Artificial Intelligence in Remedial Education: Evidence from a Cluster Randomized Trial

Philip Oreopoulos* Nina Low†

May 13, 2026

1 Introduction

For nearly fifty years, research on individualized instruction has emphasized the large achievement gains associated with one-to-one tutoring, often summarized as the "two-sigma" result. Despite this evidence, scaling individualized instruction within public education remains difficult due to the high cost of skilled human tutoring. As a result, most school systems rely on tiered support models, such as Response to Intervention (RTI), which provide supplemental small-group instruction to students performing below grade level.

While RTI offers a structured mechanism for academic support, its effectiveness is limited by the constraints of instructional labor. In practice, RTI instructors must simultaneously diagnose heterogeneous skill gaps and deliver targeted remediation to multiple students at once, often without deep subject-specific expertise for all content areas. This can lead to instruction that is necessarily standardized rather than fully individualized, particularly for students with the largest learning deficits.

This paper studies whether recent advances in large language models can improve the effectiveness of RTI by providing scalable, adaptive instructional support. We evaluate the KWiK (Khoaching with Khanmigo) program, which integrates the Khanmigo AI tutor into middle school RTI mathematics instruction in Hamilton County, Tennessee. The platform provides real-time, context-specific feedback intended to support student problem-solving through adaptive hints and guided questioning.

We estimate the effects of assigning schools to AI-augmented RTI instruction using a two-year, grade-within-school cluster-randomized controlled trial across 20 middle schools. The design allows us to estimate intent-to-treat (ITT) effects on both term-level achievement (MAP mathematics percentile ranks) and annual state standardized test performance (TCAP). Because the intervention is implemented over multiple academic years, we additionally examine how effects vary with continued exposure and across the achievement distribution. Together, the results provide evidence on whether AI-based instructional tools can improve student learning in structured remedial settings and whether such effects persist beyond initial implementation.

2 Experimental Design and Sample

2.1 Experimental design

The study is a two-year cluster-randomized controlled trial conducted in 20 middle schools in Hamilton County, Tennessee during the 2024–2025 and 2025–2026 academic years.

The unit of randomization is the grade-within-school cell. Within each school, one or two grades (Grades 6–8) were randomly assigned to receive the KWiK intervention. Assignment was conducted once prior to the start of Year 1 and held fixed across both years. Treated grades use Khan Academy with the Khanmigo AI tutor during RTI mathematics periods, while control grades receive business-as-usual RTI instruction using

*Department of Economics, University of Toronto
†Department of Economics, University of Toronto

district-approved materials. The comparison group therefore represents standard RTI practice within the same school environment rather than the absence of supplemental instruction.

RTI eligibility is determined by pre-existing district rules based on MAP mathematics percentile thresholds and teacher referral. These rules are applied uniformly across treated and control grades within each school and were fixed prior to randomization.

2.2 Study population

The study population is students scheduled into RTI mathematics under pre-existing district assignment rules based on prior achievement and teacher referral. These scheduling rules were determined independently of treatment assignment and applied uniformly across treated and control grades. Estimated effects are therefore internally valid for this administratively defined RTI population but may not generalize to general education students.

2.3 Sample construction

The analysis sample is constructed using a once-in, always-in cohort rule. Students enter the sample in the first term in which they are scheduled into RTI mathematics and remain eligible for inclusion in all subsequent observed terms regardless of later RTI scheduling status. Student-term observations require non-missing baseline and endline MAP assessments within the corresponding term. RTI placement is determined by pre-specified district rules based on prior achievement and teacher referral, applied uniformly across treated and control grades prior to randomization. Teachers making referrals are not informed of grade-level treatment assignments. The primary ITT estimand is defined within this administratively determined RTI population.

RTI scheduling is updated each term based on the same pre-specified district achievement rules, applied uniformly across treatment and control grades. Transitions into and out of scheduled RTI placement reflect the administrative RTI placement process rather than behavioral responses to treatment. As a pre-specified diagnostic, we verify that scheduled RTI entry rates and entrant characteristics are balanced across treatment arms; details are provided in Appendix A.

The primary analysis sample comprises 2,810 unique students across 56 grade-within-school clusters, with 2,576 students observed with both baseline and endline MAP scores across four completed terms. Term 5 (Spring 2026) data collection is ongoing.

2.4 Sample exclusions

Three school-level exclusions are applied based on protocol integrity:

- **East Lake Academy:** excluded due to post-randomization changes in treated-grade composition that violated assignment integrity.
- **Tyner Middle (Year 2):** excluded due to absence of RTI implementation infrastructure, preventing both treatment delivery and valid control conditions.
- **Howard Connect (Year 2):** excluded due to school closure dynamics and loss of within-school control structure.

In addition, Fall 2024 is excluded from term-level analyses due to district-wide inconsistencies in assessment timing and rostering, which prevent a clean pre/post structure. Fall 2024 MAP scores are retained only for annual gain specifications.

Student-level exclusions include transient enrollment (insufficient exposure within a term), missing MAP assessments, and one case of ambiguous multi-school enrollment.

All exclusion rules were determined based on implementation records and prior to outcome analysis, and are applied uniformly across treatment arms.

3 Empirical Strategy

The primary analysis is a single intent-to-treat regression estimating the effect of KWiK assignment on MAP mathematics percentile rank, identified by grade-within-school randomization. All remaining analyses (dosage effects, heterogeneity, and mechanisms) are secondary and clearly labeled as such.

3.1 Primary Analysis: Intent-to-Treat (ITT)

The primary estimand is the effect of assignment to a KWiK-enabled RTI grade on end-of-term MAP mathematics percentile rank, regardless of actual platform usage.

We estimate:

$$Y_{igsty} = \alpha + \beta D_{gsy} + \gamma' X_{igsty} + \theta_s + \phi_g + \lambda_{ty} + \varepsilon_{igsty} \tag{1}$$

where Y_{igsty} is student i's MAP mathematics percentile rank in grade g, school s, term t, and year y; D_{gsy} is an indicator equal to one if grade g in school s is assigned to KWiK in year y; X_{igsty} includes baseline MAP score and pre-specified covariates; θ_s, ϕ_g, and λ_{ty} are school, grade, and term-by-year fixed effects.

