



Teacher-Mediated Multimodal Platform and Math Achievement: Evidence from Goblins' Yearlong District Implementation

Anika Alam

University of Maryland,
College Park

Nicolas Dominguez-Mel

University of Maryland,
College Park

Jing Liu

University of Maryland,
College Park

Despite the rapid adoption of AI-powered learning tools in K–12 classrooms, there is limited evidence about their effectiveness in improving student learning. This paper examines the usage of Goblins—a teacher-mediated multimodal AI math platform that allows students to type, draw, and speak while receiving personalized feedback—and how that usage associates with middle school math achievement. The study took place in a large Arizona school district where Goblins was offered to all middle grade math teachers in the 2025–26 school year. Using a weighted propensity design, we find that Goblins usage is associated with a 0.11 standard deviation increase in MAP math scores within the district where students in the highest usage group demonstrate a growth of 0.31 standard deviations. These results contribute to emerging research that examines the extent to which sustained engagement with AI-powered tutoring can improve math achievement.

VERSION: October 2026

Suggested citation: Alam, Anika, Nicolas Dominguez-Mel, and Jing Liu. (2026). Teacher-Mediated Multimodal Platform and Math Achievement: Evidence from Goblins' Yearlong District Implementation. (EdWorkingPaper: 26-1611). Retrieved from Annenberg Institute at Brown University: <https://doi.org/10.26300/vcdh-yc23>

Teacher-Mediated Multimodal Platform and Math Achievement: Evidence from Goblins' Yearlong District Implementation*

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

Center for Educational Data Science and Innovation (EDSI)

University of Maryland, College Park

October 7, 2026

ABSTRACT: Despite the rapid adoption of AI-powered learning tools in K–12 classrooms, there is limited evidence about their effectiveness in improving student learning. This paper examines the usage of Goblins—a teacher-mediated multimodal AI math platform that allows students to type, draw, and speak while receiving personalized feedback—and how that usage associates with middle school math achievement. The study took place in a large Arizona school district where Goblins was offered to all middle grade math teachers in the 2025–26 school year. Using a weighted propensity design, we find that Goblins usage is associated with a 0.11 standard deviation increase in MAP math scores within the district where students in the highest usage group demonstrate a growth of 0.31 standard deviations. These results contribute to emerging research that examines the extent to which sustained engagement with AI-powered tutoring can improve math achievement.

KEYWORDS: education technology, artificial intelligence, computer assisted learning, math achievement

*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. We thank Deer Valley Unified School District’s leaders, participating teachers, and students for making this study possible. The study was preregistered, 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.

[†]Authors are listed alphabetically. Correspondence regarding this paper should be sent to Jing Liu (jliu28@umd.edu).

Introduction

Expanding access to effective, individualized instruction is a central challenge for K–12 education. 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; Robinson and Loeb, 2021). Yet delivering high-impact tutoring at scale remains difficult: recruiting, training, and compensating tutors requires substantial resources, limiting schools’ and districts’ capacity to reach all students who could benefit (Kraft and Falken, 2021; Bhatt et al., 2024). These constraints motivate a broader question: how can schools expand access to individualized learning support while preserving the instructional quality that makes tutoring effective?

Educational technology (edtech) tools offer a potential alternative by delivering individualized learning support at a substantially lower cost per student while remaining highly scalable (Escueta et al., 2020; Muralidharan et al., 2019). Recent advances in adaptive learning platforms and generative AI have further expanded the ability of edtech tools to provide personalized instruction, immediate feedback, and targeted practice (VanLehn, 2011; Kulik and Fletcher, 2016; Wang et al., 2025; Kestin et al., 2025). As a result, schools have increasingly adopted these technologies as a means of supplementing classroom instruction and addressing unfinished learning (Diliberti and Schwartz, 2024). Despite the rapid deployment of edtech tools in K–12 settings, rigorous evidence on their effectiveness remains relatively limited (Escueta et al., 2020; Fesler et al., 2026; Turano et al., 2026). 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 (Holt, 2024; Eames et al., 2026; Oreopoulos et al., 2024).

