NBER WORKING PAPER SERIES

VIRTUAL TUTORING WITH COMPUTER-ASSISTED LEARNING: AN EXPERIMENT IN TAKE-UP AND LEARNING

Philip Oreopoulos
Ruochong Dong
Nina Low

Working Paper 35622
http://www.nber.org/papers/w35622

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

We are grateful to the Toronto District School Board, participating teachers, school leaders, and students for making this study possible. We thank Amie Presley and Bryce Archer for their partnership and support throughout the project. We are also grateful to Mehwish Saeed, Vanessa Jaipersaud, Keya Patel, Tishma Joarder, Ridwan Iqbal, Hiba Jamshedi, Gia Khanh Luong, Farhan Karepilakkiyil, Khurshid Phirozmand, and Marcus Damiani for their dedication to supporting their teachers, and to the many volunteer tutors who contributed their time to support students throughout the study. This research was supported by the Smith Richardson Foundation, the Social Sciences and Humanities Research Council of Canada (SSHRC), and the Abdul Latif Jameel Poverty Action Lab (J-PAL). This study was registered with the AEA RCT Registry under trial number AEARCTR-0011639. All errors are our own. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

NBER working papers are circulated for discussion and comment purposes. They have not been peer-reviewed or been subject to the review by the NBER Board of Directors that accompanies official NBER publications.

© 2026 by Philip Oreopoulos, Ruochong Dong, and Nina Low. All rights reserved. Short sections of text, not to exceed two paragraphs, may be quoted without explicit permission provided that full credit, including © notice, is given to the source.

Virtual Tutoring with Computer-Assisted Learning: An Experiment in Take-Up and Learning
Philip Oreopoulos, Ruochong Dong, and Nina Low
NBER Working Paper No. 35622
August 2026
JEL No. C93, D91, I21, I24, I28

ABSTRACT

We explore the efficacy of an online after-school tutoring program that complements teacher-provided computer-assisted learning practice. The aim is to scaffold students struggling with online Khan Academy practice at school with low-cost, one-on-one virtual supervision for about an hour a week at home. In a two-year randomized trial with the Toronto District School Board, teachers identified struggling Grade 4–8 students, whose parents were then randomly offered additional online tutoring to support their practice at home. The main obstacle was participation. Only 45% of assigned students reached a first session in Year 1. After we reframed the invitation and simplified enrollment, initial take-up in Year 2 rose to 83%, though weekly attendance remained below half—not because students dropped out, but because they attended intermittently. On average, the offer raised Khan Academy practice by about 10 minutes per week. This translated to an average gain in math assessment scores of 0.06 SD, with a confidence interval too wide to rule out null or medium-sized gains (ITT 0.055 SD; 95%CI [0.09, 0.20]). We find a marginally significant gain in math report-card marks (about 0.08 SD; 95% CI [0.01, 0.17]) and improvements in math satisfaction and enjoyment, with effects larger in Year 2 than in Year 1, consistent with the higher take-up that year. We therefore conclude that the program holds promise, but more is needed to ensure students actually engage with tutors on a regular basis.

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

Ruochong Dong
University of Toronto
Economics
ruochong.dong@mail.utoronto.ca

Nina Low
Charles River Associates
nlow@crai.com

A randomized controlled trials registry entry is available at
https://www.socialscienceregistry.org/trials/11639

1 Introduction

Two of the most promising tools for helping students who have fallen behind in math pull in opposite directions on cost and scale. A large body of randomized evidence suggests tutoring can raise achievement substantially: recent meta-analyses put the average effect in the range of roughly 0.3 standard deviations, with the largest gains when tutoring is frequent, staffed by trained adults, and delivered during the school day (Nickow et al., 2024; Guryan et al., 2023; Kraft et al., 2026). But those same features make tutoring expensive—the intensive, in-school model studied by Guryan et al. (2023) cost several thousand dollars per student per year—and the evidence suggests effects tend to shrink as programs grow, with pooled estimates for the largest-scale programs closer to 0.15–0.20 SD (Kraft et al., 2026). Computer-assisted learning (CAL) has close to the opposite profile. Platforms like Khan Academy are free, adaptive, and available to anyone, but access does not appear to be the binding input. Whether CAL raises achievement seems to depend on whether students actually practice on it, with enough regularity and support to make the practice productive: randomized evaluations range from null results where that practice never materializes, to small gains under ordinary low-intensity classroom use, to gains approaching those of tutoring itself where schools protect repeated, supervised practice time (Escueta et al., 2020; Muralidharan et al., 2019; Eames et al., 2026; Oreopoulos et al., 2024, 2026).

This paper studies a hybrid meant to capture some of the tutoring benefit at closer to the CAL cost. Students are assigned practice on Khan Academy in class. Some keep up on their own; others, typically the ones furthest behind, do not. For those students, we add a tutor—but the tutor's job is not to teach a separate curriculum. It is to sit with the student once a week over Zoom, and work through the same Khan Academy practice the teacher has already assigned. Because the platform supplies the content and the feedback, the tutor needs little preparation or subject training; the main value they add is supervision, encouragement, and a reason to show up and keep going. Because the invitation comes from the student's own teacher and points to work the student is already meant to do, it should be easier to trust and act on than an offer from an unfamiliar outside provider.

We call the program TWiK (Tutoring WIth Khan Academy), and it is layered on a classroom CAL program called KWiK (Khoaching WIth Khan Academy), in which coaches help teachers assign and monitor Khan Academy practice (Oreopoulos et al., 2024). We evaluate the tutoring layer in a two-year randomized trial with the Toronto District School Board (TDSB), Canada's largest school district. In participating classrooms, teachers nominated Grade 4–8 students struggling on Khan Academy, and nominated students were randomly assigned to receive the tutoring offer or not. Both arms received the classroom CAL pro-

gram, so the experiment identifies the effect of adding the home tutoring layer, not the effect of the platform itself.

We find that getting students to actually use a tutor outside of school is hard—and keeping them coming back is harder. Teachers and parents consistently told us the program was a good idea—it was free, it came recommended by the student's own teacher, and it supported work the student was already assigned—and yet in Year 1 only about 45% of assigned students attended even a single session. Reaching that first session required a family to clear a series of steps outside the school day—notice the invitation, express interest, complete enrollment on an outside platform, get matched to a tutor, and show up at the scheduled time—and students were lost at every one.

Much of what we learned is about how to shorten that chain. In Year 1, families had to act on a general invitation, create an account on a third-party tutoring platform, and schedule sessions themselves; in Year 2, the teacher's invitation asked directly for a weekly time, enrollment and scheduling were handled internally, and a recurring session was set up on the family's behalf so that expressing interest was nearly enough to be enrolled. Collapsing the steps this way—fewer handoffs, no separate platform account, a standing appointment rather than a fresh decision each week—raised first-session take-up sharply, from about 45% to 83%. But getting students through the door, however, was not enough to keep them there. Attendance stayed intermittent, near two-fifths in a typical week even among students who had started, and practice among those who attended rose by only about 17 minutes a week. What looks like attrition is mostly absence: students attended some weeks while missing others.

The program hints at being effective, and could be more so if sustained take-up were higher. The intent-to-treat effect on a short achievement test covering the Khan Academy topics assigned by the teacher is 0.055 SD, with a wide confidence interval of about $[-0.09, 0.20]$—positive, but too wide to separate a null from the 0.03–0.09 SD gain that real-world Khan Academy use at this dose would predict (Eames et al., 2026). However, we do find that report-card marks rise by about 0.08 SD ($p < 0.10$), and treated students report enjoying math more and finding the platform more helpful. These gains grow between Year 1 and Year 2, as take-up rose and the program reached more students. No single result is decisive, but they point the same way and strengthen as the program reached more students—what we would expect if the program produces the modest gains its dose would predict.

The paper makes three contributions. First, it shows that a low-cost tutoring layer built on an existing classroom platform can produce a coherent set of positive signals—in engagement, attitudes, and school performance—consistent with modest learning gains, though a light realized dose leaves the achievement effect imprecise and in need of further

study. Second, it documents where a voluntary, home-based offer breaks down on the way to a first session, and shows that simple design changes—inviting families through their own teacher and stripping the sign-up down to a single step—can move take-up from under a half to over four-fifths. Third, it shows that solving entry is not the same as solving participation. The low session rate reflects intermittent absence rather than dropout: students miss weeks and come back. The tutoring was already recommended by the student's own teacher and tied to the work they were assigned in class, so the missing ingredient does not seem to be a reason to attend. Perhaps more reminders to parents and students would help.

The rest of the paper is organized as follows. Section 2 places TWiK in the context of research on tutoring, computer-assisted learning, and the take-up of voluntary education programs. Section 3 describes the setting, the program, and the experimental design. Section 4 studies take-up, engagement, and repeated participation—from the offer, to a first session, to sustained weekly practice. Section 5 reports program effect estiates on achievement and attitudes. Section 6 concludes.

2 Background

TWiK sits at the intersection of three literatures: the effectiveness of tutoring, the effectiveness of computer-assisted learning, and the take-up of voluntary education programs.

Tutoring is among the best-evidenced ways to help students who have fallen behind. A meta-analysis of nearly a hundred randomized trials puts the average effect near 0.3 standard deviations, with the largest gains from programs that are frequent, staffed by consistent and well-supported tutors, and run during the school day (Nickow et al., 2024). The intensive, in-school "high-dosage" model can raise adolescent math achievement substantially (Guryan et al., 2023). The difficulty is cost and scale. High-dosage programs run several thousand dollars per student per year, and effects tend to shrink as programs grow: aligning meta-analytic estimates to large, real-world programs lowers the expected effect to roughly 0.15–0.20 SD, driven by declining effectiveness at scale and the staffing and logistical strains that come with it (Kraft et al., 2026). Two levers dominate the cost of a tutoring program—who the tutors are and how much they must be trained—and both are hard to relax without giving up the personal attention that makes tutoring work.

Computer-assisted learning offers the opposite trade-off. Adaptive platforms like Khan Academy are free or inexpensive, scale effortlessly, and can deliver individualized practice to any student with a device. But access is not the binding input. Reviews of the evidence find that CAL raises achievement only under some conditions, with effects ranging from null to sizable depending on how the technology is used (Escueta et al., 2020; Muralidharan

et al., 2019). Recent evidence using thousands of classrooms and variation in the timing of teachers introducing or taking away Khan Academy practice in their classrooms, makes the pattern concrete. In ordinary classroom use, more weekly practice is associated with small but real test-score gains (Eames et al., 2026). When schools protect and supervise practice time, effects are much larger. In one program that this paper builds on, coaches help teachers build platform practice into the school day and monitor student progress, raising achievement well above the levels seen under unsupported classroom use (Oreopoulos et al., 2024); in another, in-class support that ensured students actually put in the minutes produced test score gains larger than 0.40 SD (Oreopoulos et al., 2026). The lesson across these studies is consistent: the platform supplies the content, but someone still has to get the student to practice regularly and with enough support for the practice to stick. That missing ingredient—supervised, sustained practice—is what a tutor can supply when motivation in school is not enough.

Delivering tutoring online is a natural way to relax the cost constraint: it removes travel, widens the pool of potential tutors, and lets volunteers rather than paid professionals staff sessions. A small but growing body of randomized evidence finds online tutoring can work. A trial of virtual tutoring for young readers, delivered during the school day, found positive and significant gains of 0.05–0.08 SD, larger for one-on-one sessions and for students furthest behind (Robinson et al., 2024). Online tutoring by college volunteers has shown consistently positive but imprecise effects on middle-school math and reading (Kraft et al., 2022). Closest to our setting, the Italian Tutoring Online Project (TOP) paired volunteer university students with disadvantaged middle-schoolers and raised math achievement by about 0.2 SD—an effect that held both during the pandemic and after schools reopened (Carlana and La Ferrara, 2025). This is encouraging for a low-cost, volunteer-staffed model, but in TOP, as in the school-day programs above, participation was secured before the experiment began: schools enrolled eligible students onto a beneficiary list, families consented up front, and only then were enrolled students randomized to receive a tutor or not. The experiment thus measures the effect of a tutor among families who had already opted in. TWiK reverses this order. We randomize the *offer* itself, sent to teacher-nominated families who have made no prior commitment, and a family must then notice the invitation, respond, enroll, and keep attending week after week—during a normal school year, outside school hours, with nothing compelling them to take part. Take-up is therefore inside our experiment rather than a precondition of the sample, and the binding constraint shifts from what happens inside a session to whether the student takes up the offer at all and keeps coming back.

That shift matters because voluntary education programs often have low take up, even when they are free and valuable. Small frictions—an extra form, an account to create, a

decision to schedule—sharply reduce participation in settings ranging from college financial aid to program enrollment (Bettinger et al., 2012; Lavecchia et al., 2016; Herd and Moynihan, 2018; Oreopoulos, 2020). The problem is especially acute for opt-in academic resources: reaching families through students alone often does nothing, and even active outreach to parents leaves take-up low, so that opt-in resources can widen rather than narrow gaps because they fail to reach the students who would benefit most (Robinson et al., 2025). A home-based tutoring offer inherits all of these frictions at once. It asks a family to notice an invitation, decide to act, complete an enrollment step, and then return week after week—each a point at which a willing family can quietly fall away. TWiK's design tries to blunt these frictions by routing the invitation through the student's own teacher, tying it to work the student is already assigned, and, in its second year, collapsing enrollment to a single step. Whether that is enough is an empirical question, and it is the first one we take up.

3 Setting, Program, and Design

Setting

We study TWiK in the Toronto District School Board (TDSB), Canada's largest school district, over two school years (2023–24 and 2024–25). The students in our sample are in Grades 4–8 and are already using Khan Academy in class as part of KWiK, a classroom program in which coaches help teachers assign platform practice, monitor progress, and build regular practice time into the school day (Oreopoulos et al., 2024). TWiK is a tutoring layer added on top of this classroom program: it targets the students within a KWiK classroom who are falling behind on their assigned practice and offers them additional one-on-one help at home.

Near the start of the school year, teachers identify students who are completing little of their assigned Khan Academy work or passing few of the exercises, and the family of each is offered free weekly online tutoring by email. Because both the tutored and the comparison students remain in the same classroom program, the experiment measures the effect of *adding* the home tutoring layer, not the effect of Khan Academy or of the classroom program itself.

What the tutor does

The tutoring works through the practice the teacher has already set. Each week, drawing on curriculum-based defaults, the teacher assigns a small set of practice exercises—typically two or three—which the teacher can adjust, add to, or tailor to individual students. Each skill pairs a short instructional video with a practice exercise of either four or seven questions, and

a student "levels up" by answering enough correctly—at least three of four, or five of seven, roughly 70%—to clear the skill and move on. What the tutor supplies is an hour a week of undivided attention in Year 2 (90 minutes in Year 1), encouragement through mistakes, and a standing reason to sit down and do the work. The tutor does not prepare lessons or teach a separate curriculum. They watch student's open their Khan Academy account and work through the exercises the teacher has already assigned, starting with whatever has gone unfinished.

A student who has fallen behind arrives at a session with the current week's exercises plus a possible backlog of earlier ones left incomplete, and the tutor helps clear that backlog one skill at a time. The goal of a session is to work through the assigned exercises toward mastery, leveling up at least one skill. Because the platform carries the content and grades the work, a volunteer with no special training can run a productive session—the value they add is supervision and persistence, not subject expertise, though they are free to offer additional support. In effect, the tutor "looks over the student's shoulder" on Zoom, for an hour a week of supervised, on-task practice for exactly the students least likely to generate it on their own.

Tutors are university student volunteers. They are not paid, though most of them from the University of Toronto could receive official extracurricular credit on their transcript, and all tutors could request a reference letter—modest incentives that, together with the appeal of the work, were enough to staff the program (often using outreach emails from instructors of large introductory and second-year courses). Tutors completed a short orientation to the platform and session routine and were then asked to complete a third-party police record check. The process generally required tutors to verify their address and upload identification and could typically be completed within a few minutes. In some cases, an inactive request expired before completion, in which case TWiK program staff sent reminders and issued a new request. Tutors under age 18 could not complete the standard Triton screening process and therefore required separate handling. Training was intentionally light because tutors were expected primarily to supervise platform-directed practice rather than deliver a curriculum of their own.

On the classroom side, teachers set up and adjusted the Khan Academy assignments with the support of KWiK coaches (Oreopoulos et al., 2024), and nominated students they thought could benefit from additional tutoring. Nominated students were randomly assigned either to receive the tutoring offer or to remain in the comparison group. Among families assigned to the offer who enrolled, program staff then matched students with available tutors. Online tutoring sessions were automatically recorded for safeguarding and quality-monitoring purposes, with parents and tutors informed of the recording practice and providing con-

Table 1: Delivery regimes in Year 1 and Year 2

	Year 1	**Year 2**
Invitation	General invitation from the teacher to express interest	Teacher invitation asking directly for weekly availability
Account	Family creates an account and password on an external tutoring platform, using the student's school email	No account required
Enrollment	A separate interest form, then account setup and session selection on the platform	A single interest-and-availability form; expressing interest is effectively enrolling
Matching	Handled through the external platform	Handled internally
Scheduling	Family books sessions themselves	A recurring weekly session is set up on the family's behalf
Access	Log in to the external platform each week and navigate to the session	Open a standing calendar invitation and click an embedded video link
Session length	Target of about 90 minutes	Target of about 60 minutes

Notes: The tutoring itself was the same in both years: a volunteer working through the student's assigned Khan Academy practice, roughly once a week. The rows describe the operational steps required of families and the program between the offer and a sustained weekly session. The Year 1 external platform required each family to create and log into an account; the Year 2 process required no account, replacing platform login with a standing calendar invitation, remind and embedded video link.

sent. This is what makes the model inexpensive: the largest cost in a conventional tutoring program—recruiting, paying, and training skilled tutors—is removed when volunteers follow a platform that supplies the curriculum.

