

Assessing the Impact of Goblins AI Math Tutor*

Anika Alam, Nicolas Dominguez-Mel, Jing Liu[†]

The Center for Educational Data Science and Innovation (EDSI)

University of Maryland, College Park

August 27, 2026

KEYWORDS: education technology, computer assisted learning, student learning

*This study is funded by Accelerate, a national nonprofit that helps states scale proven strategies in public schools. We are grateful to the Goblins team, Sawyer Altman and Alp Karavil, for their partnership and helpful collaboration. The study was pre-registered, and the pre-analysis plan is publicly available at <https://osf.io/5kwbp/overview>. The views expressed herein are those of the authors and do not necessarily reflect the views of the funder.

[†]Corresponding Author (jliu28@umd.edu).

Executive Summary

This pre-registered quasi-experimental study¹ estimates the effect of using Goblins, an online math learning platform, that was offered to all grades 6 to 8 math teachers at Deer Valley Unified School District during the 2025–26 school year. We implement a weighted propensity score design to compare changes in math achievement between Goblins users in Deer Valley Unified School District and observationally similar students who either did not use Goblins or used it below the minimum threshold for meaningful engagement. The evaluation yielded several key findings regarding the effectiveness and implementation of Goblins:

1. Adoption of Goblins substantially exceeded that of other computer-assisted learning (CAL) tools, including Khan Academy, Zearn, and IXL. Of the 6,065 students in the analytic sample, more than 50.9% (3,088) averaged at least two problems solved per week during the implementation period.
2. After adjusting for fall achievement and observed characteristics, students at or above the threshold of 2 problems solved per week scored 0.109 SD higher on the within-district standardized spring MAP outcome and 0.237 SD higher on NWEA’s Conditional Growth Index (CGI), a nationally normed measure of fall-to-spring growth.
3. There is a strong positive relationship between Goblins usage level and math achievement, where the strongest effects are observed among the highest-use groups: students who solved 10 to 19 problems per week and those who solved 20 or more problems per week demonstrated math achievement gains of 0.18 SD and 0.31 SD, respectively, on the within-district standardized MAP score, and 0.40 SD and 0.69 SD, respectively, on CGI.
4. Students with an IEP, Hispanic students, and English language learners appear to

¹The pre-analysis plan is publicly available at <https://osf.io/5kwbp/overview>

benefit much more from Goblins than their counterparts, showing additional growth by 0.122 SD, 0.084 SD, and 0.191 SD, respectively. One possible explanation is that Goblins’ multimodal features help lower barriers to engaging with the platform for these students.

Together, these findings provide initial evidence that generative AI can extend beyond expanding access to instructional support to produce measurable improvements in student learning when implemented at meaningful levels of use. Future studies should examine platform use at a more granular level, including whether specific features contribute to observed effects—for example, the effectiveness of the tool when the AI tutor is disabled during assessment mode. Future evaluations should also collect user identifiers that are not based on self-reported information, such as district-provided student IDs or SSO identifiers, to facilitate more complete linkage between district records and platform data.

Prior Research

A substantial body of evidence demonstrates that tutoring is one of the most effective interventions for improving academic achievement in K–12 education ([Guryan et al., 2023](#); [Nickow et al., 2024](#)). In the wake of the COVID–19 pandemic, concerns over widening achievement gaps have intensified, increasing the need for learning interventions that can be delivered equitably and at scale ([Backes et al., 2026](#)). Although high-impact tutoring has demonstrated potential to meaningfully improve student achievement, it is also resource intensive. The costs associated with recruiting, training, and compensating tutors make it difficult for schools and districts to provide tutoring to all students who could benefit from it. Educational technology (edtech) offers a potential alternative by delivering individualized learning support at a substantially lower cost per student while remaining highly scalable. Advances in adaptive learning platforms and artificial intelligence (AI) have further expanded the ability of edtech tools to provide personalized instruction, immediate feedback, and

targeted practice. As a result, schools have increasingly adopted these technologies as a means of supplementing classroom instruction and addressing unfinished learning. This section summarizes the existing evidence on the effectiveness of edtech tools for supporting K–12 math learning and highlights important gaps in the current evidence base.

Evidence of the effectiveness of math education technology tools

Despite the rapid deployment of edtech tools in K–12 settings, rigorous evidence on their effectiveness remains relatively limited. Moreover, much of the existing literature evaluates overall program effectiveness without examining whether students use these tools at levels sufficient to produce meaningful learning gains.

[Khan Academy](#) is one of the earliest and most widely adopted online learning platforms, providing free instructional resources in math and other core academic subjects. Despite its widespread use since its launch in 2008, relatively few rigorous evaluations have examined its impact on student achievement. One of the most notable findings is the “5 Percent Problem” by [Weatherholtz et al. \(2022\)](#) that reported an effect size of 0.26 standard deviations after restricting the analysis to 5% of students who used the platform for at least 30 minutes per week. A more recent evaluation adopted a less restrictive inclusion criterion and found that the estimated effect declined substantially to 0.03 standard deviations, reflecting average usage of approximately 11 minutes per week ([Eames et al., 2026](#)). In contrast, a randomized evaluation of Khan Academy conducted in India reported an effect size approaching 0.50 standard deviations when the program was implemented with high fidelity ([Oreopoulos et al., 2026](#)). Collectively, these findings suggest that the effectiveness of Khan Academy is closely tied to implementation fidelity and sustained student engagement.

Experimental evaluations of other math edtech tools have produced similarly mixed findings. A two-year randomized evaluation of Cognitive Tutor Algebra I (CTAI) found no effects in the first year and an improvement of 8 percentile points in the second year for high school students ([Pane et al., 2014](#)). A recent study comparing DreamBox Learning and Zearn

Math as supplements to classroom instruction found promising evidence for DreamBox in improving early elementary math skills (Foster, 2024). There is mixed experimental evidence of AI tutoring tools, where Eedi and Tutor CoPilot are found to have promising impacts on math learning (Team et al., 2025; Wang et al., 2024), while other studies have found that human tutors are needed to sustain student engagement on AI platforms compared to students engaging alone with AI platforms (Robinson et al., 2026).

Evidence from non-experimental studies generally reports positive effects for students who use programs as intended; however, these studies frequently define treatment based on relatively high levels of usage achieved by only a small share of students. For example, Marple et al. (2019) estimated an effect size of 0.22 standard deviations for students who used i-Ready at the recommended level, although these students represented only 12% of the study sample. Similarly, a study reported an effect size of 0.14 standard deviations for IXL, but only about 5% of students met the minimum usage threshold required for inclusion in the analysis (An, 2023). Low implementation fidelity has also been documented for other widely used platforms, where an analysis of DreamBox found that only 6% and 13% of students in two partner districts met the recommended levels of usage (CEPR, 2016). Moreover, Szatrowski et al. (2022) found that the small proportion of students achieving recommended usage levels for Zearn tended to be higher performing and more economically advantaged than their peers, raising concerns about the generalizability of estimated treatment effects.

Although some attention has been dedicated to evaluating the effectiveness of edtech tools, relatively little attention has been devoted to identifying the levels of usage necessary to produce meaningful improvements in student learning. Most evaluations estimate impacts for students who meet developer recommended usage thresholds, but few provide empirical evidence to justify those recommendations or examine how learning outcomes vary across different levels of engagement. Consequently, there remains limited evidence to inform educators and policymakers about what constitutes effective implementation. This study addresses these gaps by estimating both the overall effectiveness of an AI-powered math tutor

tool and the relationship between usage intensity and student achievement, providing evidence to inform implementation recommendations for future use in K–12 classrooms.

The Tool: Goblins

[Goblins](#) by Aha Moments is an online math learning platform designed for students in grades 5 to 12. One unique feature of Goblins is that it engages students through multiple modalities, enabling them to speak through problems, write on a virtual whiteboard, or type responses - all while providing real-time and personalized feedback on math work. It is designed to be highly accessible by providing voice-based interaction in seven languages for students with diverse learning and language needs.

Goblins is licensed at the school or district level and can be made available to all teachers within participating schools or districts. To begin using the platform, teachers must first activate their individual license. Once activated, teachers can create virtual classes and generate unique class codes that students use to enroll. Students can only access Goblins through enrollment in a teacher-created class. Thus, while a student-facing tool, Goblins' usage is directly mediated by teachers. This is important for our empirical decisions for evaluating Goblins, which we detail in the Methods section.

Goblins can be integrated into instruction by teachers in multiple ways. [Figure 1](#) summarizes the complete set of pathways through which teachers and students engage with Goblins, from teacher created content and assignments to student practice and feedback. The platform provides a wide range of standards- and curriculum-aligned math content across grade levels and topic areas. Teachers can assign pre-built instructional materials and assessments, including assignments, exit tickets, diagnostics, unit tests, and practice activities. In addition to using pre-made content, teachers can also create and customize their own instructional materials through a process called “goblinizing.” Goblins provides curriculum packets for each topic area that include practice sets, cool-down problems, and question banks that

teachers use to adapt and align with their instructional needs.

Goblins enables flexible implementation where teachers can assign content to an entire class, small groups of students, or individual students. Teachers also have control over how students interact with the platform. Assignments can be delivered in either AI tutor mode, where students receive interactive, step-by-step support while solving problems; or assessment mode, where students complete problems independently without AI assistance. When the AI tutor mode is enabled, students can engage with the AI tutor to guide students through the problem-solving process by prompting math reasoning, supporting procedural fluency, and providing feedback as students work through problems.

After enrolling in a class, students can engage with teacher assigned activities as well as independently explore math content through the platform’s “Infinite Mode.” Infinite Mode allows students to search for math topics and practice additional content independently, providing opportunities for self-directed learning and supplemental practice beyond assigned classroom activities.

Goblins Implementation at Deer Valley Unified School District

This evaluation took place in DVUSD, a large district in Arizona that serves over 32,000 students across 42 campuses. During the 2025-26 school year, DVUSD partnered with Goblins to offer the tool to all math teachers teaching middle school grades. To enhance take up of the tool, Goblins provided professional development for district teachers and instructional leaders in fall 2025. As discussed above, teachers had to activate their licenses, create virtual classes, and share class codes for student enrollment. Student access is mediated through teacher-created classes, after which students can independently explore content using features such as Infinite Mode.

The district provided guidance to teachers that Goblins should *not* be used for Tier 1 instruction that is provided to all students in a classroom; rather, it should be used more selectively as a supplemental support. However, both data collected throughout the school

year and anecdotal evidence suggest that some classrooms nonetheless incorporated Goblins into Tier 1 instruction. For example, we see many classrooms where teachers assigned problems to all students in Goblins. There are no known district wide policies that would affect the implementation of the tool and interpretation, such as staffing turnover, closures, or policy changes.