This quasi-experimental study¹ examines the usage of *Goblins*, an online math tool with an AI math tutor that differentiates in real-time by delivering immediate, personalized feed-

¹The pre-analysis plan is publicly available at <https://osf.io/5kwbp/overview>. We explain how we deviated from this plan in making our empirical decisions in the text.

back to student users, and how that usages associates with student learning outcomes. Two distinct features of Goblins make it a worthwhile tool to study. First, the tool usage is mediated by teachers, who shape students’ access and instructional use. After activating their Goblins account, teachers can create Goblins classes, choose which students can enroll in these classes, and assign content to enrolled students. Enrolled students could also independently explore math content that was not assigned by their teacher. Second, the tool is highly accessible. Goblins allows students to speak through a problem, show their work on a whiteboard, or type their responses, and is supported in seven languages.

This study took place in Deer Valley Unified School District (DVUSD), a large district in Arizona that serves over 32,000 students across 42 campuses. During the 2025–26 school year, Goblins was offered to all math teachers teaching middle school grades and yielded substantial adoption: 72.6% of teachers activated their Goblins license. Of the 6,065 students in our analytic sample, 50.9% averaged at least two problems solved per week during the implementation period. Using a machine learning-based weighted propensity score design, we compare changes in math achievement between Goblins users and observationally similar students who either did not use Goblins or used it below the minimum threshold of meaningful engagement after a full year of implementation.

This study yielded several key findings. First, students at or above the threshold of 2 problems solved per week scored 0.109 SD higher on the within-district standardized spring NWEA MAP math assessment. Second, there is a strong positive relationship between Goblins usage level and math achievement, where the strongest associations 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. Third, students with an IEP, Hispanic students, and English language learners appear to 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 carefully designed, generative AI-powered instruction tools can extend beyond expanding access to instructional support to produce measurable improvements in student learning when implemented at meaningful levels of use.

It is important to note several caveats when interpreting these findings. First, this study measures students’ usage of the Goblins platform broadly, and does *not* isolate usage based on their engagement with Goblins’ AI tutor. This distinction is important because teachers can disable the AI tutor when assigning content to students. Thus, the estimates should not be interpreted as measuring the effects of AI tutor engagement. Second, this is an observational study that compares voluntary users of Goblins with an observationally similar group identified using propensity scores. As we only have cross-sectional data and there is no exogenous variation we can leverage to fully address selection bias due to students’ self selection into using Goblins, our estimates are still vulnerable to bias due to unobservables. For these reasons, the reported estimates should be interpreted cautiously.

Our study contributes to the growing evidence on whether and how edtech tools, particularly those powered by AI, improve K–12 student achievement. Research on Khan Academy illustrates why effectiveness must be considered alongside implementation. [Weatherholtz et al. \(2022\)](#) reported an effect size of 0.26 standard deviations (SD), but only for the 5% of students who used the platform for at least 30 minutes per week - coining the issue of low implementation fidelity as the “5 Percent Problem.” Using broader inclusion criteria, a subsequent evaluation estimated an effect of 0.03 SD, with average usage of approximately 11 minutes per week ([Eames et al., 2026](#)). An evaluation of Khan Academy’s AI tutor, Khan-migo, found limited engagement and achievement gains of 0.06–0.08 SD, similar to those observed without AI assistance ([Oreopoulos and Low, 2026](#)). In contrast, a randomized evaluation in India found an effect approaching 0.50 SD under high-fidelity implementation ([Oreopoulos et al., 2026b](#)). These findings motivate examining not only achievement outcomes, but also how broadly students engage and how outcomes vary across usage levels.

Beyond implementation intensity, the instructional capabilities and delivery models of these tools warrant distinction. Earlier computer-assisted learning (CAL) systems established the potential of structured practice and adaptive support: a two-year randomized evaluation of Cognitive Tutor Algebra I found no effects in the first year but an 8-percentile-point improvement among high school students in the second (Pane et al., 2014). Generative AI extends these capabilities through conversational interaction and dynamically generated feedback, but its educational applications differ substantially in the roles assigned to students and educators. For example, Tutor CoPilot supports human tutors with real-time pedagogical guidance, increasing students’ likelihood of mastering lesson topics by 4 percentage points (Wang et al., 2024). Eedi and NUMI instead engage students directly in AI-supported learning activities, with emerging experimental evidence of benefits for math learning (Team et al., 2025; Oreopoulos et al., 2026c). As emphasized in a recent report released by Stanford University, evidence should therefore be interpreted in relation to the specific instructional model rather than generalized across all AI tutoring tools (Turano et al., 2026; Fesler et al., 2026).