Identification comes from within-school variation in treatment assignment across grades. Grade fixed effects absorb average differences across grade levels and are not collinear with treatment because no grade is always treated or always control across all schools. Identification comes from variation in treatment assignment across grades within schools; repeated observations over time improve precision but do not add independent treatment variation. Treatment is assigned at the grade-within-school level and held fixed across both years. The regression is estimated on a stacked student-term panel. Standard errors are clustered at the grade-within-school level, the unit of randomization. The coefficient β is the primary intent-to-treat estimand of the study. The inclusion of baseline MAP scores implicitly assumes that conditioning on current achievement adequately captures prior treatment-related differences in prior achievement across repeated student-term observations. This assumption is assessed in the appendix using the cumulative gain specification and the Year 1-only analysis.

3.2 Inference

Primary inference uses cluster-robust standard errors clustered at the grade-within-school level. The study includes 56 grade-within-school clusters (29 treated, 27 control), supporting standard asymptotic inference. As an additional robustness check, we report wild cluster bootstrap p-values (Rademacher weights, 999 replications). Student-level clustered standard errors are reported as a further robustness check.

3.3 Secondary Analysis: Dosage Effects (LATE via IV)

To characterize the behavioral mechanism underlying the ITT, we estimate the causal effect of platform engagement among compliers. We instrument realized usage with randomized assignment. The primary dosage measure is total learning path minutes. Three secondary exploratory dosage measures (skills leveled up to proficiency, Khanmigo chat interactions, and consistency of engagement across RTI weeks) are analyzed separately and are not compared in magnitude against the primary.

Each dosage measure is instrumented in a distinct 2SLS regression and should not be interpreted as a system of endogenous regressors with a single instrument.

First stage:

$$D^k_{igsty} = \pi_0 + \pi_1 Z_{gsy} + \gamma' X_{igsty} + \theta_s + \phi_g + \lambda_{ty} + \nu_{igsty} \tag{2}$$

Second stage:

$$Y_{igsty} = \alpha + \beta^k_{LATE} \widehat{D}^k_{igsty} + \gamma' X_{igsty} + \theta_s + \phi_g + \lambda_{ty} + \varepsilon_{igsty} \tag{3}$$

where D^k_{igsty} is dosage measure k, Z_{gsy} is assignment to a KWiK grade, and $\widehat{D}^k_{igsty}$ is predicted usage from the first stage. This LATE identifies the effect of assignment-induced engagement intensity among compliers, a bundle that includes both direct platform usage and any complementary instructional changes induced by treatment, and should not be interpreted as the isolated technological effect of platform usage alone.

3.4 Secondary Analysis: Heterogeneous Treatment Effects

The following heterogeneity analyses are pre-specified and exploratory.

3.4.1 Baseline achievement

We estimate:

$$Y_{igsty} = \alpha + \beta D_{gsy} + \delta \left(D_{gsy} \times \text{BaselineMAP}_i \right) + \gamma' X_{igsty} + \theta_s + \phi_g + \lambda_{ty} + \varepsilon_{igsty} \tag{4}$$

to assess whether effects vary with prior achievement.

3.4.2 Demographic heterogeneity

We estimate interactions of treatment assignment with gender, race/ethnicity, and economic disadvantage (FARMS eligibility). Each interaction is estimated separately. We adjust for multiple testing across heterogeneity specifications using Romano–Wolf step-down procedures.

3.4.3 Grade-level heterogeneity

We estimate heterogeneous treatment effects by grade level using interactions between treatment assignment and grade indicators. Because treatment is assigned at the grade-within-school level, grade-specific effects are identified from the same randomized assignment as the primary ITT, using variation across schools in whether a given grade is assigned to treatment or control. These estimates are therefore best interpreted as descriptive heterogeneity across grade cohorts and may reflect both differences in treatment response and differences in cohort composition across grades.

4 Primary Outcomes

MAP mathematics percentile rank (term-level) is the primary confirmatory outcome. TCAP mathematics scale score (annual) is the co-primary outcome, providing an external validity anchor on a high-stakes state assessment administered independently of the intervention. All other outcomes are secondary.

To address multiplicity across confirmatory outcomes, we pre-specify the following hierarchy: MAP mathematics percentile rank is the primary outcome and is evaluated unconditionally. TCAP mathematics scale score is evaluated as a confirmatory co-primary outcome; a statistically significant MAP NPR effect is not required for TCAP to be interpreted, but divergence between the two outcomes will be explicitly addressed in interpretation. MAP annual gain is treated as descriptive and is not subject to hypothesis testing. No family-wise correction is applied across these outcomes given the pre-specified hierarchy.

4.1 MAP mathematics percentile rank (term-level)

The primary outcome is student performance on the NWEA MAP mathematics assessment, expressed as a nationally normed percentile rank. For each academic term (Fall, Winter, Spring), the outcome is the end-of-term MAP percentile rank. All specifications control for the corresponding beginning-of-term MAP score.

We use percentile ranks rather than RIT scores to ensure comparability across grades and testing windows, given the vertically scaled nature of the MAP assessment and the stacked panel structure spanning multiple grades. As a pre-specified robustness check, we replicate all primary ITT specifications using RIT scale scores and within-grade standardized z-scores to confirm that results are not sensitive to outcome scaling.

4.2 TCAP mathematics scale score (annual)

The co-primary outcome is student performance on the Tennessee Comprehensive Assessment Program (TCAP) mathematics exam, measured as a scale score and administered at the end of each academic year. TCAP scores are linked at the student-year level. Specifications control for prior achievement using lagged MAP.

5 Mechanisms

The primary ITT estimate identifies the effect of assignment to AI-augmented RTI instruction but does not explain through which channels the platform affects achievement. We conduct two complementary analyses: a descriptive analysis documenting associations between engagement intensity and achievement gains within the treated sample, and an IV analysis recovering the causal effect of engagement among compliers. The OLS associations are not causal and serve to characterize variation in engagement within the treated group. Neither analysis redefines the primary ITT estimand.