Data

This study links two data sources from the 2025-26 academic year: user-level activity data from Goblins' platform and administrative student records provided by DVUSD.

Goblins data consist of detailed, clickstream measures that capture platform engagement at the user level. The data include detailed measures such as language preferences, time spent on the platform, problems and assignments attempted and completed, modules visited, and the frequency of different input modalities (e.g., drawing, speech, and typing). The platform also logs students' use of instructional support features, including hints and scaffolds (e.g., conceptual guidance and calculation corrections), as well as measures of problem accuracy and content mastery.

DVUSD administrative data include student-level math achievement data from the NWEA MAP assessment administered at three time points during the school year (beginning, middle, and end of year). In addition, the dataset contains demographic information for all students in grades 6 to 8, along with teacher background information.

As there is no shared individual-level IDs across the two datasets, we link them using student names available in both sources. This approach is overall successful but not perfect. Specifically, because students self entered their names when creating or accessing their Goblins accounts, the platform data contained inconsistencies in name spelling, abbreviations, nicknames, and typographical errors relative to the district administrative records. To account for these discrepancies, we used a fuzzy matching procedure. This process successfully

matched approximately 87% of Goblins users identified by the platform as belonging to the Deer Valley district. Relative to matched users, unmatched users skew toward lighter platform use ($p < 0.001$; see [Table A1](#) for a full comparison). In the next section, we describe the characteristics of the matched analytic sample through a series of descriptive statistics, establishing the distribution of demographic, academic, and platform engagement measures used in subsequent analyses.

Sample Characteristics

We construct the final analytic sample by applying three inclusion criteria. First, we only include students with complete math achievement records for the 2025–26 school year, defined as having NWEA MAP assessment scores at all three testing occasions (beginning, middle, and end of the year). Second, we retain students who remained enrolled with the same math teacher throughout the school year to avoid confounding effects associated with teacher changes. Third, the teacher identified in the district administrative records must be a classroom math teacher rather than a school administrator or other non-instructional staff member. Starting from the full population of grades 6-8 students in DVUSD ($n = 7,083$), the inclusion criteria retain 86% of students, resulting in a final analytic sample of 6,065 students.

[Table A2](#) provides baseline information on the analytic sample of students and teachers, including a breakdown of the sample based on the activation of a Goblins account. 80% of all grades 6 to 8 students at DVUSD had an activated Goblins account. The remaining 20% of students may come from classrooms where teachers did not opt into Goblins (so no student in these classrooms had access), classrooms where teachers did not assign problems to these students or they chose not to use it, or could capture students who had a Goblins account that could not be successfully merged with district records. [Table A5](#) provides school-level characteristics of the analytic sample.

Overall, Goblins student users are more likely to be in grades 6 and 7, Hispanic, English

language learners, identified as having special needs, and having lower baseline math ability. The achievement difference is especially stark; the average baseline MAP test scores for non-Goblin user students are 0.4 SD higher than Goblin users, suggesting that teachers primarily used Gobblins for Tier 2 and 3 instruction. At the teacher level, we observe that female, non-White, and less experienced teachers are more likely to use Gobblins. However, we do not observe significant differences in teachers' prior year value-added scores, a measure of their contribution to student math learning estimated using prior students' achievement growth. These value-added scores are available only for teachers who taught in the district during the previous school year and whose students had observable math achievement in both fall 2024 and spring 2025.

Gobblins Usage Statistics

We characterize student engagement on Gobblins using three measures: *active minutes*, *problems attempted*, and *problems solved*. *Active minutes* captures the total time students spend on the platform, including interacting with the AI tutor, working on assignments or independent practice problems, and navigating other features within Gobblins. *Problems attempted* refers to the number of problems that students open, regardless of whether they ultimately submit a response. *Problems solved* represents the number of problems students answer correctly. Because the AI tutor requires students to solve the current problem before progressing to the next one, this measure is equivalent to the total number of correctly completed problems. For all engagement measures, we aggregate activity across both teacher-assigned and independent practice modes and do not distinguish between these two channels of platform use. This approach captures students' overall engagement with Gobblins irrespective of the context in which the activity occurred.

Another potential measure of student engagement that we ended up not using is the number of assignments completed. The reason is that teacher-created assignments vary substantially in length and content, making assignments not directly comparable. Furthermore,

students may engage with Goblins outside of teacher-created assignments through independent practice, which provides an unlimited sequence of problems. As a result, assignment completion does not comprehensively capture students’ overall use of the platform.

Figure 2 reports student-level Goblins usage reported separately by grade and for the full analytic sample. “Teacher assigned usage” refers to Goblins content that was explicitly assigned by a student’s math teacher, whereas “Independent usage” refers to Goblins activity that was not assigned by the teacher. Across all grade levels, students spent more time using Goblins in teacher assigned mode than in independent mode. 8th grade students exhibited the lowest overall usage. Among the 4,833 students matched to platform accounts, students averaged about 318 minutes on Goblins and solved about 136 problems across teacher-assigned and independent activity during the implementation period.

Table A3 provides patterns of engagement across the three measures with usage classified into groups. The patterns are broadly similar across these measures, and we thus focus on the average problems solved per week as the primary measure that assume 27 weeks of instruction in the 2025-26 school year. The distribution of solved problems closely mirrors attempted problems, indicating that most students completed problems that they initially attempted. There is a substantial portion of students (46%) who either attempted fewer than two problems per week or did not have an activated Goblins account. In Table A3, 16.41% of the analytic sample attempted 5 to fewer than 10 problems per week, 15.48% solved 5 to fewer than 10 problems per week, and 18.68% recorded 5 to fewer than 10 active minutes per week. Figure 2 reports attempted problems per week among the same matched-user denominator. The highest-usage group across the three measures generally includes three to six percent of the analytic sample.

Defining Treatment

One challenge for this evaluation is how to define treatment. A convenient solution is that students with a Goblins account would be considered as the treatment group. However, we

found substantial variation in Goblins usage across classrooms and students. For example, a student who had an account might only be logged into the system once with a few minutes of usage. Considering these students is useful if our goal is to evaluate the impact of having *access* to Goblins. However, in our context, we are more interested in estimating the impact of *using* Goblins.

For the reason above, we explored treatment thresholds based on a minimum level of usage rather than account activation alone. We investigated implementation patterns that characterize student engagement with Goblins across the same three measures described earlier: active minutes, problems attempted, and problems solved.

The implementation patterns provided in [Table A3](#) support the hypothesis that account activation alone does not reflect meaningful engagement on Goblins. This informs our approach to include students with low levels of Goblins usage in the comparison group.

Based on our conversations with the provider, we operationalize “Goblins usage” as a student’s actual engagement with the platform. Among the three measures of student engagement on Goblins, we also define Goblins usage at a minimum level total of 54 problems solved during the implementation period (i.e., the 2025-26 school year). This threshold is conservative, assuming that we would observe meaningful improvement in student achievement if a student on average solved two problems per week in the Goblins platform. Approximately 51% of all students in the analytic sample had met this threshold. Using the other two measures to operationalize Goblins usage does not meaningfully change our findings reported in this article.

[Table A4](#) reports the composition of students above and below the minimum threshold. Descriptively, we observe that students who met the threshold have lower incidences of disability status and are more likely to be on grade level at baseline ($p < 0.05$). We also see an overrepresentation of novice teachers included in the treated group ($p < 0.05$).

To understand how usage varies among users who had met the minimum threshold (i.e., the treated group), we disaggregate the distribution of total problems solved in [Table 4](#). We

create the following six dosage categories based on their distributions: (1) at least 2 but fewer than 3 problems per week, (2) 3 to fewer than 4 problems per week, (3) 4 to fewer than 5 problems per week, (4) 5 to fewer than 10 problems per week, (5) 10 to fewer than 20 problems per week, and (6) 20 or more problems per week during the implementation period. These categories will also be applied to understand heterogeneous treatment effects.

Outcome Measures

We measure math achievement using spring 2026 NWEA MAP math scores in two ways: (1) creating a standardized measure within the district, a typical approach in the literature and (2) using NWEA’s national Conditional Growth Index ([Weatherholtz et al., 2022](#)). The within-district measure standardizes students’ spring math scores by calculating a z -score within each grade-level distribution in DVUSD. This serves as our primary outcome as it allows us to directly benchmark our estimates relative to effects reported in the education intervention literature ([Kraft, 2020](#)).

In contrast, NWEA computes Conditional Growth Index (CGI), a measure of student growth that compares a student’s progress over the full school year with that of a national reference group of students who were tested in the same period (fall 2025 to spring 2026), are in the same grade, and began the year at a similar achievement level ([NWEA, 2026](#); [Thum, 2014](#)).

We retain the within-district standardized spring score as the pre-registered primary outcome. It answers a within-district question: *How much higher did threshold students score in spring, relative to district peers, after adjustment for fall achievement?* This scale supports comparison with standardized effects in the education-intervention literature ([Kraft, 2020](#)).

We report CGI as a complementary growth outcome because it answers a different question: *How did each student’s realized fall-to-spring progress compare with nationally normed expected progress for students in the same grade with similar starting achievement and test intervals* ([NWEA, 2026](#); [Thum, 2014](#))? CGI is familiar to MAP-using districts and permits

metric-matched context with prior MAP Accelerator and district-facing research [Weatherholtz et al. \(2022\)](#); [Schechter and Chase \(2025\)](#). Because the two outcomes use different standardizations and constructs, their numerical magnitudes are not comparable.

Research Questions

This report examines the following research questions:

Implementation Question

- How did teachers incorporate Goblins in their instruction (e.g., universal supports versus targeted interventions)?

Descriptive Question

- How did students engage on the Goblins platform? What engagement patterns were predictive of content mastery and learning outcomes?

Outcomes Questions

- What is the effect of Goblins’ multimodal math tutor on student mathematics achievement? Do the effects differ by student subgroup (such as race/ethnicity, gender, and economic disadvantage) or by student baseline proficiency level?
- What is the relationship between the level of usage and student achievement growth? How does it vary across classroom and school-level contexts?

[Table 1](#) provides a detailed breakdown of how each research question will be measured, the data source, and the timing of data collection. All data received from the district partner and Goblins were stored in a secure Box folder approved by the University of Maryland, with access limited to authorized research team members. Data were reviewed upon receipt

for completeness and consistency, and any discrepancies were addressed as needed. Goblins stores user activity data on secure, encrypted cloud servers to protect against unauthorized access.

Methods

We follow the study preregistration and employ a weighted least squares regression with propensity weights to estimate the effect of Goblins usage on student math achievement. The analytic approach comprises of four stages: estimating each student’s *propensity to treatment* scores, computing weights, ensuring covariate balance, and estimating the average treatment effect on the treated (ATT) on the balanced sample.