Our study advances this literature by evaluating a student-facing math platform whose access and instructional use are mediated by classroom teachers. Teachers can shape which students participate, what content they encounter, and how platform activities connect to classroom instruction. Experimental research supports the importance of this context: human support can increase engagement with AI tutoring, while teacher coaching can strengthen computer-assisted learning implementation (Robinson et al., 2026; Oreopoulos et al., 2026a). On Goblins, students gain access through teacher-created classes where teachers select and assign activities while students receive individualized AI feedback. We examine this model over a full school year in a district where all middle school math teachers were offered the tool and retained discretion over adoption and use. This setting extends evidence on AI tutoring beyond access alone to its use within teacher-directed instructional practice.

Accessibility provides another important dimension for evaluating edtech tools. Multi-

modal and multilingual interfaces may enable students to engage with instructional support in ways that text-based interaction alone does not accommodate. Goblins allows students to speak, draw, or type their responses and provides voice-based support in seven languages, features that may be particularly relevant for students with disabilities and English learners. Our exploratory subgroup analyses examine whether achievement associations differ across these populations, providing initial evidence on whom this model may serve. Although these analyses cannot isolate the effects of accessibility features, they broaden the evaluation beyond average achievement to consider learners with different instructional needs.

Finally, we address implementation concerns illustrated by the “5 Percent Problem” by documenting engagement across the district sample and examining achievement across multiple usage levels. Approximately 51% of students met our treatment threshold of two problems solved per week. Together, our analyses provide yearlong evidence on the reach and achievement associations of a teacher-mediated, multimodal platform, while identifying questions about teacher involvement and accessibility for future experimental research.

The remainder of the paper proceeds as follows. We first describe Goblins and its implementation in DVUSD. We then introduce the data, characterize the analytic sample and platform usage, and define the treatment and outcome measures. Next, we present the empirical strategy, followed by the main results, subgroup and dosage analyses, and robustness checks. The final section discusses the findings, limitations, and implications for future research.

Goblins and Its Implementation

[Goblins](#) by Aha Moments is an online math learning platform designed for students in grades 5 to 12. 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 describe further in the study design.

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.

Implementation at Deer Valley Unified School District

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. 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 are no shared individual-level identifiers across the two datasets, we link them using student names available in both sources. This approach is overall successful but not perfect. 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.

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-Goblins user students are 0.4 SD higher than Goblins users. At the teacher level, we observe that female, non-White, and less experienced teachers are more likely to use Goblins. 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.

Goblins Usage Statistics

We characterize student engagement on Goblins 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 Goblins. *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.

We do not describe measures related to assignment type and teacher engagement with the platform. Assignment-level measures do not provide a consistent or comprehensive measure of implementation since 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.

We cannot definitively determine whether Goblins content was assigned to a full district class or to selected students, and therefore cannot conclusively classify implementation as Tier 1, Tier 2, or Tier 3. As described earlier, 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 observe enrollment sizes of Goblins classes, 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.

We 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. 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. The highest-usage group across the three measures generally includes three to six percent of the analytic sample.

Figure 2 reports student-level Goblins usage 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. Eighth-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.

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, 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. In our context, we are more interested in estimating the association between *using* Goblins and student achievement.

For the reason above, we explored treatment thresholds based on a minimum level of usage rather than account activation alone. While active minutes and problems attempted share similar distributions, 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 and assumes 27 weeks of instruction during the school year.

The current approach of estimating the average treatment effect on the treated (ATT) departs from the preregistration plan which specified an intent-to-treat (ITT) estimation to capture the effects of being offered Goblins through teacher activation. We do not estimate the ITT for two reasons. First, Goblins usage statistics show that student usage varied considerably, even among students with the same teacher. Among the 4,833 students with an activated Goblins account who would have comprised the ITT treatment group, engagement was limited for a meaningful share of student users. Over the 27-week implementation period (i.e., the 2025-26 school year), 1% of Goblins student users solved 0 problems; 5% solved 3 or fewer problems; and 10% solved 12 or fewer problems. Additionally, 15% spent less than 60 minutes on the platform across the entire school year. This distinction is particularly relevant because the implementation plan did not prescribe a minimum level of Goblins usage or require teachers to meet a specified dosage threshold.

