	-0.463	-0.086
		(0.058)		(0.129)		(0.067)		(0.079)		(0.108)		(0.146)
Math Baseline (pop. SD)	-0.917	0.018	-0.538	-0.152	-0.742	-0.041	-0.867	0.095	-0.662	0.134	-0.811	-0.014
		(0.043)		(0.096)		(0.049)		(0.067)		(0.093)		(0.081)
N: full RTI sample (Control / Treated)	556	721***	101	91	421	342***	120	135	94	65**	1292	1354
N: endline only (Control / Treated)	511	623***	93	85	400	335**	111	125	84	53***	1199	1221
N: baseline & endline (Control / Treated)	505	596***	86	80	348	290**	109	123	79	49***	1127	1138

Notes: Treated-versus-control covariate balance measured once per student, at the term of first entry into the estimating sample. Columns and construction follow Table 2: for each entry term, the first column reports the control-arm mean and the second the treated-minus-control difference, with its standard error in parentheses; baselines are the prior MAP screeners standardized to the full grade-level population. Per-term differences use heteroskedasticity-robust standard errors; pooled differences cluster at the grade-within-school level, the unit of randomization. $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table A2: Practice and Khanmigo Intensity by Term

	Winter 2025	Spring 2025	Fall 2025	Winter 2026	Spring 2026
Learning time (min/wk)	29.3***	22.3***	33.1***	33.4***	29.8***
	(3.8)	(2.8)	(4.1)	(2.8)	(2.7)
Skills worked /wk	3.9***	2.7***	4.4***	4.9***	4.4***
	(0.5)	(0.4)	(0.5)	(0.5)	(0.3)
Skills to proficient /wk	2.5***	1.7***	3.2***	3.5***	2.9***
	(0.4)	(0.3)	(0.4)	(0.3)	(0.2)
Skills worked /hr	8.1***	5.1***	4.9***	5.9***	5.5***
	(1.5)	(1.2)	(0.4)	(0.6)	(0.5)
Skills to proficient /hr	3.7***	2.6***	3.4***	4.2***	3.5***
	(0.5)	(0.5)	(0.3)	(0.4)	(0.2)
Khanmigo chats /wk	1.6***	1.2***	2.3***	1.8***	1.5***
	(0.2)	(0.2)	(0.3)	(0.1)	(0.2)
Min RTI weeks	4	7	4	8	8
N	1101	1173	1419	1616	1593

Notes: Intent-to-treat effects of KWiK grade assignment on platform use, by term, on the sample of student-terms with both a baseline and an endline MAP score. Each cell is the treated-minus-control difference in that term's engagement measure, from a regression with school and grade fixed effects; standard errors, in parentheses, are clustered at the grade-within-school level, the unit of randomization. Per-hour measures are set to zero when no learning time is logged, and weekly measures exclude holiday weeks. The platform was available only in treated grades, so control means are near zero and the differences approximately equal treated levels. Min RTI weeks is each term's protected delivery window. $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table A3: Khanmigo analysis sample

	Count
Panel A. Sample construction	
Activated Year 2 treatment students in included schools	573
Matched to Khan Academy exercise data with at least one recorded exercise session	563
Panel B. Scope of the analysis sample	
Students	563
Schools	16
Khan Academy active student-days	25,286
Distinct exercises worked	1,086
Exercise sessions	100,017
of which: mistake sessions	56,362
Panel C. Khanmigo use observed during exercise sessions	
Students ever sending a Khanmigo message	539 (95.7%)
Student messages sent to Khanmigo	94,485
of which: messages during mistake sessions	61,313

Notes: The analysis sample consists of the 563 activated Year 2 treatment students in included schools who were successfully matched to Khan Academy exercise records and had at least one recorded exercise session. East Lake Academy, Tyner Middle, and Howard Connect are excluded because of incomplete Khanmigo data capture. The observation window covers the three terms of the 2025–26 school year, from August 2025 through May 2026. A Khan Academy active student-day is a student–calendar day with at least one recorded exercise attempt. An exercise session is defined as the interval between the start of one exercise and the start of the next. A mistake session is an exercise session containing either a submitted incorrect answer or a start-over. Khanmigo use is measured from student messages to Khanmigo that fall inside an exercise-session window. Message counts refer to student messages only and exclude Khanmigo responses. The percentage in parentheses is calculated relative to the 563 students in the analysis sample.