We estimate the minimum detectable effect size (MDES) based on the analytic sample. The power analysis assumes teacher-level assignment to treatment, with all 124 grades 6-8 math teachers in the district included in the sample and an average of 80 students per teacher. We use the actual number of teachers whose students met the threshold of 2 problems solved per week (90 teachers) as the number assigned to the treatment condition. Following the MDES calculation for a two-level cluster-randomized design with treatment assigned at the teacher level (Level 2), we assume that student-level covariates explain 60% of the variance in the outcome ($R^2 = 0.60$), teacher-level covariates explain 60% of the between-teacher variance ($R^2 = 0.60$), and the teacher-level intraclass correlation coefficient (ICC) is 0.10. Calculations assume a two-tailed hypothesis test with a significance level of 0.05 and 80% statistical power. Under these assumptions, the estimated MDES is 0.119 standard deviations.

We did not compute an intent-to-treat (ITT) estimate for two reasons. First, an ITT estimator intends to capture the effects of being offered Goblins through teacher activation. With Goblins being offered to the entire district, the vast majority of teachers (73%) activated Goblins, leaving only a small portion of students to serve as comparisons for an ITT

estimator. Given our design is a matching approach, using such a small number of students to construct the comparison group provide very little common support, resulting in extreme weights for certain students. Second, even within a teacher, student usage of Goblins varied considerably. Among the 4,833 students with an activated Goblins account who would have comprised the ITT treatment group, usage data indicate that engagement was very limited for a meaningful share of students. As teachers were not provided a specific dosage threshold, over the 27-week implementation period, almost half of students whose teachers activated a Goblins account solved fewer than 2 problems per week, a level considered the minimum engagement with Goblins by the provider. Thus, an ATT would more genuinely capture the impact of using Goblins at a minimum dosage level on student learning outcomes.

While Goblins was available to all teachers in the district, the sample characteristics presented in [Table A2](#) suggest that users and non-users differed systematically in their baseline characteristics. To account for these pre-existing differences observed, the analysis follows the guidelines from the [What Works Clearinghouse 5.0 Procedures and Standards Handbook](#). We summarize our analytic approach below and provide additional methodological details in the Technical Appendix.

Estimation of Propensity Scores

All stages of the analysis are executed using the **twang** (Toolkit for Weighting and Analysis of Nonequivalent Groups) software package developed by [Ridgeway et al. \(2022\)](#). We first estimate individual propensity scores for the full analytic sample. Each propensity score represents student i 's probability of using Goblins, conditional on a set of observed covariates. We observe covariates at the student, teacher, and school level to help account for systematic differences between students who did and did not use Goblins.

Student level covariates include baseline math achievement, measured using fall 2025 NWEA MAP math scores that are standardized within district and grade level, current grade level, gender, disability status (i.e., having an Individualized Education Plan), English

learner (EL) status, and a set of binary indicators for race and ethnicity (American Indian, Asian, Black, Hispanic, Native Hawaiian, White, and multiracial).

Teacher characteristics include teacher gender, whether the teacher is White, holding a master’s degree, lagged teacher value added scores, and indicators for teaching experience. Experience is represented using three mutually exclusive categories: novice (fewer than 5 years of experience), experienced (5–15 years), and veteran (more than 15 years). Teacher value added is intended to capture a teacher’s contribution to student achievement growth. In this analysis, value added is estimated using student math achievement from a single prior school year lagged-score value-added model following [Bacher-Hicks and Koedel \(2023\)](#)’s two-step (“average residuals”) approach. Additional details regarding the estimation of teacher value added are provided in the Technical Appendix.

Finally, propensity score estimations include school IDs that are operationalized as a binary indicator for each school in the sample. Including school indicators ensures that the propensity score model accounts for differences in the distribution of Goblins users across schools. Although this approach does not restrict comparisons to students within the same school, it adjusts for school level differences in treatment assignment, helping ensure that the subsequent treatment effect estimates appropriately account for school context.

Propensity scores are estimated using a gradient boosted machine (GBM) framework developed by [Friedman \(2002\)](#). GBM is an ensemble tree-based method that combines many decision trees to estimate the probability of treatment assignment. Each decision tree recursively partitions the data using a series of binary “if-then” rules based on the input variables, creating increasingly homogeneous groups with respect to the outcome (i.e., treatment assignment). GBM uses boosting, a sequential learning algorithm that builds trees iteratively rather than independently. At each iteration, a new tree is fit to reduce the prediction errors of the previous tree, thereby improving overall model performance. Through this process, many relatively simple trees (often referred to as “weak learners”) are combined to produce a more accurate estimate of the propensity score. The algorithm

continues until a prespecified number of iterations is reached or another stopping criterion is satisfied (Friedman, 2002; Polikar, 2012; James et al., 2023). The GBM model estimates the log-odds of treatment assignment denoted by

$$g(x) = \log \left(\frac{p(x)}{1 - p(x)} \right),$$

where $p(x)$ denotes the probability of treatment conditional on x observed covariates.

The propensity score estimation uses the following arguments to control the gradient boosting for the ATT estimand: a maximum of 10,000 boosting iterations; an interaction depth of 1; a shrinkage (learning rate) of 0.01; and a minimum terminal node size of 10 observations. The propensity score model uses the (`es.mean`) stopping criterion, which selects the iteration with the smallest mean absolute standardized mean difference (SMD) across covariates (see the later subsection for details on the how SMD is computed). The same arguments are used in propensity score estimations of robustness checks as well.

Compared to logistic regression and other machine learning approaches, GBM offers several advantages for propensity score estimation. First, unlike other tree-based methods such as classification and regression trees (CART) and random forests, GBM employs a boosting algorithm in which each successive tree is fit to the residual errors of the previous iteration, resulting in progressively improved model performance. Second, GBM provides flexible stopping rules that optimize covariate balance according to different criteria, such as minimizing the maximum standardized mean difference (`es.max`) or the mean Kolmogorov–Smirnov statistic (`ks.mean`). This flexibility allows researchers to tailor model selection to their preferred balance metric. Finally, unlike logistic regression, GBM automatically captures non-linear relationships and higher-order interactions among covariates without requiring these terms to be specified a priori (Keller and Tipton, 2016; Leite et al., 2025).

Inclusion of Propensity Weights

After estimating propensity scores, we generate propensity score weights to adjust for differences in students' probabilities of receiving treatment, conditional on observed covariates. We estimate weights corresponding to the average treatment effect on the treated (ATT), which assign a weight of 1 to all treated units and a weight of $w(\mathbf{x})$ to each comparison unit. Following [Hirano and Imbens \(2001\)](#), these weights can be expressed by applying Bayes' Theorem as

$$w(\mathbf{x}) = \frac{f(x \mid t = 1)}{f(x \mid t = 0)}$$

where $f(x \mid t = 1)$ and $f(x \mid t = 0)$ denotes the distribution of features for the treatment units and comparison units, respectively. This approach is analogous to the use of survey weights in that it assigns unique weights to each unit in the comparison group so that their covariate distribution more closely matches that of the treated group ([Wooldridge, 2010](#); [Ridgeway et al., 2022](#)).

Intuitively, comparison group units who are underrepresented relative to the treated group receive larger weights, whereas those who are overrepresented receive smaller weights. For example, suppose 20% of treated students have both an IEP and a fall math score of -0.1 standard deviations, such that $f(fall = -0.1, iep = Y \mid t = 1) = 0.20$. In the same instance, if only 4% of comparison students share these same characteristics (i.e., $f(fall = -0.1, iep = Y \mid t = 0) = 0.04$). Because this comparison student profile is five times less common in the comparison group than in the treated group ($0.20/0.04 = 5$), each comparison student with these characteristics receives a weight of 5. In the weighted sample, these students would collectively represent 20% of the comparison group rather than 4%, making the weighted comparison group more similar to the treated group with respect to these two observed characteristics.

Weighted means have higher sampling variance relative to that of unweighted means. This increase in variance can be observed by calculating [Kong et al. \(1994\)](#)'s formula for

weighted comparison group’s effective sample size (ESS) using the following formula:

$$\text{ESS} = \frac{(\sum_{i \in C} w_i)^2}{\sum_{i \in C} w_i^2}$$

The effective sample size (ESS) represents the size of a simple random sample that would produce an estimate with the same sampling variance of the weighted comparison group. As a result, the ESS provides an estimate of the number of comparison participants who contribute meaningful information because they are comparable to the treatment group after weighting ([Ridgeway et al., 2022](#); [McCaffrey et al., 2004](#)).

The ESS can be an accurate measure of precision under two assumptions: (1) that the weights are fixed, and (2) weights are uncorrelated with the outcome variable. However, because propensity score weights are typically correlated with outcomes, the ESS often underestimates the true effective sample size ([Little and Vartivarian, 2005](#); [Ridgeway et al., 2022](#)). As a result, the ESS should be interpreted as a conservative lower bound on the effective sample size. Despite this limitation, it remains a useful diagnostic for comparing alternative propensity score models and evaluating the overall quality of the weighting procedure, as it provides a practical measure of the potential loss in precision due to weighting.

Assessment of covariate balance

The causal warranty of a propensity score matching approach rests on the covariate balance between the treatment and control groups. We use the `twang` package to do this assessment. For each covariate included in the propensity score estimation, we report the group means and standardized mean differences (SMDs) after weighting. The SMD is a widely used measure of covariate balance and is calculated as:

$$d = \frac{\bar{X}_T - \bar{X}_C}{SD_{\text{pooled}}}$$

where $\bar{X}_T$ and $\bar{X}_C$ denote the mean values of the treatment and comparison groups, respectively, and SD_{pooled} is the pooled standard deviation. SMD is analogous to “standardized” bias reported in the propensity score literature and is consistent with the reporting recommendations in *What Works Clearinghouse 5.0 Procedures and Standards Handbook*.

Table 2 reports treatment and comparison group sizes and SMD before and after weighting. After applying the stopping rules and balance measures described earlier, the weighting procedure reduces the effective sample size of the comparison group. Although the original comparison group consisted of 2,977 observations, the propensity score weights result in an effective sample size (ESS) of only 1,256 comparison observations. While this may appear to represent a substantial loss of information, it instead reflects the fact that many of the original observations were dissimilar to the treated observations and therefore contributed little to estimating the treatment effect. Comparable or even larger reductions in effective sample size are common with other propensity score methods, such as matching or stratification, which similarly discard or downweigh observations that provide limited information for estimation (Ridgeway et al., 2022; McCaffrey et al., 2004). Table 2 also shows that the introduction of propensity score weights substantially improves covariate balance between the treated and control groups, where the maximum SMD decreases from 1.38 to 0.24 (a reduction of approximately 83%), while the mean SMD falls from 0.15 to 0.05 (a reduction of approximately 67%). These reductions suggest that the weighted sample is considerably more balanced than the unweighted sample.