Second, we do not estimate the ITT because we do not have sufficient administrative data to detect the indirect or spillover effects that are typically relevant to ITT estimates of learning interventions. Although Goblins was available to all teachers in the district, students without an activated account may have indirectly benefited from Goblins, including through alternative forms of math support provided by teachers who use Goblins for other students within the same classroom. Observing the ITT within a district classroom would therefore require classroom rosters to observe such indirect exposure to Goblins, as is generally examined when estimating ITT effects. As described previously, we do not have district classroom rosters that would allow us to determine whether teachers intended for some students to indirectly benefit from Goblins while omitting them from access to the platform.

We aggregate activity across both teacher-assigned and independent practice modes and do not distinguish between these two channels of platform use. Although independent mode allows students to work autonomously, teachers can still monitor students’ activity, progress, and content mastery. Thus, independent mode use may remain teacher mediated, complicating the interpretation of results. Our approach to aggregate student use across both modes captures students’ overall engagement with Goblins irrespective of the context in which the activity occurred.

Based on our conversations with the provider, we operationalize “Goblins usage” at a minimum total of 54 problems solved during the implementation period (i.e., the 2025-26 school year), which translates to an average of 2 problems solved per week. Approximately 51% of all students in the analytic sample had met this threshold. [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 3](#) with the following six dosage categories: (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 score is our 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 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 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.

Study Design

We follow the study preregistration and employ a weighted least squares regression with propensity weights to estimate the association between Goblins usage and student math achievement. The analytic approach comprises 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.

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, this analysis follows guidelines from the [What Works Clearinghouse 5.0 Procedures and Standards Handbook](#). 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 summarize our analytic approach below and provide additional methodological details in the [Technical Appendix](#).

Estimation of Propensity Scores

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 MAP math scores that are standardized within district and grade level, current grade level, gender, disability status (i.e., having an Individualized Education Program), 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 Gobblins 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 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 $f(x \mid t = 0)$ denote 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, let's assume that 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 the weighted comparison group's effective sample size (ESS):

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

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 propensity score matching approach rests on the achievement of 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 mean and standardized mean difference (SMD) 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 the *What Works Clearinghouse 5.0 Procedures and Standards Handbook*.

Table 1 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 stratifi-

cation, which similarly discard or downweight observations that provide limited information for estimation (Ridgeway et al., 2022; McCaffrey et al., 2004). Table 1 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 2 compares the distributions of all covariates after weighting. 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_{its} 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 Goblines relative to low or zero use of it. Heteroskedasticity-robust standard errors are clustered at the teacher level.

In addition to measuring the overall associations between Goblines usage and math achievement, we examine whether and how dosage (i.e., Goblines usage intensity) makes a difference. This model extends the primary specification by replacing the single binary treatment indicator with a set of mutually exclusive dosage categories, $D_{d1}...D_{d6}$. The six dosage categories correspond to the cutoffs provided in [Table 3](#). 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, $\gamma_{d1}... \gamma_{d6}$ 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.

Results

Main Findings

We report the main outcome findings in [Table 4](#) and [Table 5](#). Each table reports the ATT of using Goblins. 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 (CGI) as a secondary outcome.

[Table 4](#) 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 coefficient size is more than twice as large, corresponding to an overall association of 0.24 standard deviations in MAP math achievement growth ($p < 0.001$).

The subgroup model findings are reported in [Table 5](#). We find no differential associations of Goblins usage and achievement 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.

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.

Fourth, following the recommendations of (Imbens and Xu, 2025), we conduct a placebo check using a placebo math achievement outcome that predates the introduction of Gobblins in DVUSD: students’ spring 2025 MAP scores. This placebo analysis is limited to students in the analytic sample who have a spring 2025 MAP test record at DVUSD. A drawback of this exercise is that our available data from the prior school year (2024–25) include MAP scores only for students in grades 6–8. As a result, we cannot observe spring 2025 math scores for sixth-grade students in the analytic sample, as these students were in fifth grade during the 2024–25 school year, for which we do not have MAP data. The resulting placebo-check sample consists of 3,468 students, representing approximately 57% of the analytic sample.

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 A5](#)), suggesting that the alternative models were successful in creating comparable treatment and control groups. We do not describe the covariate balance for the placebo check since model specification is identical to the main specification, where the only adjustment is made to the outcome.

