

Evaluating Goblins: A Quasi-Experimental Study of an AI-Powered Math Tutoring Tool

Amber Wang, Ph.D.
Leanlab Education

Executive Summary

In spring 2026, Leanlab Education conducted a quasi-experimental design (QED) study evaluating the impact of Goblins—a web-based AI math tutoring platform—on student outcomes in Grades 8 and 9. The study took place across one high school and two middle schools within a single school district over the course of the spring semester. The math assessment analytic sample comprised 127 students (treatment: $n = 96$; control: $n = 31$) across 8 teachers; the post-survey sample comprised 105 students (treatment: $n = 64$; control: $n = 41$). Quantitative analyses examined student math achievement and math-related attitudes and beliefs; qualitative data were collected through a teacher midpoint survey and end-of-study focus groups.

Key Findings:

- **Student Math Achievement:** Treatment students showed a favorable but statistically non-significant adjusted difference in posttest Regents scores compared to control students (2.08 points; Hedges' $g = 0.20$). Though non-significant, an effect of this magnitude may still be educationally meaningful, particularly in the context of a relatively brief study period.
Student Math-Related Attitudes and Beliefs: No statistically significant differences emerged between treatment and control students on any of the eight survey subscales, at either the conventional or Bonferroni-corrected threshold. Data collection limitations prevented a pre-to-post comparison, so this study cannot speak to whether attitudes changed over time.
- **Teacher Perceptions and Implementation:** Teachers generally rated Goblins as fitting well into their instructional routines and valued its data for targeting instructional support, though device access limitations, technical reliability

issues, and variation in implementation fidelity shaped how consistently the tool was used.

Given baseline non-equivalence between conditions, limited statistical precision due in part to the small number of teacher clusters, and the other data limitations described in this report, achievement findings should be interpreted with caution. Findings offer preliminary signals that could inform a future study with a larger sample and more consistent implementation.

Overview

In spring 2026, Leanlab Education conducted a quasi-experimental design study evaluating the impact of Goblins on math student outcomes. This research study is the final phase of a multiphase research project in partnership with NewSchools Venture Fund:

1. Late summer/early fall 2025: Baseline evaluation of Goblins using an LLM-as-a-judge benchmarking approach (The Learning Agency) and research-ready audit (Leanlab Education)
2. Fall 2025: Rapid cycle evaluation usability study
3. Spring 2026: Quasi-experimental design impact study and final benchmarking

About the Tool

Goblins is a web-based AI math tutoring platform that allows students to solve problems on a digital canvas while writing, drawing, and explaining their reasoning aloud. The platform is designed to identify where mistakes or misconceptions arise in real time and respond to students with Socratic questions, brief explanations, and visual supports rather than simply providing the answer. Teachers can select from thousands of standards-aligned skills and pre-made assignments or upload their own worksheets. The teacher-facing dashboard provides information about students' problem-solving processes, including formative and error-analysis insights, conversation transcripts, session replays, standards tracking, and recommendations for small-group instruction or intervention. The platform supports spoken interaction in multiple languages. Goblins states that it complies with COPPA, FERPA, and student data privacy laws in all 50 states. This evaluation focused on its use in Grade 8 and 9 math tutoring instruction.

Product Change Recommendations from the RCE

Prior to the spring 2026 pilot QED study, Leanlab conducted a three-cycle Rapid Cycle Evaluation (RCE) with five teachers between October and December 2025. Across cycles, teachers raised concerns about the usability of trackpad-based writing and drawing input, an unreliable microphone feature, gaps in content

coverage for certain topics, and a desire for stronger differentiation and scaffolding for struggling students.

Goblins made iterative improvements between cycles, including a new assignment-creation and upload workflow, a live view of student engagement, and avatar customization. However, the microphone feature remained non-functional across all three cycles, and Leanlab's RCE report flagged fixing it as a required product update before the spring 2026 pilot; other recommendations— including lightweight gamification, the ability to edit assignments after creation, and expanded file-upload and Google Drive integration—were categorized as recommended future enhancements rather than prerequisites for the pilot. By the start of the spring pilot, the microphone issue was substantially resolved, though one treatment teacher continued to report intermittent issues during the study period (see Qualitative Findings).

Research Questions

1. To what extent do students who use Goblins differ in mathematics achievement at posttest from students who do not use Goblins, after accounting for pretest achievement?
2. How do students who use Goblins compare to students who do not use Goblins on math-related attitudes and beliefs—including self-confidence, positive and negative valence toward math, math anxiety, math self-efficacy, intrinsic motivation, and liking of math—at the end of the study period?
3. How do teachers experience implementing Goblins in their classrooms, and what factors support or challenge its effective use?

Methodology

Sample

The spring 2026 evaluation began with 11 participating educators—6 assigned to the treatment group and 5 to the control group—within a single school district in New

York. Participants included both classroom teachers and administrators involved in Algebra I instruction or tutoring. For simplicity, the report refers to all participating educators as teachers throughout. Of the original 11 participants, 3 (1 treatment and 2 control) did not continue in the study because they were ultimately not assigned a student group or faced availability constraints. The remaining 8 participants—5 treatment and 3 control—contributed complete student mathematics assessment data and constituted the analytic sample (**Table 1**), which spanned one high school and two middle schools within the district. The treatment group consisted of four classroom teachers and one administrator, whereas all three control participants were administrators.