Two delivery regimes

The tutoring itself—a volunteer working an hour a week through a student's assigned Khan Academy practice—was the same across both years. What changed was how families were invited, how they enrolled, how they were matched to a tutor, and how sessions were scheduled. Year 1 ran through an external tutoring platform and asked families to complete several steps on their own. Year 2 handled enrollment, matching, and scheduling internally and collapsed those steps into one. Table 1 summarizes the contrast.

Each of these steps is a point at which a willing family can quietly drop out, so the Year 1 pipeline had many more places to lose a student than the Year 2 one; how much this mattered is a central question of Section 4. The redesign was not a clean experiment,

however: Year 2 differed from Year 1 in cohort, calendar, invitation wording, session length, and back-end operations all at once. We therefore treat the within-year randomization as our source of causal estimates and read the cross-year comparison as descriptive.

From offer to practice

Randomization assigns the *offer* of tutoring, but an offer raises achievement only if it survives a sequence of steps. Let Z_i denote assignment to the tutoring offer, D_i an indicator for whether student i attends at least one session, H_i the additional Khan Academy practice the offer generates, and Y_i end-of-year math achievement. The offer can raise Y_i only by first moving D_i and then H_i:

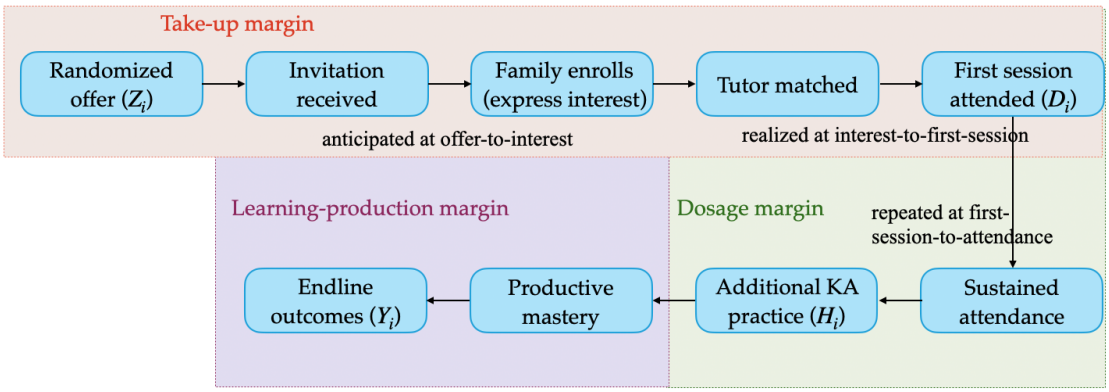

Figure 1: From offer to practice. The randomized offer Z can raise achievement Y only by first moving first-session receipt D and then additional Khan Academy practice H.

A family must notice the invitation and express interest; interest must convert into a first attended session; a first session must turn into sustained weekly participation; and that participation must generate enough productive practice to move learning. Each link can fail, and a failure anywhere breaks the chain: a compelling offer that no one acts on, or a first session that no one returns to, produces no learning no matter how effective the tutoring would have been. This framework organizes the analysis that follows. Section 4 works along the chain from offer to sustained practice, and Section 5 asks whether the practice that resulted moved achievement and attitudes.

Design, sample, and measurement

Within each participating classroom, and within each grade, nomination round, and year, teachers' nominated students were randomly assigned either to receive the tutoring offer or

not. We use the classroom × grade × nomination-round × year cell as the unit of randomization, absorb it with fixed effects, and cluster standard errors at that level. Randomization details specific to each wave are described in Appendix B. Table A1 in Appendix C reports sample sizes and shows that assigned and control students were balanced on baseline characteristics, with one exception—a small difference in prior-year math marks—that we note where it is relevant.

Our outcomes follow the offer-to-practice chain. The first stage is Khan Academy practice itself—the mechanism through which the offer could raise achievement. We measure it with weekly learning minutes and session attendance, drawn from platform activity logs and tutoring records, and treat it both as an outcome of the offer and as the input to everything downstream.

Our primary achievement outcome is an end-of-year math assessment built around the Khan Academy topics students practiced during the year. To make the test sensitive to that practice, we drew its items from the platform content tutored students worked on; because the relevant content was only known once the year's assignments were set, item selection took place after randomization. (Appendix M documents the procedure). The test was administered in class in the final weeks of the school year and was designed to take about thirty minutes, with the same form given to treatment and control students. Students worked through the problems with paper and pencil and then submitted their final answers on a computer through a Qualtrics form[1], which logged each student's responses. TWiK program staff scored the submitted responses with student names masked, so scorers did not know whose work they were grading; scoring was therefore blind to treatment status. This construction makes the assessment *favorable* to detecting a treatment effect: it is aligned to the practiced material, so if the offer had raised learning on these skills, the test should register it. We adjust scores relative to the control-group distribution and report effects in standard-deviation units.

Two further outcomes let us examine whether any gains extended beyond the practiced items. We obtain end-of-year mathematics report-card marks from administrative records, and, in both years, we field a short student survey on attitudes toward math, with identical items across the two years. Report-card marks, graded between 1 and 4, reflect a teacher's overall judgment of performance rather than a single test, and the survey captures how students felt about the subject; neither is a clean achievement measure, but both are informative, and neither is confined to the specific skills the assessment covers.

We estimate intent-to-treat effects of the offer throughout, and, where useful, scale them by first-session take-up to express effects per participating student. Appendix M describes

[1]Qualtrics is an online survey platform; see https://www.qualtrics.com/

each outcome and its construction in more detail.

4 From Offer to Practice: Take-Up and Participation

We now follow the offer along the chain of Figure 1: from the randomized offer, to a first session, to sustained weekly practice. Three facts organize the section. Take-up was low in Year 1 and rose sharply after we simplified the pipeline in Year 2. The practice the offer generated was real but light, and even among students who started, weekly participation stayed intermittent.

Take-up and the first stage

Table 2 reports the effect of assignment on attending a first tutoring session. Assignment raised the probability of attending at least one session 0.45 in Year 1 and 0.83 in Year 2. Because almost no control students attended, these coefficients closely track raw take-up among assigned students. The offer was free, endorsed by the student's own teacher, and tied to work the student was already assigned—and still, in Year 1, more than half of assigned students never attended a single session.

These figures condition on the teacher having delivered the invitation. Restoring the cells where no invitation reached the family lowers take-up to 0.41 in Year 1 and 0.79 in Year 2, so in Year 1 roughly one nominated student in ten never had the offer reach them at all. Teacher delivery is thus a zeroth stage in the pipeline; our main estimates condition on it and isolate the family-facing margins that follow.

Where the offer leaked

Reaching a first session requires two steps: an assigned family must express interest—respond to the email saying they want in—and that interest must convert into an attended session. Writing take-up T as interest I times conversion C, $T = I \times C$, locates which step lost students. Measured interest rose from 0.56 to 0.83 and conversion from 0.82 to near one, and an accounting decomposition places most of the Year 1 take-up shortfall at the expressed interest margin—before families completed any enrollment step (Table 3; the full decomposition is in Appendix J).

In Year 1, expressing interest and enrolling were distinct steps—an interest form, then account creation with a third party-platform, session selection, and a manual match—so conversion was a meaningful quantity below one. In Year 2, a single form sent from the teacher collected interest and availability and fed directly into matching, so an interested

Table 2: Effect of assignment on invitation response and first-session receipt

	Year 1	Year 2
Effect on received tutoring	0.453***	0.825***
	(0.024)	(0.043)
Effect on parent responded to offer	0.543***	0.832***
	(0.028)	(0.043)
Observations	1,096	502
Randomization-cell FE	Yes	Yes

Notes: Delivered-invitation sample. Each row is the coefficient on assignment from a regression of the listed outcome on assignment, with randomization-cell fixed effects and standard errors clustered at the cell. "Parent responded to offer" equals one if a family returned the interest/availability form in response to the invitation; it measures whether the family acted on the offer, not the strength of their underlying interest. "Received tutoring" equals one if the student attended at least one session. $^{*}p < .10$, $^{**}p < .05$, $^{***}p < .01$.

family was nearly an enrolled family *by construction*. Part of the measured rise in conversion is therefore this collapse of two steps into one, not a separately measured drop in the cost of a second step. Another part may have been due to greater trust in sign-up through a teacher-provided link (google forms) rather than through a third party site that was not part of the district.

Much of the Year 1 leakage occurred early, before families completed enrollment. Partway through Year 1, the steps that follow a family's yes—enrollment, matching, and scheduling—were simplified while the invitation stayed the same. Because this coincided with a tutor shortage pushing the other way, it cannot cleanly separate the conversion step from tutor supply, but it is revealing in one respect: in this period, families who responded but never reached a session were mostly waiting for an available tutor, not withdrawing after saying yes. And a test of whether Year 2's gains were concentrated among teachers who had run the program before—the natural "families trusted it more" story—finds no such pattern, if anything the reverse (Appendix L). The leak was real, but it sat partly at teacher delivery, partly at the invitation, partly in operations and tutor supply, and partly in how the two years defined the steps—not from lack of family enthusiasm.

What the offer bought: a light, supervised dose

More students reached a first session in Year 2. What did they do once there? When a tutored student attends, the session is substantial: about an hour of one-on-one work that registers as roughly 60 minutes of platform activity that week, against about 17 minutes for

Table 3: Take-up chain and decomposition of the Year-1 to Year-2 gain

Panel A: Take-up chain (assigned students)		
	Year 1	Year 2
Responded to offer (I)	0.560	0.826
Reached a session, given a response (C)	0.816	0.992
Received tutoring ($T = I \times C$)	0.457	0.819
Panel B: Decomposition of the take-up gain		
	Contribution	Share
Total gain (Year 2 − Year 1)	0.362	100.0%
Via response margin (I)	0.240	66.3%
	(0.045)	(5.8)%
Via conversion margin (C)	0.122	33.7%
	(0.019)	(5.8)%

Notes: Take-up decomposes as $T = I \times C$, where I is the share of assigned families who responded to the invitation and C is the share of those who then reached a first session. Panel B is a symmetric, order-averaged accounting decomposition of the Year-1-to-Year-2 change in take-up into these two margins. Parentheses report cluster-bootstrap standard errors (2,000 replications, resampling randomization cells).

a control student—some 40 minutes of additional practice in a session week. But sessions are intermittent and take-up is incomplete, so averaged across all offered students and all weeks the effect is smaller. Table 4 reports that assignment raised weekly learning minutes by about 8 in Year 1 and 14 in Year 2 (pooled 10), against control means near 19–20 minutes. Scaled by first-session take-up, the per-complier effect is about 18 minutes in Year 1 and 17 in Year 2—essentially flat across years. The larger Year 2 offer effect therefore reflects greater *reach*, not deeper practice. Year 2 reached more students without reducing captured practice per participant.

This practice is best read as a platform-practice dose to compare with the computer-assisted-learning literature, not as total tutoring time: the learning-minutes field records active Khan Academy time, not tutor explanation or off-platform work, so it undercounts what happens in a session. And most of what it does capture is the tutoring session itself, relocated onto the platform. In Year 2, where we observe session realization rate (share of scheduled tutoring sessions that were actually held) week by week, we can split the offer's effect on weekly minutes by whether the week contained a session (Table 5). In session weeks, offered students log about 60 minutes against a control mean of 17; in weeks without a session, about 21, an intent-to-treat effect on non-session weeks of +4.4 ($p < .01$). The offer's effect splits algebraically into a session part and a non-session part: about five-sixths

Table 4: Effect of assignment on weekly Khan Academy learning minutes

	Control mean [SD]	ITT (SE)	TOT (SE)
Year 1 (2023–24)	19.09 [18.37]	8.23*** (1.44)	17.70*** (2.95)
Year 2 (2024–25)	19.99 [20.10]	13.75*** (1.73)	16.71*** (1.78)
Pooled	19.50 [19.17]	10.24*** (1.15)	17.20*** (1.71)
Observations (pooled)		1,131	

Notes: Matched Khan Academy accounts. Weekly minutes are averaged over non-break instructional weeks in each cohort's post window. Each cell is a separate regression with randomization-cell fixed effects and standard errors clustered at the cell; the TOT (treatment-on-the-treated) column instruments first-session receipt with assignment. $^{*}p < .10$, $^{**}p < .05$, $^{***}p < .01$. Effects on levels completed, skills to proficiency, and skills worked on are similar and reported in Appendix H, along with an event study of minutes around each student's start.

(13.8 minutes, 85%) is relocated supervised session time, and only about one-sixth (2.4 minutes, 15%) is independent practice. The offer did raise self-directed study, through assigned between-session work and routine, but modestly.

Platform-engagement dashboards are the standard way to judge whether an online tool is working, but much of the engagement a tutoring offer generates is the supervised session showing up on the platform. The same metric measures supervised practice far more than it measures added independent study—and, as here, it can make a real weekly tutoring session look like a handful of extra minutes once averaged across a calendar of mostly quiet weeks.

Over the school year, this pattern accumulates to a modest platform total: about six hours per assigned student and about ten hours per first-session complier (Appendix H). The gap between an intensive session—a full hour of supervised instruction when a student attends—and this light cumulative platform dose is the central fact the achievement analysis must weigh.

Getting in the door is not staying in the room

Reaching a first session did not secure the weeks that followed. Tutoring is not a one-time decision: a student must come back each week, and each week is a fresh chance not to show up. Figure 2 shows the share of eligible student-weeks in which a tutoring session was held from February through June of Year 2.[2] Among students who received tutoring, a session is held in about two-fifths of eligible weeks; among all assigned students the rate runs near one-third. Attendance does not collapse over the term. It simply never rises.

The low average attendance rate could arise two ways. A committed minority might

[2]Systematic session tracking in Year 2 begins in February, so the analysis covers February through June 2025. Comparable session-level records for Year 1 are not available.

Table 5: Where the engagement effect occurs: relocated session time versus independent practice (Year 2, February–June)

	Weekly KA minutes
Control mean (all weeks)	17.1
Offered, weeks *with* a session	60.3
Offered, weeks *without* a session	20.6
ITT, weeks without a session (independent practice)	4.41***
	(1.23)
*Decomposition of the offer's weekly effect (*16.2 *min):*	
via session weeks (relocated supervised time)	13.8 (85%)
via non-session weeks (independent practice)	2.4 (15%)
Session-week share, offered student-weeks	32%
among receivers' post-start weeks	40%

Notes: Year-2 weekly minutes over February–June 2025, the period for which attendance is observed. Because sessions occur in only about a third of offered student-weeks (about 40% of started weeks), the 60-minute session-week figure averages down to the offer's roughly 16-minute weekly effect over this window, and to the smaller full-window effects in Table 4. The decomposition writes the offer's effect on average weekly minutes as the session-week share times its mean plus the complementary share times its mean (the law of total expectation), splitting it into relocated supervised time and independent practice. Standard errors clustered at the randomization cell. $^{*}p < .10$, $^{**}p < .05$, $^{***}p < .01$.

attend most weeks while everyone else quietly drops out; or most students might attend intermittently, missing weeks here and there. The data point to the second. Student-level session rates spread across a broad middle rather than piling up at zero: the median receiver attends about half of their eligible weeks, and fewer than one in ten attends none at all (Appendix G). Low attendance is pervasive rather than concentrated.

The missed weeks, in turn, are mostly not exits. A session is held in about 67% of weeks following a week with a session, and in about 61% of weeks following a missed week—a gap of under six percentage points. A student who misses one week is only slightly less likely to attend the next. Some students do leave: about one in ten is formally withdrawn, and roughly a third of those who ever attend have stopped by early June. But the larger share of missed weeks comes from students who are still in the program and return shortly afterward. What looks like attrition is mostly absence.

Part of the pattern is due to statutory holidays, the Easter weekend, and Professional Activity days, and roughly a quarter of missed weeks coincide with one. These weeks were not centrally cancelled—the session was left to each family to keep or skip—and none were rescheduled. A week lost this way was simply lost.