5.1 Descriptive Engagement Analysis

Within the treated sample, we estimate OLS associations between end-of-term MAP gains and four engagement dimensions: total learning path minutes (primary), skills leveled up to proficiency, Khanmigo chat interactions, and consistency of engagement across RTI weeks. Because these measures are highly correlated, coefficients are interpreted as conditional associations rather than independent causal channels. We control for multiple hypothesis testing across the engagement outcomes using the procedure in Anderson (2008). These associations are not interpreted as mediation estimates; clean mediation is not identified due to post-treatment confounding in engagement.

5.2 Causal Effect of Dosage

The descriptive associations in Section 5.1 are subject to selection: students who engage more with the platform may differ systematically from low-engagers in ways that independently predict achievement. To recover the causal effect of platform engagement, we instrument realized dosage with randomized grade-level assignment, following the 2SLS specifications in Section 3.3. The resulting LATE identifies the effect of increased engagement among compliers and should be interpreted alongside, not instead of, the primary ITT.

Total learning path minutes is the primary pre-specified dosage measure. Three secondary measures (skills leveled up to proficiency, Khanmigo chat interactions, and consistency of engagement) are analyzed in parallel and interpreted as complementary evidence.

6 Power Analysis

The study consists of 56 grade-within-school clusters, with a harmonic mean of approximately 46 student-term observations per cluster among students with both baseline and endline MAP scores. Power calculations assume a two-sided test with $\alpha = 0.05$ and 80 percent power and account for clustering at the grade-within-school level. These calculations reflect ex-post precision given realized sample characteristics; Term 5 (Spring 2026) data collection is ongoing at the time of registration.

Precision is improved through inclusion of baseline achievement (start-of-term MAP scores) in an ANCOVA specification. Power calculations use an ICC of 0.095 and an ANCOVA R^2 of 0.580, estimated from the realized sample. These design parameters were calculated from the full sample without reference to treatment effects or outcome differences across arms. The outcome standard deviation is 23.37 percentile points.

For the primary ITT estimand, the minimum detectable effect is 3.83 percentile points (0.16 SD). This is somewhat larger than effect sizes reported for one-to-one and small-group tutoring programs in Nickow, Oreopoulos, and Quan (2020), but consistent with detecting economically meaningful effects of a scalable classroom-based intervention.

Power is more limited for secondary estimands due to reduced effective cluster sizes in subset analyses, yielding MDEs of 0.18 to 0.25 standard deviations across dynamic specifications (Table 1). Power calculations are based on the primary stacked panel sample; the baseline cohort robustness sample is not used to define power. Given these design constraints, null results for dynamic, subgroup, and other secondary estimands should not be interpreted as evidence of no effect; power is limited for these specifications and null results should be interpreted cautiously.

Table 1: Minimum Detectable Effects by Estimand

Estimand	Clusters	MDE (percentile)	MDE (SD)
Primary ITT	56	3.83	0.164
First-year ITT	56	4.43	0.190
Cohort stability	27	5.85	0.250
Persistence	53	4.17	0.179
Cumulative two-year ITT	53	4.17	0.179
Baseline cohort robustness	56	4.43	0.190

Notes: ICC = 0.095. ANCOVA $R^2 = 0.580$. Outcome SD = 23.37 NPR points. Harmonic mean cluster size = 46.00 (primary ITT); 15.56 (first-year ITT, Y1 terms only); 26.98 (Y2 estimands). Power calculations reflect realized sample design parameters across four completed terms; Term 5 data collection is ongoing.

7 Pre-Specified Analyses

The following table summarizes the full set of pre-specified analyses, organized by estimand hierarchy. The primary estimand is the ITT effect of KWiK assignment on MAP NPR. All other analyses are secondary or exploratory and do not redefine the primary conclusion.

Table 2: Pre-Specified Analyses by Estimand Hierarchy

Hierarchy	Analysis	Outcome	Status
Primary	ITT, stacked OLS	MAP NPR	Confirmatory
Co-primary	Annual ITT	TCAP scale score	Confirmatory
Secondary	Annual ITT	Fall–Spring MAP NPR gain	Secondary
Secondary	LATE, learning path minutes	MAP NPR	Secondary
Exploratory	LATE, skills leveled up	MAP NPR	Exploratory
Exploratory	LATE, Khanmigo chat interactions	MAP NPR	Exploratory
Exploratory	LATE, consistency of engagement	MAP NPR	Exploratory
Exploratory	Heterogeneity by baseline achievement	MAP NPR	Exploratory
Exploratory	Heterogeneity by gender, race/ethnicity, FARMS	MAP NPR	Exploratory
Exploratory	Heterogeneity by grade level	MAP NPR	Exploratory
Exploratory	Descriptive engagement analysis	MAP NPR	Exploratory
Secondary	RTI exit test	RTI exit	Secondary
Secondary	Post-baseline RTI entry test	RTI entry	Secondary
Secondary	Entrant covariate balance test	Baseline covariates	Secondary
Robustness	Baseline cohort ITT (Term 1 only)	MAP NPR	Robustness
Robustness	Entry-timing cohort analysis	MAP NPR	Diagnostic
Robustness	Student-level cumulative gain specification	MAP NPR	Robustness
Robustness	Spillover diagnostic (share treated within school)	MAP NPR	Diagnostic

Notes: Confirmatory analyses are pre-specified with a single designated primary outcome. Exploratory analyses are pre-specified but not used to define the primary conclusion. Robustness and diagnostic analyses test interpretability assumptions of the primary estimand and are reported in the appendix. The hierarchy is fixed prior to analysis and is not updated based on results.

Pre-Specified Tables

All tables in this section are pre-specified diagnostic and descriptive tables. They are not used to modify estimation samples, specification selection, or outcome definitions.