Table 1: Final Teacher Participants

Teacher ID	Condition	Grade(s) Taught for Study	Teacher's District Role
1	Treatment	9	Teacher
2	Treatment	9	Teacher
3	Treatment	9	Teacher
4	Treatment	8	Administrator
5	Treatment	8	Teacher
6	Control	9	Administrator
7	Control	8	Administrator
8	Control	8	Administrator

The **student math assessment analytic sample** comprised 127 students (treatment: $n = 96$; control: $n = 31$) with complete pretest and posttest data, drawn from the 8 teachers described above (see **Table 2** in Results for sample size by grade level). Thirty-one students (24%) had an individualized education program (IEP; treatment: $n = 29$; control: $n = 2$). Among Grade 9 students, all 24 students with IEPs were in the treatment condition. Among Grade 8 students, 5 students with IEPs were in the treatment condition and 2 were in the control condition.

The **student survey analytic sample** comprised 105 students at post-survey (treatment, $n = 64$; control, $n = 41$), slightly smaller than the math assessment analytic sample.

Participant Onboarding

Prior to the start of the study, all participating teachers attended an onboarding session covering the study's goals, timeline, and their roles and responsibilities. To accommodate scheduling constraints, teachers had the option to attend a live session or watch an asynchronous recording. Treatment teachers additionally had access to Goblins onboarding videos to familiarize themselves with the platform before implementation.

Measures and Data Collection

Math Assessment

Pretest math achievement scores were drawn from two of the district's Regents-aligned interim assessments, administered to students in Grades 8 and 9 across the school year and designed to closely resemble the length and content of the actual Algebra I Regents exam. Grade 8 students' pretest scores were drawn from one such interim assessment, administered in late January 2026. Grade 9 students' pretest scores were drawn from a different assessment, also administered in late January. Posttest math achievement was measured using the Algebra I Regents exam itself—New York State's end-of-course standardized exam for Algebra I, administered in June, independently of the district's interim assessment series.

Because the Grade 8 and Grade 9 pretest instruments are related but distinct measures without formal vertical equating, pretest percent-correct scores were standardized within grade and assessment form prior to analysis.

Student Survey

A 34-item student survey was administered online at the beginning and end of the study period¹ to assess math-related attitudes and beliefs. The survey assessed eight constructs drawn from three validated instruments (see **Appendix A**):

¹ Pre-survey data was collected, but not used in analysis (see Analytic Approach).

Self-Confidence (6 items, 1–5 scale), Positive Valence (4 items, 1–5 scale), Negative Valence (6 items, 1–5 scale, reverse-scored), Time (2 items, 1–5 scale), Math Anxiety (4 items, 1–6 scale, reverse-scored), Math Self-Efficacy (4 items, 1–6 scale), Intrinsic Motivation (4 items, 1–6 scale), and Liking of Math (4 items, 1–6 scale).

Internal consistency reliability at post-survey ranged from $\alpha = .22$ to $.92$ across the eight subscales. Seven subscales met or exceeded the conventional threshold ($\alpha \geq .70$); Time fell substantially below it ($\alpha = .22$), reflecting that only 2 of the source instrument's original 7 items were included in this survey.

Teacher Midpoint Survey

A brief midpoint survey was administered online to treatment teachers at approximately the halfway point of the study period to gather early implementation feedback on their use of Goblins. A parallel survey was also administered to control teachers to gather information about their instructional context and business-as-usual math instruction during the study period. See **Appendix B**.

Teacher Focus Groups

Qualitative data were collected from treatment teachers at the end of the study period using a structured question guide (see **Appendix C**) covering teachers' experiences implementing Goblins, including perceived impacts on student learning, engagement, and mathematical reasoning, as well as implementation challenges and recommendations. Two treatment teachers participated in individual live sessions with a Leanlab researcher; the remaining three teachers were unable to attend a live session and completed the same guide as a written questionnaire. Data were analyzed thematically and are reported in the Qualitative Findings section.

Implementation Timeline

The study was intended to run from late February through the end of April 2026. In practice, implementation start and end dates varied by teacher. Some treatment teachers did not begin using Goblins until May, due to the time the school needed to identify and assign student tutoring groups. This same school-level assignment process left one treatment teacher with a notably shortened implementation period—approximately two weeks—a deviation discussed further as a limitation.

Some control teachers continued their business-as-usual instruction through mid-June, later than the study's intended end date.

Despite this variation in implementation timing, the sequencing of the math assessment data held consistently across teachers: pretest data were drawn from district interim assessments administered in late January, before any teacher's implementation period began, and posttest data came from the Algebra I Regents exam, administered statewide on June 17, after every teacher's implementation period had ended.

Implementation Context

The primary intent of Goblins use in this study was preparation for the Algebra I Regents exam. Goblins was used primarily during intervention blocks or designated independent practice time. One teacher incorporated it as a whole-class practice activity, selecting a set of priority standards with identified gaps and assigning a target number of problems each class period. Another teacher incorporated it as homework to supplement in-class Algebra I instruction.