Platform practice tells the same story and is available in both years. Among tutored

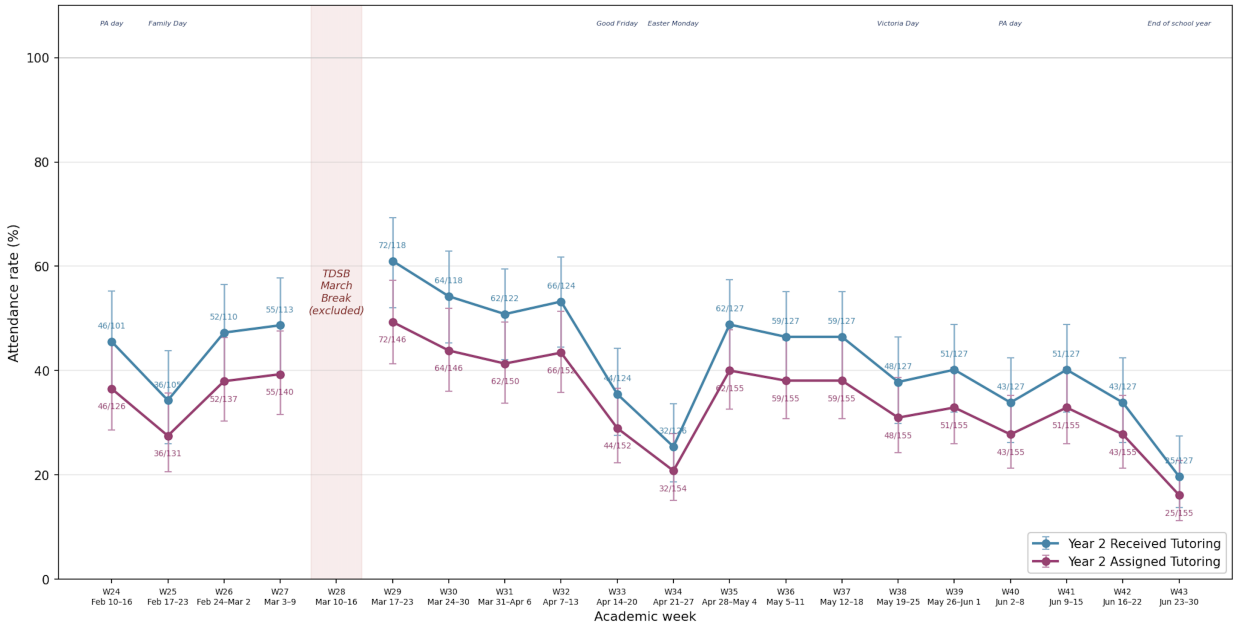

Figure 2: Weekly Year 2 tutoring session rates by calendar week (February to June 2025). The figure shows the share of eligible student-weeks in which a tutoring session was held, separately for the received cohort and the full assigned cohort. A student-week is a session week if a held tutoring session is recorded. Counts at each marker give sessions held over eligible student-weeks; eligibility varies as students enter the program. Vertical bars show 95% confidence intervals. Statutory holidays and Professional Activity days are noted in the figure. TDSB March Break is excluded. Appendix G decomposes these weeks by participation state.

students, some Khan Academy activity is recorded in 62% of Year 1 weeks and 64% of Year 2 weeks, but a substantive half-hour of practice—the kind the program was built around—in only 41% and 45% of weeks. The shortfall is not that most weeks are blank; it is that fewer than half clear a meaningful practice threshold.

Entry and retention are separate problems, and only the first was solved from recruitment changes. Simplifying the sign-up brought students in, but nothing in the redesign made the weekly session more likely to happen: sessions were not tracked as they occurred, a session that did not take place went unreported, holiday weeks were not rescheduled, and a silent miss prompted no follow-up. The session records show whether a session was held, not who failed to appear, so we cannot attribute any particular absence to the student, the family, or the tutor. What we can say is that the students missing those weeks have not given up—they miss, and they come back. That makes the weeks recoverable in principle, and it points to reminders, rescheduling, and follow-up as the natural place to look. The tutoring was recommended by the student's own teacher and tied to the work they were assigned in class, so what seems to be missing is not a reason to attend, but factors outside of school that held the session in place. How much reminders and follow-up would recover is an open

question.

5 From Practice to Learning: Achievement and Attitudes

Assignment moved students along the chain: more reached a first session, and they practiced more on Khan Academy. Did that translate into measured learning? This section reports three downstream outcomes—the end-of-year assessment, report-card marks, and student attitudes. The short answer is that the program looks promising where it reached students, but the achievement test is too imprecise, with the light dose delivered, to confirm a learning effect on its own. The clearer positive signals are in report cards and attitudes.

The achievement estimate is positive but imprecise

Our primary achievement outcome is the end-of-year distributed assessment. We built it to be sensitive to the material that at least some tutored students actually practiced (Appendix M documents the procedure). This makes the test favorable to detecting a treatment effect on practiced skill.

The estimated effects are imprecise. The pooled intent-to-treat effect is 0.055 SD, with a 95% confidence interval of about $[-0.09, 0.20]$ (Table 6). The point estimate is positive, and larger in Year 2 (0.081 on the full sixteen-item score) than in Year 1 (0.047), tracking that year's higher take-up—but none of the estimates is statistically distinguishable from zero. The interval is too wide to separate a null from the modest gains this dose would plausibly produce: the design has 80% power only for effects of about 0.21 SD or larger, several times what the realized dose would generate.

The offer raised cumulative Khan Academy practice by about six hours per assigned student and ten per participant over the school year. At that scale, the best real-world evidence on Khan Academy—Eames et al. (2026), using a panel of over 200,000 students and idiosyncratic classroom practice variation—implies a gain of roughly 0.03–0.06 SD (0.005–0.010 SD gains for each hour of total school year Khan Academy practice). Our confidence interval comfortably contains that range and lies far below the 0.21 SD the design could detect. Read against these benchmarks, we can rule out the 0.30–0.50 SD gains of intensive, professionally staffed, school-day tutoring, but not the small gains that Khan Academy use at this weekly dose would predict.

Table 6: Effects on end-of-year math achievement (standardized)

	ITT	95% CI	TOT
Year 1 (8-item)	0.047	$[-0.12, 0.22]$	0.096
	(0.087)		(0.177)
Year 2 (16-item)	0.081	$[-0.22, 0.38]$	0.100
	(0.154)		(0.190)
Pooled (primary)	0.055	$[-0.09, 0.20]$	0.094
	(0.074)		(0.126)
Minimum detectable effect (pooled, 80% power)		0.21	

Notes: Effects in within-form experimental-control SD units. The pooled primary specification combines Year 1 with the first eight Year 2 questions (the competence-selected items); the more treatment-favoring second Year 2 section is reported in Appendix M. Standard errors parentheses, clustered at the randomization cell; randomization-cell fixed effects throughout. The MDE is the effect detectable with 80% power at a two-sided 5% test. $^{*}p < .10$, $^{**}p < .05$, $^{***}p < .01$.

Table 7: Effects on end-of-year mathematics report-card marks (1–4 scale)

	Control mean	ITT (SE)	TOT (SE)
Year 1 (2023–24)	2.891	0.058 (0.047)	0.128 (0.102)
Year 2 (2024–25)	2.860	0.102 (0.073)	0.124 (0.089)
Pooled	2.880	0.070* (0.040)	0.125* (0.070)

Notes: Full-sample effects of the tutoring offer on end-of-year TDSB mathematics marks, computed by TDSB on restricted records and returned in aggregate. The ITT uses the same randomization-cell specification as the rest of the paper and is exact; the TOT (treatment-on-the-treated) column is an approximate per-complier effect, obtained by scaling the ITT by the first-session take-up rate rather than by a fitted instrumental-variables regression. Marks are teacher-assigned on the 1–4 provincial scale. $^{*}p < .10$, $^{**}p < .05$, $^{***}p < .01$.

Report-card marks show a small positive effect

TDSB end-of-year mathematics report-card marks, on the provincial 1–4 scale, offer a second read on performance—one that, unlike the short assessment, reflects a teacher's overall assessment of a student's math across the year (Table 7). Marks rise by about 0.07 points, or roughly 0.08 SD, in the pooled full sample (marginally significant, $p < .10$), with a per-complier estimate near 0.13 points. As with the assessment, the effect is larger in Year 2 than in Year 1. Teachers assign these marks, so they may reflect participation and effort alongside tested skill—we read the estimated treatment effect estimates on these outcomes as corroborating rather than decisive. But they are a broad, curriculum-wide measure of school-reported performance, and they move in the same direction as everything else.

Table 8: Effects on student attitudes (Year 2)

	Control mean	ITT (SE)	TOT (SE)
Math enjoyment (1–7)	3.538	0.573** (0.249)	0.693** (0.301)
KA helpfulness (1–7)	4.004	0.590*** (0.206)	0.709*** (0.248)
Weekly math homework time (1–6)	1.954	0.368** (0.160)	0.446** (0.195)

Notes: Year-2 endline survey. Randomization-cell fixed effects; standard errors clustered at the cell. All three effects survive a Benjamini–Hochberg correction across the three outcomes at the 5% level. Year-1 estimates (Appendix M) are small and insignificant. $^{*}p < .10$, $^{**}p < .05$, $^{***}p < .01$.

Students liked math more

The clearest downstream effects are on how students felt about math. The endline survey was completed on the same occasion as the main assessment – in class, in the final weeks of the school year—by treatment and control students alike – so it captures how students felt after a year of the program. In Year 2, the offer raised math enjoyment by 0.57 points on a seven-point scale, perceived helpfulness of Khan Academy by 0.59, and self-reported weekly math homework time by 0.37 points—all statistically significant (Table 8). These are self-reported, and students knew they had been offered tutoring, so part of the effect may be the good feeling of working with a tutor rather than a change in skill. In Year 1, when fewer students were actually tutored, the same three effects estimates were null. The point estimates therefore suggest that attitudes improved in-line with higher take-up—the same pattern that runs through the engagement and achievement estimates.

6 Discussion

A free, teacher-endorsed tutoring offer reached fewer than half of assigned students in Year 1; redesigning the offer raised that to about 83%. But moving more students to a first session did not, on its own, deliver learning. The value of tracing the program as a chain—from offer, to a first session, to weekly practice—is that it shows where the promise of a low-cost home tutoring model is realized and where it leaks away.

Read across the two years, the program traces a moving bottleneck. In Year 1 the binding constraint was take-up, and simplifying the offer largely removed it. But once most students were reaching a first session, the constraints past entry came into view: weekly sessions were held in only about two-fifths, most of the added practice was the tutoring session itself rather than new independent study, and the cumulative dose stayed light—about ten hours of platform practice per participant, well below the level at which either our own design or the external Khan Academy benchmarks would expect a visible achievement gain given

our sample size. Solving the sign-up bottleneck revealed another one in attendance. The achievement result is best read not as "the program does not work" but as "with take-up solved, dose and sustained practice are the next margins to move."

On its own, the achievement estimate looks like a null—0.055 SD, an interval that includes zero. But it is a noisy null and it does not sit alone. Report-card marks rise modestly, attitudes improve, and across all three outcomes the estimates are larger in Year 2 than in Year 1, growing as the program reached more students. No one of these is decisive, but they lean the same way, and they strengthen where the program did more. On balance, taken together, there seems to be something there that this design was too noisy to pin down.

Two lessons generalize beyond this program. The first is that in a voluntary, home-based program, participation is not one problem but several, and each must be secured separately. A randomized offer is not a treatment: it has to survive teacher delivery, a family's response, enrollment, a first session, and then a return every week after. These are distinct margins, and they fail for distinct reasons. Getting families in the door is largely a matter of *reducing friction*—a clear invitation, few steps, a standing appointment—the kind of small design choices that shape take-up across many public programs (Lavecchia et al., 2016; Oreopoulos, 2020; Herd and Moynihan, 2018).

But reducing friction, we found, buys entry and not retention. Once a student has started, keeping them coming back each week is a different problem, and a simpler sign-up does nothing for it. What the Year-2 session records show is that the students who miss weeks are mostly not students who have quit: session rates spread across a broad middle rather than splitting into regulars and dropouts, and a missed week barely predicts another. A quarter of missed weeks fall on holidays and professional-activity days that were never rescheduled. That points to reminders, rescheduling, and follow-up on sessions that silently fail to happen as the natural things to try. The tutoring was recommended by the student's own teacher and tied to the work they were assigned in class, so the missing ingredient does not seem to be a reason to attend. How much reminders and follow-up would recover is an open question, but the missed weeks look recoverable in principle, which is a more encouraging position than a program whose students have simply left.

The second lesson is that a low-cost hybrid can move students into supervised practice, but supervised practice at home, on a voluntary basis, tends to be light and irregular—and it is dose, not the design of any single session, that appears to bind. The tutoring followed a real mastery routine; the shortfall was not in what happened inside a session but in how few sessions and how little sustained practice a home-based offer produced. Whether that practice, delivered at a heavier and steadier dose, would raise learning is the question this design leaves open.

Several limitations bound these conclusions. The setting is a single large urban district over two years, and the sample is teacher-nominated Grade 4–8 students. Because the control group also received the coach-supported classroom Khan Academy program, the experiment identifies the marginal effect of adding tutoring, not the effect of the platform program. The cross-year redesign is not randomized, and the response and conversion margins are not defined identically across the two years. The achievement assessment is short, platform-aligned, and partly selected on treatment-influenced performance; the report-card marks are teacher-assigned and not blind; and the attitude results are secondary and self-reported.

This study shows a home-based offer can move students into sessions and raise platform practice; it cannot show which features of tutoring turn that practice into durable learning. A stronger design would raise the dose, use a validated common assessment, measure cost, and record instructional time more completely—linking platform timestamps, attendance, tutor logs, and session video—so that platform practice, tutor explanation, and independent study can be told apart.

An alternative is to move TWIK into the classroom, where attendance and practice time could be protected rather than left to home routines. An in-class version keeps the core of what we studied—a student working through platform math while a tutor watches for where they struggle—but sets it in a place where the practice block is scheduled and support can be targeted in real time. That shifts the question from whether families take up a home offer to whether supervised, well-targeted practice produces learning, and it makes the quality of tutoring easier to study: with students present and practice time fixed, one can vary the timing of help and the pacing toward mastery. The catch is the greater difficulty in scheduling tutors to meet and how to handle multiple conversations in the classroom at the same time.

In offer-based education programs, invitation, conversion, retention, and learning production are separate design problems, and a program is only as strong as its weakest link. What this study shows is that a home-based tutoring offer built on a classroom platform can clear the first links cheaply and at scale—families can be brought in, and students can be moved into supervised practice, at close to volunteer cost. The last link is where the model is still unproven: the practice it generates is light and irregular, and at that dose neither our assessment nor the platform benchmarks can say whether it becomes durable learning.

References

Bettinger, Eric P., Bridget Terry Long, Philip Oreopoulos, and Lisa Sanbonmatsu, "The Role of Application Assistance and Information in College Decisions: Results from the H&R Block FAFSA Experiment," *Quarterly Journal of Economics*, 2012, *127* (3), 1205–1242.

Carlana, Michela and Eliana La Ferrara, "Apart but Connected: Online Tutoring, Cognitive Outcomes, and Soft Skills," *American Economic Review*, 2025, *115* (10), 3487–3513.

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* (1), e2507708123.

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.

Guryan, Jonathan, Jens Ludwig, Monica P. Bhatt, Philip J. Cook, Jonathan M. V. Davis, Kenneth A. 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.

Herd, Pamela and Donald P. Moynihan, *Administrative Burden: Policymaking by Other Means*, Russell Sage Foundation, 2018.

Kraft, Matthew A., Beth E. Schueler, and Grace T. Falken, "What Impacts Should We Expect from Tutoring at Scale? Exploring Meta-Analytic Generalizability," *Review of Educational Research*, 2026. Advance online publication.

_ , John A. List, Jeffrey A. Livingston, and Sally Sadoff, "Online Tutoring by College Volunteers: Experimental Evidence from a Pilot Program," *AEA Papers and Proceedings*, 2022, *112*, 614–618.

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.

Lee, David S., "Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects," *Review of Economic Studies*, 2009, *76* (3), 1071–1102.

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), 74–107.

Oreopoulos, Philip, "Nudging and Shoving Students Toward Success," *Education Next*, 2020, *20* (2), 60–67.

_ **, Chloe 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.

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

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. Advance online publication.

_ **, Cynthia Pollard, Sarah Novicoff, Sara White, and Susanna Loeb**, "The Effects of Virtual Tutoring on Young Readers: Results from a Randomized Controlled Trial," *Educational Evaluation and Policy Analysis*, 2024. Advance online publication.

Shapley, Lloyd S., "A Value for n-Person Games," in Harold W. Kuhn and Albert W. Tucker, eds., *Contributions to the Theory of Games II*, Princeton: Princeton University Press, 1953, pp. 307–317.

Shorrocks, Anthony F., "Decomposition Procedures for Distributional Analysis: A Unified Framework Based on the Shapley Value," *The Journal of Economic Inequality*, 2013, *11* (1), 99–126.

Appendix

A Implementation Details

This appendix documents the operational details summarized in Section 3. The main text focuses on the offer-to-practice margins used in the empirical analysis. The material here records how teachers used Khan Academy, how KWiK program coaches—referred to as "Khoaches"—supported classroom adoption, and how the tutoring layer was added to the classroom program, and how the implementation chain differed between the third-party platform and internally managed regimes.

A.1 Classroom Khan Academy support and the khoach model

The Khoaching WIth Khan Academy program supported classroom Khan Academy adoption before the tutoring offer was added. Participating teachers were paired with an instructional assistant, called a "khoach," who helped teachers align Khan Academy assignments with the class's current mathematics unit. At the beginning of the year, teachers shared information with their khoaches about grade level, curriculum coverage, instructional pacing, and classroom constraints. Khoaches used this information to help set up the platform and identify appropriate Khan Academy content.

The support model followed a weekly cycle. Each week, khoaches sent suggested Khan Academy assignments for the following instructional week. Teachers reviewed the recommendations and selected the assignments they wished to use. Once approved, khoaches helped publish the assignments so that students could begin working on them at the start of the subsequent week and continue through the end of that week. This weekly process reduced the teacher's cost of finding relevant platform content while preserving teacher discretion over classroom implementation.