A. Sample Definition and Cohort Construction

Table 3: Pre-Specified Sample Construction and Analysis Samples

	Has Endline		Has End and Baseline	
	Control	Treated	Control	Treated
Panel A: Full Sample				
Term 1	651	692	644	664
Term 2	724	759	685	691
Term 3	686	765	622	680
Term 4	761	844	745	826
Panel B: Imbalanced Sample				
Term 1	651	679	644	656
Term 2	724	634	685	578
Term 3	686	638	622	578
Term 4	761	794	745	777
Panel C: Balanced Trim Sample				
Term 1	651	686	644	661
Term 2	721	756	683	689
Term 3	0	0	0	0
Term 4	0	0	0	0

Notes: Panel A (Primary Analysis Sample) includes all students satisfying RTI enrollment and attendance criteria, regardless of treatment compliance. Panel B (Implementation Sensitivity Sample) excludes treated students with zero recorded Khan Academy usage. Panel C (Attendance-Restricted Sample) restricts to students with attendance rates above 55%. Panel A is the primary estimation sample. Panels B and C are pre-specified robustness checks. "Has Endline" indicates availability of endline MAP scores. "Has Both" indicates availability of both baseline and endline assessments.

Table 4: Cohort Definition and Tracking of Year 1 RTI Students

Status of Y1 RTI Students	Term 2	Term 3	Term 4	Term 5
Active in RTI (%)	81.5%	53.9%	52.9%	.%
Spillover (Exited RTI) (%)	18.4%	26.9%	27.9%	.%

Notes: This table tracks students entering the RTI analysis sample over time under the study's once-in, always-in retention rule. Students remain in the estimation sample after first scheduled RTI placement regardless of later RTI scheduling status. Students are classified as Active (currently enrolled in RTI), Spillover (no longer enrolled in RTI but retained in the analysis cohort), or Attrited (no longer observed in outcome data). Term 5 data collection is ongoing.

Table 5: Distribution of Student Exposure Histories

Trajectory	Role	Students (N)	Clusters
(00) Pure Control	Counterfactual	924	37
(01) Movers-In	Replication ($\hat{\beta}_{1m}$)	275	29
(10) Formerly-Treated	Persistence ($\hat{\delta}$)	740	46
(11) Second-Year Treated	Cumulative ($\hat{\beta}_2$)	219	24
Excluded from primary trajectory analyses			
New Y2 Entrants	Robustness check only	414	22
Total (primary sample)		2,158	56

Notes: Trajectories are defined by ITT grade-level randomization assignment (`d_KWiK_grade`) rather than individual platform compliance. Sample restricted to `sample_all` = 1 and `has_bothlines` = 1. Cluster counts are not additive across trajectories: a single grade-within-school cluster may contain students from multiple trajectory groups, so the cluster column does not sum to the total. New Y2 Entrants are students who appear in a treated grade in Year 2 with no Year 1 RTI presence; they are excluded from the primary trajectory analyses per the PAP movers-in definition (Section 2) and included in the Section 9.3 robustness check.

Notes: Exposure categories: (00) never assigned to KWiK, (11) assigned in both years, (10) assigned in Year 1 only, (01) assigned in Year 2 only. Purely descriptive.

B. Implementation and Usage (Descriptive Statistics)

Table 6: School-Level Enrollment and Platform Usage, Year 1

		Treatment						Control					
		Fall 2024			Winter 2025			Fall 2024			Winter 2025		
School	Grade	N	Avg. Time	Avg. N Active	N	Avg. Time	Avg. N Active	N	Avg. Time	Avg. N Active	N	Avg. Time	Avg. N Active
Brown Middle School	Grade 6	75	30	48.8	81	35	39.1						
	Grade 7							90	0	0.2	91	.	.
	Grade 8	88	35	68.5	90	29	56.1						
Chattanooga School For The Arts and Sciences Upper	Grade 6	6	69	6.0	7	34	3.9						
	Grade 7							3	.	.	3	.	.
	Grade 8							3	8	0.5	3	.	.
Chattanooga School for Liberal Arts	Grade 6							7	22	0.5	11	.	.
	Grade 7	5	39	3.0	5	23	1.2						
	Grade 8							6	10	0.2	5	.	.
Dalewood Middle School	Grade 6							82	.	.	71	.	.
	Grade 7	74	12	2.5	70	32	5.1						
	Grade 8	72	30	36.2	63	27	6.1						
East Hamilton Middle School	Grade 6	4	121	3.8	4	98	3.3						
	Grade 7	6	93	6.0	8	102	5.0						
	Grade 8							3	.	.	5	.	.
East Ridge Middle School	Grade 6							25	.	.	21	.	.
	Grade 7							17	.	.	11	70	0.1
	Grade 8	15	46	11.2	8	41	6.0						
Hamilton County Virtual School	Grade 6	2	40	1.0	5	77	2.4						
	Grade 7	5	76	3.5	4	84	0.9						
	Grade 8							6	.	.	6	.	.
Hixson Middle School	Grade 6	4	22	3.8	4	19	2.0						
	Grade 7	1	38	1.0	1	13	0.9						
	Grade 8							7	.	.	7	.	.
Howard Connect Academy	Grade 7	5	65	4.5	7	36	4.3						
	Grade 8							11	.	.	13	.	.
Loftis Middle School	Grade 6							4	.	.	11	.	.
	Grade 7							11	.	.	10	11	0.3
	Grade 8	12	86	11.8	6	69	5.2						
Lookout Valley Middle/High School	Grade 6							7	2	0.2	11	.	.
	Grade 7	6	99	6.0	7	43	1.6						
	Grade 8	2	75	1.8	5	26	2.1						
Normal Park Museum Magnet	Grade 6							14	.	.	14	.	.
	Grade 7	7	64	2.8	12	62	4.9						
	Grade 8	6	65	5.8	8	38	4.3						
Ooltewah Middle School	Grade 6							29	1	0.2	30	.	.
	Grade 7	29	50	24.8	29	49	14.4						
	Grade 8							23	.	.	18	.	.
Orchard Knob Middle School	Grade 6							67	.	.	83	.	.
	Grade 7	54	23	10.8	56	75	18.9						
	Grade 8	42	16	11.2	40	48	4.6						
Red Bank Middle School	Grade 6	33	70	8.0	60	77	9.6						
	Grade 7							30	9	0.2	50	7	0.1
	Grade 8							25	.	.	48	12	0.7
Sale Creek Middle/High School	Grade 6	6	34	6.0	10	37	3.9						
	Grade 7	7	54	6.8	10	58	6.9						
	Grade 8							10	.	.	3	.	.
Signal Mountain Middle/High School	Grade 6							12	.	.	12	.	.
	Grade 7							7	.	.	8	.	.
	Grade 8	14	38	12.5	10	40	6.4						
Soddy Daisy Middle School	Grade 6	7	92	7.0	7	75	3.0						
	Grade 7	9	115	8.0	7	50	4.8						
	Grade 8							6	.	.	5	11	0.2
Tyner Middle Academy	Grade 6							66	.	.	65	.	.
	Grade 7	68	28	51.2	67	27	20.0						
	Grade 8							73	.	.	70	.	.
Total		664	56	12.9	691	49	8.5	644	7	0.3	685	22	0.3