Teachers' existing curricula varied, with Illustrative Mathematics commonly reported; one participating teacher did not use a separate core curriculum for their intervention group. Teachers supplemented Goblins with a range of other tools, including Khan Academy, Desmos, Amplify, Zearn, DeltaMath, and traditional materials such as graphing paper and calculators. When asked how well Goblins fit their existing curriculum and instructional routines, treatment teachers' ratings skewed toward the higher end: 3 of 5 rated the fit as *very well*, 1 as *moderately well*, and 1 as *slightly well*; none rated the fit as *not at all well* or *extremely well*.

One control teacher reported having used Goblins school-wide earlier in the year, before the start of the study, but discontinued use due to shifting school priorities during that time. This teacher's control-group instruction during the study period reflected business-as-usual practice without active Goblins use. However, the prior exposure is noted as a potential contamination consideration (see Limitations).

Dosage Recommendations

Goblins' recommended dosage is 20 minutes, 2–3 times per week; teachers were asked to aim for this benchmark for the study. Teacher-reported usage from the midpoint survey varied considerably relative to this benchmark. One treatment teacher reported using Goblins once per week, below the recommended frequency, while another reported 5 or more sessions per week. The remaining three treatment teachers reported approximately 2 sessions per week, within or near the lower end of the recommended range, assuming session length approximately matched the recommended 20 minutes.

Platform usage data showed considerable variability across teachers, generally consistent with the self-reported patterns from the midpoint survey and with known differences in implementation timing—most notably, one teacher's compressed engagement window aligned with a much lower overall volume of platform activity compared to teachers with a full implementation period. However, this usage data reflected all Goblins activity at the participating schools rather than activity limited to study participants, since the tool was used more broadly across the school beyond this study; for this reason, platform usage data are described only in general terms here rather than as precise, sample-specific dosage figures.

Analytic Approach

Math Assessment

The primary analysis examined the effect of Goblins on students' Algebra I Regents posttest scores using a multilevel model that accounted for students being grouped within teachers, the lowest-level clustering unit available in the analytic data. Because Grade 8 and Grade 9 students completed different, non-vertically-equated pretest instruments (see Measures and Data Collection), pretest percent-correct scores were standardized (z-scored) within grade and assessment form prior to analysis; the Regents posttest, a scale score, was analyzed on its original metric.

The analytic model compared treatment and control students' posttest scores after adjusting for pretest performance and grade level. The model was estimated using restricted maximum likelihood (REML), with Kenward–Roger small-sample

corrections applied to the significance tests and confidence intervals because only eight teacher clusters were available. Effect sizes were measured using Hedges' g .²

Student Survey

Prior to analysis, negatively worded items were reverse-coded so that higher scores consistently reflected more positive responses within each subscale. Post-survey data were analyzed using independent-samples t -tests comparing treatment and control students on each of the eight subscales. To account for the eight simultaneous comparisons, a Bonferroni-corrected significance threshold of $p < .00625$ was applied.

A pre-survey using the same items was also administered at the start of the study period. However, teachers could not be confidently identified for a large portion of control student responses due to technical issues. As a result, pre-survey data were not analyzed, and no pre-to-post comparison is reported.

Qualitative Data

Teacher midpoint survey responses and focus group data were analyzed to capture implementation experiences and perceptions of Goblines across the study period. Midpoint survey responses were examined descriptively to identify early patterns in implementation feedback. Focus group data were analyzed thematically; themes were identified inductively across all responses, regardless of whether they were collected through a live session or a written questionnaire, and represent cross-cutting patterns in teachers' reported experiences with Goblines. Quotes are drawn from the midpoint survey, focus group sessions, and written questionnaire responses, and have been lightly edited to remove filler words for readability.

Results

Math Assessment

Preliminary Analyses

² Hedges' g is a standardized measure of the difference between two groups. It expresses the difference in standard deviation units and includes a correction to improve accuracy when sample sizes are small.

Treatment and control students differed significantly at baseline on standardized pretest scores overall (Hedges' $g = -0.53$, treatment lower; $p < .05$; **Table 3**). This difference exceeded the What Works Clearinghouse (WWC) baseline equivalence threshold for differences that can be addressed through statistical adjustment,³ limiting the strength of evidence this study can support (see Limitations). This overall gap was driven primarily by Grade 8 students, who differed substantially at baseline ($g = -1.13$, $p < .001$).

The intraclass correlation coefficient (ICC) indicated that approximately 16% of the variation in posttest scores was attributable to differences between teachers (95% CI [.037, .482]), a wide interval reflecting the small number of teacher clusters.

Table 3: Pretest Scores by Condition and Grade (N = 127)

Grade/Sample	Treatment		Control		Hedges' g
	n	$M (SD)$	n	$M (SD)$	
Grade 8	21	32.30 (14.44)	27	52.09 (19.03)	-1.13
Grade 9	75	46.57 (18.60)	4	47.87 (12.96)	-0.07
Standardized pretest (overall)	96	-0.13 (0.99)	31	0.39 (0.94)	-0.53

Note. Grade 8 and Grade 9 students completed different Regents-aligned pretest assessments. Original percent-correct scores are therefore reported separately by grade and assessment form. For the pooled baseline-equivalence analysis, pretest scores were standardized within grade and assessment form across conditions. Hedges' g was calculated as treatment minus control; negative effect sizes indicate lower baseline performance in the treatment group.

Treatment Effects

Regents posttest means by condition are shown in **Table 4**.

³ What Works Clearinghouse. (2022). *What Works Clearinghouse procedures and standards handbook, Version 5.0*. Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance (NCEE). This report is available on the What Works Clearinghouse website at <https://ies.ed.gov/ncee/wwc/Handbooks>.