Overall, the estimated association of Goblins usage and math achievement was robust across the first three alternative models (see [Table A6](#)). Relative to the main specification, which estimated an effect of 0.11 SD, the first three alternative models produced estimates within 0.01 standard deviations of the primary estimate, indicating a high degree of consistency across alternate specifications. The fourth alternative model, a placebo check, yielded a marginally significant difference of 0.049 SD ($p < 0.10$) between treated and comparison students. Given we do not observe six graders' prior-year test scores, this test only captures part of our analytic sample. However, this placebo estimate suggests that while our propensity score matching method is able to achieve balance on student characteristics and baseline scores, it might still leave important differences between the treatment and control groups unaddressed.

Dosage

[Table 6](#) 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 dosages above 3 problems per week were statistically significant ($p < 0.01$).

Conclusion

The central challenge in educational technology is no longer simply whether schools can provide students with access to individualized digital instruction, but whether such instruction can be delivered at sufficient scale and intensity to meaningfully improve learning. This study provides evidence that Goblins—a teacher-mediated, AI-enabled math tool—can reach a substantial share of students who engaged with Goblins at consistent levels. The estimated association 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. This exceeds the proportion of students who met the treatment threshold in evaluations of other edtech tools, notably Khan Academy ([Weatherholtz et al. \(2022\)](#)) and IXL ([An \(2023\)](#); [Marple et al. \(2019\)](#)), in which the reported effect sizes were examined for 5% and 12% of the sample, respectively.

These findings contribute to the emerging literature on AI-supported instruction in two important ways. First, they suggest that the educational potential of AI-powered tools depends not only on the capabilities of the technology itself, but also on whether students engage with it sufficiently and whether that engagement is embedded within instructional practice. The substantial variation in outcomes across usage levels underscores the importance of moving beyond simple measures of program adoption or access when evaluating edtech tools; sustained use may be an important condition for realizing its instructional potential. The fact that approximately half of the students in this study sample reached relatively modest engagement threshold suggests that the implementation fidelity can improve beyond what was presented in the “5 Percent Problem”, although differences in usage definitions and study designs make direct comparisons difficult.

Second, the results provide preliminary evidence that Goblins may have particular relevance for students who face greater barriers to conventional instructional supports. Students with IEPs, Hispanic students, and English language learners exhibited larger achievement

associated with Goblins use than their respective comparison groups. Although these subgroup differences should be interpreted as exploratory and cannot establish that Goblins' accessibility features caused the differential patterns, they are consistent with the possibility that multimodal and multilingual forms of instructional support can reduce barriers to engagement for some students. Identifying the mechanisms underlying these differences is an important direction for future research.

Despite our efforts to leverage all available data and employ a weighted propensity score design, this study has several inherent limitations. The first is the incomplete linkage between the Goblins and the district dataset. 13% of Goblins student users could not be matched to a student record in the district administrative data. This is likely attributable, at least in part, to inconsistencies arising from students' self-entry of their names on the Goblins platform. The unmatched records may reflect two scenarios. One, some Goblins 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). The second alternative is that 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 sample of comparison students.

The second limitation concerns the potential for spillover effects. Although Goblins was available to all teachers in the district, some teachers may have benefited from access to Goblins without providing their students direct access to the platform. For example, teachers could have used Goblins to generate or extract practice problems, assignments, or worksheets for use in their regular classroom instruction. As a result, students classified as untreated may have been indirectly exposed to instructional materials developed through Goblins. 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 implemen-

tation period. For example, the assessment mode feature, which allows the AI tutor to be turned off when teachers assign content, was introduced in the middle of the implementation period. It was not available when Goblins 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.

The final consideration is the lack of a conventional business-as-usual condition in which students received no comparable technology-based instructional support. Based on our conversations with DVUSD personnel, many teachers and schools who did not use Goblins had adopted another edtech tool with similar functionalities. Consequently, students in the comparison group may not represent a typical business-as-usual population, which could attenuate the estimated effect of Goblins. The reported findings may therefore underestimate Goblins' impact and should be interpreted as conservative estimates of its effects.

Taken together, this study provides initial evidence that AI-powered edtech tools can move beyond the provision of scalable access to individualized support and be associated with meaningful differences in student learning. The findings also suggest that the relevant question for schools and districts is not simply whether to integrate such tools, but to understand what conditions are needed to achieve meaningful student engagement and how to effectively integrate such tools into teaching and learning. Establishing these conditions and determining whether the achievement patterns observed here persist under experimental conditions will be critical to assessing the broader role of AI in expanding effective, individualized instruction in K–12 education.