Notes: Descriptive statistics on Khan Academy usage in Year 1 by school-grade-term cell. *N* denotes students with both baseline and endline MAP data. Average usage computed among active users. Thanksgiving week (Week 4) excluded from Semester 1 averages.

Table 7: School-Level Enrollment and Platform Usage, Year 2

		Treatment						Control					
		Fall 2025			Winter 2026			Fall 2025			Winter 2026		
School	Grade	N	Avg. Time	Avg. N Active	N	Avg. Time	Avg. N Active	N	Avg. Time	Avg. N Active	N	Avg. Time	Avg. N Active
Brown Middle School	Grade 6	44	56	27.0	52	62	46.5						
	Grade 7							84	.	.	88	.	.
	Grade 8	90	49	20.0	89	62	26.6						
Chattanooga School For The Arts and Sciences Upper	Grade 6	4	30	4.0	7	32	5.5						
	Grade 7							7	.	.	6	.	.
	Grade 8							3	.	.	4	.	.
Chattanooga School for Liberal Arts	Grade 6							8	.	.	12	.	.
	Grade 7	10	.	.	6	.	.						
	Grade 8							7	.	.	3	.	.
Dalewood Middle School	Grade 6							70	.	.	76	65	1.8
	Grade 7	71	.	.	76	17	22.8						
	Grade 8	67	.	.	76	22	31.8						
East Hamilton Middle School	Grade 6	5	93	4.7	12	98	6.5						
	Grade 7	9	127	1.0	12	86	8.8						
	Grade 8							11	.	.	7	.	.
East Ridge Middle School	Grade 6							25	.	.	40	.	.
	Grade 7							58	.	.	64	.	.
	Grade 8	19	82	10.0	54	66	22.8						
Hamilton County Virtual School	Grade 6	4	76	4.0	5	85	1.6						
	Grade 7	2	.	.	2	.	.						
	Grade 8							2	.	.	3	.	.
Hixson Middle School	Grade 6	41	62	33.7	54	41	25.6						
	Grade 7	44	79	43.0	50	55	27.4						
	Grade 8							38	57	1.0	49	.	.
Howard Connect Academy	Grade 7												
	Grade 8												
Loftis Middle School	Grade 6							10	.	.	19	.	.
	Grade 7							14	.	.	10	.	.
	Grade 8	14	73	13.0	9	69	8.4						
Lookout Valley Middle/High School	Grade 6							7	.	.	7	.	.
	Grade 7	12	54	7.7	14	52	6.6						
	Grade 8	4	59	2.0	6	34	2.4						
Normal Park Museum Magnet	Grade 6							16	.	.	19	.	.
	Grade 7	14	86	9.0	23	82	17.2						
	Grade 8	13	78	12.0	14	71	11.0						
Ooltewah Middle School	Grade 6							15	.	.	19	.	.
	Grade 7	29	46	21.3	30	42	18.5						
	Grade 8							35	.	.	38	.	.
Orchard Knob Middle School	Grade 6							19	.	.	17	.	.
	Grade 7	32	36	5.7	30	61	11.8						
	Grade 8	23	28	7.0	26	59	8.4						
Red Bank Middle School	Grade 6	47	38	37.0	57	57	25.1						
	Grade 7							76	.	.	86	.	.
	Grade 8							73	.	.	81	.	.
Sale Creek Middle/High School	Grade 6	3	134	2.0	5	95	3.4						
	Grade 7	7	94	3.0	4	102	3.0						
	Grade 8							5	.	.	9	.	.
Signal Mountain Middle/High School	Grade 6							11	.	.	13	.	.
	Grade 7							13	.	.	8	.	.
	Grade 8	5	80	3.3	3	41	2.0						
Soddy Daisy Middle School	Grade 6	33	82	32.3	54	60	43.9						
	Grade 7	34	75	31.3	56	57	43.0						
	Grade 8							15	.	.	67	.	.
Tyner Middle Academy	Grade 6												
	Grade 7												
	Grade 8												
Total		680	70	14.5	826	60	17.2	622	57	1.0	745	65	1.8

Notes: Descriptive statistics on Khan Academy usage in Year 2 by school-grade-term cell. Average usage computed among active users.