Table 4: Posttest Regents Scores by Condition and Grade (N = 127)

Grade/Sample	Treatment		Control	
	<i>n</i>	<i>M (SD)</i>	<i>n</i>	<i>M (SD)</i>
Grade 8	21	64.86 (11.96)	27	69.59 (9.90)
Grade 9	75	67.45 (9.66)	4	65.00 (13.29)
Overall	96	66.89 (10.20)	31	69.00 (10.25)

Note. Grade-level descriptives are provided for transparency only. The primary model's treatment effect estimate is based on the pooled sample.

The primary model estimated an adjusted posttest difference between treatment and control students of 2.08 points, favoring the treatment group, after accounting for pretest performance and grade level ($SE = 2.86$; $t(6.78) = 0.73$; $p = .491$; 95% CI $[-4.72, 8.88]$; Hedges' $g = 0.20$).⁴ This difference was not statistically significant.⁵

Student Survey

Independent-samples *t*-tests comparing treatment and control students at post-survey found no statistically significant differences on any of the eight subscales, whether evaluated at the conventional $\alpha = .05$ threshold or the Bonferroni-corrected threshold of $\alpha = .00625$ applied for multiple comparisons (p -values ranged from .30 to .92 across constructs). Post-survey subscale scores by condition are shown in **Table 5**.

⁴ Hedges' g was calculated using the model-adjusted mean difference as the numerator and the unadjusted pooled posttest standard deviation as the denominator, with a small-sample correction applied.

⁵ Because IEP status differed substantially between conditions, a supplementary model included IEP status as an additional covariate to assess the robustness of the primary finding. The estimated treatment-control difference remained non-significant and consistent with the primary model ($b = 2.39$, $SE = 2.87$, $t = 0.83$, $p = .432$, 95% CI $[-4.43, 9.21]$).

Table 5: Post-Survey Subscale Scores by Condition

Subscale	Scale Range	Treatment <i>n</i> = 64 Post <i>M</i> (<i>SD</i>)	Control <i>n</i> = 41 Post <i>M</i> (<i>SD</i>)
Self-Confidence	1–5	3.11 (0.95)	3.29 (0.78)
Positive Valence	1–5	3.20 (0.96)	3.21 (0.82)
Negative Valence (reversed)	1–5	3.15 (1.01)	3.20 (0.94)
Time	1–5	2.41 (1.01)	2.35 (0.97)
Math Anxiety (reversed)	1–6	3.70 (1.39)	3.82 (1.32)
Math Self-Efficacy	1–6	3.55 (1.27)	3.63 (1.27)
Intrinsic Motivation	1–6	3.77 (1.17)	3.73 (1.17)
Liking of Math	1–6	3.52 (1.48)	3.37 (1.36)

Note. Negative Valence and Math Anxiety were reverse-scored so that higher values indicate more favorable outcomes across all subscales, consistent with the direction of the other constructs.

Teacher Perceptions and Implementation Experiences: Qualitative Findings

Qualitative data were collected through a teacher midpoint survey and end-of-study focus groups. Five treatment teachers completed the midpoint survey at approximately the halfway point of the study period; at the end of the study, two treatment teachers participated in individual live sessions and three completed the same structured guide as a written questionnaire. Findings from both data sources are integrated and organized into five cross-cutting themes that reflect teachers' perceptions and experiences implementing Goblines across the study period.

Theme 1: Teachers Identified Device Access and Tutoring Structure as Implementation Factors

Teachers identified two implementation conditions beyond the platform itself that affected their experiences with Goblins: access to appropriate input devices and the structure of the tutoring schedule.

Four of five teachers described input-device limitations that affected students' ease of use, pace, or engagement. Writing or graphing with a Chromebook trackpad could be slow and frustrating, while supplemental trackpads, styluses, mice, or tablets made it easier for students to show their work. One teacher reported that initially low engagement increased substantially after students received plug-in trackpads, while another observed that students worked more quickly once pads and styluses were available.

One teacher also identified tutoring structure as a possible factor. After implementing Goblins during an intensive 10-day block, she believed that shorter, recurring sessions over a longer period may have supported greater growth. She contrasted her approach with two colleagues' more sustained tutoring models, although she acknowledged having limited information about exactly how they used Goblins. The available data therefore do not establish that schedule structure itself produced differences in outcomes.

Together, these findings suggest that some implementation challenges may have reflected hardware and scheduling conditions, rather than the platform alone.

"Engagement was low in the beginning when students did not have access to a touchscreen or additional plug-in to their computer to help them show work on the screen. Their Chromebooks alone are not enough to fully engage with the Goblins platform. After we ordered the plug-in trackpads, engagement increased exponentially." — Teacher 1, Focus Group.

"At first, I noticed students were frustrated with the slow pace in terms of writing on the laptop and having to get used to the pace of interaction with the AI guidance; but especially with pad and stylus pen, they picked up a pace and also getting the hang of showing each step of the work in order to proceed." — Teacher 2, Focus Group.

"I would approach Goblins not as intensive tutoring, but instead as a tutoring platform that you use over time. Because what I'm noticing is [...] their tutoring sessions, which was over time, had more growth." — Teacher 4, Focus Group.