To further assess covariate balance, Table 3 compares the distributions of all covariates after weighting. For each variable, the table reports the weighted sample means, standardized mean difference (SMD), and the corresponding p-value from a t-test of mean differences between the treated and comparison group for each variable. The table does not display balance along school identifiers, though these identifiers were used in the propensity estimation and weighting estimations. After weighting there remains some imbalance for baseline math achievement, grade level, gender, disability status, American Indian ethnicity, Asian

ethnicity, Hispanic ethnicity, multiple races, teacher gender, and teacher value add score. Although balance is not achieved for every covariate (including for most schools), the absolute SMD for all variables remains below 0.25. This satisfies WWC’s baseline equivalence standard that requires absolute SMD to fall between 0.05 and 0.25 standard deviations.

Estimation of the treatment effect on the treated

We estimate the ATT of using Goblins’ AI tutor on math achievement using a weighted least squares regression. Math achievement is measured using the two standardized measures of spring 2026 MAP scores described earlier. Following the standard practice established in [Hirano and Imbens \(2001\)](#), we fit the weighted sample of treated and comparison students and integrate propensity weights derived from the GBM model into the following specification:

$$Y_{its} = \beta_0 + \beta_1 T_i + \beta_2 X_{its} + \tau_s + \varepsilon_{its} \quad (1)$$

where Y_{ist} is spring 2026 standardized math test score for student i in teacher t ’s classroom at school s . The model includes the treatment indicator, T_i , equal to 1 if the student solved at least 54 problems during the implementation period (an average of 2 problems per week), the full set of covariates used in the propensity score model (X_{its}), and school fixed effects (τ_s) to ensure that treatment effects are identified by comparisons among students within the same school. The regression is estimated using the propensity score weights, allowing the weighted comparison group to approximate the treated population. The coefficient of interest, β_1 , captures the estimated difference in standardized spring math achievement attributable to active use of Goblins relative to low or zero use of it. Heteroskedasticity-robust standard errors are clustered at the teacher level.

In addition to measuring the overall effect of using Goblins’ AI math tutor, we examine whether and how dosage (i.e., Goblins usage intensity) relates to math achievement. This model extends the primary specification by replacing the single binary treatment indicator

with a set of mutually exclusive dosage categories, $D_{d1} \dots D_{d6}$. The six dosage categories correspond to the cutoffs provided in [Table 4](#). These dosage indicators are each interacted with the binary treatment indicator, T_i in the below specification:

$$Y_{its} = \beta_0 + \sum_{d=1}^6 \gamma_d (D_d T_{its}) + \phi X_i + \tau_s + \epsilon_{its} \quad (2)$$

Interacting the dosage indicators with the treatment indicator ensures that only students meeting the minimum treatment threshold (i.e., those averaging at least 2 problems solved per week) are assigned to one of the dosage categories. This function produces six coefficients, $\gamma_{d1} \dots \gamma_{d6}$ that each represent the estimated difference in standardized spring 2026 math achievement between students in a given dosage group and the common reference group consisting of students whose average usage remained below the treatment threshold of 2 problems per week.

Third, we estimate treatment effects for student subgroups by interacting the treatment indicator with each student characteristic of interest. Separate models are estimated for each subgroup characteristic to assess whether the effect of Goblins varies across different student populations.

Fourth, we conduct moderation analyses examining classroom- and school-level characteristics. The moderation analyses examine four moderators: whether more than 50% of a teacher's roster was below grade level at baseline, whether the teacher had fewer than five years of teaching experience, whether more than 50% of students in the school were below grade level at baseline, and whether the school was classified as Title I schools that have 40% or more of their students receiving free or reduced meals. Similar to the dosage and student subgroup models, we interact the treatment indicator with each moderating characteristic of interest. There are two limitations to note. First, we do not include FRM eligibility as a moderator because 19% of observations in the analytic sample have missing FRM data. Second, as we do not have classroom IDs or another identifier linking students to specific classrooms, we cannot cleanly measure classroom context. Instead, we can construct

teacher-level rosters and use teacher characteristics as proxies for classroom context.

Robustness checks

We assess the robustness of our primary findings through a series of sensitivity analyses. These specifications retain the full set of student characteristics and school identifiers in the propensity score estimation and included school fixed effects in the outcome regression. First, we examine an alternate treatment threshold where students must have solved an average of at least 3 problems per week. This alternative specification evaluates whether the estimated treatment effects are sensitive to the choice of threshold.

Second, we incorporate an additional pre-treatment characteristic to approximate economic disadvantage. This teacher level indicator identifies teachers for whom more than 50% of enrolled students are eligible for free or reduced-price lunch, allowing us to account for differences in the socioeconomic composition of teachers' classrooms that may not be captured by student level covariates alone.

Third, we impose a minimum support requirement to strengthen within school comparisons. Under this restriction, the analytic sample includes only students attending schools with at least five students in each treatment condition (i.e., both the treatment and comparison groups). This sensitivity analysis addresses whether the findings are robust to excluding schools with limited overlap between treatment assignments.

Results

RQ1: How did teachers incorporate Goblins in their instruction (e.g., universal supports versus targeted interventions)?

Ideally, we would like to classify teachers' use of Goblins as Tier 1, Tier 2, or Tier 3. As described earlier in the report, teachers can create multiple classes on Goblins, and students

join these classes using a unique invitation code. Teachers also have flexibility in how they use these classes - they may enroll different groups of students, assign content to all students in a class, or assign content to selected students within the same class.

However, we do not have district classroom rosters or classroom IDs that would allow us to link students to specific math classes or class periods. While we can link students to their math teachers and therefore observe the total number of students taught by each teacher, we cannot determine whether a particular Goblins class represents an entire classroom, a subset of students receiving targeted support, or students drawn from multiple class periods. Because these platform-based classes cannot be validated against district classroom rosters, enrollment size alone cannot be used to definitively classify implementation as Tier 1 versus Tier 2/3.

To provide the closest available proxy for implementation context, we added descriptive analyses of teacher-created Goblins classes. Specifically, we examine: (1) the number of student users per Goblins teacher, (2) the number of Goblins classes created per teacher, and (3) enrollment sizes within Goblins classes.

Using district-provided records, we identified 124 grades 6 to 8 math teachers in the analytic sample. On average, teachers were associated with 49 students, although the number varied substantially. The 25th percentile was 12 students, the median was 44 students, and the 75th percentile was 80 students. Among all student Goblins users in the analytic sample ($n = 4,833$), 10.6% were enrolled in more than one Goblins class.

The patterns observed in the Goblins data are broadly consistent with these district-level teacher rosters. As shown in [Table 5](#), Goblins teachers had an average of 54 student users, with 25%, 50%, and 75% of teachers having 22, 50, and 79 students, respectively. Among the 90 Goblins teachers, over 70% created more than one class, and 40% created four or more classes. Across the 273 teacher-created Goblins classes, enrollment sizes also varied considerably: more than half of classes enrolled 20 or more students, while approximately 12% enrolled fewer than 6 students.

These findings do not allow us to determine whether larger Goblins classes correspond to a single district math class or whether students were drawn from multiple classes taught by the same teacher. However, the fact that more than half of Goblins classes enrolled at least 20 students suggests that Goblins was sometimes implemented with groups comparable in size to a typical classroom, while smaller classes may reflect more targeted implementation. We therefore present these analyses as descriptive evidence of variation in implementation rather than as a definitive classification of Tier 1 versus Tier 2/3 use. Given these data limitations, we are unable to estimate separate treatment effects for Tier 1 and Tier 2/3 implementation.

We also considered measures related to assignment type and teacher engagement with the platform. However, as described in the “Goblins Usage Statistics” section, assignment-level measures do not provide a consistent or comprehensive measure of implementation. Teachers can customize pre-made assignments or create their own using Goblins’ question bank, resulting in substantial variation in assignment length, content, and structure. Moreover, students can use Goblins independently outside of teacher-created assignments through an unlimited sequence of practice problems. As a result, assignment completion does not capture students’ overall engagement with the platform.

Appendix [Table A6](#) disaggregates the total number of problems assigned by teachers according to the enrollment size of teacher-created Goblins classes. We categorize classes with 1–2 enrolled students as a proxy for Tier 3 implementation, classes with 3–15 students as a proxy for Tier 2 implementation, and classes with 16 or more students as a proxy for Tier 1 implementation. Teachers with Tier 1-sized classes assigned the largest number of problems, followed by those with Tier 2- and Tier 3-sized classes.

These findings should be interpreted with caution. Although class enrollment size provides a useful proxy for implementation tier, teachers can assign content to individual students or subsets of students within the same Goblins class. As a result, we cannot definitively determine whether content was assigned to the full class or to selected students, and therefore

cannot conclusively classify implementation as Tier 1, Tier 2, or Tier 3.

RQ2: How did students engage on the Goblins platform? What engagement patterns were predictive of content mastery and learning outcomes?

We investigate implementation patterns that characterize student engagement with Goblins across the same three continuous measures described earlier: active minutes, problems attempted, and problems solved. Correlational analyses presented in [Table 6](#) indicate that the three measures of Goblins usage are highly correlated with one another, particularly problems attempted and problems solved ($R = 0.98$). In contrast, correlations between spring MAP scores and each measure of Goblins usage are relatively small. Total problems solved has the strongest correlation with spring math achievement ($R = 0.12$). Together, these findings inform our decision to use problems solved as the primary measure of Goblins usage, as it demonstrates the strongest association with mastery of math concepts measured at endline.

We further examine engagement patterns using a minimum level of usage rather than account activation alone. The implementation patterns presented in [Table A3](#) describe student engagement across the same three measures. Among active Goblins users, the largest share of students solved and attempted 5 to 10 problems per week during the implementation period and spent 5 to 10 minutes per week on the platform. While active minutes and problems attempted are highly correlated with problems solved, each has important limitations. Active minutes include idle time spent on Goblins' platform, while problems attempted captures only problems opened by the student and does not indicate whether the student actually provided an answer. For this reason, the remainder of the analysis focuses on problems solved as a primary measure.

To understand how usage varies among users who had met the minimum threshold (i.e.,

the treated group), we disaggregate the distribution of total problems solved in [Table 4](#). We use the six dosage categories defined in the [section](#) Defining Treatment. These categories will also be applied to understand heterogeneous treatment effects.

RQ3: What is the effect of Goblins’ multimodal math tutor on student mathematics achievement? Do the effects differ by student subgroup (such as race/ethnicity, gender, and economic disadvantage) or by student baseline proficiency level?