C. Balance and Randomization Checks

Table 8: Baseline Balance Checks, Year 1

	Has Endline		Has End and Baseline	
	Control	Diff	Control	Diff
Panel A: Term 1				
Female	0.521	-0.053*	0.520	-0.061**
		(0.027)		(0.028)
White	0.189	0.047**	0.189	0.052**
		(0.022)		(0.023)
Black	0.596	-0.031	0.599	-0.038
		(0.027)		(0.027)
Hispanic	0.198	-0.019	0.194	-0.016
		(0.021)		(0.022)
Other Race	0.017	0.003	0.017	0.002
		(0.007)		(0.007)
FARMS	0.728	-0.020	0.727	-0.020
		(0.025)		(0.025)
Economic Disadvantage	0.476	-0.002	0.474	-0.004
		(0.027)		(0.028)
Dyslexia	0.169	0.058***	0.169	0.055**
		(0.022)		(0.022)
SPED Disability	0.155	-0.008	0.155	-0.009
		(0.020)		(0.020)
ELA Baseline	36.846	-3.571***	36.928	-3.327**
		(1.363)		(1.377)
Math Baseline	29.376	-1.829	29.376	-1.829
		(1.239)		(1.239)
Panel B: Term 2				
Female	0.523	-0.041	0.528	-0.039
		(0.026)		(0.027)
White	0.191	0.032	0.197	0.037*
		(0.021)		(0.022)
Black	0.583	0.001	0.582	-0.014
		(0.026)		(0.027)
Hispanic	0.206	-0.033	0.200	-0.023
		(0.020)		(0.021)
Other Race	0.021	0.000	0.020	-0.000
		(0.007)		(0.008)
FARMS	0.728	-0.012	0.726	-0.016
		(0.023)		(0.024)
Economic Disadvantage	0.479	0.003	0.480	-0.009
		(0.026)		(0.027)
Dyslexia	0.167	0.053***	0.172	0.049**
		(0.020)		(0.021)
SPED Disability	0.153	0.021	0.152	0.022
		(0.019)		(0.020)
ELA Baseline	35.915	-1.315	35.953	-0.948
		(1.393)		(1.421)
Math Baseline	29.491	-1.053	29.491	-1.053
		(1.218)		(1.218)

Notes: Each row reports coefficients from regressions of baseline covariates on treatment assignment with school fixed effects. Standard errors clustered at the grade-within-school level.

Table 9: Baseline Balance Checks, Year 2

	Has Endline		Has End and Baseline	
	Control	Diff	Control	Diff
Panel C: Term 3				
Female	0.529	0.003	0.531	0.000
		(0.026)		(0.028)
White	0.335	-0.044*	0.302	0.002
		(0.024)		(0.026)
Black	0.448	0.062**	0.481	0.024
		(0.026)		(0.028)
Hispanic	0.190	-0.016	0.193	-0.025
		(0.020)		(0.021)
Other Race	0.028	-0.003	0.024	-0.001
		(0.008)		(0.008)
FARMS	0.638	0.036	0.645	0.020
		(0.025)		(0.026)
Economic Disadvantage	0.345	0.081***	0.344	0.071***
		(0.025)		(0.027)
Dyslexia	0.092	0.011	0.093	0.008
		(0.016)		(0.016)
SPED Disability	0.150	0.018	0.150	0.021
		(0.019)		(0.020)
ELA Baseline	35.910	-0.670	36.078	-0.594
		(1.341)		(1.361)
Math Baseline	26.177	-0.422	26.177	-0.422
		(1.045)		(1.045)
Panel D: Term 4				
Female	0.536	-0.011	0.536	-0.011
		(0.025)		(0.025)
White	0.334	-0.020	0.336	-0.022
		(0.023)		(0.024)
Black	0.447	0.034	0.443	0.036
		(0.025)		(0.025)
Hispanic	0.194	-0.018	0.196	-0.018
		(0.019)		(0.020)
Other Race	0.025	0.003	0.026	0.004
		(0.008)		(0.008)
FARMS	0.628	0.042*	0.626	0.046*
		(0.024)		(0.024)
Economic Disadvantage	0.340	0.076***	0.336	0.080***
		(0.024)		(0.024)
Dyslexia	0.092	0.006	0.094	0.007
		(0.015)		(0.015)
SPED Disability	0.160	0.007	0.161	0.006
		(0.018)		(0.019)
ELA Baseline	40.767	-0.915	40.887	-0.862
		(1.175)		(1.180)
Math Baseline	32.526	-0.891	32.526	-0.891
		(1.059)		(1.059)

Notes: Same specification as Year 1 balance checks.

D. Compliance and Instrument Strength Diagnostics

Table 10: Pre-Treatment Usage and Compliance by Treatment Status

	Treatment		Control	
	Mean	SD	Mean	SD
Panel A: Term 1				
N	664		644	
Avg. Weekly Learning Time (mins)	30.5	31.4	0.1	1.6
Avg. Weekly Skills Worked On	4.0	5.2	0.0	0.1
Avg. Weekly Skills Leveled Up To Proficient	2.4	4.4	0.0	0.1
Avg. Weekly Khanmigo Chats	1.4	2.4	0.0	0.4
Avg. Weekly Khanmigo Interactions	3.3	7.4	0.0	0.6
Panel B: Term 2				
N	570		576	
Avg. Weekly Learning Time (mins)	26.4	27.4	0.3	3.4
Avg. Weekly Skills Worked On	3.5	4.2	0.0	0.6
Avg. Weekly Skills Leveled Up To Proficient	2.3	3.5	0.0	0.6
Avg. Weekly Khanmigo Chats	1.4	2.4	0.0	0.2
Avg. Weekly Khanmigo Interactions	4.7	9.1	0.0	0.6
Panel C: Term 3				
N	555		538	
Avg. Weekly Learning Time (mins)	43.3	36.9	0.1	2.5
Avg. Weekly Skills Worked On	5.9	6.9	0.0	0.1
Avg. Weekly Skills Leveled Up To Proficient	4.2	5.5	0.0	0.1
Avg. Weekly Khanmigo Chats	2.9	4.0	0.0	0.2
Avg. Weekly Khanmigo Interactions	10.7	17.5	0.0	0.6
Panel D: Term 4				
N	597		516	
Avg. Weekly Learning Time (mins)	47.8	31.9	0.4	5.9
Avg. Weekly Skills Worked On	7.2	5.9	0.1	0.7
Avg. Weekly Skills Leveled Up To Proficient	5.1	4.5	0.0	0.6
Avg. Weekly Khanmigo Chats	2.7	2.8	0.0	0.3
Avg. Weekly Khanmigo Interactions	11.2	16.7	0.1	0.9

Notes: Descriptive usage statistics by treatment status. For each student, usage is averaged over weeks of positive engagement and then aggregated to the student level.