Theme 2: Teachers Identified Motivation, Proficiency, and AI Tone as Engagement Factors

Teachers attributed variation in student engagement with Goblins to students' motivation, proficiency level, and reactions to the AI's conversational style. Across accounts, however, no consistent profile emerged of the students who engaged with or used Goblins most frequently.

One teacher believed self-starters and intrinsically motivated students were more likely to benefit from the platform. Another reported greater engagement among both higher-performing and below-grade-level students, while students performing in the middle engaged less—a pattern she attributed to students' stamina for independent work rather than to the platform itself. A third teacher reported the opposite pattern among higher-performing students: because they needed less intervention, they used Goblins less often, while students near or below proficiency used it more consistently and, in his view, benefited most.

Teachers also described mixed reactions to the AI's informal style. One teacher said some students found the casual, "Gen Z" language and avatar off-putting, describing the AI as trying too hard to sound relatable. Another reported that his students especially enjoyed the AI-driven feedback and the platform's relatable language and terminology. These accounts suggest that the AI's style may resonate differently across student groups and contexts.

"Many of my students are work avoidant in general. However, I've seen an increase in student usage from my higher performers and those that are below grade level. Students performing in the middle of the pack aren't engaging with the platform as much but that's not due to the accessibility of the platform but rather our student's stamina to complete work outside of class." — Teacher 1, Focus Group.

"Scholars performing at higher proficiency levels generally used the platform less frequently because they required less targeted intervention. Scholars who were

considered 'cusp' scholars or performing below proficiency benefited from the platform the most and utilized it more consistently." — Teacher 5, Focus Group.

"My students said they didn't like the Gen Z language [...] I asked why, and they're like, 'oh, because the robot is trying too hard.'" — Teacher 4, Focus Group.

"They especially enjoyed the AI-driven feedback and the relatable language and terminology the platform used." — Teacher 5, Focus Group.

Theme 3: Teachers Valued Work-Showing but Wanted Greater Flexibility and Assessment Alignment

Teachers generally valued Goblins' requirement that students show their work, noting that it encouraged students to follow mathematical procedures rather than jump directly to an answer. One teacher agreed that explanation was valuable but found repeated explanations less practical during an intensive 10-day Regents-preparation block, where they slowed instructional pace and created an additional challenge for English language learners. Her account suggests that the appropriate level of explanation may depend on the instructional context.

Three teachers independently recommended allowing students to photograph or scan paper-based work for AI feedback. This option could preserve the platform's emphasis on mathematical reasoning while reducing the time and difficulty involved in reproducing work on-screen.

Teachers also differed in their views of how well the problem bank aligned with high-stakes assessments. One described the questions as well aligned with Regents and Illustrative Mathematics. At the same time, another found the problems less rigorous than Regents items, supplemented Goblins with released exam questions, and wanted more conceptual and application-based problems rather than predominantly procedural ones. These differing perspectives may reflect variation in grade level, instructional goals, or how teachers used the platform. A separate teacher reported that students sometimes received repeated problems, raising a related concern about content variety and differentiation.

"Students were frustrated that they couldn't just list the answer and move on without showing work. As a teacher, I love that they cannot move on without showing work." — Teacher 1, Focus Group.

"If I'm talking about the beginning of the year [...] I want them to explain. [...] But now, when I'm talking about intensive tutoring [...] I need to get through more material. And the 'why' is important, but I only need it once, and that's it; and then let them power through some problems." — Teacher 4, Focus Group.

"I would add a feature that allows scholars to scan or upload paper-based work and receive AI-generated feedback directly from their written solutions." — Teacher 5, Focus Group.

"I had to mix in Regents problems, because the Goblins word problems were not as rigorous. [...] It didn't match up with what they would typically see on the Regents." — Teacher 4, Focus Group.

"I like the problems Goblins comes up with for practice problems that seem procedural focused, but I think it could expand the types of problems it creates and include more variety, such as conceptual and application-based questions." — Teacher 4, Midpoint Survey.

Theme 4: Teachers Used Platform Data and Feedback to Target Instructional Support

Teachers described using Goblins-generated data and feedback to target instruction and concentrate their attention on students who needed more support. They generally viewed the platform as complementing, rather than replacing, the teacher's role.

Three teachers described using platform information to inform instruction or student support. One used identified procedural misconceptions to target small-group reteaching. In contrast, another used weekly completion, mastery, and student-level gap data to tailor subsequent lessons and prioritize students with greater needs during independent practice. A third used weekly data to check in with students about their progress and behavior patterns.

Teachers also described Goblins' individualized feedback as allowing them to redistribute their attention. One likened the platform to an assistant teacher that could support self-motivated students while she circulated and worked with others. Another similarly reported that students' ability to work independently with the AI allowed him to devote more time to small groups and students requiring intensive support.

One teacher offered a more cautious perspective, emphasizing that Goblins worked best as a supplement to a teacher who could provide alternative explanations or representations when the AI's approach did not meet a student's needs. Together, these accounts suggest that teachers perceived Goblins as helping them target support more efficiently, although the study did not directly measure changes in how they allocated instructional time.

"One of the most valuable aspects of Goblins was its ability to highlight procedural misconceptions. I was able to identify exactly where scholars struggled within multi-step processes and use that information to provide more targeted small-group instruction and reteaching." — Teacher 5, Focus Group.