Table A4: Distribution of engagement characteristics

	N	Mean	SD	P10	P25	Median	P75	P90	P95
Panel A. Practice, per student (counts)									
Distinct exercises worked	563	64	57	7	20	51	90	139	172
Exercise sessions	563	178	163	13	44	143	248	387	510
Submitted exercises	563	128	114	10	36	104	180	279	358
Mistake sessions	559	101	96	7	25	79	138	228	298
KA-active days	563	45	30	7	17	43	69	85	94
Panel B. Khanmigo use, per student (share of activity, %)									
Share of active days with Khanmigo	563	37.1	23.7	9.1	18.5	33.3	51.0	71.3	82.3
Share of sessions with Khanmigo	563	18.2	15.8	2.8	6.7	14.0	26.2	40.0	50.0
Share of mistake sessions with Khanmigo	559	21.9	18.1	3.8	8.2	17.1	31.2	50.0	53.2
Panel C. Messages per exercise session (messages within the window)									
Khanmigo messages, all sessions	100,017	0.9	4.3	0.0	0.0	0.0	0.0	2.0	6.0
non-mistake sessions	43,655	0.8	3.8	0.0	0.0	0.0	0.0	1.0	4.0
mistake sessions	56,362	1.1	4.6	0.0	0.0	0.0	0.0	2.0	7.0
mistake sessions with any message	9,362	6.5	9.7	1.0	1.0	3.0	8.0	16.0	22.0

B Robustness

Table A5: Inference Robustness: Primary Estimate

Inference approach	ITT estimate	SE	p-value
Primary inference			
Cluster-robust SE (grade-within-school, J=53)	1.264	(0.597)	0.0391
Wild cluster bootstrap (Rademacher, 999 reps)	1.264	—	0.0511
Sensitivity			
Cluster-robust SE (school, J=18); asymptotics weak	1.264	(0.538)	0.0312
Leave-one-school-out range (18 schools)	[1.0558, 1.6105]	—	—
Sign flips	0 of 18	—	—
Observations	6,902		

Notes: Alternative inference procedures for the pooled intent-to-treat estimate of Table 5 (baseline specification, percentile-rank outcome). The point estimate is identical in every row; only the inference procedure changes. Wild cluster bootstrap uses Rademacher weights with 999 replications. School-level clustering leaves 18 clusters, so its large-sample justification is weak and it is reported as a sensitivity check. The leave-one-school-out exercise re-estimates the effect dropping each school in turn and reports the resulting range and the number of sign changes.

Table A6: Alternative Fixed-Effects Structures

	(1) Primary	(2) Saturated	(3) Drop Grade	(4) + Demographics
KWiK Assignment	0.040**	0.043**	0.041**	0.039**
	(0.019)	(0.019)	(0.019)	(0.018)
Baseline (pop SD)	0.843***	0.845***	0.843***	0.804***
	(0.021)	(0.021)	(0.021)	(0.022)
School	Y		Y	Y
Grade	Y	Y		Y
Term	Y		Y	Y
School×Term		Y		
Demographics				Y
Observations	6,902	6,902	6,902	6,902
Clusters	53	53	53	53

Notes: The pooled intent-to-treat estimate of Table 5 under alternative fixed-effects structures, on the outcome standardized by the full study-population standard deviation. Column 1 reproduces the primary specification with school, grade, and term fixed effects; column 2 replaces school and term effects with school-by-term effects, retaining grade effects; column 3 drops grade effects; column 4 adds demographic controls. All columns condition on the prior-term baseline score; standard errors, in parentheses, are clustered at the grade-within-school level, the unit of randomization. $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table A7: Effects per Unit of Alternative Engagement Measures

	Year 1		Year 2		Stacked	
	Base	Full	Base	Full	Base	Full
Panel A: Learning-path minutes/week						
Per min/week	0.0009	0.0008	0.0029*	0.0028*	0.0023*	0.0023*
	(0.0015)	(0.0015)	(0.0015)	(0.0014)	(0.0013)	(0.0013)
Panel B: Skills to proficiency/week						
Per skill/week	0.0112	0.0103	0.0286**	0.0280**	0.0253*	0.0254*
	(0.0195)	(0.0189)	(0.0141)	(0.0136)	(0.0136)	(0.0131)
Panel C: Khanmigo chats/week						
Per chat/week	0.0162	0.0150	0.0482*	0.0470*	0.0405*	0.0404*
	(0.0287)	(0.0278)	(0.0244)	(0.0236)	(0.0225)	(0.0217)
Panel D: Engagement consistency (share of weeks with any practice)						
Per unit (0–1 share)	0.0531	0.0491	0.2047*	0.2000*	0.1559*	0.1562*
	(0.0943)	(0.0915)	(0.1046)	(0.1010)	(0.0876)	(0.0847)
Panel E: Platform exposure hours (total, term calendar through pre-testing)						
Per exposure hour	0.0034	0.0031	0.0079**	0.0077**	0.0072*	0.0072*
	(0.0059)	(0.0057)	(0.0039)	(0.0038)	(0.0040)	(0.0038)
First-stage F	60.7	60.7	129.6	129.6	143.6	143.6
Observations	1,214	1,214	1,545	1,545	2,759	2,759
Clusters	53	53	51	51	53	53