Table 11: First-Stage Effects of Assignment on Platform Usage

	Treatment	Control	Difference	SE
Panel A: Term 1				
N	664	644		
Avg. Weekly Learning Time (mins)	30.5	0.1	30.3***	(4.5)
Avg. Weekly Skills Worked On	4.0	0.0	4.0***	(0.6)
Avg. Weekly Skills Leveled Up To Proficient	2.4	0.0	2.4***	(0.4)
Avg. Weekly Khanmigo Chats	1.4	0.0	1.4***	(0.3)
Avg. Weekly Khanmigo Interactions	3.3	0.0	3.3***	(0.7)
Panel B: Term 2				
N	691	685		
Avg. Weekly Learning Time (mins)	21.9	0.3	21.7***	(3.1)
Avg. Weekly Skills Worked On	2.9	0.0	2.9***	(0.4)
Avg. Weekly Skills Leveled Up To Proficient	1.9	0.0	1.9***	(0.4)
Avg. Weekly Khanmigo Chats	1.2	0.0	1.2***	(0.2)
Avg. Weekly Khanmigo Interactions	3.9	0.0	3.9***	(0.8)
Panel C: Term 3				
N	680	622		
Avg. Weekly Learning Time (mins)	35.8	0.1	35.7***	(7.5)
Avg. Weekly Skills Worked On	4.9	0.0	4.9***	(1.0)
Avg. Weekly Skills Leveled Up To Proficient	3.5	0.0	3.5***	(0.7)
Avg. Weekly Khanmigo Chats	2.4	0.0	2.4***	(0.5)
Avg. Weekly Khanmigo Interactions	8.7	0.0	8.7***	(1.9)
Panel D: Term 4				
N	826	745		
Avg. Weekly Learning Time (mins)	35.0	0.3	34.7***	(4.1)
Avg. Weekly Skills Worked On	5.2	0.0	5.2***	(0.7)
Avg. Weekly Skills Leveled Up To Proficient	3.7	0.0	3.7***	(0.5)
Avg. Weekly Khanmigo Chats	2.0	0.0	1.9***	(0.2)
Avg. Weekly Khanmigo Interactions	8.1	0.0	8.1***	(1.0)

Notes: First-stage regressions of platform usage measures on treatment assignment. Standard errors clustered at the grade-within-school level.

A RTI Scheduling Diagnostics

A.1 Administrative scheduling process

The study population is students scheduled into RTI mathematics under pre-existing district assignment rules based on prior achievement and teacher referral. These scheduling rules were determined independently of treatment assignment and applied uniformly across treated and control grades. The analysis sample is defined using a once-in, always-in retention rule: once a student enters the district RTI placement process in any term, they are retained in the sample for all subsequent terms regardless of later scheduling status.

RTI scheduling is updated each term based on the same pre-specified district achievement rules applied uniformly across treatment and control grades. Scheduling decisions are made prior to the start of each term using pre-treatment screening data, and teachers making referrals are not informed of grade-level treatment assignments. Transitions into and out of scheduled RTI placement reflect the administrative district placement process and are expected to be balanced across arms.

All primary ITT specifications are defined within this administratively determined RTI population. Once students enter scheduled RTI mathematics placement, they remain in the analysis sample for all subsequent observed terms regardless of later RTI scheduling status.

A.2 Pre-specified balance diagnostics

Test 1: Scheduled RTI entry rates. We estimate a linear probability model for first scheduled RTI placement after Term 1 among students not yet scheduled for RTI at Term 1, with school, grade, and term-by-year fixed effects and standard errors clustered at the grade-within-school level. The treatment share among post-baseline entrants is approximately 48 percent, compared to 53 percent at baseline, consistent with balance.

Test 2: Entrant covariate balance. Among post-baseline scheduled entrants, we test whether baseline observable characteristics differ by treatment status, including prior MAP scores and demographic characteristics. Observable characteristics show some differences by treatment status, particularly in race and ethnicity composition.

Test 3: Baseline-defined cohort robustness. We estimate all primary specifications on a fixed cohort restricted to students first scheduled for RTI mathematics in Term 1 (Year 1), comprising 1,402 students with both MAP scores across 56 clusters. This cohort serves as a robustness sample and should not be interpreted as a preferred estimand.

A.3 East Lake Academy supplemental analysis

East Lake Academy is excluded from the primary analysis sample due to post-randomization changes in treated-grade composition: Grade 6 (the treated grade) enrolled additional students mid-implementation based on a teacher's assessment of platform efficiency, selectively expanding the treated group beyond the intended RTI-eligible population. Grades 7 and 8 served as control grades and their RTI rosters were unaffected.

To recover a valid treated sample for an exploratory analysis, we reconstruct the intended RTI-eligible population in Grade 6 using a pre-specified propensity score prediction approach. RTI scheduled placement is modeled as a function of baseline math and ELA screener scores and demographic characteristics (dyslexia diagnosis, SPED disability status, FARMS eligibility). Two logit models are estimated on the control grades and applied to Grade 6:

- **Model A:** Trained on Grade 7 control, validated on Grade 8 control, applied to Grade 6 treated.
- **Model B:** Trained on Grade 8 control, validated on Grade 7 control, applied to Grade 6 treated.

The reconstructed treated sample is the intersection of Grade 6 students predicted RTI-eligible by both models. This intersection criterion ensures that inclusion in the reconstructed sample is robust to the choice of training grade. The control grade requires no adjustment, as the original scheduling protocol was followed correctly there.

This comparison estimates the effect of KWiK relative to business-as-usual RTI instruction using a treated sample whose composition is validated against the observed eligibility rule in control grades. It is pre-specified as exploratory, reported separately from all primary results, and not used to inform the primary ITT estimand.

B Estimand Weighting and Additional Robustness Checks

B.1 Dynamic weighting in the stacked panel

The primary stacked OLS specification weights student-term observations rather than students equally. Students observed for more terms contribute more weight to the stacked ITT estimate. The estimand is closer to the average effect per student-term observation than to an equal-weighted student-level effect. This is not incorrect but affects interpretation, particularly if persistent RTI students systematically differ from students with shorter RTI spells.