"It gave me weekly, live data on who is completing work and their level of mastery. [...] It gave me and my teachers specific gaps that it noticed with student examples. It helped me make my content more catered when they were in class the following day." — Teacher 1, Focus Group.

"Students' individual feedback to what they wrote is the power in AI. It is as if I had an assistant teacher with me. And while I can circulate and give meaningful feedback to a course of students, Goblins helps another subset of students who are self-motivators." — Teacher 4, Focus Group.

"I really think Goblins is a great program if you have a physical teacher in the room. [...] I'm not sure if the AI is good at finding alternate routes towards an answer." — Teacher 3, Focus Group.

Theme 5: Teachers Reported Technical and Data-Reliability Issues

Four of five teachers described apparent platform or feature-reliability issues, including lost student work, loading failures, inconsistent microphone functionality, and correct answers being marked incorrect.

Two teachers reported that student work or progress was sometimes lost. One said completed assignments occasionally failed to save, requiring students to redo their work. Another described reports from multiple students that their work had disappeared, including one student whose assignment was marked complete after he worked on it at home with a parent but was missing when he arrived at school.

Other concerns involved specific platform functions. One teacher reported repeated loading failures that required a full-page refresh. Another said some students' microphones worked while others' did not, despite students using the same Chromebooks with the same permissions; she was unable to identify a cause. She also relayed student feedback suggesting that the feature for photographing written work did not function effectively. A separate teacher reported occasional instances in which the system marked correct student answers as incorrect.

Not all technical friction was necessarily attributable to the platform. One teacher said the issues observed in that classroom were related more to students' handling of their laptops than to Goblins itself. Together, these accounts indicate that technical and data-reliability issues arose across several implementations, while suggesting that platform behavior, device use, and local setup may have contributed differently across contexts.

"Goblins would often have difficulty loading at times. I had to refresh the whole page to make it load on several occasions." — Teacher 1, Focus Group.

"I had a kid tell me that he did his homework with his parents, and, like, Goblin said that it was complete, but when he got to school, his data was lost. — Teacher 3, Focus Group.

"Some people said their microphones worked, some people said that they didn't. [...] All the students had the same Chromebook. [...] They have all the same permissions." — Teacher 4, Focus Group.

“There were occasional moments when scholars entered correct answers that were incorrectly marked wrong by the system.” — Teacher 5, Focus Group.

Discussion

This study examined the relationship between Goblins use and students' math achievement and math-related attitudes and beliefs over the course of a spring 2026 pilot, while also documenting teachers' experiences implementing the tool as a tutoring supplement across a single district. Taken together, findings offer preliminary, non-significant evidence regarding Goblins' relationship to math achievement and attitudes, alongside a richer set of implementation findings that point to specific conditions under which the tool appeared to work well or less well.

Math Achievement

Goblins' theory of change anticipates that students working through problems on the platform—supported by Socratic-style AI questioning that identifies misconceptions and builds conceptual understanding—will show short-term gains in math proficiency. The primary analysis found a favorable but non-significant adjusted difference in posttest Regents scores between treatment and control students (2.08 points; Hedges' $g = 0.20$). In educational contexts, an effect of this magnitude can still be substantively meaningful, particularly over a relatively short implementation period. However, there was considerable uncertainty around the estimate, reflecting in part the analysis's reliance on only eight teacher clusters. This finding neither clearly supports nor contradicts the theory of change's proficiency pathway and should also be interpreted alongside the baseline non-equivalence between treatment and control students and the considerable variation in how consistently teachers used Goblins relative to the recommended dosage.

A replication study with a larger sample and more teacher clusters, along with stronger implementation fidelity—for example, more consistent usage in line with the recommended dosage across all participating treatment teachers—would provide a more precise test of Goblins' effect on student achievement. Qualitative evidence offers some corroborating context for how the tool's intended mechanism operated in practice: several teachers described using Goblins' dashboard data to target

instruction toward students with the greatest need (Qualitative Findings, Theme 4), a pattern consistent with the theory of change's assumption that real-time insight into student struggles would inform teachers' instructional decisions. At the same time, device limitations and technical reliability issues (Themes 1 and 5) may have introduced friction into this intended pathway for at least some students and classrooms.

Math-Related Attitudes and Beliefs

Goblins' theory of change anticipates that AI-driven feedback and Socratic-style engagement will initially strengthen students' confidence in mathematics and their perceptions of themselves as capable math learners, potentially contributing over time to greater interest in and motivation to pursue mathematics. At the end of the study, however, treatment and control students reported similar scores across all eight mathematics-related attitude and belief constructs, with no statistically significant differences at either the conventional or Bonferroni-corrected threshold. Because a group-level pre-survey comparison could not be conducted (see Analytic Approach), the study cannot determine whether students' attitudes changed over the course of the study or whether the groups differed before implementation. The findings indicate only that treatment and control students reported similar attitudes and beliefs at the end of the study period.

Implementation and Implications for Practice

Teachers generally rated Goblins as fitting well into their existing instructional routines, with fit ratings skewing toward the higher end of the scale, and several described the platform's data and feedback as helping them target instructional support more efficiently. Consistent with the theory of change's teacher-dashboard pathway, teachers described using Goblins' insights to identify procedural misconceptions, tailor small-group reteaching, and redirect their attention toward students who needed the most support, generally describing the platform as complementing rather than replacing their own instruction.