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.
- Bhatt, M. P., Guryan, J., Khan, S. A., LaForest-Tucker, M., and Mishra, B. (2024). Can technology facilitate scale? evidence from a randomized evaluation of high dosage tutoring. Technical report, National Bureau of Economic Research.
- Diliberti, M. K. and Schwartz, H. L. (2024). Districts’ pandemic recovery efforts as COVID-relief aid expires: Selected findings from the spring 2024 american school district panel survey. Technical Report RR-A956-23, RAND Corporation.
- 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.
- Escueta, M., Nickow, A. J., Oreopoulos, P., and Quan, V. (2020). Upgrading education with technology: Insights from experimental research. *Journal of Economic Literature*, 58(4):897–996.
- Fesler, L., Martinez Claeys, J., Agnew, C., and Loeb, S. (2026). The evidence base on ai in k-12: A 2026 review.
- 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.
- Holt, L. (2024). The 5 percent problem. *Education Next*, 24(4):26–30.
- Imbens, G. W. and Xu, Y. (2025). Comparing experimental and nonexperimental methods: What lessons have we learned four decades after lalonde (1986)? *Journal of Economic Perspectives*, 39(4):173–201.
- 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.
- Kestin, G., Miller, K., Klales, A., Milbourne, T., and Ponti, G. (2025). AI tutoring outperforms in-class active learning: An RCT introducing a novel research-based design in an authentic educational setting. *Scientific Reports*, 15.
- 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.
- Kraft, M. A. and Falken, G. T. (2021). A blueprint for scaling tutoring and mentoring across public schools. *Aera Open*, 7:23328584211042858.
- Kulik, J. A. and Fletcher, J. D. (2016). Effectiveness of intelligent tutoring systems: A meta-analytic review. *Review of Educational Research*, 86(1):42–78.
- 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.
- Muralidharan, K., Singh, A., and Ganimian, A. J. (2019). Disrupting education? experimental evidence on technology-aided instruction in india. *American Economic Review*, 109(4):1426–1460.
- 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., Gibbs, C., Jensen, M., and Price, J. (2024). Teaching teachers to use computer assisted learning effectively: Experimental and quasi-experimental evidence. Working Paper 32388, National Bureau of Economic Research.
- Oreopoulos, P., Gibbs, C. R., Jensen, M., and Price, J. P. (2026a). Teaching teachers to use computer assisted learning effectively: Experimental and quasi-experimental evidence. *Journal of Human Resources*.
- Oreopoulos, P., Keyes-Kryszakowski, O., and Agarwal, D. (2026b). 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.
- Oreopoulos, P., Liut, M., Sungu, A., and Low, N. (2026c). Making ai tutoring productive: Evidence from a mastery-based math practice experiment. Technical report, National Bureau of Economic Research.
- Oreopoulos, P. and Low, N. (2026). One click away: Ai tutoring with khanmigo in a two-year school experiment. 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*.
- Robinson, C. D. and Loeb, S. (2021). High-impact tutoring: State of the research and priorities for future learning. *National Student Support Accelerator*, 21(284):1–53.
- 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*.
- 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*.

- Turano, C., Pihl, V., Agnew, C., Ziegler, L., and Loeb, S. (2026). AI tutoring is not a monolith: What we actually know. Research brief, National Student Support Accelerator, SCALE Initiative at Stanford University.
- VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist*, 46(4):197–221.
- 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*.
- Wang, R. E., Ribeiro, A. T., Robinson, C. D., Loeb, S., and Demszky, D. (2025). Tutor CoPilot: A human-AI approach for scaling real-time expertise. EdWorkingPaper 24-1054, Annenberg Institute at Brown University.
- 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