Khan Academy provided immediate feedback to students as they worked through problems. On the back end, KWiK obtained platform data on student activity and performance, including assignment attempts, correctness, and progress through skills. These data were processed into weekly progress reports for teachers. The reports summarized engagement and performance measures such as the share of assignments attempted, the share passed, the number of attempts, and the total number of assigned tasks. The reports were intended to give teachers timely information about student progress and to help identify students who might benefit from additional support.

Teachers were encouraged to allocate dedicated classroom time for Khan Academy practice, for example by using one class period per week for platform-based learning. Actual

implementation varied across classrooms. Some teachers used Khan Academy primarily during class time, while others assigned the platform work as homework. This variation is part of the real-world implementation setting in which the tutoring offer was evaluated. Because both treated and control students were exposed to this khoach-supported classroom program, the randomized contrast isolates the incremental effect of the tutoring layer.

A.2 Student nomination and the tutoring layer

The tutoring layer was added on top of classroom Khan Academy adoption. Teachers were encouraged to nominate students who appeared to be struggling with the assigned platform material. Guidance emphasized students who made repeated attempts but had not yet reached a mastery benchmark, though nomination ultimately remained at the teacher's discretion. Some teachers prioritized contemporaneous Khan Academy performance, while others incorporated prior achievement, classroom knowledge, or other information.

Students nominated by teachers formed the experimental sample. Nomination occurred separately by implementation round. Within classroom-grade-round-year cells, nominated students were randomly assigned either to receive a tutoring offer or to remain in the control condition. Students assigned to the offer were invited to receive free weekly online tutoring over Zoom. University volunteers served as tutors and worked with students on the Khan Academy material already assigned in class. Tutoring was voluntary, and families could discontinue participation at any time.

The research team did not directly contact parents or guardians. Teachers served as the intermediary between TWiK and families: after randomization, teachers received information about which nominated students had been assigned to the offer and distributed the invitation to those families. Moving from assignment to a first session therefore required several steps: the teacher had to send the invitation, the family had to express interest, the family and tutor had to complete enrollment or availability steps, program staff had to form a match, and the student had to attend a first session.

A.3 Year 1 Schoolhouse-mediated pipeline

In Year 1, the tutoring was delivered through Schoolhouse.world, a free online tutoring platform that handled tutor accounts, family enrollment, matching, and session hosting. Families and tutors interacted with the program through this external platform rather than directly with TWiK.

Figure A1 documents the resulting Year 1 workflow. Time flows from left to right. Rows identify the main actors: parent, student, Schoolhouse, TWiK, and tutor. Solid arrows show

within-actor steps, and cross-lane arrows show handoffs across actors.

Under the Schoolhouse-mediated regime, families first received the tutoring invitation from the teacher and expressed interest. Families then received Schoolhouse-specific instructions. To attend tutoring, they had to create an account on Schoolhouse.world using the student's TDSB school email, select or confirm a session, locate the scheduled session on the platform, and access Zoom through Schoolhouse. Weekly session access also required returning to Schoolhouse.world, locating the session, and clicking through to Zoom. Because expressing interest and completing enrollment were distinct steps in this regime, conversion conditional on interest is a meaningful sub-one quantity in Year 1.

Tutors faced a parallel platform-mediated process. University volunteers expressed interest, created and validated a Schoolhouse account using a university email, waited for manual approval, completed training and police-check requirements, and listed their availability on Schoolhouse. Matching required both the family and the tutor to clear their respective platform steps. Schoolhouse generated match confirmations; TWiK then pulled match information and sent separate confirmation messages.

The Year 1 chainlink therefore involved several one-time entry steps and several recurring access steps. On the family side, the external platform introduced learning costs associated with understanding Schoolhouse and compliance costs associated with account creation, session selection, and weekly navigation. On the tutor side, external account creation, manual approval, training, police-check, and availability-listing steps could delay tutor availability and slow matching.

A.4 Year 2 internally managed pipeline

Figure A2 documents the internally managed pipeline used after the late-Year 1 transition and throughout Year 2. The conceptual stages were the same as in Year 1: invitation, interest expression, matching, first session, and repeated weekly participation. The operational steps were substantially simplified.

Families completed a TWiK interest form (A Google Form linked in the teacher's email to a parent) that collected availability. TWiK then matched the family with a tutor internally and sent a confirmation message with session dates and a recurring calendar invitation. To attend the first session, the student or parent opened the calendar invitation and clicked the embedded Zoom link. The same calendar invitation was used in subsequent weeks. Families no longer needed to create a Schoolhouse account, log into an external platform, or navigate through a platform session page. Because a single form collected interest and availability and fed directly into internal matching, an interested family was effectively an enrolled family in

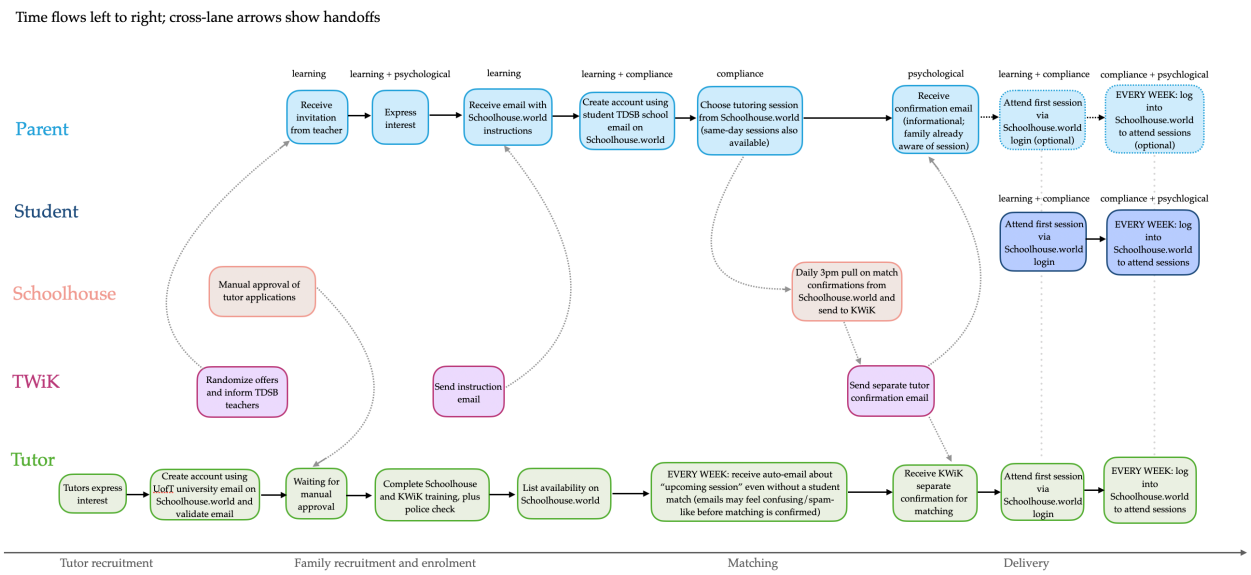

Figure A1: Year 1 Schoolhouse-mediated implementation pipeline. Time flows from left to right; rows identify actors and cross-lane arrows indicate handoffs. The figure documents the external-platform steps required for family enrollment, tutor onboarding, matching, first-session access, and weekly access.

this regime, which is why conversion conditional on interest approaches one in Year 2 and why the cross-regime split of the take-up gain into interest and conversion margins is partly definitional.

Tutors also moved through an internal process. Volunteers signed up through a TWiK form, reported their availability, completed TWiK-managed training and police-check requirements, and then became available for internal matching. Once matched, tutors received the same recurring calendar invitation as families. The internal regime therefore reduced the procedural depth of enrollment and matching, although it did not eliminate the need for students and tutors to show up each week.

A.5 Operational transition timeline

The timing of the operational transition matters for interpreting the within-Year 1 comparison in Section 4. Year 1 began under the Schoolhouse regime. Round 1 randomization began on October 4, 2023; invitations began on October 6; and the first Schoolhouse-delivered sessions began on October 17. The transition unfolded across three distinct stages. First, the *randomization and invitation* boundary occurred on November 22, when the last Schoolhouse-era randomizations occurred; students randomized after that date entered the internal pipeline. Second, the *family-interest* boundary occurred on November 26, when Schoolhouse stopped accepting new family interest forms; families who expressed interest af-

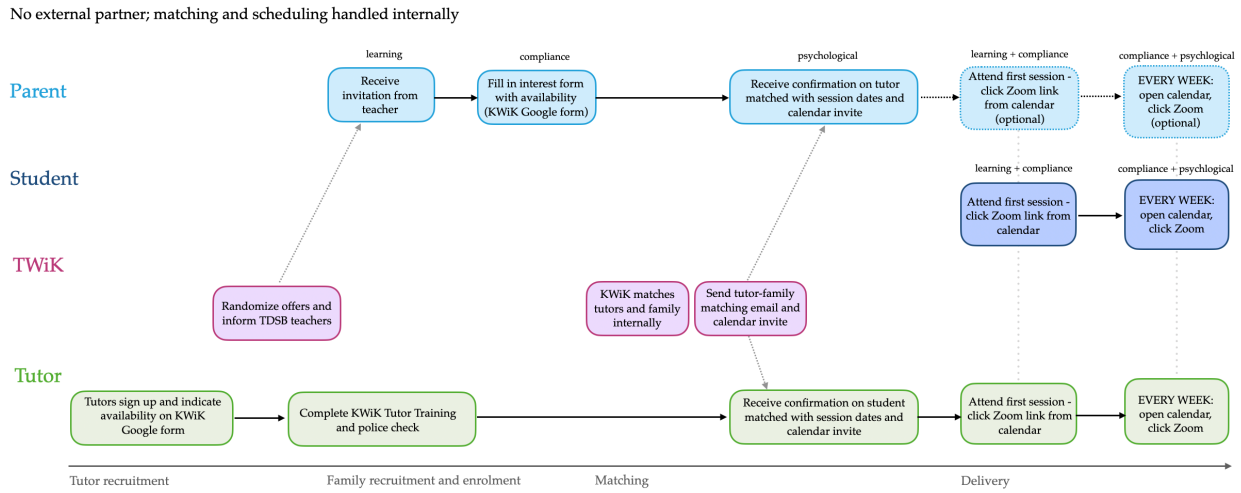

Figure A2: Internally managed implementation pipeline used after the late-Year 1 transition and throughout Year 2. Schoolhouse is removed from the workflow: families and tutors submit TWiK forms, TWiK performs matching internally, and weekly sessions are accessed through standing calendar invitations with embedded Zoom links.

ter that date were onboarded internally even if their classroom had been randomized earlier. Third, the *session-delivery* boundary occurred in early December: transition emails were sent to tutors and families on December 1 and December 3, the last Schoolhouse-delivered tutoring session occurred on December 5, and subsequent sessions ran internally over Zoom. Round 2 randomization, beginning February 11, 2024, took place entirely under the internal regime.

These staggered cutoffs explain the operational categories used in Appendix K. Some families expressed interest during the Schoolhouse period and reached a first session through Schoolhouse. Others expressed interest during the Schoolhouse period but were effectively rescued by the internal process, attending only after the December switch. A third group expressed interest during the Schoolhouse period but never reached a first session. Families expressing interest after November 26 entered the internal process directly. Thus the within-Year 1 transition is useful for documenting where families and tutors were in the pipeline, but it is not a clean experiment in which only one implementation feature changed.

Year 2 used the internal regime from the start. Round 1 randomization began on October 20, 2024, invitations began on October 21, and the first tutoring session began on November 24. Round 2 randomization began on January 16, 2025.

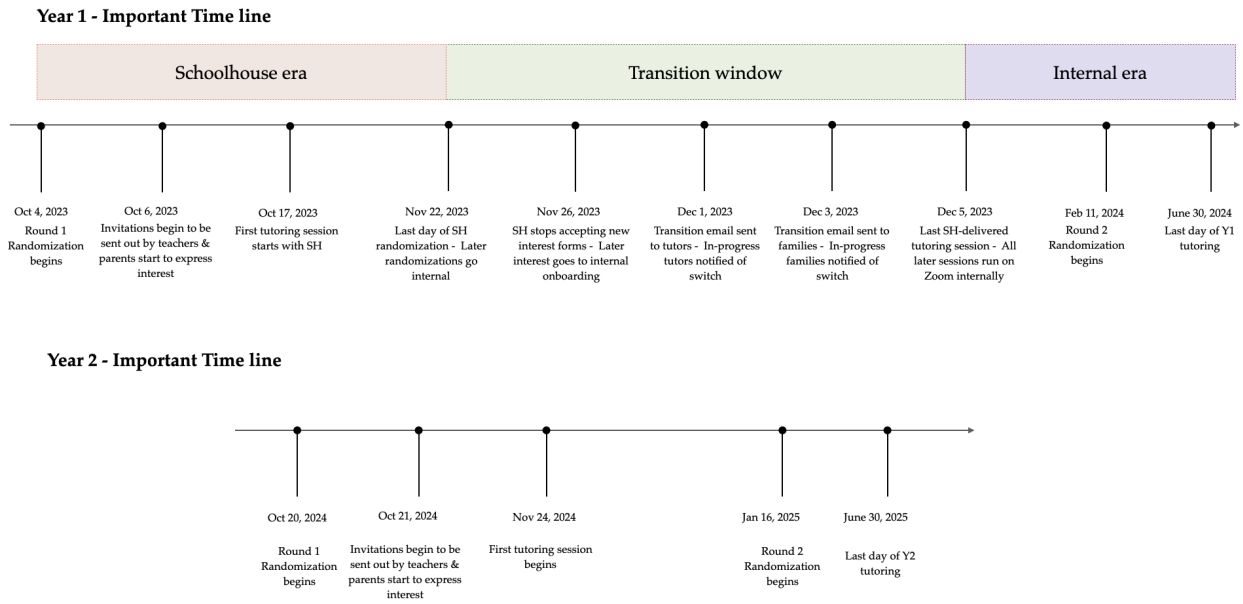

Figure A3: Operational timing in Year 1 and Year 2. Year 1 begins under the Schoolhouse regime, transitions in late November and early December 2023, and then runs internally. Year 2 uses the internal regime from the start.

B Randomization Details

This appendix documents the wave-specific randomization rules. The main text uses the classroom × grade × nomination round × year cell as the fixed-effect absorber and clustering unit. The details here explain how treatment shares were implemented within the operational constraints of each wave.

B.1 Nomination and eligible sample

Teachers nominated students whom they believed were likely to benefit from additional instructional support. Guidance emphasized students who had made repeated attempts on Khan Academy assignments but had not yet reached a mastery benchmark, although teachers retained discretion over nominations. Treatment assignment was restricted to nominated students. The nominated students form the experimental sample.

A teacher could contribute more than one classroom to the experiment, and randomization was carried out separately by classroom and nomination round. In the analysis, randomization cells are further defined by grade because grade-specific curricula and assignments may differ within a classroom. Cells with no within-cell variation after fixed effects are absorbed contribute no identifying information to fixed-effect treatment-control comparisons and are mechanically dropped from those regressions.

B.2 Year 1, Round 1

Year 1 Round 1 was implemented while TWiK partnered with Schoolhouse. The design included three arms: a control group, a group-tutoring arm, and a one-on-one tutoring arm. The intended classroom-level shares were 50 percent control, 15 percent group tutoring, and 35 percent one-on-one tutoring. In the main analysis, the two tutoring arms are pooled and coded as assigned to the tutoring offer. This pooling reflects the paper's focus on the offer-to-practice pipeline: both arms required families to receive the invitation, express interest, complete enrollment, and reach a session. The paper does not use this arm distinction to draw separate conclusions about group versus one-on-one tutoring.

When classroom cell sizes did not yield exact integer allocations, residual students were assigned using the protocol used at the time of randomization. If a residual student was assigned to treatment, the student was further allocated between group tutoring and one-on-one tutoring using the intended relative probabilities for the two tutoring modalities.

B.3 Year 1, Round 2

By Year 1 Round 2, the Schoolhouse partnership had ended and the program was operated internally. The design was simplified to two arms: control and treatment. For most classrooms with even numbers of nominated students, assignments were split evenly between treatment and control. When the number of nominated students was odd, the residual student was assigned to control. When the number of nominated students exceeded tutoring capacity, treatment assignment was capped within the classroom and remaining nominated students were assigned to control.

B.4 Year 2

Both Year 2 rounds were delivered internally. The design retained two arms, with an intended allocation of roughly 65 percent control and 35 percent treatment. In classrooms with two nominated students, one student was assigned to control and one to treatment. In larger classrooms, assignments followed the intended 65/35 rule, with integer rounding handled by assigning residual students to control. All random assignments were generated using fixed random seeds.

B.5 Cleaning and analysis assignment

The main analysis uses the cleaned offer assignment after excluding classroom-grade-round cells in which no invitation was delivered and after applying documented corrections to the

assignment file. The sample construction table reports both the initial randomized roster and the final analysis sample. All ITT, first-stage, TOT, engagement, achievement, and attitude regressions use the cleaned analysis assignment. The cleaning procedure preserves the intention-to-treat logic within the set of cells where an offer was actually delivered.

B.6 Inference

Identification uses within-cell variation. Standard errors are clustered at the cell level, matching the level of assignment and allowing arbitrary within-cell correlation in residuals.

C Baseline Balance

This appendix reports baseline balance for the experimental sample. For privacy reasons, TDSB did not release the restricted student-level administrative records to the research team. Instead, the research team supplied TDSB with the experimental student roster, which was assembled from teacher-provided nomination lists and reconciled with the student identifiers available in the downloaded Khan Academy platform records. TDSB then linked this roster to its restricted demographic, attendance, EQAO, and report-card files and returned aggregate SPSS output.