Notes: Two-stage least squares estimates in which each engagement measure, in turn, is instrumented by grade assignment; construction otherwise follows Table 8. Because a single instrument scales the same reduced-form effect by different first stages, the panels are alternative expressions of one experimental contrast, not separate causal channels. Panels A–D instrument weekly measures defined within protected delivery windows (Section 3): average learning-path minutes per week (A), skills brought to proficient mastery per week (B), Khanmigo chat sessions per week (C), and engagement consistency, the share of delivery weeks in which the student logged any platform activity (D), so that a value of one denotes practicing every week. Panel E instruments total platform learning hours over the term's calendar weeks through the last week before endline testing, the exposure measure of Table 8; its coefficient is the effect of one additional hour of realized practice, and its first-stage F is reported. First stages for the other measures are weaker. The outcome is end-of-year RIT standardized by the full study-population standard deviation. Standard errors, in parentheses, are clustered at the grade-within-school level, the unit of randomization. $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table A8: Restricting to Term-1 Entrants: Term-Stacked Estimates

	No Controls	Baseline	Full Controls
KWiK Assignment	0.069	0.058***	0.053***
	(0.048)	(0.019)	(0.019)
Baseline (pop SD)		0.834***	0.792***
		(0.027)	(0.029)
Observations	3943	3943	3943
Clusters	53	53	53

Notes: The pooled intent-to-treat estimate restricted to students who entered the sample in the first study term, followed across all five terms, removing any influence of later entry on sample composition. The outcome is standardized by the full study-population standard deviation. Columns add controls in sequence as in Table 5; standard errors, in parentheses, are clustered at the grade-within-school level. $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table A9: Stacked Term Effects by Entry Cohort

	Y1-T1	Y1-T2	Y2-T3	Y2-T4	Y2-T5
Panel A: control SD, pooled by grade					
KWiK Assignment	0.072***	0.001	-0.005	0.064	0.398***
	(0.023)	(0.071)	(0.029)	(0.072)	(0.054)
Panel B: full-population SD, pooled by grade					
KWiK Assignment	0.058***	0.004	-0.004	0.049	0.320***
	(0.019)	(0.057)	(0.023)	(0.058)	(0.043)
Students	1,186	172	750	236	124
Student-terms	3,943	455	1,930	446	124
Clusters	53	40	47	40	28

Notes: Intent-to-treat estimates computed separately for each entry cohort: students whose first term in the sample is the column term, followed as student-terms over all their subsequent terms under the standard retention rules. Coefficients are per-term effects, comparable across cohorts and to the pooled estimate of Table 5; the first cohort's estimate is the Term-1-entrant restriction of Appendix Table A8. Panels standardize the outcome by the control-group and full-population standard deviations. Specifications condition on the baseline score with school and grade fixed effects; standard errors, in parentheses, are clustered at the grade-within-school level. Cohort samples shrink sharply in later terms. $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table A10: Including Reconstructed East Lake

	No Controls		Baseline		Full Controls	
	Primary	Incl. EL	Primary	Incl. EL	Primary	Incl. EL
KWiK Assignment	0.003	-0.009	0.040**	0.043**	0.039**	0.043***
	(0.046)	(0.040)	(0.019)	(0.016)	(0.018)	(0.016)
Baseline RIT (pop-SD)			0.843***	0.848***	0.804***	0.812***
			(0.021)	(0.019)	(0.022)	(0.020)
Observations	6902	7712	6902	7712	6902	7712
Clusters	53	56	53	56	53	56