To assess sensitivity to this weighting, we estimate a collapsed student-level cumulative gain specification:

$$\Delta MAP_i = MAP_{final,i} - MAP_{baseline,i} \tag{5}$$

where $MAP_{baseline,i}$ is the student's MAP score in their first observed RTI term and $MAP_{final,i}$ is their MAP score in their last observed RTI term. We then estimate:

$$\Delta MAP_i = \alpha + \beta D_i + \gamma' X_i + \theta_s + \phi_g + \varepsilon_i \tag{6}$$

This specification equal-weights students regardless of the number of terms observed, providing a direct robustness check on whether the stacked panel results are driven by disproportionate weighting of persistent RTI students. Convergence with the primary ITT estimate reduces concern about dynamic weighting.

B.2 Entry-timing cohort analysis

The dynamic RTI scheduling process means students enter the analysis sample at different points during the study. Under the study's once-in, always-in sample rule, students remain in the estimation sample for all subsequent observed terms after first scheduled RTI placement, regardless of later RTI scheduling status.

To assess whether treatment effects differ by entry timing, we partition students according to the term of first scheduled RTI placement and estimate separate ITT specifications within each entry cohort. For example, students first scheduled in Term 1 contribute outcomes through Terms 1–5, while students first scheduled in Term 2 contribute outcomes through Terms 2–5.

This decomposition assesses whether the primary pooled ITT estimate masks heterogeneity associated with timing of entry into the RTI population or changing cohort composition across terms. Results are interpreted as descriptive diagnostics rather than causal heterogeneity estimates.

B.3 Spillover and diffusion diagnostic

If KWiK affects teacher behavior or school-wide instructional norms, control-grade students in schools with more treated grades may experience indirect exposure. We test for spillovers by estimating:

$$Y_{igsty} = \alpha + \rho \text{ ShareTreated}_{sy} + \gamma' X_{igsty} + \theta_s + \phi_g + \lambda_{ty} + \varepsilon_{igsty} \tag{7}$$

where ShareTreated_{sy} is the fraction of grades assigned to KWiK in school s in year y, estimated on the control-grade sample only. This specification is a descriptive diagnostic and should not be interpreted as causal spillover estimation. In particular, with only one or two treated grades per school, ShareTreated_{sy} is mechanically tied to school-level treatment saturation and therefore cannot separately identify indirect exposure effects from other school-level differences associated with treatment intensity. Any estimated relationship should be interpreted as suggestive evidence consistent with spillovers or diffusion effects rather than a causal estimate of peer or teacher-mediated spillovers.

C Exposure Heterogeneity and Specification Validity

This appendix examines whether the baseline stacked ITT specification used in the main text is a valid reduced-form summary of potentially heterogeneous treatment exposure patterns over time. The goal is not to estimate alternative causal effects, but to assess whether pooling across exposure histories is consistent with the structure of the randomized design.

C.1 Exposure history framework

Although treatment assignment is fixed at the grade-within-school level, students may experience different exposure histories over the two-year study window due to cohort progression across grades and years.

We define four mutually exclusive exposure histories:

- (00): never assigned to KWiK
- (11): assigned to KWiK in both years
- (10): assigned in Year 1 only (formerly treated in Year 2)
- (01): assigned in Year 2 only (movers-in)

Let D^1_{igsty} and D^2_{igsty} denote assignment in Year 1 and Year 2 respectively. These exposure histories are determined by the random assignment of grade-level treatment and the fixed grade structure over time.

C.2 Saturated exposure specification

We estimate the following fully flexible exposure model:

$$Y_{igsty} = \alpha + \beta_1 D^1_{igsty} + \beta_2 D^2_{igsty} + \beta_3 (D^1_{igsty} \times D^2_{igsty}) + \gamma' X_{igsty} + \theta_s + \phi_g + \lambda_{ty} + \varepsilon_{igsty} \tag{8}$$

This specification allows outcomes to vary arbitrarily across exposure histories without imposing additivity across years of assignment. The interaction term β_3 captures deviations from additivity in exposure effects across the two-year treatment horizon.

C.3 Additivity of exposure effects

The main specification in the paper implicitly assumes that the effect of exposure in multiple years is additive in levels. This corresponds to the restriction:

$$H_0 : \beta_3 = 0 \tag{9}$$

Under this restriction, the stacked ITT estimator in the main text can be interpreted as a pooled average effect of assignment to a KWiK-enabled grade, independent of exposure history. We test this restriction directly in the saturated model as a pre-specified diagnostic. Failure to reject is consistent with the validity of pooling exposure histories in the main specification, though this check is not a structural proof of additivity and should be interpreted accordingly.

C.4 Persistence after exposure

To assess whether treatment effects persist after active exposure ends, we compare students formerly assigned to KWiK in Year 1 but not in Year 2 to students never assigned to KWiK.

We estimate:

$$Y_{igsty} = \alpha + \delta \mathbf{1}\{D^1_{igsty} = 1, D^2_{igsty} = 0\} + \gamma' X_{igsty} + \theta_s + \phi_g + \lambda_{ty} + \varepsilon_{igsty} \tag{10}$$

The coefficient δ captures reduced-form differences in outcomes for formerly exposed students relative to never-treated students, conditional on baseline covariates and fixed effects. This parameter should not be interpreted as a structural decay process.

C.5 Interpretation and relationship to main specification

The ITT estimand remains well-defined as long as assignment is randomized; exposure heterogeneity affects interpretation of the pooled estimate but not its causal validity. The specifications in this appendix do not replace the main ITT estimator. Instead, they serve as diagnostic checks on the validity of pooling exposure histories in a single stacked regression.

The key question is whether a parsimonious additive representation provides a sufficient summary of the randomized exposure structure. If the additivity restriction in Section B.3 is not rejected and persistence effects in Section B.4 are small relative to sampling uncertainty, then the main specification provides a reasonable reduced-form approximation of the treatment effect of assignment to KWiK-enabled RTI instruction.

If these restrictions fail, the main ITT estimator remains valid as an average effect of assignment, but its interpretation as a stable pooled effect across exposure histories would be limited, and heterogeneous exposure-specific effects become the appropriate objects of inference.