We report the main outcome findings in [Table 7](#) and [Table 8](#). Each table reports the ATT of using Goblins’ AI math tutor. The first column reports estimates where the outcome is the within-district standardized MAP score, while the second column reports results using conditional growth index as a secondary outcome.

[Table 7](#) demonstrates that, conditional on pre-treatment characteristics, students who used Goblins — defined as meeting the minimum usage threshold of solving at least 2 problems per week — exhibited 0.11 standard deviations (SD) greater growth in their MAP math scores compared to their peers at DVUSD who either did not use Goblins or used the platform below the minimum threshold. When benchmarked against national norms, the estimated effect size is more than twice as large, corresponding to an overall effect of 0.24 standard deviations in MAP math achievement growth ($p < 0.001$).

The subgroup model findings are reported in [Table 8](#). We find no differential effects of Goblins usage by grade level, baseline proficiency, gender, and most racial subgroups ($p > 0.05$). We do find that when compared to White students, Hispanic students benefit more from Goblins by an additional 0.084 SD ($p < 0.05$). Importantly, students with an IEP also scored higher than their non-IEP peers by 0.122 SD ($p < 0.05$). While only marginally significant, students designated as English language learners scored higher by 0.191 SD compared to their peers ($p < 0.10$). We suspect that the multimodal features of

Goblins reduced barriers for students with an IEP and ELL students to interact with the tool compared to unimodal tools.

The results of the sensitivity analyses are presented in the Appendix. The covariate balance statistics indicate that the three alternative model specifications achieved similar balance to that of the main model (see [Table A7](#)), suggesting that the alternative models were successful in creating comparable treatment and control groups.

Overall, the estimated treatment effect of Goblins' AI math tutor was robust across three alternative models (see [Table A8](#)). Relative to the main specification, which estimated an effect of 0.11 SD, these alternative models produced effect size estimates within 0.01 standard deviations of the primary estimate, indicating a high degree of consistency across alternate specifications.

RQ4: What is the relationship between the level of usage and student achievement growth? How does it vary across classroom and school-level contexts?

[Table 9](#) presents results from an analysis that disaggregates Goblins usage into six dosage categories. The findings reveal a positive relationship between usage intensity and math achievement growth, with larger estimated effects among students with higher levels of engagement ($p < 0.001$). The strongest effects are observed among the highest-use groups: students who solved 10 to 19 problems per week and those who solved 20 or more problems per week demonstrated adjusted differences of 0.18 SD and 0.31 SD, respectively, on the within-district standardized MAP score and 0.40 SD and 0.69 SD, respectively, on CGI. All four estimates were statistically significant ($p < 0.001$).

Results from the moderation analyses are presented in [Table 10](#). Although statistically significant, the estimated effects are smaller among students whose teachers or schools have more than 50% of students below grade level at baseline, ranging from 0.07 to 0.08 SD

compared with the overall estimated effect of 0.11 SD. In contrast, Goblins usage is associated with a 0.14 SD increase in math achievement among students with novice teachers (i.e., fewer than five years of teaching experience), relative to students with novice teachers who did not use Goblins. The estimated treatment effect is also larger among Goblins users attending Title I schools, who exhibit a 0.18 SD increase in math achievement.

Conclusion

This study provides initial evidence on the implementation and effectiveness of Goblins, an AI-powered math tutor tool, during its pilot launch in the 2025–26 school year. The estimated effects are directionally positive and provide promising evidence regarding the potential of AI-based tutoring tools to support student learning when implemented at scale. Importantly, a substantial proportion of students engaged with Goblins at levels consistent to generalize usage patterns. The estimated treatment effects of 0.11 SD within the district and 0.24 SD on CGI reflect outcomes for approximately 50% of students in the district sample who met the minimum engagement threshold of 2 problems solved per week. Notably, the proportion of students meeting the treatment threshold exceeds that reported in prior evaluations of edtech interventions, including Khan Academy ([Weatherholtz et al. \(2022\)](#)) and IXL ([An \(2023\)](#)), in which approximately 5% of students met the treatment threshold, and [Marple et al. \(2019\)](#), which examined usage for only 12% of the study sample.

This study might also under-estimate Goblins’ impact, thus the results should be interpreted as conservative estimates. Based on our conversation with DVUSD personnel, many teachers in the district who did not engage with Goblins were using another edtech tool that had similar functionalities as Goblins. Thus, the comparison group students do not represent a typical business-as-usual group, lowering the effect size we estimate in this study.

As Goblins was made available to all middle school teachers in DVUSD, this study inherently has several limitations despite our best efforts in using all the data available with

a propensity score matching design. The first is the incomplete linkage between the Gobblins and the district datasets. Approximately 13% of Gobblins student users could not be matched to a student record in the district administrative data. The incomplete match rate is likely attributable, at least in part, to inconsistencies introduced through students self-entering their names on the Gobblins platform. The unmatched records may reflect one of two scenarios. First, some Gobblins users may not have been enrolled in the district during the study period despite being associated with the district in the platform data (e.g., due to account assignment issues or changes in enrollment). Second, some unmatched records may correspond to students who were enrolled in the district but could not be successfully linked to their administrative records. As a result, these students would be included in the comparison sample of students in the district.

The second limitation concerns the potential for spillover effects. Although Gobblins was made available to all teachers in the district, some teachers may have benefited from access to Gobblins without providing their students direct access to the platform. For example, teachers could have used Gobblins to generate or extract practice problems, assignments, worksheets, or formative assessments for use in their regular classroom instruction. As a result, students classified as untreated may have been indirectly exposed to instructional materials developed through the platform. While this form of indirect use is not observable in the data, it represents a potential source of treatment contamination that could attenuate the estimated treatment effects by reducing the distinction between the treatment and comparison groups.

The third limitation is that the product introduced new features during the implementation period. For example, the assessment mode feature, which allows the AI tutor to be turned off, was introduced after implementation began and therefore was not available when Gobblins was first introduced to the district in August 2025. The problems solved measure includes all problems completed during the implementation period, regardless of whether the AI tutor was available.

Implications for Practice

Given these findings, practitioners and decision makers should consider Goblins a potentially beneficial tool for supplementing math instruction. The platform’s flexibility allows teachers to tailor its use to different instructional needs and student groups. In particular, Goblins’ ability to create unlimited classes may facilitate the delivery of Tier 1, Tier 2, and Tier 3 instruction within a single platform, which may be an important feature for supporting effective implementation.

Practitioners should consider the extent to which these findings generalize beyond the study setting. Because the evaluation was conducted in a single district over one school year, the results may not generalize to districts with different student populations, instructional contexts, implementation practices, or levels of familiarity with the platform. Similarly, the findings should not be assumed to apply equally across grade bands, as students’ developmental characteristics, curricular content, and patterns of technology use may influence the effectiveness of the intervention. Finally, this evaluation captures effects over a single school year and therefore does not speak to whether effects would be sustained, increase, or diminish with multiple years of use. Replication across districts, grade levels, and longer implementation periods would provide stronger evidence regarding the conditions under which the observed effects are likely to persist and generalize.

Future Research

We have several recommendations for future studies. First, evaluations of edtech tools should collect user identifiers that are not based on self-reported information. Access to identifiers such as single sign-on (SSO) login IDs or student ID numbers obtained from the district can facilitate more complete linkage between district records and platform data. This, in turn, increases the likelihood that evaluations capture the full population of platform users, which is particularly important in experimental studies where attrition and incomplete

outcome data are key concerns.

Second, evaluations of edtech tools should consider “freezing” the product version introduced at the beginning of the implementation period. Substantial changes to the product’s pedagogy or student-facing functionality should be minimized during the study period. Maintaining a consistent product version would create a more stable treatment condition and facilitate clearer interpretation of the study’s findings.

References

- An, X. (2023). The impact of ixl on math and ela learning in new york state.
- Bacher-Hicks, A. and Koedel, C. (2023). Estimation and interpretation of teacher value added in research applications. *Handbook of the Economics of Education*, 6:93–134.
- Backes, B., DeArmond, M., Dizon-Ross, E., Goldhaber, D., and Salazar, A. (2026). A widening chasm: The divergent paths of high-and low-achieving students in algebra i after the pandemic. working paper no. 339-0526. *National Center for Analysis of Longitudinal Data in Education Research (CALDER)*.
- CEPR (2016). DreamBox Learning Achievement Growth in the Howard County Public School System and Rocketship Education. <https://cepr.harvard.edu/sites/g/files/omnuum9881/files/cepr/files/dreambox-key-findings.pdf>.
- Eames, T., Brunskill, E., Yamkovenko, B., Weatherholtz, K., and Oreopoulos, P. (2026). Computer-assisted learning in the real world: How khan academy influences student math learning. *Proceedings of the National Academy of Sciences*, 123(1):e2507708123.
- Foster, M. E. (2024). Evaluating the impact of supplemental computer-assisted math instruction in elementary school: A conceptual replication. *Journal of Research on Educational Effectiveness*, 17(1):94–118.
- Friedman, J. H. (2002). Stochastic gradient boosting. *Computational statistics & data analysis*, 38(4):367–378.
- Guryan, J., Ludwig, J., Bhatt, M. P., Cook, P. J., Davis, J. M., Dodge, K., Farkas, G., Fryer Jr, R. G., Mayer, S., Pollack, H., et al. (2023). Not too late: Improving academic outcomes among adolescents. *American Economic Review*, 113(3):738–765.
- Hirano, K. and Imbens, G. W. (2001). Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization. *Health Services and Outcomes research methodology*, 2(3):259–278.
- James, G., Witten, D., Hastie, T., Tibshirani, R., and Taylor, J. (2023). Statistical learning. In *An introduction to statistical learning: With applications in Python*, pages 15–67. Springer.
- Kane, T. J., McCaffrey, D. F., Miller, T., and Staiger, D. O. (2013). Have we identified effective teachers? validating measures of effective teaching using random assignment. *Research Paper. MET Project. Bill & Melinda Gates Foundation*.
- Keller, B. and Tipton, E. (2016). Propensity score analysis in r: A software review. *Journal of Educational and Behavioral Statistics*, 41(3):326–348.
- Kong, A., Liu, J. S., and Wong, W. H. (1994). Sequential imputations and bayesian missing data problems. *Journal of the American statistical association*, 89(425):278–288.