At the same time, several implementation conditions appeared to shape how well this worked in practice, including device access limitations, recurring technical and data-reliability issues, mixed student reception of the AI's

conversational style, and differing perceptions of how well the problem bank aligned with Regents-level rigor (Themes 1, 3, and 5). One teacher's very short implementation window may have further limited how fully she could evaluate the tool's effect on her students.

Together, these findings suggest that Goblins was generally well received by teachers as a complement to their own instruction, particularly for using platform data to target support, while surfacing implementation considerations related to device access, technical reliability, and problem-bank alignment with high-stakes assessments. Future studies with more consistent implementation fidelity would help clarify whether Goblins' relationship to student math achievement and attitudes reflects a genuine and generalizable effect.

Limitations

Several limitations should be considered when interpreting the findings from this study.

First, treatment and control students were not equivalent at baseline on standardized pretest scores, particularly among Grade 8 students, exceeding the WWC threshold for differences addressable through statistical adjustment. The conditions also differed in grade composition, with most treatment students in Grade 9 and most control students in Grade 8. The treatment group was composed primarily of classroom teachers, while all control participants were administrators—an additional asymmetry between conditions. The primary model's treatment effect estimate also relied on only 8 teacher-level clusters, further limiting its precision. Given these concerns, math achievement findings should be interpreted with caution.

Second, Grade 8 and Grade 9 students completed different pretest instruments without formal vertical equating, and pretest and posttest scores were reported on different metrics (percent-correct versus Regents scale score). Standardizing pretest scores within grade and assessment form addressed differences in scale but did not

fully equate the two instruments' content, which may affect the comparability of pretest scores across grades.

Third, implementation fidelity varied across teachers. Reported and platform-observed usage fell short of Goblins' recommended dosage for some teachers, and implementation start and end dates varied, including one treatment teacher whose implementation period was compressed to approximately two weeks. Additionally, one control teacher reported prior exposure to Goblins earlier in the school year, before the study began, which may have influenced this teacher's business-as-usual instruction during the study period.

Finally, the student survey analyses were limited by data collection issues that prevented a group-level pre-to-post comparison; teachers could not be confidently identified for a large portion of control student pre-survey responses. As a result, only post-survey group comparisons are reported, and these do not control for baseline differences. One survey subscale, Time, also fell below the conventional threshold for internal consistency reliability ($\alpha = .22$), limiting confidence in findings specific to that construct.

Appendices

Appendix A: Student Survey

The student survey comprised 34 items assessing eight constructs related to math attitudes and beliefs. Items were selected and adapted from validated scales; not all items from the original instruments were included. Items are organized below by construct. Asterisks (*) indicate items that were reverse-coded before analysis so that higher scores consistently reflect more positive attitudes across all subscales.

Attitudes Toward Mathematics (ATMI): Self-Confidence Subscale

Tapia, M., Marsh, G. (2000). *Attitudes Toward Mathematics Instrument: An Investigation with Middle School Students*. Paper presented at the Annual Meeting of the Mid-South Educational Research Association (Bowling Green, KY, November 15-17, 2000).

Response scale: 1–5

Self-Confidence (6 items)

1. I have a lot of self-confidence when it comes to mathematics.
 2. Studying mathematics makes me feel nervous.*
 3. My mind goes blank and I am unable to think clearly when working with mathematics.*
 4. I learn mathematics easily.
 5. I am confident that I could learn advanced mathematics.
 6. When I hear the word mathematics, I have a feeling of dislike.*
-

Mathematics Interest Inventory (MII)

Stevens, T., & Olivárez, A. (2005). Development and evaluation of the Mathematics Interest Inventory. *Measurement and Evaluation in Counseling and Development*, 38(3), 141–152. <https://doi.org/10.1080/07481756.2005.11909775>

Wei, T., Chesnut, S. R., Barnard-Brak, L., Stevens, T., & Olivárez, A. (2014). Evaluating the Mathematics Interest Inventory using Item Response Theory: Differential item functioning across gender and ethnicities. *Journal of Psychoeducational Assessment*, 32(8), 747–761. <https://doi.org/10.1177/0734282914540449>

Response scale: 1–5

Positive Valence (4 items)

1. Knowing a lot about math is helpful.
2. I feel good when it comes to working on math.
3. I feel excited when a new math topic is announced.
4. I choose to work on math.

Negative Valence (6 items)

1. I am wasting my time on math.*
2. I am bored when working on math.*
3. I give up easily when working on math.*
4. I get mad easily when working on math.*
5. I have difficulty paying attention when working on math.*
6. I spend as little time as possible working on math.*

Time (2 items)

1. I want to talk about math with my friends.
2. I prefer easy math over math that is hard.*

Math Minds Measures

Sinclair, H. C., Kim, Y. K., & Terrio, S. R. (2025). *The Math Mind Measures: A User Guide*. Social Research and Evaluation Center at the Louisiana State University College of Human Sciences and Education.

Response scale: 1–6

Math Anxiety (4 items)

1. Just being in math class makes me feel stressed.*
2. I start to feel stressed-out as soon as I begin my math work.*
3. My mind goes blank when doing math work.*
4. I get very tense when I do math homework.*

Math Self-Efficacy (4 items)

1. I can usually solve any math problem.
2. I'm confident that I can do an excellent job on my math tests.
3. I'm certain I can understand the most difficult material presented in math books.
4. I'm certain I can master the skills being taught in my math class.