Some student identifiers were incomplete or inconsistent across the teacher rosters, Khan Academy accounts, and TDSB administrative systems. In addition, certain administrative measures are available only for particular grades, cohorts, or school years. Consequently, not every student is observed for every baseline measure, and the non-missing sample size varies across variables. We therefore report the non-missing N separately for every variable and analysis sample.

For each variable, we also report p_{miss}, the two-sided p-value from a pooled two-proportion z-test of whether the missing-data rate differs between the randomly assigned offer and control groups. The denominators are the treatment and control counts in the corresponding TDSB-linked Has-Endline or Full-Sample roster, and the numerators are the numbers of students without a non-missing value for the variable. We refer to this as our assignment-based missing-at-random check. Across the reported baseline variables, we find no detectable differential missingness by randomized assignment. This result means that the available TDSB records pass our test for equal missingness rates across the randomized arms.

Prior Khan Academy engagement is available in the research-team analysis file and is reported separately below. EQAO Grade 3 mathematics is a provincial standardized assessment administered at the end of Grade 3. We include it only where the administration

precedes the treatment year, which restricts it to the Grade 4 and Grade 5 cohorts who sat the Grade 3 test one to two years before the study.

We report two kinds of balance evidence. The first table uses TDSB-returned aggregate regression output for demographic, attendance, EQAO, and prior report-card variables. These variables are useful for checking randomization, but the underlying student-level records are not released to the research team. The second table uses the student-level baseline measures available in our analysis file: prior Khan Academy engagement.

The tables report control-group means and treatment-control differences. Estimates are shown for the delivered-invitation analysis roster and for the subsample with an end-of-year assessment. All specifications include randomization-cell fixed effects and cluster standard errors at the randomization-cell level. The large majority of covariates—sex, special-education and English-learner status, immigrant status, attendance, and prior report-card marks in most subjects—are balanced. Two baseline mathematics measures are not: treated students score modestly higher on EQAO Grade 3 mathematics (pooled +0.131, $p < .05$; Year 1 +0.148, $p < .10$) and on prior-year mathematics report-card marks in Year 2 (+0.137, $p < .10$; Table A1). Among the roughly fourteen covariates tested across three panels and two samples, a handful of significant cells is consistent with chance; but because both imbalances run in the *same* direction—treated students slightly stronger in baseline mathematics—any resulting bias in the achievement estimate would run upward. Because the achievement instrument is in any case too imprecise, and too selected, to test learning in either direction, we do not lean on this; we treat the imbalances only as a further reason to read the achievement estimates cautiously rather than as evidence of randomization failure, and note that controlling for prior Khan Academy engagement leaves the achievement estimate essentially unchanged.

Table A1: Student-Level Balance Tests, by panel

	Has Endline		**Full Sample**	
Variable	Control Mean	T-C Diff. (SE)	Control Mean	T-C Diff. (SE)
Panel A: Year 1 (2023–2024)				
Female	0.466	0.035	0.474	-0.003
	[0.500]	(0.034)	[0.500]	(0.032)
	[$N = 722$; $p_{\text{miss}} = 0.830$]		[$N = 962$; $p_{\text{miss}} = 0.521$]	
Age (years)	10.494	-0.041	10.567	-0.012
	[1.173]	(0.043)	[1.221]	(0.039)
	[$N = 723$; $p_{\text{miss}} = 0.725$]		[$N = 963$; $p_{\text{miss}} = 0.467$]	
Grade	5.872	—	5.978	—
	[1.041]	(—)	[1.101]	(—)
	[$N = 413$]		[$N = 591$]	
Special Education	0.279	-0.009	0.283	0.003
	[0.449]	(0.029)	[0.451]	(0.027)
	[$N = 723$; $p_{\text{miss}} = 0.725$]		[$N = 963$; $p_{\text{miss}} = 0.467$]	

	Has Endline		Full Sample	
	Control Mean	T-C Diff.	Control Mean	T-C Diff.
Immigrant	0.256	0.004	0.246	0.022
	[0.437]	(0.030)	[0.431]	(0.027)
	[$N = 723$; $p_{\text{miss}} = 0.725$]		[$N = 963$; $p_{\text{miss}} = 0.467$]	
Previous Year Absence Rate, pct. pts.	10.154	-0.147	11.070	-0.327
	[7.427]	(0.510)	[8.751]	(0.552)
	[$N = 684$; $p_{\text{miss}} = 0.497$]		[$N = 905$; $p_{\text{miss}} = 0.265$]	
ELL/ESL	0.119	-0.024	0.119	-0.005
	[0.324]	(0.025)	[0.324]	(0.022)
	[$N = 723$; $p_{\text{miss}} = 0.725$]		[$N = 963$; $p_{\text{miss}} = 0.467$]	
EQAO Gr. 3 Math	2.493	0.148*	2.476	0.138*
	[0.792]	(0.085)	[0.793]	(0.077)
	[$N = 255$; $p_{\text{miss}} = 0.751$]		[$N = 320$; $p_{\text{miss}} = 0.846$]	
EQAO Gr. 3 Reading	2.855	0.014	2.880	-0.029
	[0.647]	(0.067)	[0.647]	(0.057)
	[$N = 208$; $p_{\text{miss}} = 0.873$]		[$N = 268$; $p_{\text{miss}} = 0.822$]	
EQAO Gr. 3 Writing	2.627	0.026	2.624	-0.007
	[0.539]	(0.086)	[0.542]	(0.073)
	[$N = 207$; $p_{\text{miss}} = 0.941$]		[$N = 266$; $p_{\text{miss}} = 0.829$]	
Previous Year Report Card Math	3.101	-0.022	3.081	-0.047
	[0.789]	(0.051)	[0.788]	(0.048)
	[$N = 674$; $p_{\text{miss}} = 0.612$]		[$N = 886$; $p_{\text{miss}} = 0.209$]	
Previous Year Report Card Reading	3.125	0.038	3.124	-0.015
	[0.716]	(0.052)	[0.709]	(0.049)
	[$N = 676$; $p_{\text{miss}} = 0.619$]		[$N = 889$; $p_{\text{miss}} = 0.186$]	
Previous Year Report Card Writing	2.974	0.056	2.974	-0.001
	[0.727]	(0.046)	[0.722]	(0.044)
	[$N = 676$; $p_{\text{miss}} = 0.476$]		[$N = 888$; $p_{\text{miss}} = 0.120$]	
Previous Year Report Card Science	3.194	0.062	3.176	0.033
	[0.705]	(0.048)	[0.718]	(0.050)
	[$N = 672$; $p_{\text{miss}} = 0.605$]		[$N = 882$; $p_{\text{miss}} = 0.116$]	
Missing Endline	—	—	0.301	-0.008
			[0.459]	(0.022)
			[$N = 1,103$]	
Panel B: Year 2 (2024–2025)				
Female	0.520	-0.096	0.505	-0.072
	[0.501]	(0.076)	[0.501]	(0.060)
	[$N = 327$; $p_{\text{miss}} = 0.482$]		[$N = 425$; $p_{\text{miss}} = 0.409$]	
Age (years)	10.551	-0.028	10.580	-0.028
	[1.285]	(0.064)	[1.297]	(0.054)
	[$N = 327$; $p_{\text{miss}} = 0.482$]		[$N = 425$; $p_{\text{miss}} = 0.409$]	
Grade	5.946	—	6.017	—
	[1.196]	(—)	[1.237]	(—)
	[$N = 259$]		[$N = 354$]	
Special Education	0.289	-0.013	0.321	-0.020
	[0.454]	(0.053)	[0.468]	(0.048)
	[$N = 327$; $p_{\text{miss}} = 0.482$]		[$N = 425$; $p_{\text{miss}} = 0.409$]	
Immigrant	0.240	0.034	0.232	0.021
	[0.428]	(0.051)	[0.423]	(0.043)
	[$N = 327$; $p_{\text{miss}} = 0.482$]		[$N = 425$; $p_{\text{miss}} = 0.409$]	

	Has Endline		Full Sample	
	Control Mean	T-C Diff.	Control Mean	T-C Diff.
Previous Year Absence Rate, pct. pts.	9.265	-1.044	10.196	-1.098
	[7.215]	(0.694)	[8.245]	(0.789)
	[$N = 323$; $p_{\text{miss}} = 0.672$]		[$N = 421$; $p_{\text{miss}} = 0.543$]	
ELL/ESL	0.084	-0.004	0.082	-0.008
	[0.279]	(0.027)	[0.275]	(0.026)
	[$N = 327$; $p_{\text{miss}} = 0.482$]		[$N = 425$; $p_{\text{miss}} = 0.409$]	
EQAO Gr. 3 Math	2.363	0.011	2.321	0.120
	[0.820]	(0.127)	[0.814]	(0.107)
	[$N = 175$; $p_{\text{miss}} = 0.576$]		[$N = 229$; $p_{\text{miss}} = 0.911$]	
EQAO Gr. 3 Reading	2.785	-0.102	2.789	-0.063
	[0.687]	(0.093)	[0.717]	(0.080)
	[$N = 151$; $p_{\text{miss}} = 0.622$]		[$N = 185$; $p_{\text{miss}} = 0.854$]	
EQAO Gr. 3 Writing	2.542	-0.057	2.555	0.004
	[0.648]	(0.121)	[0.625]	(0.105)
	[$N = 151$; $p_{\text{miss}} = 0.622$]		[$N = 185$; $p_{\text{miss}} = 0.854$]	
Previous Year Report Card Math	2.922	0.116	2.886	0.137*
	[0.842]	(0.090)	[0.863]	(0.077)
	[$N = 319$; $p_{\text{miss}} = 0.263$]		[$N = 413$; $p_{\text{miss}} = 0.151$]	
Previous Year Report Card Language	3.032	0.005	2.967	0.072
	[0.726]	(0.071)	[0.769]	(0.062)
	[$N = 320$; $p_{\text{miss}} = 0.303$]		[$N = 418$; $p_{\text{miss}} = 0.279$]	
Previous Year Report Card Science	3.186	-0.013	3.136	0.033
	[0.694]	(0.082)	[0.734]	(0.070)
	[$N = 318$; $p_{\text{miss}} = 0.227$]		[$N = 413$; $p_{\text{miss}} = 0.151$]	
Missing Endline	—	—	0.268	-0.009
			[0.444]	(0.035)
			[$N = 508$]	
Panel C: Pooled (Year 1 + Year 2)				
Female	0.486	-0.001	0.485	-0.022
	[0.500]	(0.033)	[0.500]	(0.029)
	[$N = 1,049$; $p_{\text{miss}} = 0.442$]		[$N = 1,387$; $p_{\text{miss}} = 0.815$]	
Age (years)	10.515	-0.037	10.572	-0.017
	[1.214]	(0.036)	[1.248]	(0.032)
	[$N = 1,050$; $p_{\text{miss}} = 0.495$]		[$N = 1,388$; $p_{\text{miss}} = 0.862$]	
Grade	5.900	—	5.993	—
	[1.103]	(—)	[1.153]	(—)
	[$N = 672$]		[$N = 945$]	
Special Education	0.283	-0.010	0.296	-0.003
	[0.451]	(0.026)	[0.457]	(0.024)
	[$N = 1,050$; $p_{\text{miss}} = 0.495$]		[$N = 1,388$; $p_{\text{miss}} = 0.862$]	
Immigrant	0.250	0.012	0.241	0.021
	[0.433]	(0.026)	[0.428]	(0.023)
	[$N = 1,050$; $p_{\text{miss}} = 0.495$]		[$N = 1,388$; $p_{\text{miss}} = 0.862$]	
Previous Year Absence Rate, pct. pts.	9.818	-0.403	10.745	-0.544
	[7.354]	(0.416)	[8.572]	(0.457)
	[$N = 1,007$; $p_{\text{miss}} = 0.851$]		[$N = 1,326$; $p_{\text{miss}} = 0.492$]	
ELL/ESL	0.106	-0.019	0.106	-0.006
	[0.308]	(0.019)	[0.308]	(0.017)
	[$N = 1,050$; $p_{\text{miss}} = 0.495$]		[$N = 1,388$; $p_{\text{miss}} = 0.862$]	

	Has Endline		Full Sample	
	Control Mean	T-C Diff.	Control Mean	T-C Diff.
EQAO Gr. 3 Math	2.430	0.101	2.401	0.131**
	[0.807]	(0.072)	[0.806]	(0.063)
	[$N = 430$; $p_{\text{miss}} = 0.452$]		[$N = 549$; $p_{\text{miss}} = 0.457$]	
EQAO Gr. 3 Reading	2.820	-0.028	2.837	-0.041
	[0.667]	(0.056)	[0.681]	(0.047)
	[$N = 359$; $p_{\text{miss}} = 0.389$]		[$N = 453$; $p_{\text{miss}} = 0.631$]	
EQAO Gr. 3 Writing	2.585	-0.004	2.591	-0.003
	[0.596]	(0.071)	[0.583]	(0.060)
	[$N = 358$; $p_{\text{miss}} = 0.348$]		[$N = 451$; $p_{\text{miss}} = 0.616$]	
Previous Year Report Card Math	3.034	0.018	3.009	0.006
	[0.814]	(0.045)	[0.821]	(0.041)
	[$N = 993$; $p_{\text{miss}} = 0.755$]		[$N = 1,299$; $p_{\text{miss}} = 0.699$]	
Previous Year Report Card Science	3.191	0.040	3.161	0.033
	[0.701]	(0.042)	[0.723]	(0.041)
	[$N = 990$; $p_{\text{miss}} = 0.725$]		[$N = 1,295$; $p_{\text{miss}} = 0.494$]	
Missing Endline	—	—	0.289	-0.008
			[0.453]	(0.019)
			[$N = 1,611$]	

Notes. Pre-treatment baseline balance (TDSB administrative records, aggregate SPSS output). Control mean over treatment–control difference; [N; p_{miss}] gives the non-missing N and the differential-missingness p-value (missing-at-random check). * $p < .10$, ** $p < .05$, *** $p < .01$.

We also report balance on prior Khan Academy engagement, averaged over the pre-intervention period.

Table A2: Baseline balance: prior Khan Academy engagement

	Has Endline		Full Sample	
Variable	Control Mean [SD]	T–C Diff. (SE)	Control Mean [SD]	T–C Diff. (SE)
Panel A: Year 1 (2023–2024)				
Avg. Learning Minutes (Prior)	11.654	0.675	10.249	1.155
	[19.170]	(0.746)	[19.634]	(0.838)
Avg. Levels Up (Prior)	1.204	0.039	1.058	0.197
	[2.092]	(0.098)	[2.276]	(0.130)
Avg. Proficiency (Prior)	0.893	-0.012	0.789	0.145
	[1.782]	(0.087)	[1.924]	(0.117)
Avg. Worked On (Prior)	1.445	0.037	1.266	0.201
	[2.335]	(0.105)	[2.562]	(0.135)
Observations (students)	774		1103	
Panel B: Year 2 (2024–2025)				
Avg. Learning Minutes (Prior)	30.030	-0.691	27.328	0.845
	[26.339]	(1.894)	[25.243]	(1.494)
Avg. Levels Up (Prior)	2.684	-0.257	2.456	-0.130
	[2.341]	(0.223)	[2.353]	(0.171)
Avg. Proficiency (Prior)	1.810	-0.103	1.681	-0.017
	[1.870]	(0.190)	[1.865]	(0.142)
Avg. Worked On (Prior)	3.351	-0.309	3.044	-0.146
	[2.659]	(0.261)	[2.670]	(0.200)
Observations (students)	375		510	
Panel C: Pooled (Year 1 + Year 2)				
Avg. Learning Minutes (Prior)	18.292	0.300	16.188	1.074
	[23.717]	(0.748)	[23.208]	(0.729)
Avg. Levels Up (Prior)	1.739	-0.042	1.544	0.111
	[2.297]	(0.096)	[2.396]	(0.107)
Avg. Proficiency (Prior)	1.224	-0.037	1.099	0.103
	[1.865]	(0.082)	[1.950]	(0.094)
Avg. Worked On (Prior)	2.133	-0.058	1.884	0.110
	[2.620]	(0.106)	[2.733]	(0.114)
Observations (students)	1149		1613	
Randomization Group FE		✓		✓
SE Clustered at Rand. Group		✓		✓

Notes: Each row reports control-group means (SD in brackets) and treatment–control differences (SE in parentheses) from regressions of baseline Khan Academy engagement on assignment. The *Has Endline* columns restrict to students with an end-of-year assessment; the *Full Sample* columns include all nominated students. Prior measures are averaged over the pre-intervention weeks. All specifications include randomization-cell fixed effects (classroom × grade × round × year), SE clustered at the cell. Demographic, EQAO, and prior report-card balance variables are reported separately using TDSB-returned aggregate output. * $p < .10$, ** $p < .05$, *** $p < .01$.

D Outcome Observability and Analysis Samples

The main text uses one experimental roster but outcome-specific observation rules. Table A3 reports the number of students observed for each outcome family (Panel A) and, in Panel B,

regresses an indicator for being observed on assignment. Assessment and survey observability do not differ detectably by assignment (except for Khan Academy record matching does differ in Year 2: assigned students are approximately 4.5 percentage points more likely to have a matched record. Appendix E therefore examines alternative codings and Lee bounds.) Missingness is not differential by arm largely, and the treatment–control comparison within each outcome sample is not biased by selective attrition—though, as the counts make clear, the observed sample need not represent the full delivered roster, and endline, engagement, and attitude estimates are not all estimated on the same set of students.