- Kraft, M. A. (2020). Interpreting effect sizes of education interventions. *Educational researcher*, 49(4):241–253.
- Leite, W., Zhang, H., Collier, Z., Chawla, K., Kong, L., Lee, Y., Quan, J., and Soyoye, O. (2025). Machine learning for propensity score estimation: A systematic review and reporting guidelines. *Psychological Methods*.
- Little, R. J. and Vartivarian, S. (2005). Does weighting for nonresponse increase the variance of survey means? *Survey Methodology*, 31(2):161.
- Marple, S., Jaquet, K., Laudone, A., Sewell, J., and Liepmann, K. (2019). i-ready in 7th grade math classes: A mixed methods case study. Technical report, Technical Report. WestEd, San Francisco, CA. [https://www . . .](https://www...)
- McCaffrey, D. F., Ridgeway, G., and Morral, A. R. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological methods*, 9(4):403.
- Nickow, A., Oreopoulos, P., and Quan, V. (2024). The promise of tutoring for prek–12 learning: A systematic review and meta-analysis of the experimental evidence. *American Educational Research Journal*, 61(1):74–107.
- NWEA (2026). NWEA Connection — [connection.nwea.org. https://connection.nwea.org/s/article/{C}onditional-{G}rowth-{I}ndex?language=en_{U}{S}](https://connection.nwea.org/s/article/{C}onditional-{G}rowth-{I}ndex?language=en_{U}{S}).
- Oreopoulos, P., Keyes-Krysakowski, O., and Agarwal, D. (2026). How in-school supervised ed-tech support produces massive learning gains: A khan academy field experiment in india. Technical report, National Bureau of Economic Research.
- Pane, J. F., Griffin, B. A., McCaffrey, D. F., and Karam, R. (2014). Effectiveness of cognitive tutor algebra i at scale. *Educational Evaluation and Policy Analysis*, 36(2):127–144.
- Parsons, E., Koedel, C., and Tan, L. (2019). Accounting for student disadvantage in value-added models. *Journal of Educational and Behavioral Statistics*, 44(2):144–179.
- Polikar, R. (2012). Ensemble learning. In *Ensemble machine learning*, pages 1–34. Springer.
- Ridgeway, G., McCaffrey, D. F., Morral, A. R., Cefalu, M., Burgette, L. F., Pane, J. D., and Griffin, B. A. (2022). *Toolkit for weighting and analysis of nonequivalent groups: a tutorial for the R TWANG package*. Rand Santa Monica, Calif.
- Robinson, C. D., Gormley, D., Ribeiro, A. T., and Loeb, S. (2026). Access is not enough: Human support improves engagement with ai tutoring. edworkingpaper no. 26-1451. *Annenberg Institute for School Reform at Brown University*.
- Schechter, R. L. and Chase, P. (2025). Ebli teacher and student lessons efficacy study with whole class, core instruction: Grades k-5 using nwea map growth. *Online Submission*.
- Szatrowski, A., Rickel, J., and Rosemond, C. (2022). Efficacy analysis of zearn math in dc public schools. *Online Submission*.

- Team, L., Wang, A., Rysbek, A., Huber, A., Nambiar, A., Kenolty, A., Caulfield, B., Lilley-Draper, B., Groot, B., Veprek, B., et al. (2025). Ai tutoring can safely and effectively support students: An exploratory rct in uk classrooms. *arXiv preprint arXiv:2512.23633*.
- Thum, Y. M. (2014). The effective use of student and school descriptive indicators of learning progress: From the conditional growth index to the learning productivity measurement system. *Online Submission*.
- Wang, R. E., Ribeiro, A. T., Robinson, C. D., Loeb, S., and Demszky, D. (2024). Tutor copilot: A human-ai approach for scaling real-time expertise. *arXiv preprint arXiv:2410.03017*.
- Weatherholtz, K., Grimaldi, P. J., and Hill, K. M. (2022). Use of map accelerator associated with better-than-projected gains in map growth scores. mountain view, ca: Khan academy. *Online Submission*.
- Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data*. MIT press.

Tables and Figures

Figure 1: Overview of Goblins' Platform

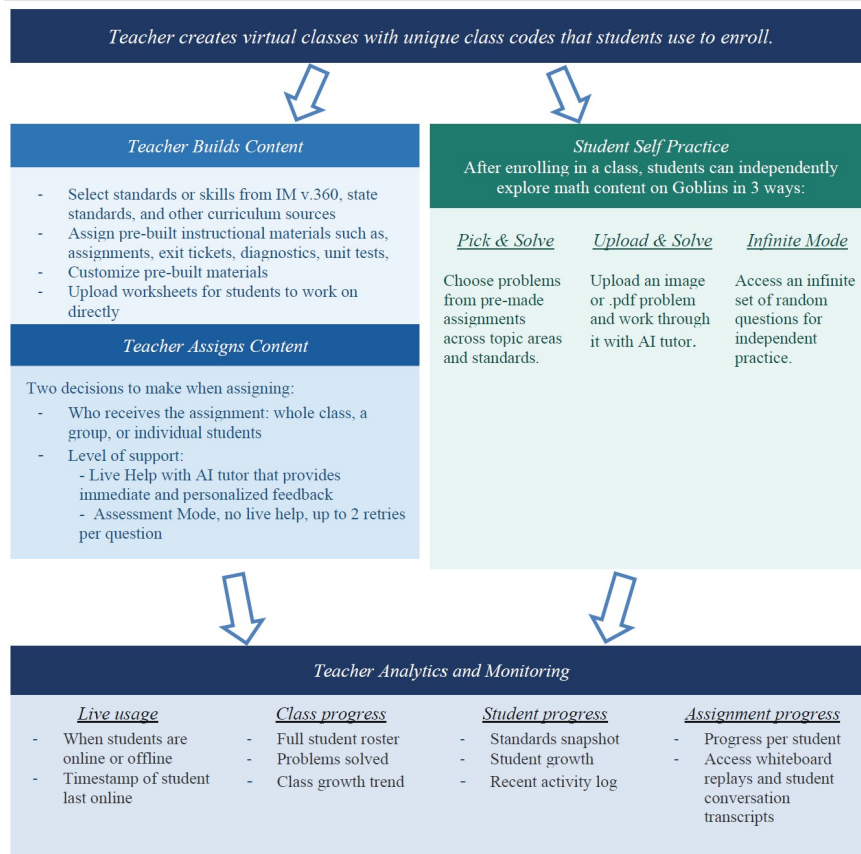
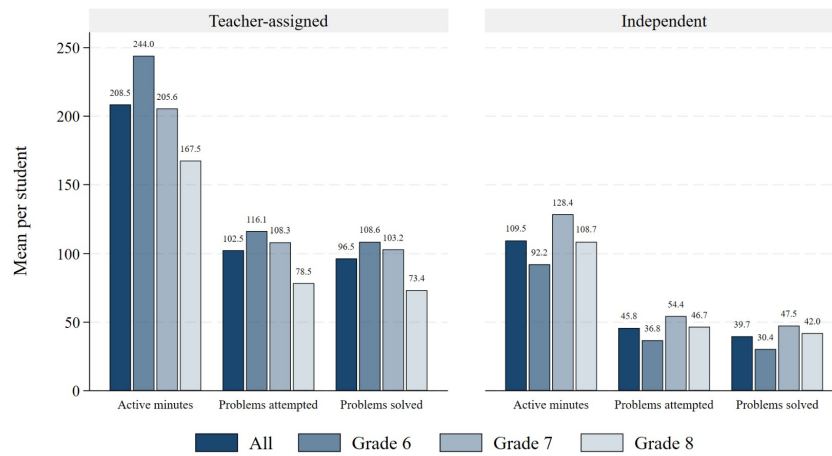


Figure 2: Average Goblins usage by mode and by grade



NOTES: This figure presents average student-level usage, reported separately by grade and for the matched user sample ($n = 4,833$). “Teacher-assigned usage” refers to Goblins activity that was explicitly assigned by a student’s math teacher, whereas “independent usage” refers to Goblins activity that was not assigned by the teacher.

Table 1: Research Question Summary

Research Question	Realized Measure(s)	Data Source	Timing of Data Collection
How did teachers implement Gobblins in their instruction (e.g., universal supports versus targeted interventions)?	Gobblins usage metrics such as • Student users per Gobblins teacher • Gobblins classes created per teacher • Enrollment size per Gobblins class • Share of students enrolled in more than one class • Problems assigned per student, by class size band (1–2, 3–15, 16+) as proxies for Tier 3, Tier 2, Tier 1	Gobblins usage data & district teacher-student records	Continuous collection of Gobblins usage from September 2025 – May 2026
How did students engage on the Gobblins platform? What engagement patterns were predictive of content mastery and learning outcomes?	Gobblins usage metrics such as – • Active minutes • Problems attempted • Problems solved (as a continuous measure and binned into six dosage categories) • Correlation among usage measures and with spring 2026 MAP math score Each measure is reported as totals and weekly averages over 27 weeks, by grade and by mode (teacher-assigned vs. independent)	Gobblins usage data	Continuous collection of Gobblins usage from September 2025 – May 2026
What is the effect of Gobblins’ multimodal math tutor on student mathematics achievement? Do the effects differ by student subgroup or by student baseline proficiency level?	Spring 2026 MAP math score; fall 2025 MAP math score	Gobblins usage data & district student achievement data	Continuous collection of Gobblins usage from September 2025 – May 2026. District will provide MAP data (baseline and endline) in late spring 2026
What is the relationship between the level of usage and student achievement growth? How does it vary across different contexts?	Spring 2026 MAP math score; fall 2025 MAP math score	Gobblins usage data & district student achievement data	Continuous collection of Gobblins usage from September 2025 – May 2026. District will provide MAP data (baseline and endline) in late spring 2026

NOTES: All Gobblins usage data and district-provided achievement data will be checked for completeness, consistency, and accuracy prior to analysis, including range checks, deduplication, and reconciliation of student, teacher, and school identifiers across data sources. Data will be stored on encrypted, access-controlled servers, with access limited to approved research staff under the terms of the data transfer use agreement with Deer Valley USD. Personally identifiable information will be handled in accordance with FERPA and the district’s data privacy requirements, and any data shared or reported externally will be de-identified or aggregated as appropriate.

Table 2: Comparison of unweighted vs. weighted statistics

Method	Treated group	Control group	ESS Control group	Maximum SMD	Mean SMD	Iterations
Unweighted	3,088	2,977	2,977	1.38	0.15	NA
Weighted	3,088	2,977	1,256	0.24	0.05	1,131

NOTES: SMD denotes standardized mean difference; ESS denotes effective sample size. There are no iterations reported for the unweighted comparison as this uses raw treated and comparison means and not propensity score estimations.