Notes: The pooled intent-to-treat estimate with and without East Lake Academy, the school excluded because staff reassigned students across grades in violation of the randomized assignment. Scheduled RTI placement is modeled with a logit on baseline MAP mathematics and reading scores and indicators for dyslexia, special-education status, and economic disadvantage, estimated on the school's correctly implemented control grades and applied to all grades; a student-year is classified as RTI-eligible when its predicted probability exceeds the control-grade placement rate under models trained on each control grade. Both arms are defined by predicted eligibility, so they are constructed symmetrically. Reconstructed East Lake enters as its own school with three grade-within-school clusters. The outcome is standardized by the full study-population standard deviation; standard errors, in parentheses, are clustered at the grade-within-school level. $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table A11: Cross-Grade Spillovers

	Control-RTI students
Share of Grades Treated (ρ)	-3.601
	(3.095)
Baseline MAP (NPR)	0.776***
	(0.024)
Observations	3389
Clusters	38

Notes: Tests whether treatment in a school's assigned grades affected control-grade RTI students in the same school. The sample is control-grade RTI student-terms with both scores; the outcome is the end-of-term MAP percentile rank. The regressor of interest is the share of the school's grades assigned to KWiK, fixed across years, identified from cross-school variation; the specification conditions on the baseline score and demographics with grade and term fixed effects, and standard errors are clustered at the grade-within-school level. The treated share may proxy for other school characteristics, so the estimate is descriptive. $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table A12: Effects by Two-Year Exposure History

	No Controls	Baseline	Full Controls
Assigned Year 1 (β_1)	-0.013	0.058	0.046
	(0.103)	(0.064)	(0.069)
Assigned Year 2 (β_2)	0.196*	0.219***	0.208***
	(0.106)	(0.061)	(0.056)
Year 1 × Year 2 (β_3)	-0.241*	-0.184**	-0.172**
	(0.128)	(0.072)	(0.073)
Observations	1,030	1,030	1,030
Clusters	34	34	34

Notes: Year-2 outcomes by two-year assignment history, on the panel of students observed in both years. The specification regresses the Year-2 annual outcome on indicators for Year-1 assignment, Year-2 assignment, and their interaction, conditioning on the fall Year-2 baseline score, with school and grade fixed effects; the omitted category is students assigned in neither year. Because the baseline absorbs any retained Year-1 gains, all coefficients measure new Year-2 growth: the implied growth effects relative to never-assigned students are β_1 for Year-1-only, β_2 for newly assigned, and $\beta_1 + \beta_2 + \beta_3$ for twice-assigned students. The outcome is standardized by the full study-population standard deviation. Columns add controls in sequence; standard errors, in parentheses, are clustered at the grade-within-school level. $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

C Heterogeneity

Table A13: Heterogeneity by Student Characteristics

	Female	Black	Hispanic	FARMS	SPED
KWiK Assignment	0.029	0.043**	0.040**	0.037*	0.038**
	(0.022)	(0.020)	(0.020)	(0.022)	(0.017)
KWiK × Female	0.021				
	(0.021)				
KWiK × Black		-0.008			
		(0.022)			
KWiK × Hispanic			-0.004		
			(0.021)		
KWiK × FARMS				0.005	
				(0.020)	
KWiK × SPED					0.002
					(0.029)
Observations	6902	6902	6902	6902	6902
Clusters	53	53	53	53	53

Notes: Pooled intent-to-treat estimates interacted with pre-specified student characteristics, on the outcome standardized by the full study-population standard deviation. FARMS denotes eligibility for free and reduced-price meals; SPED denotes a special-education designation. Each column reports the main assignment effect and its interaction with the indicated characteristic from a separate regression conditioning on the baseline score with school, grade, and term fixed effects; standard errors, in parentheses, are clustered at the grade-within-school level. Interaction standard errors of 0.02 to 0.03 mean the design can rule out only differential effects larger than roughly 0.05 SD; the estimates indicate no large heterogeneity rather than its absence. Heterogeneity analyses were pre-registered as exploratory. $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table A14: Heterogeneity by Grade

	Grade heterogeneity
KWiK Assignment	0.010
	(0.041)
KWiK × Grade 7	0.005
	(0.054)
KWiK × Grade 8	0.087
	(0.074)
Observations	6902
Clusters	53

Notes: Pooled intent-to-treat estimates interacted with grade level, constructed as in Table A13. $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table A15: Heterogeneity by Baseline Achievement

	Baseline interaction
KWiK Assignment	0.0487
	(0.0325)
KWiK × Baseline RIT	0.0134
	(0.0401)
Baseline RIT (pop-SD)	0.8355***
	(0.0259)
Observations	6902
Clusters	53

Notes: Pooled intent-to-treat estimates interacted with baseline mathematics achievement, constructed as in Table A13. $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.