Table A3: Outcome observability and analysis samples

	Year 1	Year 2	Pooled
Panel A: Observed students, count (share of delivered roster)			
Delivered-invitation roster	1,103 (100.0%)	510 (100.0%)	1,613 (100.0%)
Endline assessment score	774 (70.2%)	375 (73.5%)	1,149 (71.2%)
Matched Khan Academy record	683 (61.9%)	448 (87.8%)	1,131 (70.1%)
Endline score and matched KA record	511 (46.3%)	337 (66.1%)	848 (52.6%)
Math enjoyment response	737 (66.8%)	354 (69.4%)	1,091 (67.6%)
Khan Academy helpfulness response	736 (66.7%)	348 (68.2%)	1,084 (67.2%)
Homework-time response	737 (66.8%)	347 (68.0%)	1,084 (67.2%)
Panel B: Assignment differential in observability			
Endline assessment score	0.009	0.012	0.010
	(0.022)	(0.034)	(0.019)
Matched Khan Academy record	-0.010	0.045	0.006
	(0.014)	(0.015)	(0.011)
Endline score and matched KA record	0.003	0.032	0.011
	(0.019)	(0.033)	(0.017)
Math enjoyment response	0.009	0.003	0.007
	(0.024)	(0.036)	(0.020)
Khan Academy helpfulness response	0.007	0.010	0.008
	(0.024)	(0.035)	(0.020)
Homework-time response	0.009	0.012	0.010
	(0.024)	(0.034)	(0.020)

Notes. The denominator is the delivered-invitation analysis roster used for assignment-control comparisons. *Panel A* reports outcome availability by outcome family. Matched Khan Academy records are students with any captured post-window platform activity; they define the main engagement sample. Attitude response rows treat raw zeros as item non-response because the documented response scales begin at one. *Panel B* reports coefficients from regressions of each observation indicator on assignment, with randomization-cell fixed effects and standard errors clustered at the cell. These diagnostics assess differential observability by assignment; they do not imply that observed students are representative of the full delivered roster.

E Observability of Khan Academy Engagement

Some records were deleted from the Khan Academy platform before the research team completed data collection and therefore could not be recovered. In addition, some platform records could not be linked reliably to the experimental roster because they lacked consistent student identifiers. Consequently, outcome availability differs across students, and not every outcome is observed for every student on the experimental roster.

Engagement estimates use the matched sample (students with any captured post-window Khan Academy activity), because in Year 1 a large share of accounts record zero activity all year, which we treat as incomplete data capture rather than true non-use. Conditioning on being matched conditions on a post-randomization outcome, so this appendix decomposes the missingness and shows the engagement effect is robust to how the zeros are handled. Table A4 collects the evidence.

The missingness has two layers. Panel A shows that the all-zero accounts are overwhelmingly a *cell-level* phenomenon. In Year 1, 72% of the zero accounts sit in cells we call *fully unobserved*: no nominated student in the cell—treated or control—has any recorded Khan Academy activity all year, which reflects class-level loss in account linkage and extraction rather than student behavior. A quarter of Year-1 cells (32 of 130) are fully unobserved, and 57 of 130 (43.8%) are already unobserved in the pre-treatment window, before any student could have acted on the offer. This is a fixed property of the cell, not a treatment response: all 32 post-period unobserved cells are among the pre-period unobserved set, and about 23% of students in unobserved cells have a recorded tutoring session despite no captured platform activity—enrolled, attending students whose platform use was simply not captured. Cell-level loss of this kind removes whole cells from the observed sample but, because randomization is within cell, cannot select treatment relative to control inside any surviving cell. The share falls sharply in Year 2 (11% of cells unobserved), consistent with the account-linkage improvements between the two years.

Is observability moved by the offer? The smaller, individual-level layer—a student with no captured activity inside an otherwise-observed cell—is the one that could bias a conditional estimate, because it could in principle be a treatment response. Panel B tests this directly with an intent-to-treat regression of the match indicator on assignment, with cell fixed effects. In Year 1, the high-unmatched year, the offer has no effect on observability, whether estimated on all delivered cells (-0.010), within observed cells (-0.013), or within cells observable in the pre-period (-0.011), all insignificant. In Year 2 a small positive effect appears ($+0.045$ to $+0.051$ across the three samples, all $p < .01$). Because an un-

Table A4: Khan Academy observability and robustness of the engagement effect

Panel A: Where the all-zero (unmatched) accounts are

	Year 1	Year 2
All-zero / unmatched accounts	420 (38.1%)	62 (12.2%)
in fully-unobserved cells	304 (72.4%)	38 (61.3%)
in partially-observed cells	116 (27.6%)	24 (38.7%)
Fully-unobserved cells (of all)	32/130 (24.6%)	9/84 (10.7%)
Cells unobserved in the pre-period (of all)	57/130 (43.8%)	14/84 (16.7%)

Panel B: Is being matched affected by the offer? ITT on the match indicator

	Year 1	Year 2
All delivered cells	-0.010	0.045***
	(0.014)	(0.015)
Within observed cells	-0.013	0.049***
	(0.018)	(0.016)
Within pre-observed cells	-0.011	0.051***
	(0.018)	(0.017)

Panel C: Weekly learning-minutes ITT under three codings of the zeros

	Year 1	Year 2	Pooled
(1) Matched accounts *(main)*	8.234***	13.752***	10.235***
	(1.443)	(1.733)	(1.148)
(2) Delivered roster, zeros = 0 *(naive)*	4.767***	13.208***	7.151***
	(1.044)	(1.680)	(0.968)
(3) Cell-observed (pre), zeros = 0	6.752***	14.495***	9.722***
	(1.598)	(1.802)	(1.300)
Lee (2009) bounds, Year 2		[9.77, 14.86]	
lower-bound 95% CI (cluster bootstrap)		[5.63, 13.90]	

Notes. Delivered-roster sample. *Panel A*: a randomization cell (classroom × grade × round × year) is *fully unobserved* if no nominated student in it—treated or control—records any captured Khan Academy activity; such cells indicate class-level data loss in linkage and extraction, not behavior. *Panel B*: intent-to-treat regressions of the match indicator on assignment, with randomization-cell fixed effects and SE clustered at the cell, on all delivered cells, on cells with any post-window activity, and on cells with any *pre-period* activity (a pre-determined retention set). *Panel C*: weekly learning-minutes ITT (cell FE, cell-clustered SE) under (1) the matched sample used in the main text; (2) the delivered roster coding all-zero/unmatched accounts as zero (mechanically attenuated in Year 1 because entire unobserved cells enter as zeros); and (3) the cell-observed specification, which drops fully-unobserved cells but codes the remaining within-cell zeros as zero, using the pre-period observability flag so the retained set is not selected on post-treatment activity. Lee (2009) bounds trim the over-represented (assigned) arm by the differential matching share; the interval excludes zero. * $p < .10$, ** $p < .05$, *** $p < .01$.

matched account predominantly reflects an observability or account-linkage gap rather than verified inactivity, we read this cautiously: it is consistent with either a genuine extensive-margin response—the offer bringing a few otherwise-dormant students onto the platform—or a treatment-correlated improvement in account linkage, since enrolled students are somewhat more likely to have a matched, scrapeable account. That it survives within *pre-observed* cells makes a pure scrape-timing artifact unlikely, but we do not lean on this margin: it is small, and the matched intensive-margin estimate is robust to coding all unmatched accounts as zero and to the Lee bounds, both of which still exclude zero.

The engagement effect is robust to every coding of the zeros. Panel C reports the weekly-learning-minutes ITT under three treatments of the all-zero accounts: the matched sample used in the main text; the delivered roster with all zeros coded zero (the most conservative coding, which mechanically attenuates Year 1 because entire unobserved cells enter as zeros); and the cell-observed specification, which drops fully-unobserved cells but codes the remaining within-cell zeros as zero, using the *pre-period* observability flag so the retained set is pre-determined and not selected on post-treatment activity. The estimate is positive and significant under all three in both years and pooled. As expected, the naive delivered-roster Year 1 estimate (+4.8) is attenuated by the unobserved cells and rises to +6.8 once they are removed, partway back toward the matched estimate (+8.2); Year 2, which has few unobserved cells, is stable near +13–14.5 throughout. Finally, Lee (2009) bounds on the Year-2 estimate—which trim the over-represented assigned arm by the differential matching share to bound the worst case of individual selection—lie in $[9.8, 14.9]$ weekly minutes and exclude zero. Because the bounds estimator reports only i.i.d. standard errors, we cluster-bootstrap the bounds over randomization cells (500 replications); the lower bound's cluster-robust 95% confidence interval, $[5.6, 13.9]$, remains bounded away from zero. The matched-sample restriction therefore improves cross-year comparability of levels without manufacturing the engagement effect.

F Year-2 Weekly Sessions and the Session-vs-Practice Decomposition

This appendix documents the attendance measure and the decomposition behind Table 5.

Held-session measure. In Year 2, tutoring sessions were delivered over Zoom and automatically recorded. For the period in which recordings are available—February 3 to June 30, 2025, excluding the TDSB March break—we treat a recording as evidence that a session was

held. Each recorded session is dated and attributed to a student, and students are linked to the Khan Academy panel through their platform account identifier. A student-week is then a *session week* if a held session is recorded for that student that week. Eligibility is dynamic: a student enters the attendance denominator only from the week of their own first session, so the eligible pool grows through the term as more students begin. On this basis the weekly rate of attending a session, among eligible (already-started) receivers, is about 40%; pooled over all offered student-weeks—which also include students who never started and weeks before late-starters began—the session-week share is about 32%.

Window. Because held sessions are observed from February–June, the relocation analysis uses that window, a subset of the main engagement post-window (November–June; Table 4). February–June is the active tutoring stretch, so the offer's effect on weekly Khan Academy minutes over this window (about 16.5 on a cell-fixed-effect basis) is larger than the full-window Year-2 estimate of about 13.8; the two are not comparable levels.

Decomposition. Let lm denote weekly Khan Academy learning minutes, s the session-week indicator, P_s the share of offered student-weeks that are session weeks, and m_s, m_{ns}, c the offered session-week mean, offered non-session-week mean, and control mean. Control students never have sessions, so $\mathbb{E}[\text{lm} \mid \text{control}] = c$. By the law of total expectation, $\mathbb{E}[\text{lm} \mid \text{offered}] = P_s\, m_s + (1 - P_s)\, m_{ns}$, so the offer's effect on average weekly minutes splits exactly into a session-week and a non-session-week part:

$$\underbrace{\mathbb{E}[\text{lm} \mid \text{offered}] - c}_{\text{offer's weekly effect}} = \underbrace{P_s\,(m_s - c)}_{\text{relocated session time}} + \underbrace{(1 - P_s)\,(m_{ns} - c)}_{\text{independent practice}}. \tag{1}$$

Evaluating equation (1) with the Year-2 means—$m_s = 60.3$, $m_{ns} = 20.6$, $c = 17.1$, and a session-week share of about 0.32—gives a session-week contribution of 13.6 minutes (85% of the 16.2-minute effect) and a non-session-week contribution of 2.4 minutes (15%). This raw decomposition is exact—the two contributions sum to the 16.2-minute average offered-minus-control gap—and the cell-fixed-effect regression estimate of the same effect, 16.5 minutes, is slightly larger but divides in the same 85/15 proportion. The non-session term is the part of the engagement effect that is genuine independent practice; the corresponding cell-fixed-effect, cluster-robust intent-to-treat effect on minutes in weeks without a session is +4.4 (SE 1.2, $p < .01$). The decomposition conditions on the realized, post-randomization attendance pattern and is therefore descriptive accounting rather than a separate causal estimate; the headline engagement and achievement estimates do not depend on it. A companion analysis of the recordings' instructional content and tutor quality, and of the Year-1 sessions, is

reserved for separate work.

G Weekly Participation in Year 2

This appendix decomposes the Year-2 weekly session rate reported in Section 4. The sample is the 127 Year-2 students who received tutoring (attended at least one session), consistent with the 83% take-up rate among the 155 assigned students in the Year-2 analysis sample. The observation window runs from February 3 to June 30, 2025, the period for which session records are available, excluding the TDSB March break. A student's eligible weeks run from the week of their own first session through the end of the window.

The distribution of session rates. Student-level session rates are broadly distributed rather than concentrated at zero. Panel (a) of Figure A5 shows that 29.1% of receivers hold sessions in 50–75% of their eligible weeks and a further 24.4% in 25–50%. The median receiver's session rate is 47.4% and the mean is 42.8%; only 9.4% of receivers hold no session at all during the window.[3] The low average is therefore not driven by a small group of students with no sessions; most student–tutor pairs meet in some, but not most, of their eligible weeks.

Participation states. Figure A4 decomposes each week of the assigned cohort into one of seven mutually exclusive participation states. Two margins contribute to the low average. The first is discontinuation: 14 of the 127 receivers are formally withdrawn during Year 2, and a further group has no session for an extended period and never returns within the window, so that by early June roughly a third of students who ever attended have stopped. The second is intermittent absence—weeks without a session among pairs that resume afterward—which is the larger single component of missed weeks. The figure distinguishes inferred withdrawal (four or more consecutive weeks with no later return) from weeks that are simply truncated by the end of the observation window, so that short gaps near the end of the term are not misread as exits.

Are missed weeks shocks or spells? To ask whether missed weeks cluster into withdrawal episodes or arrive as isolated shocks, we compute the probability that a session is held in week $t + 1$ conditional on the state of week t, restricted to weeks within each stu-

[3] The student-level mean (42.8%) and the pooled student-week rate reported in the main text (about two-fifths) are the same phenomenon measured with different weights: the former averages across students, the latter across student-weeks.

dent's active window.[4] A session is held in 66.7% of weeks following a session-held week and in 60.9% of weeks following a missed week: a persistence gap of 5.8 percentage points. This gap is small relative to the dispersion in session rates across students in Panel (a), so between-student differences in engagement dominate within-student week-to-week dependence. Missed weeks are better read as moderately persistent one-week shocks than as the onset of multi-week withdrawal: a student who misses a week usually returns, though a miss does modestly raise the chance of another.

The calendar. The deepest weekly troughs in Figure 2 coincide with Family Day, the Easter weekend, Victoria Day, and Professional Activity days—days on which students are home but school staff are working. These weeks were not centrally cancelled; only the Christmas and March breaks were program-wide cancellations. Roughly 28% of missed weeks among active pairs fall on a statutory holiday or PA day, and none were rescheduled.

Grade. Panel (b) of Figure A5 shows session rates by grade among receivers. Grades 4 through 7 have similar medians, near 50%. Grade 8 is a clear outlier, with a median near 10%. One reading is a terminal-year effect: Grade 8 is the final elementary year in Ontario, and families facing the September transition to secondary school may deprioritize elementary tutoring in the spring or leave attendance to the student. That the drop appears only at the terminal grade, rather than rising gradually with age, is more consistent with the transition than with age itself. We report this as a description rather than a finding: Grade 8 receivers are the oldest students as well as the terminal-year students, the two explanations cannot be separated here, and the grade-specific cells are small.

[4]The active window runs from a student's first observed session to their last observed session. Every consecutive-week transition is treated as one observation and the pooled ratio is computed at the end. Transitions spanning the excluded March-break week are dropped, so all transitions correspond to one calendar week of separation.

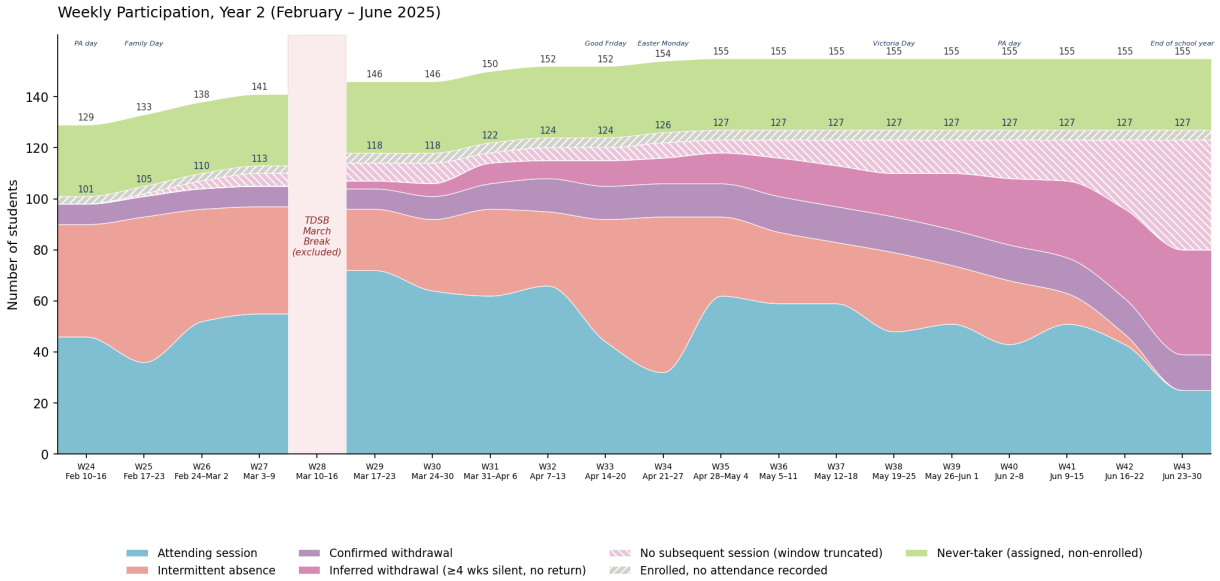

Figure A4: Year 2 weekly session-status composition of the assigned cohort (February–June 2025). Each week's assigned cohort is decomposed into seven mutually exclusive participation states. *Attending session*: a tutoring session is held that week. *Intermittent absence*: no session that week, but at least one later session. *No subsequent session (window truncated)*: no later observed session, but fewer than four weeks remain in the window, so withdrawal cannot be inferred. *Inferred withdrawal*: no session for at least four consecutive weeks, no later return, and no formal withdrawal record. *Confirmed withdrawal*: the student was formally withdrawn or unmatched from their tutor. *Enrolled, no session held*: students in the received cohort with no session observed during the window. *Never-taker*: assigned students who never entered the received cohort.