Table 3: Pre-treatment balance check

Covariate	Treated Mean	Comparison Mean (Weighted)	Standardized Mean Difference
Fall math score (z)	0.06	-0.04	0.12**
Grade 6	0.38	0.36	0.04
Grade 7	0.37	0.35	0.05
Grade 8	0.25	0.29	-0.09*
Male	0.51	0.55	-0.09*
Has an IEP	0.08	0.13	-0.17***
English Language Learner	0.06	0.07	-0.06
American Indian	0.01	0.01	0.05*
Asian	0.04	0.03	0.08**
Black	0.03	0.03	-0.01
Hispanic	0.24	0.28	-0.09*
Multiple Races	0.06	0.04	0.07*
Native Hawaiian	0.00	0.00	0.02
White	0.62	0.61	0.00
Teacher is female	0.73	0.66	0.16***
Teacher is White	0.90	0.91	-0.02
Teacher has master's	0.30	0.30	0.00
Teacher Experience < 5 Years	0.28	0.29	-0.04
Teacher Experience 5–15 Years	0.55	0.52	0.05
Teacher Experience > 15 Years	0.18	0.19	-0.02
Teacher value added	0.00	0.01	-0.13***

NOTES: SD denotes standard deviation. The comparison group is reweighted using propensity weights selected with the es.mean stopping criterion, which minimizes the mean absolute standardized mean difference across covariates. Statistical significance is denoted with $^+p < 0.10$, $*p < 0.05$, $**p < 0.01$, $***p < 0.001$.

Table 4: Goblins dosage levels

Average problems per week	Count	Percent	Cumulative
Fewer than 2	2,977	49.08	49.08
2 but fewer than 3	598	9.86	58.94
3 to fewer than 4	516	8.51	67.45
4 to fewer than 5	359	5.92	73.37
5 to fewer than 10	939	15.48	88.85
10 to fewer than 20	512	8.44	97.30
20 or more	164	2.70	100.00
Total	6,065	100.00	

NOTES: Usage levels represent average weekly engagement for all grades 6 to 8 students at Deer Valley ($n=6,065$) assuming 27 weeks of instruction. Students who did not activate a Goblins account were included in the lowest category of usage (i.e., solving fewer than 2 problems per week).

Table 5: Descriptive statistics of teacher-created classes on Goblins

Panel A: Number of student users per Goblins teacher	
Student users (N)	4,833
Minimum	1
Mean	53.5
Standard deviation	34.7
Median	50
25th percentile	22
75th percentile	79
Maximum	125
Panel B: Number of Goblins-created classes per Goblins teacher	
Teacher (N)	90
Teachers with 1 Goblins class (%)	31.1
Teachers with 2 Goblins classes (%)	20.0
Teachers with 3 Goblins classes (%)	8.9
Teachers with 4+ Goblins classes (%)	40.0
Panel C: Enrolled students per Goblins class	Value
Goblins classes (N)	273
Goblins classes with 1–2 students (%)	6.2
Goblins classes with 3–5 students (%)	5.9
Goblins classes with 6–9 students (%)	7.3
Goblins classes with 10–19 students (%)	29.3
Goblins classes with 20–29 students (%)	40.7
Goblins classes with 30+ students (%)	10.6

Table 6: Correlation Matrix Comparing Student Engagement

	Total problems solved	Total problems attempted	Total active minutes	Spring test score (Z)
Total problems solved	1.0000			
Total problems attempted	0.9861	1.0000		
Total active minutes	0.9070	0.9078	1.0000	
Spring test score (Z)	0.1241	0.0960	-0.0182	1.0000

Table 7: Estimated treatment effects

	Within-District Z-Scored EOY	Conditional Growth Index (CGI)
Treatment	0.109*** (0.021)	0.237*** (0.047)

NOTES: All models control for all covariates used in propensity score estimation and include school fixed effect. Heteroskedasticity-robust standard errors are clustered at the teacher level and are reported in parentheses. Statistical significance levels: $^+p < 0.10$, $*p < 0.05$, $**p < 0.01$, $***p < 0.001$.

Table 8: Estimated treatment effects for subpopulations

	Within-District Z-Scored EOY	Conditional Growth Index EOY
Panel A: By grade level (reference group is grade 6)		
Grade 7 ($n = 1,150$)	-0.028 (0.047)	-0.020 (0.106)
Grade 8 ($n = 778$)	-0.023 (0.055)	-0.016 (0.119)
Panel B: By race/ethnicity (reference group are White students)		
Asian ($n = 127$)	0.101 (0.076)	0.219 (0.168)
Black ($n = 94$)	0.117 (0.113)	0.254 (0.247)
Hispanic ($n = 735$)	0.084* (0.041)	0.184* (0.090)
Multiple races ($n = 187$)	0.003 (0.065)	0.019 (0.143)
Panel C: By baseline proficiency (reference group are students at or above grade level)		
One grade level below ($n = 190$)	-0.029 (0.040)	-0.051 (0.092)
Two grade levels below ($n = 150$)	0.064 (0.055)	0.102 (0.118)
Panel D: By gender		
Male ($n = 1,579$)	-0.008 (0.040)	-0.029 (0.087)
Panel E: By disability status		
Has an IEP ($n = 259$)	0.122* (0.047)	0.273* (0.106)
Panel F: By English proficiency status		
English Language Learner ($n = 170$)	0.191+ (0.114)	0.389 (0.248)

NOTES: n captures the number of student observations in the subgroup who met or exceeded the threshold of 2 problems solved/week. All models control for all covariates used in propensity score estimation and include school fixed effect. Heteroskedasticity-robust standard errors are clustered at the teacher level and are reported in parentheses. Statistical significance levels: $^+p < 0.10$, $*p < 0.05$, $**p < 0.01$, $***p < 0.001$.

Table 9: Estimated dosage effects

Average problems solved per week	Within-District Z-Scored EOY	Conditional Growth Index
At least 2 but fewer than 3	0.045+ (0.026)	0.099+ (0.057)
3 to fewer than 4	0.084** (0.027)	0.180** (0.059)
4 to fewer than 5	0.080** (0.028)	0.176** (0.063)
5 to fewer than 10	0.116*** (0.028)	0.247*** (0.063)
10 to fewer than 20	0.184*** (0.032)	0.400*** (0.071)
20 or more	0.312*** (0.049)	0.686*** (0.115)

NOTES: Estimates represent the treatment effect for each category when interacted with the treatment indicator. All models control for all covariates used in propensity score estimation and include school fixed effect. Heteroskedasticity-robust standard errors are clustered at the teacher level and are reported in parentheses. Statistical significance levels: $^+p < 0.10$, $^*p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Table 10: Estimated treatment effects for moderators

	Teacher has >50% of students below grade level (baseline) (1)	Teacher has <5 years of experience (2)	School has >50% of students below grade level (baseline) (3)	School receives Title I funding (4)
Outcome: Within-District EOY				
Treatment	0.074*** (0.020)	0.138** (0.042)	0.080*** (0.020)	0.176** (0.054)
Outcome: Conditional Growth Index				
Treatment	0.157*** (0.047)	0.309** (0.093)	0.176*** (0.046)	0.391** (0.117)

NOTES: Estimates represent the treatment effect for each moderation characteristic interacted with the treatment indicator. The models control for all covariates used in propensity score estimation and include school fixed effect. Heteroskedasticity-robust standard errors are clustered at the teacher level and are reported in parentheses. Statistical significance levels: $^+p < 0.10$, $^*p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Technical Appendix

Teacher value added in propensity scores

The propensity score estimations include teacher value added (TVA) scores as a covariate. TVA is an effort to parse a teacher’s individual contribution to student learning. Following value-added model (VAM) literature, we estimate TVA as a linear regression of math achievement in period t (i.e., 2024-25 school year) with a lagged-score specification of 1 prior achievement period, $t - 1$ (i.e., 2023-24 school year).² Following [Parsons et al. \(2019\)](#), we estimate the “average residual” framework where the estimation occurs in two stages:

$$Y_{ijst} = \beta_0 + Y_{it-1}\beta_1 + X_{it}\beta_2 + S_{jg}\beta_3 + \eta_{ijst} \quad (3)$$

$$\eta_{ijst} = T_{jt}\theta + v_{ijst} \quad (4)$$

where the first stage Y_{ijst} captures the NWEA MAP Spring math score for student i with teacher j at school s at the end of year t . Y_{it-1} is a vector of NWEA Fall math achievement. X_{it} is a vector of student characteristics and S_{jg} a vector of leave-one-out peer characteristics aggregated at the teacher-by-grade level. T_{jt} is a vector of teacher-by-year indicators where the value for the teacher j assigned to student s in year t equals 1 and all other indicators are set to zero.

The second stage calculates individual teacher j ’s value-added in vector θ when it is regressed on the residuals estimated in the first stage. The two stages approach uses both between- and within- teacher variation ([Bacher-Hicks and Koedel, 2023](#); [Parsons et al., 2019](#); [Kane et al., 2013](#)).

[A2](#) presents aggregate teacher value-added scores across the treatment and comparison groups. Approximately 20 percent of teachers in the analytic sample have missing value-added scores, with no statistically significant difference in missingness between teachers

²TVA calculations are made on data prior to the 2025-26 school year, when the intervention was rolled out.

using Goblines and those not using Goblines ($p > 0.05$). To avoid losing observations through listwise deletion, we impute missing value-added scores as zero. This approach preserves the full analytic sample while accounting for the absence of observed value-added measures.

Appendix Tables and Figures

Table A1: Comparing matched and unmatched Goblins users

	Matched (1)	Unmatched Goblins Users (2)	Difference (3)
Grade 5	0.00	0.04	−0.04***
Grade 6	0.29	0.47	−0.17***
Grade 7	0.31	0.23	0.08***
Grade 8	0.24	0.16	0.09***
Teacher assigned activity			
Active minutes	208.53	135.85	72.68***
Problems attempted	102.46	68.53	33.93***
Problems solved	96.49	64.87	31.62***
Independent activity			
Active minutes	109.52	82.12	27.40***
Problems attempted	45.77	31.05	14.72***
Problems solved	39.67	27.88	11.79***
Observations	4,833	725	5,558

NOTES: Columns (1)–(3) report means. Statistical significance levels: $^+p < 0.10$, $*p < 0.05$, $**p < 0.01$, $***p < 0.001$.