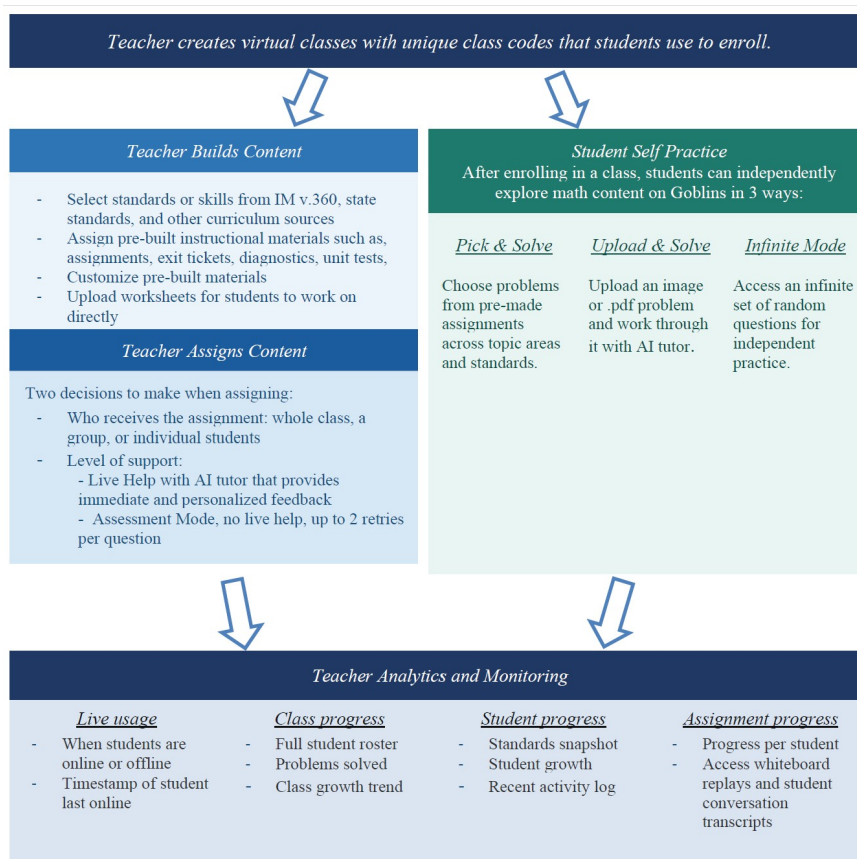
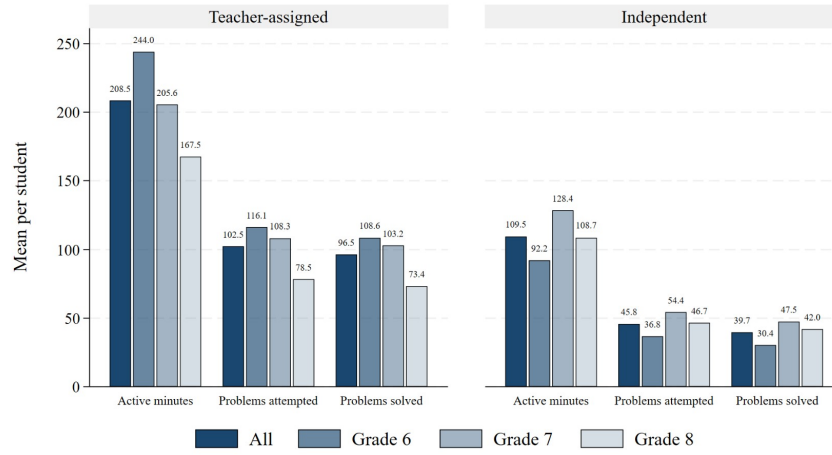


Figure 2: Average Goblins usage by mode and by grade



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: 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 2: 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 3: Goblins dosage levels

Average problems solved 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., who solved fewer than 2 problems per week).

Table 4: 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 5: 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 6: 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: $^+p < 0.10$, $^*p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

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](#)).

[Table 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 using Goblins and those not using Goblins ($p > 0.05$). To avoid losing observations through

²TVA calculations are made using data from the 2024-2025 school year.

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	Unmatched Goblins Users	Difference
	(1)	(2)	(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)–(2) report means. Column (3) reports the difference in means between reported values in Columns (1) and (2). 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: 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. We do not include covariate balance for the fourth sensitivity model as it uses the same model specification from the main model and only adjusts the outcome.

Table A6: Estimated treatment effects with sensitivity models

	Treatment is 3 problems solved/week (1)	Including economic disadvantage (2)	With minimum support rule (3)	Placebo outcome of spring 2025 math score (4)
Outcome: Within-District EOY				
Treatment	0.118*** (0.021)	0.118*** (0.025)	0.111*** (0.021)	0.049+ (0.025)
Outcome: Conditional Growth Index				
Treatment	0.254*** (0.046)	0.254*** (0.057)	0.240*** (0.047)	-0.071 (0.057)

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: $^+p < 0.10$, $^*p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$