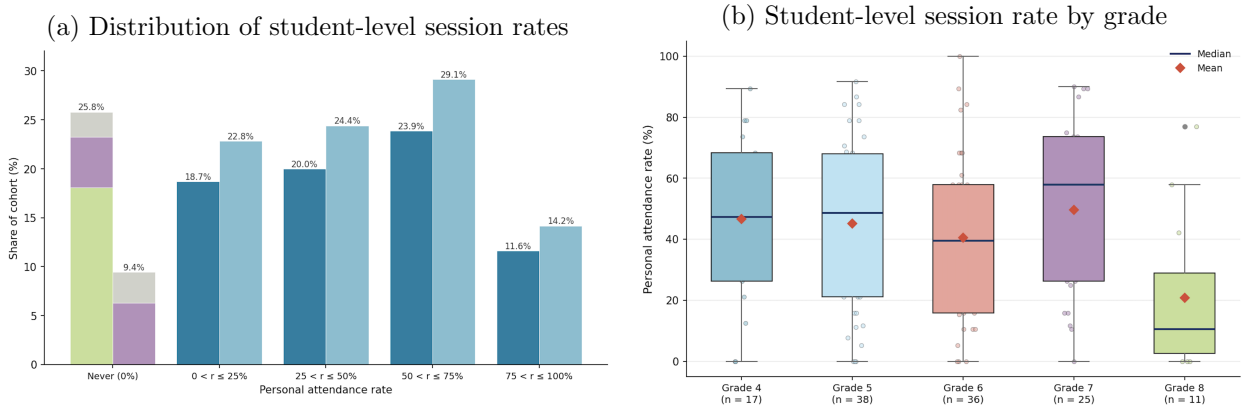

Figure A5: Year 2 weekly participation (February–June 2025). For each student, the session rate is the share of eligible weeks in which a tutoring session was held, where eligible weeks run from the student's first observed session through the end of the window, excluding the TDSB March break. Panel (a) shows the distribution for the assigned cohort (left bar) and the received cohort (right bar). Panel (b) shows the distribution among receivers by grade; boxes give the interquartile range, horizontal lines the median, diamonds the mean, and dots individual students.

H Repeated Weekly Participation

This appendix reports the cross-year platform-practice measure referenced in Section 4. Among tutored students with matched Khan Academy records, Table A5 gives the share of instructional weeks reaching each activity threshold. The main text uses the half-hour (30-minute) threshold as a substantive practice-week bar; the table shows the pattern is similar across thresholds. Because tutoring-session attendance is observed only in Year 2, this platform measure is the consistent cross-year participation proxy.

Table A5: Weekly Khan Academy participation among tutored students, by activity threshold

Weekly activity threshold	Year 1	Year 2	Pooled
Any activity (> 0 min)	0.619	0.642	0.628
≥ 10 minutes	0.557	0.589	0.570
≥ 20 minutes	0.481	0.530	0.501
≥ 30 minutes *(prior threshold)*	0.407	0.450	0.425
Student-weeks	5400	3744	9144

Notes. Each cell is the share of tutored-student post-window instructional weeks in which Khan Academy learning minutes reached the stated threshold (TDSB winter/March-break weeks excluded). Year-specific columns are primary because Year 1 advertised 90-minute tutoring sessions and Year 2 advertised 60-minute sessions; the pooled column is descriptive. The thresholds are common Khan Academy platform-practice thresholds, not fractions of scheduled tutoring time. The weekly participation rate falls steeply as the bar rises, so the table reports the full threshold menu; the repeated-participation figure uses the ≥ 30-minute cutoff as a substantive practice-week bar, and continuous post-window minutes remain the primary dosage measure. Tutoring-session attendance is recorded only in Year 2 (external Schoolhouse delivery in Year 1), so this KA-platform measure is the consistent cross-year participation proxy.

Cumulative dosage in hours. Converting weekly learning minutes into cumulative hours summarizes the platform dose generated by the offer before turning to downstream outcomes. Over the implemented instructional post windows, the offer raised cumulative Khan Academy practice by roughly 6 hours on an intent-to-treat basis and by about 10 hours per first-session complier, with a larger Year 2 ITT gain (+7.3 hours) than Year 1 (+4.9). Year 2 therefore generates a larger assigned-student cumulative platform dose because it reaches more students, while the per-first-session-complier platform dose is not larger. Like every learning-minutes measure in the paper, these hours exclude accounts that record no activity all year, which almost always reflect data-capture gaps rather than genuine non-use (Appendix E). These hours are cumulative Khan Academy practice hours, not total tutoring hours. They may understate total instructional exposure if tutors spend time explaining off-platform, and they may overstate independent practice if the platform work occurs dur-

ing tutoring. The achievement section therefore uses them as the observed platform dosage benchmark, not as a complete measure of tutoring dosage.

Table A6: Cumulative Khan Academy practice over instructional post weeks (total learning hours, matched accounts)

	Control mean [SD]	ITT (SE)	TOT (SE)
Year 1 (2023–24)	11.453	4.940***	10.620***
	[11.024]	(0.866)	(1.768)
Observations	683		
Year 2 (2024–25)	10.662	7.334***	8.911***
	[10.719]	(0.924)	(0.950)
Observations	448		
Pooled	11.092	5.808***	9.762***
	[10.885]	(0.657)	(0.996)
Observations	1131		

Notes. Cumulative Khan Academy learning hours over instructional post weeks (total instructional-week learning minutes ÷60); TDSB winter/March-break weeks are excluded from outcomes. Matched accounts only: students with any captured post-window platform activity; the match flag may use break-week activity, but the hours reported here use non-break instructional weeks only. All-zero and all-missing accounts are excluded as scraping gaps rather than true zeros, consistently with every other learning-minutes measure in the paper. The Observations row gives the number of matched students. Each pair of rows is two regressions with randomization-cell fixed effects and SE clustered at the cell; the TOT column instruments first-session receipt with assignment. Because the offered session length and post-window length differ across years, year-specific rows are the primary dose estimates and pooled rows are descriptive summaries across implemented program-years. * $p < .10$, ** $p < .05$, *** $p < .01$.

I Tutoring Session Protocol

The tutoring sessions followed a structured, mastery-oriented protocol rather than open-ended help. Year 1 sessions were advertised as 90 minutes: the first 5–10 minutes for a check-in or icebreaker and follow-up on independent practice, the middle 70–80 minutes for supporting the student's progress on Khan Academy assignments, and the last 5 minutes for setting expectations for independent practice. Year 2 sessions were advertised as 60 minutes, with the same structure compressed into a roughly 5–50–5 schedule. Tutors worked the student's assigned Khan Academy material rather than a separate curriculum. During the middle block, students were expected to use their own Khan Academy account to watch assigned videos and complete corresponding exercises; these actions can register as active learning minutes in the platform reports. Tutor explanations, whiteboard work, and other

off-platform support do not register in the learning-minutes measure. The pedagogy emphasized process over outcome: tutors withheld answers, gave hints tied to the assigned video rather than to the specific item, and had a student redo an exercise—even one already passed at the platform's 70% mastery threshold—if the student had leaned on hints to complete it, with a working target of roughly two skills brought to mastery per hour.

Two features of this protocol matter for the analysis. First, because the middle block of each session uses Khan Academy, an attended session can generate logged learning minutes in the week it occurs. These are active platform minutes, not all scheduled minutes and not total tutoring time: the platform excludes navigation, tutor talk, whiteboard work, and other off-platform instruction. Section 4 uses this distinction, and the Year-2 session-recording decomposition, to show that the measured engagement gain is mostly session-linked platform practice—about five-sixths relocated supervised time and only about one-sixth added independent practice—so it should not be read as a direct measure of added self-directed study. Second, the protocol is a genuine mastery model, so the imprecise achievement evidence in Section 5 is not naturally read as a failure of session design—the binding constraints are the dose and reach of supervised practice, not the structure within a session.

J Decomposition of the Take-Up Gain

Let ι denote interest expression and c conversion conditional on interest, with subscripts indexing Year 1 and Year 2. Because take-up is $t = \iota \times c$, the cross-year change $t_2 - t_1$ is split by averaging over the two orderings of the two factors. The offer-to-interest contribution is

$$\Delta_\iota = \tfrac{1}{2}\left[(\iota_2 - \iota_1)c_1 + (\iota_2 - \iota_1)c_2\right], \tag{2}$$

and the interest-to-first-session contribution is

$$\Delta_c = \tfrac{1}{2}\left[\iota_1(c_2 - c_1) + \iota_2(c_2 - c_1)\right]. \tag{3}$$

The two sum exactly to $t_2 - t_1$. This order-averaged rule is the symmetric decomposition that, in this two-factor case, coincides with the Shapley value (Shapley, 1953; Shorrocks, 2013). We quantify sampling uncertainty with a cluster bootstrap that resamples randomization cells and recomputes the funnel rates and components over 2,000 replications. The bootstrap reflects sampling noise only; it does not address the cross-regime comparability of the interest and conversion margins discussed in the main text. Panel B of Table 3 reports the resulting contributions and shares.

K Within-Year 1 Transition Details

This appendix reports the within-Year 1 operational-transition analysis referenced in Section 4, together with the descriptive cross-tabulations that document where conversion succeeded or failed during the Schoolhouse and internal periods. In late November 2023 the program moved from the Schoolhouse back end to internal operations, simplifying account creation, matching, scheduling, and session access while leaving the family invitation unchanged. If the back-end simplification reduced realized operational burden and were otherwise comparable across students, we would expect conversion conditional on interest to rise after the transition, with little change in interest expression.

For each outcome Y (interest expression, conversion conditional on interest, or total take-up) for Year 1 assigned student i, we estimate

$$Y_i = \theta_0 + \theta_1 \text{Post}_i + \mu_{g(i)} + \varepsilon_i, \tag{4}$$

where Post_i indicates that student i was randomized after the November 22 back-end transition, $\mu_{g(i)}$ are grade fixed effects, and standard errors are clustered at the randomization cell. The sample is Year 1 assigned students for interest expression and total take-up, and Year 1 interest-expressers for conversion. The operational-burden interpretation predicts $\theta_1 > 0$ for conversion and $\theta_1 \approx 0$ for interest, since the invitation did not change.

Table A7 provides descriptive evidence. Interest expression is essentially unchanged across the transition, consistent with the invitation being held fixed. Conversion conditional on interest rises by 13.9 percentage points, the direction the operational-burden interpretation predicts, while the total take-up estimate is positive but imprecise. This comparison should not be read causally. "Post-Nov-22" was not randomly assigned: after conditioning on grade, the indicator still combines the internal back end with later calendar timing, a different mix of classrooms and teachers, and a different tutor-supply environment. Thus the table is best read as a process diagnostic, not as an estimate of the causal effect of internalizing operations.

We do not report a teacher-fixed-effect version as a main specification. Teacher fixed effects would ask a much narrower question: among teachers observed on both sides of the transition, did their later-round assigned students convert at different rates than their earlier-round assigned students? That comparison is thin in these data. Of 80 Year 1 teachers with assigned students, 33 have assigned students on both sides of the November boundary; for the conversion outcome, only 25 teachers have interest-expressing students on both sides, contributing 99 interest-expressing students to the within-teacher comparison. Many of those teachers have only one post-transition interest-expresser. Adding teacher fixed effects

Table A7: Within-Year-1 back-end transition: interest expression, conversion, and take-up

	Post-transition difference
Outcome: Interest expression (Y1 assigned sample)	
Post-Nov-22 (internal backend)	0.004
	(0.061)
Pre-Nov-22 mean	0.561
Observations	514
R-squared	0.006
Outcome: Conversion given interest (Y1 interest-expressers)	
Post-Nov-22 (internal backend)	0.139***
	(0.048)
Pre-Nov-22 mean	0.778
Observations	288
R-squared	0.031
Outcome: Total take-up (Y1 assigned sample)	
Post-Nov-22 (internal backend)	0.078
	(0.055)
Pre-Nov-22 mean	0.436
Observations	514
R-squared	0.009
Grade FE	Yes
SE clustered at randomization-cell level	Yes

Notes. Within-Year-1 transition specification using the cleaned Year-1 assigned sample. Each pair of rows reports the Post-Nov-22 coefficient and standard error for the indicated outcome from a descriptive regression with grade fixed effects. Post-Nov-22 indicates assignment after the back-end transition began; it was not randomly assigned. Interest expression requires both a recorded interest form and cleaned assignment to tutoring. The invitation email content was unchanged across the November 22 transition; the operational back end, calendar timing, teacher/classroom composition, and tutor availability differed. Standard errors are clustered at the randomization-cell level. * $p < .10$, ** $p < .05$, *** $p < .01$.

would therefore replace the already descriptive grade-fixed-effect comparison with a small and selected within-teacher comparison, while still not solving the calendar and tutor-supply problem. For transparency, we describe why this stricter comparison is not reported, rather than presenting an unstable column as a competing estimate.

A further complication is that the transition coincided with a period of tutor shortage. A shortage depresses conversion directly, because interested families cannot be matched without available tutors, and it works in the opposite direction to the back-end simplification. The two cannot be separated within this window. We therefore treat the transition as descriptive process evidence: it is consistent with the operational-burden interpretation but does not identify the back-end channel, and we do not rely on it for the cross-year decomposition. It does, however, reinforce a substantive point made in the main text, that conversion conditional on interest reflects tutor supply as well as procedural burden, so the realized operational margin is supply-constrained.

The remaining tables document the transition in levels. The first cross-tabulates interest expression by randomization period; since the invitation did not change at the November boundary, similar never-expressed shares before and after are consistent with the back-end transition not moving the offer-to-interest margin.

Table A8: Year-1 interest expression by randomization period

Interest timing	Rand. on/before Nov 22	Rand. after Nov 22	Total
Expressed before Nov 26	136 (36.9%)	0 (0.0%)	137 (26.7%)
Expressed on/after Nov 26	71 (19.2%)	80 (55.6%)	151 (29.4%)
Never expressed	162 (43.9%)	64 (44.4%)	226 (44.0%)

Notes. Year-1 assigned students; column percentages within randomization period. The never-expressed share is similar across periods, consistent with anticipated sludge (carried by the unchanged invitation) being flat across the Nov-22 backend switch.

The second decomposes Year 1 assigned students by operational experience. Students who expressed interest before Schoolhouse stopped accepting new interest forms are classified as Schoolhouse-period interest-expressers; some reached a first session through Schoolhouse, some were transferred into the internal process and attended only after the switch, and some became stuck before reaching any session. Students who expressed interest after the transition entered the internal process. Program records indicate that the internal-period interest-without-session cases primarily reflect tutor-supply or matching constraints rather than families declining after expressing interest. This table is descriptive and documents where conversion succeeded or failed in the observed process; it is not a separate causal estimate.

Table A9: Year-1 assigned students: granular decomposition by operational experience

Group	Description	N	% of Y1	% within regime
Expressed interest before Nov 26 (Schoolhouse period)				
1a	SH onboarding, SH first session	72	14.0	52.6
1b	SH onboarding, internal first session (rescued)	25	4.9	18.2
1c	SH onboarding, no session (stuck)	40	7.8	29.2
Expressed interest on/after Nov 26 (internal period)				
2a	Internal interest, received tutoring	138	26.8	91.4
2b	Internal interest, no session/unmatched	13	2.5	8.6
0	No interest expressed	226	44.0	—
Total		514	100.0	—

Notes. Year-1 assigned students. Schoolhouse-period = interest before Nov 26; internal-period = on/after. Rescued (1b) = Schoolhouse-onboarded students whose first session occurred after the internal switch. No-session/unmatched cases include students who expressed interest but did not reach a first session; program records indicate that the internal-period no-session cases primarily reflected tutor-supply or matching constraints rather than families declining after expressing interest. The last column uses the period subtotal as denominator.

L Teacher Reputation

This appendix reports the returning-teacher analysis used to assess one alternative explanation for the Year 2 take-up gain: that families participated more in Year 2 because some teachers had prior experience with TWiK and could describe it more credibly. A pure individual-teacher reputation story predicts that the cross-year gain should be concentrated among teachers who participated in both years, since only they accumulated experience with their own families; new Year 2 teachers did not.