Table A2: Sample Characteristics

	All Students (1)	Goblins Users (2)	Non-Goblins Users (3)	Difference (4)
Grade 6	0.34	0.36	0.23	0.13***
Grade 7	0.33	0.35	0.28	0.07***
Grade 8	0.33	0.29	0.49	-0.20***
Student Characteristics				
Female	0.47	0.48	0.45	0.02
Asian	0.04	0.04	0.05	-0.02**
Black	0.03	0.03	0.03	0.00
Hispanic	0.23	0.24	0.17	0.08***
White	0.63	0.62	0.69	-0.07***
Other race/ethnicity	0.07	0.07	0.06	0.01
EL Learner	0.05	0.06	0.03	0.03***
Has an IEP	0.12	0.12	0.11	0.02*
Beginning-of-Year Math Achievement				
MAP Math score (<i>SD</i>)	0.03 (0.98)	-0.06 (0.95)	0.34 (1.03)	-0.40*** (-12.27)
At grade level	0.66	0.63	0.78	-0.16***
Teacher Characteristics				
Female	0.73	0.74	0.70	0.04***
White	0.92	0.91	0.97	-0.07***
Has masters	0.30	0.31	0.28	0.03**
Experience < 5	0.23	0.26	0.12	0.14***
Experience 5–15	0.53	0.55	0.48	0.07***
Experience > 15	0.24	0.19	0.40	-0.21***
Value added score (<i>SD</i>)	0.00	0.00	0.00	0.00
Missing value added score	0.20	0.20	0.18	0.02
Observations	6,065	4,833	1,232	6,065

NOTES: Columns (1)–(3) report means. Column (4) reports the difference in means between reported values in Columns (2) and (3). Statistical significance levels: $^+p < 0.10$, $^*p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Table A3: Detailed Goblins usage statistics

Usage level	Count	Percent	Cumulative Percent
Average Problems Attempted per Week			
Fewer than 2	2,784	45.90	45.90
2 but fewer than 3	621	10.24	56.14
3 to fewer than 4	517	8.52	64.67
4 to fewer than 5	389	6.41	71.08
5 to fewer than 10	995	16.41	87.49
10 to fewer than 20	568	9.37	96.85
20 or more	191	3.15	100.00
Total	6,065	100.0	
Average Problems Solved per Week			
Fewer than 2	2,977	49.08	49.08
2 but fewer than 3	598	9.86	58.94
3 to fewer than 4	516	8.51	67.45
4 to fewer than 5	359	5.92	73.37
5 to fewer than 10	939	15.48	88.85
10 to fewer than 20	512	8.44	97.30
20 or more	164	2.70	100.00
Total	6,065	100.0	
Average Active Minutes per Week			
Fewer than 5	2,935	48.39	48.39
5 to fewer than 10	1,133	18.68	67.07
10 to fewer than 15	674	11.11	78.19
15 to fewer than 20	412	6.79	84.98
20 to fewer than 30	539	8.89	93.87
30 or more	372	6.13	100.00
Total	6,065	100.0	

Notes: Usage levels represent average weekly engagement for all grades 6 to 8 students at Deer Valley ($n=6,065$) when disaggregating total Goblins usage into 27 weeks of instruction. Students who did not activate a Goblins account were included in the lowest category of usage (i.e., fewer than 2 problems or fewer than 5 minutes).

Table A4: Distributions of students above and below minimum threshold

	All Students (1)	At or above threshold (2)	Below threshold (3)	Difference (4)
Student Characteristics				
Grade 6	0.34	0.38	0.30	0.08***
Grade 7	0.33	0.37	0.29	0.08***
Grade 8	0.33	0.25	0.41	-0.16***
Female	0.47	0.49	0.45	0.04***
Asian	0.04	0.04	0.04	0.00
Black	0.03	0.03	0.03	-0.00
Hispanic	0.23	0.24	0.22	0.02*
White	0.63	0.62	0.66	-0.04***
Other race/ethnicity	0.07	0.08	0.06	0.02***
EL Learner	0.05	0.06	0.05	0.00
Has an IEP	0.12	0.08	0.15	-0.07***
Beginning-of-Year Math Achievement				
MAP Math score (<i>SD</i>)	0.03	0.06	-0.01	0.07***
At grade level Fall	0.66	0.68	0.64	0.05***
Teacher Characteristics				
Female	0.73	0.73	0.73	-0.00
White	0.92	0.90	0.94	-0.04***
Has masters	0.30	0.30	0.31	-0.01
< 5 Years Experience	0.23	0.28	0.19	0.09***
5–15 Years Experience	0.53	0.55	0.52	0.03**
> 15 Years Experience	0.24	0.18	0.29	-0.11***
Value added score (<i>Z</i>)	0.00	-0.00	0.01	-0.01***
Observations	6,065	3,088	2,997	6,065

NOTES: Columns (1)–(3) report means and Column (4) reports the difference in means between reported values in Columns (2) and (3). Statistical significance levels: ⁺ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table A5: School Characteristics

School Name	BOY (Z)	EOY (Z)	ELL	IEP	Teachers (N)	Total students in grades 6–8	Students in analytic sample ($N = 6,065$)	≥ 2 problems solved/week ($N = 3,088$)
Las Brisas Elementary School	0.53	0.59	0%	8%	2	67	60	98%
Greenbrier Elementary School	-0.13	-0.06	9%	19%	4	60	57	89%
Aspire, Deer Valley's Online Academy	-0.16	-0.32	5%	21%	1	43	19	89%
Park Meadows Elementary School	-0.35	-0.29	22%	10%	2	75	67	84%
Norterra Canyon School	-0.06	0.02	2%	14%	3	192	168	82%
Copper Creek Elementary School	0.31	0.35	4%	9%	2	96	76	79%
Desert Sky Middle School	-0.37	-0.40	6%	16%	7	560	477	78%
Canyon Springs STEM Academy	0.05	0.07	1%	11%	6	313	288	78%
Deer Valley Middle School	-0.63	-0.62	19%	18%	6	400	331	78%
Desert Mountain School	0.00	0.12	1%	9%	4	156	151	77%
Village Meadows Elementary School	-0.41	-0.43	17%	28%	3	51	46	67%
Arrowhead Elementary School	-0.27	-0.30	8%	17%	3	73	64	64%
Sunrise Elementary School	-0.94	-0.54	9%	21%	3	50	47	64%
Terramar Academy of the Arts	0.23	0.15	3%	9%	5	368	235	63%

School Name	BOY (Z)	EOY (Z)	ELL	IEP	Teachers (N)	Total students in grades 6–8	Students in analytic sample	>2 problems solved/week
Union Park School	0.09	0.15	9%	7%	3	252	241	62%
Mirage Elementary School	-0.72	-0.70	8%	27%	3	82	78	60%
West Wing School	0.18	0.16	2%	8%	3	334	251	60%
Peoria Hills School	-0.47	-0.40	13%	13%	4	201	174	59%
Inspiration Mountain School	0.20	0.30	1%	8%	4	315	305	56%
Gavilan Peak School	0.23	0.18	3%	9%	4	195	174	54%
Highland Lakes School	0.15	0.17	2%	12%	9	482	422	51%
Stetson Hills School	0.28	0.39	6%	8%	6	301	289	44%
Mountain Shadows Elementary School	-0.44	0.06	10%	12%	2	53	49	39%
Desert Sage Elementary School	-0.09	-0.01	4%	14%	4	88	83	35%
Legend Springs Elementary School	0.16	0.42	7%	9%	3	91	86	34%
Esperanza Elementary School	-0.52	-0.78	20%	22%	1	65	50	30%
Anthem School	0.00	0.15	5%	12%	4	154	143	27%
Hillcrest Middle School	0.30	0.28	3%	11%	10	718	676	26%
Sunset Ridge School	0.19	0.19	3%	14%	4	177	155	21%
Sonoran Foothills School	0.08	0.21	3%	10%	3	318	188	14%
Diamond Canyon School	0.36	0.32	2%	8%	4	279	262	13%
New River Elementary School	-0.34	-0.52	0%	28%	2	20	18	11%
Sierra Verde STEAM Academy	0.27	0.32	5%	7%	5	323	280	4%

School Name	BOY (Z)	EOY (Z)	ELL	IEP	Teachers (N)	Total students in grades 6–8	Students in analytic sample	>2 problems solved/week
Traditional Academy at Bellair	0.34	0.36	0%	17%	2	32	30	0%
Vista Peak School	-1.78	-1.77	0%	100%	2	14	9	0%
Constitution Elementary School	–	–	–	–	–	61	0	0%
Boulder Creek High School	–	–	–	–	–	17	0	0%
Mountain Ridge High School	–	–	–	–	–	5	0	0%
Sandra Day O'Connor High School	–	–	–	–	–	1	0	0%
Deer Valley High School	–	–	–	–	–	1	0	0%

NOTES: Columns (1)-(5), (7) and (8) compute their values from the analytic sample of 6,065 students. Column (6) shows the values from the Deer Valley district Fall records ($N = 7,083$)

Table A6: Goblins Assigned Problems Disaggregated by Class Size

Panel A: classrooms with 1–2 students (proxy for Tier 3)	
Problems assigned per student	Value
Student enrollments in subsample (N)	24
Distinct students in subsample (N)	21
Classrooms in subsample (N)	17
Minimum	9
Mean	763.1
Standard deviation	948.6
Median	161
25th percentile	161
75th percentile	1422
Maximum	3973
Panel B: classrooms with 3–15 students (proxy for Tier 2)	
Problems assigned per student	Value
Student enrollments in subsample (N)	668
Distinct students in subsample (N)	608
Classrooms in subsample (N)	72
Minimum	36
Mean	890.8
Standard deviation	939.2
Median	735
25th percentile	238
75th percentile	1360
Maximum	5520
Panel C: classrooms with 16+ students (proxy for Tier 1)	
Problems assigned per student	Value
Student enrollments in subsample (N)	4,691
Distinct students in subsample (N)	4,330
Minimum	13
Mean	975.7
Standard deviation	1100.9
Median	659
25th percentile	356
75th percentile	991
Maximum	5520
Classrooms in subsample (N)	184

Table A7: Sensitivity Model Balance Check

Sensitivity Model	Treated group	Unweighted control group	ESS (Weighted) Control group	Maximum SMD	Mean SMD	Iterations
(1) Alternate threshold of 3 problems solved/wk	2,490	3,575	1,079	0.20	0.05	9,993
(2) Including economic disadvantage	2,463	2,437	896	0.21	0.06	1192
(3) Imposing a minimum support rule	3,010	2,919	1,213	0.24	0.05	1,342

NOTES: SMD denotes standardized mean difference; ESS denotes effective sample size. The comparison group is reweighted using propensity weights selected with the es.mean stopping criterion, which minimizes the mean absolute standardized mean difference across covariates.

Table A8: Estimated treatment effects with sensitivity models

	Treatment is 3 problems solved/week (1)	Including economic disadvantage (2)	With minimum support rule (3)
Outcome: Within-District EOY			
Treatment	0.118*** (0.021)	0.118*** (0.025)	0.111*** (0.021)
Outcome: Conditional Growth Index			
Treatment	0.254*** (0.046)	0.254*** (0.057)	0.240*** (0.047)

NOTES: ll models control for all covariates used in propensity score estimation. Models (1), (3), and (4) include school fixed effect. Heteroskedasticity-robust standard errors are clustered at the teacher level and are reported in parentheses. Statistical significance levels: $^+p < 0.10$, $*p < 0.05$, $**p < 0.01$, $***p < 0.001$