Intrinsic Motivation (4 items)

1. I like challenging myself when learning new math material.
2. I choose to try hard in math because it matters to me.
3. I want to live up to what I expect of myself in my math classes.
4. I compete with myself in math.

Liking of Math (4 items)

1. I like math.
2. I think math is fun.
3. I like problem-solving.
4. I look forward to the next time I'll get to do math.

Appendix B: Teacher Midpoint Survey

Treatment Group Questions:

Thank you for completing this brief midpoint survey. Please answer the following questions based on your experience with your math students who are part of this study.

1. [Single select option] You are in the treatment group, which means you have been using Goblins with your students. During the study period so far, approximately how many Goblins sessions per week did your students typically complete?
 - a. 0 sessions per week
 - b. 1 session per week
 - c. 2 sessions per week
 - d. 3 sessions per week
 - e. 4 sessions per week
 - f. 5 or more sessions per week
2. [Optional write-in] Is there anything else you'd like to share about the frequency or duration of your Goblins use?
3. [Short answer] For this study, in which specific grade level(s) and course(s) are you currently using Goblins? (E.g., Grade 6 regular math, Grade 7 intervention class, Grade 6 inclusion math.)
4. [Short answer] What is the core math curriculum you use in your math class(es)?
5. [Short answer] What supplemental materials (e.g., tools, resources), if any, do you also use with your students?

6. [Long answer] Please describe how Goblins is incorporated into your math instruction (e.g., when during instruction it is used, whether it is part of a regular routine, whether it occurs during independent learning time, etc.).
7. [Rating] How well does Goblins fit into your existing math curriculum and instructional routines?
 - a. Not at all well
 - b. Slightly well
 - c. Moderately well
 - d. Very well
 - e. Extremely well
8. [Rating] To what extent have you used data or insights from Goblins to guide your instruction (e.g., addressing misconceptions or adjusting lessons)?
 - a. Not at all
 - b. A little
 - c. A moderate amount
 - d. Quite a bit
 - e. A great deal
9. [Optional write-in; long answer] What challenges or usability barriers, if any, have you or your students experienced when using Goblins?
10. [Optional; long answer write-in] Is there anything else you would like to share?

Control Group Questions:

Thank you for completing this brief midpoint survey. Please answer the following questions based on your experience working with your math students who are part of this study.

1. You are in the control group for this study, which means you will NOT use Goblins with your students. Please confirm that you have not used this tool.
(*Your honest response will help us accurately understand implementation during this study.*)
 - a. I confirm that I have NOT used Goblins with students this semester.
 - b. I have used Goblins with my students this semester.

[LOGIC: if "B: I have used"] Please briefly describe how you have used Goblins.

2. Have you used any other artificial intelligence (AI) tools or platforms to support students' math learning during this study?
 - a. Yes
 - b. No

[LOGIC: If "Yes"] Please list the tool(s) and briefly describe how you use them with your students.

3. [Short answer] What is the core math curriculum you use in your math class(es)?
4. [Short answer] What supplemental materials (e.g., tools, resources), if any, do you also use with your students?
5. [Optional; long answer write-in] Is there anything else you would like to share?

Appendix C: Focus Group Questions

Goblins: Treatment Focus Group

(Bolded numbers indicate “priority” questions if time runs short.)

CONSENT TO RECORD

Is it ok with you if we record our discussion? The recording will not be shared beyond our Leanlab research team.	Yes No
---	-----------

Intro

Can you introduce yourself and how long you have been teaching math?

Usage

1. Can you briefly describe how you used Goblins in your classroom?
 - a. Probe: frequency, duration of sessions, class type, group vs. individual use
2. Walk me through a typical day when students used the tool.
3. How did you decide when and how to use it?
4. What kinds of students used it most? Least?
5. How much guidance did you provide vs. letting students use it independently?
- 6.** Did you do anything with information you learned about students from their Goblins use?

Teacher Perceptions of Student Learning and Engagement

- 7.** How would you describe student engagement when they were using Goblins?
8. What, if anything, did you notice about student learning when using the tool?

9. Do you feel like Goblins helped students understand math concepts or approach solving math problems differently?
10. Did you notice any changes in students' motivation to do math?
11. What did students do if they got stuck on a math problem in Goblins?

Multimodal Features and Student Experience

12. Which features did your students use?
13. Were there any features students gravitated toward or avoided?
14. How did these features impact students' experience with Goblins?

Teacher Experience and Instructional Impact

15. Did Goblins affect your role during math instruction?
16. Did it impact how you spent your time in class?
17. What do you think is the most optimal role of an AI-powered tool like Goblins in a math classroom?
18. Did you or your students run into any challenges or issues using Goblins?
 - a. Probes:
 - i. Technical issues
 - ii. Student misuse or confusion
 - iii. Alignment with curriculum
 - iv. Time constraints
19. If you could change one thing about the tool, what would it be?
20. What would make this more effective or easy to integrate into your classroom?
21. Did you use any other supplemental math learning tools with your students over the past few months?
22. If you were no longer able to continue using Goblins, how disappointed would you be? (Select One)
 - a. Very Disappointed
 - b. A Little Disappointed
 - c. Not Disappointed At All
23. Is there anything else you would like to share?