For outcome Y (interest expression, conversion conditional on interest, or total take-up) for assigned student i taught by teacher j in year t, we estimate

$$Y_{ijt} = \gamma_0 + \gamma_1 \text{Year2}_t + \gamma_2 \text{Returning}_j + \gamma_3 \left(\text{Year2}_t \times \text{Returning}_j\right) + \delta_{g(i)} + \varepsilon_{ijt}, \qquad (5)$$

where Returning_j equals one if teacher j participated in both years, $\delta_{g(i)}$ are grade fixed effects, and standard errors are clustered at the teacher level. The reputation hypothesis predicts $\gamma_3 > 0$: returning teachers should show a larger Year 2 gain than new teachers. Table A10 reports the estimates. The coefficients on Year 2 and on returning-teacher status are descriptive level differences: they show how outcomes differ across years and teacher

Table A10: Returning teachers and the reputation channel

	(1) Interest	(2) Conversion	(3) Take-up
Year 2	0.311***	0.187***	0.404***
	(0.103)	(0.033)	(0.102)
Returning teacher	0.146***	0.007	0.122***
	(0.054)	(0.050)	(0.046)
Year 2 × Returning	-0.128	-0.023	-0.116
	(0.116)	(0.045)	(0.114)
Constant	0.507***	0.814***	0.413***
	(0.035)	(0.036)	(0.031)
Observations	669	416	669

Full assigned sample; conversion column conditions on recorded family interest.
All columns include grade fixed effects and SE clustered at teacher.
The reputation diagnostic is Year2 x Returning: a positive interaction would mean the Year-2 gain is concentrated among teachers with prior KWiK experience.

types, but they do not by themselves identify reputation. The relevant diagnostic is γ_3, the interaction between Year 2 and returning-teacher status. It is negative and statistically insignificant for interest, conversion, and take-up, so the Year 2 gain is not concentrated among returning teachers; if anything, returning teachers show somewhat smaller gains than new teachers. We read this as weakening the individual-teacher-reputation explanation.

Two caveats bound the inference. First, returning-teacher status is not randomly assigned: teachers who participate in both years may differ from one-year teachers in motivation, school context, or student composition, so this is a descriptive comparison rather than a causal test. Second, the interaction is estimated on a modest number of returning teachers and is correspondingly imprecise, so an insignificant γ_3 cannot rule out a moderate reputation channel. The analysis therefore weakens, but does not eliminate, the reputation explanation, and it does not address school-level diffusion, peer recommendations, or changes in teacher communication, which would not operate through individual returning teachers.

M Outcome Measurement and Standardization

This appendix provides additional detail on the engagement, survey, and achievement measures. The main text reports the primary outcome definitions and refers to this appendix for measurement details, assessment construction, and standardization robustness.

M.1 Khan Academy engagement measures

The Khan Academy data contain four platform measures used in the engagement analysis. *Learning minutes* measure the time a student is actively engaged on Khan Academy, excluding navigation and non-learning pages. This is the main dosage measure. It approximates time on instructional material and practice, but it does not by itself distinguish productive from unproductive time, nor can it be consistently decomposed across both years into session and non-session minutes.

Skills worked on counts the distinct skills for which a student attempted at least one problem—a breadth measure of how many topics the student touched, regardless of mastery. *Levels completed* counts skills for which the student advanced a mastery level (for example, from attempted to familiar, or from familiar to proficient). *Skills to proficiency* counts skills the student brought to a high mastery level, classified by Khan Academy as proficient or mastered. The first two measures capture exposure and breadth; the latter two capture platform-defined progress and accumulated mastery.

M.2 Endline survey measures

The endline survey contains three student-reported secondary outcomes. *Math enjoyment* asks, "On a scale of 1 to 7, how much do you enjoy learning math?" *Perceived Khan Academy helpfulness* asks, "On a scale of 1 to 7, how helpful do you find working on Khan Academy assignments for understanding math concepts?" The *homework* item asks, "How much time do you typically spend on math homework outside of school each week?" and is coded from 1 (less than 1 hour) to 6 (more than 5 hours), with intermediate one-hour bins. We analyze these items in their original ordinal units. Because the documented scales begin at one, raw zeros are coded as skipped responses. The homework item is intentionally broad: it is useful as self-reported study effort, but it cannot separate Khan Academy from non-Khan Academy homework, or supervised in-session platform work from independent out-of-session practice.

Table A11: Effects on student attitudes (Year 1)

	Control mean	ITT (SE)	TOT (SE)
Math enjoyment (1–7)	4.008	-0.080 (0.149)	-0.159 (0.295)
KA helpfulness (1–7)	3.964	-0.053 (0.124)	-0.105 (0.244)
Weekly math homework time (1–6)	1.911	0.066 (0.082)	0.130 (0.163)

Notes: Year-1 endline survey. Randomization-cell fixed effects; standard errors clustered at the cell. All three effects survive a Benjamini–Hochberg correction across the three outcomes at the 5% level. $^{*}p < .10$, $^{**}p < .05$, $^{***}p < .01$.

M.3 End-of-year assessment construction

End-of-year achievement is measured with short mathematics assessments drawn from the Khan Academy exercise bank. The Year 1 assessment contained 8 items; the Year 2 assessment contained 16 items in two 8-item sections. The tests were built to be sensitive to the material students practiced, so items were drawn from the Khan Academy content used in tutoring. Because the covered content was only known once the year's assignments were set, item selection took place after randomization, using within-year platform performance data. We document the procedure here because it shapes how the achievement estimates should be read.

For each candidate exercise we computed average platform scores separately for compliant students (those who attended at least one tutoring session), control students, and a pool of non-nominated students. This is a low bar for "compliant"—it does not imply regular attendance—and the platform scores used in selection partly reflect tutor-assisted in-session work on the selected exercises.

Items were prioritized by two rules, both conditioning on compliant students' platform performance. A *competence* rule selected exercises on which compliant students had reached a platform-score threshold (about 70) while the non-nominated pool had not reached full mastery, identifying content where compliers had developed an apparent platform advantage. A *contrast* rule selected exercises on which compliant students' average platform score exceeded the control group's by a large margin (15 to 30 points), or on which compliant completion rates exceeded control rates. The Year 1 assessment and the first Year 2 section were drawn from the competence rule; the second Year 2 section was drawn from the contrast rule and is the most treatment-favoring by construction. For each section we selected ten candidate items, reproduced eight of them in the administered instrument (dropping any that could not be reproduced), and used those. Because both rules select on compliant students' post-treatment platform performance, the primary pooled specification uses the competence-selected items (Year 1 and the first Year 2 section), and the more favoring contrast-selected section is reported separately.

This construction has a clear implication for interpretation, and it is the reason the main text reads the achievement estimate as it does. The test is deliberately *favorable* to detecting a treatment effect: it is aligned to the practiced content, and its items were chosen on dimensions where tutored students had done relatively well. If the offer had produced a sizable gain in tested skill on this material, this instrument should have registered it. It did not register a statistically distinguishable effect—but the point estimate is positive, and the confidence interval is wide. Read together with the light realized dose, the natural conclusion is that the estimate is imprecise rather than that the true effect is zero: a favorable, short,

platform-aligned test, at a roughly ten-hour dose, simply cannot separate a null from the small-to-moderate gains such a dose would plausibly produce (Section 5). The selection cuts the same way as the imprecision, not against it. What the assessment cannot do is serve as a clean, pre-registered test of learning; for that, a next study needs a validated common assessment fixed in advance.

M.4 Standardization

Because assessment forms differed across classes, raw scores are not directly comparable across all students: a given number correct reflects both student performance and form difficulty. The main specification standardizes scores within form, using randomized control students who took that form as the reference distribution:

$$z_{if} = \frac{y_{if} - \bar{y}_f}{s_f},$$

where $\bar{y}_f$ and s_f are computed among control students who took form f, and a form is defined by the year $\times$ assessment $\times$ grade combination. The randomization design enters the estimating equation through randomization-cell fixed effects and clustered standard errors, not through the standardization denominator. Appendix robustness (Table A12) re-estimates the ITT under alternative standardization denominators; the estimates remain statistically indistinguishable from zero.

Table A12: End-of-year achievement: robustness to standardization scheme (ITT)

Standardization denominator	Year 1 (8 items)	Year 2 (all 16)	Pooled (Q1–8)	Pooled (8+16)
Within form, experimental controls *(main)*	0.047	0.081	0.055	0.057
	(0.087)	(0.154)	(0.074)	(0.076)
Observations	716	349	1062	1065
Within form, all untreated test-takers	-0.016	0.095	0.006	0.017
	(0.074)	(0.110)	(0.060)	(0.061)
Observations	745	366	1109	1111
Within year, experimental controls	0.035	0.118	0.044	0.060
	(0.063)	(0.090)	(0.051)	(0.052)
Observations	759	366	1125	1125
Within randomization cell, experimental controls	0.007	-0.028	0.006	-0.003
	(0.091)	(0.157)	(0.079)	(0.079)
Observations	694	340	1031	1034

Notes. ITT on the standardized end-of-year score under four standardization denominators; randomization-cell fixed effects, standard errors (in parentheses, reported on the line below each coefficient) clustered at the cell. Columns: Year 1 (8-item test); Year 2 (all 16 questions); pooled comparable Q1–8 score (Year 1 plus Year 2 questions 1–8); and pooled Year 1 (8 items) plus Year 2 (all 16 questions). All treatment-effect regressions use the cleaned nominated randomized sample; Observations count students with a non-missing standardized score, and the fixed-effect estimator drops singleton cells. The first row is the main specification (within assessment form, randomized experimental controls) and reproduces Table 6. The all-untreated row uses non-nominated same-form test-takers only to scale scores, never as experimental controls. The within-cell denominator is least attractive because each cell contains few controls and therefore a noisy standard deviation. The estimate is an imprecise null under every scheme, so the conclusion does not depend on the choice. * $p < .10$, ** $p < .05$, *** $p < .01$.

N Post-Start Platform Dosage Among Tutored Students

This appendix reports descriptive post-start platform dosage among students who received tutoring, referenced from Section 5. The main ITT engagement tables use a common program-year post window for all randomized students. Here the measurement window is individualized: for each tutored student, it begins in the week containing the first tutoring session and runs through the end of that cohort's post window, excluding TDSB winter and March-break weeks.

The table is descriptive. Receipt, start timing, and post-start intensity are selected, and practice intensity is endogenous to motivation, baseline achievement, family support, tutor quality, difficulty of assigned material, and the tutoring experience itself. The achievement association therefore does not identify the causal return to an additional hour of practice. Its limited purpose is to ask whether, among the selected students who began tutoring, those with more recorded post-start Khan Academy practice also had higher end-of-year achievement. The estimates are tiny and noisy: the Year 1 coefficient is slightly negative, but economically negligible and statistically indistinguishable from zero, and the pooled association is essentially zero.

Table A13: Post-start platform dosage among tutored students

	Year 1	Year 2	Pooled
Panel A: intensive margin after first tutoring session			
Mean weekly learning minutes after first session	40.811	36.814	39.060
Mean total learning hours after first session	14.705	16.159	15.342
Mean post-start share of weeks active	0.612	0.639	0.624
Observations (tutored, active)	150	117	267
Panel B: achievement association among tutored students			
Post-start KA learning hours	-0.004	0.029	0.005
	(0.012)	(0.036)	(0.013)
Observations	84	51	135

Notes. Panel A uses matched tutored students and begins the measurement window in the week containing each student's first tutoring session; it then follows that student through the end of the cohort's post window. TDSB winter/March-break weeks are excluded. Panel B regresses the within-form experimental-control standardized end-of-year score on post-start cumulative KA learning minutes among tutored students with endline scores, with randomization-cell fixed effects and standard errors clustered at the cell. The coefficient is scaled to one additional hour of recorded post-start Khan Academy learning time. This is not a tutoring-hour estimate because tutor explanation, whiteboard work, and other off-platform instruction are not captured in KA learning minutes. Receipt, start timing, and intensity are selected, so the table is descriptive and should not be read as the causal return to an additional hour of practice. * $p < .10$, ** $p < .05$, *** $p < .01$.

O Heterogeneity by Grade

This appendix reports exploratory heterogeneity by grade. The analysis is included for transparency and for generating hypotheses about population alignment, not as definitive evidence of heterogeneous treatment effects. Several grade-year cells are small, especially Grade 4 in Year 1, and the analysis considers multiple outcomes. We therefore avoid drawing strong conclusions from any single grade-specific coefficient.

O.1 Participation chain by age group

A natural hypothesis is that younger students are more parent-mediated. For younger children, a parent or guardian is more likely to notice the invitation, judge its legitimacy, complete the form, and coordinate attendance. Older students may be more able to navigate the offer themselves. This hypothesis implies that invitation-side and enrollment-side changes could matter more in lower grades.

To avoid giving inferential weight to small grade-by-year cells, we collapse grades into younger students (Grades 4–5) and older students (Grades 6–8). We estimate the same interaction diagnostic for each funnel margin: interest, conversion conditional on interest,

and take-up. For funnel outcome F_{igt} for assigned student i in grade g and year t, we estimate

$$F_{igt} = \mu_g + \delta_1 \text{Year2}_t + \delta_2 \left(\text{Year2}_t \times \text{Young}_i\right) + e_{igt}, \tag{6}$$

where μ_g are grade fixed effects. There is no separate coefficient on Young_i because grade fixed effects absorb fixed grade differences. This is the same diagnostic structure as the returning-teacher table: the interaction is the object that asks whether the Year 2 gain is larger for one subgroup. Here δ_1 is the Year 2 minus Year 1 change for older students, δ_2 is the younger-minus-older difference in that change, and $\delta_1 + \delta_2$ is the Year 2 minus Year 1 change for younger students. To make the table readable, Table A14 reports the two group-specific changes first and the interaction/difference last. The point estimates show similar gains for younger and older students: take-up rises by about 35 percentage points for older students and 38 percentage points for younger students, with a small and statistically insignificant difference. Thus the cross-year funnel improvement is not concentrated in younger students. These are descriptive cross-year differences, not causal effects of the redesign.

Table A14: Younger versus older differences in funnel improvements

	Interest	Conversion	Take-up
Year-2 change: older students (Gr. 6–8)	0.256***	0.186***	0.353***
	(0.071)	(0.023)	(0.069)
Year-2 change: younger students (Gr. 4–5)	0.297***	0.153***	0.383***
	(0.058)	(0.042)	(0.058)
Difference: younger minus older	0.041	-0.033	0.031
	(0.093)	(0.048)	(0.091)

Notes. Assigned students. For each funnel margin, the model includes grade fixed effects, a Year-2 indicator, and a Year-2 × younger-student interaction; *younger* means Grades 4–5 and older means Grades 6–8. Conversion conditions on recorded family interest. The first two rows report the Year-2 minus Year-1 change for older and younger students. The final row is the interaction coefficient; a positive value means the Year-2 gain is larger for younger students. These are descriptive cross-year differences, not randomized redesign effects. Parentheses report standard errors clustered at teacher. * $p < .10$, ** $p < .05$, *** $p < .01$.

O.2 Engagement and achievement by grade

For engagement and achievement, the variation of interest is randomized assignment rather than the cross-year redesign. For outcome Y_{ic} for student i in randomization cell c, we estimate

$$Y_{ic} = \alpha_c + \beta_O Z_{ic} + \beta_Y \left(Z_{ic} \times \text{Young}_i\right) + u_{ic}, \tag{7}$$

where Z_{ic} is assignment to the tutoring offer, α_c are randomization-cell fixed effects, and Young_i equals one for students in Grades 4–5 and zero for students in Grades 6–8. The main effect of being young is absorbed by the randomization-cell fixed effects because cells are grade-specific. Standard errors are clustered at the randomization-cell level.

Table A15 follows the same reporting logic. It first reports the ITT for older students, β_O, then the ITT for younger students, $\beta_O + \beta_Y$, and then the interaction/difference, β_Y. The difference row is the formal heterogeneity test for the younger-versus-older split: a positive value means assignment raised the outcome more for younger students than for older students. The last row comes from a more flexible version that interacts assignment with each grade separately and reports the p-value for a joint test that the treatment effect differs across grades. This is a compact way to ask whether treatment effects appear systematically different by age without giving inferential weight to very small grade-year cells. The engagement ITT is positive for both groups: about 9.3 weekly Khan Academy learning minutes for older students and 11.6 for younger students. The 2.3-minute younger-minus-older difference is not statistically significant, so the age split does not show a reliable difference in engagement effects. The achievement estimates are also too imprecise to support a grade-heterogeneity claim: the younger-student point estimate is larger, but the younger-minus-older difference is noisy and the joint grade test does not reject equality. We therefore read this appendix as showing broadly similar effects by age, not as evidence of reliable grade-specific achievement effects.

Table A15: Grade heterogeneity in engagement and achievement (pooled ITT)

	Engagement (weekly min.)	Achievement (std.)
ITT, older students (Gr. 6–8)	9.300***	-0.032
	(1.277)	(0.104)
ITT, younger students (Gr. 4–5)	11.612***	0.176
	(2.091)	(0.109)
Difference: younger minus older	2.312	0.208
	(2.450)	(0.150)
Joint test: effect varies by grade (p)	0.060	0.391
Observations	1118	1065

Notes. Pooled ITT with randomization-cell FE, SE clustered at the cell. *Younger* = Grades 4–5. The older-student ITT is the coefficient on assignment; the younger-student ITT is the linear combination of assignment and assignment × younger. The difference row is the interaction coefficient; a positive value means the assigned-control gap is larger for younger students. Parentheses report standard errors. The last row reports the p-value of a joint test that the treatment effect differs across all grade levels (treatment × grade interactions). * $p < .10$, ** $p < .05$, *** $p < .01$.