

Dosage and Dialogue: Understanding Human Support in Hybrid Human-AI Math Tutoring

Ge Gao¹, Ashish Gurung², Danielle R. Thomas², Marie Cynthia Abijuru Kamikazi², Conrad Borchers³, Erin Gatz², Shivang Gupta², Stephen Fancsali⁴, Steve Ritter⁴, Lee Branstetter², Kenneth Koedinger², and Emma Brunskill¹

¹ Stanford University, Stanford, CA, USA
{gegao, ebrun}@stanford.edu

² Carnegie Mellon University, Pittsburgh, PA, USA
{agurung, mabijuru, shivangg, branstet}@andrew.cmu.edu
{drthomas, koedinger}@cmu.edu

³ Vanderbilt University, Nashville, TN, USA
c.borchers@vanderbilt.edu

⁴ Carnegie Learning, Inc., Pittsburgh, PA, USA
{sfancsali, sritter}@carnegielearning.com

Abstract. One-on-one human tutoring is among the most effective educational interventions, yet cost and staffing constraints limit scalability. Hybrid human-AI tutoring, in which students work in math learning software while receiving as-needed one-on-one (1:1) support from a remote human tutor, offers a promising alternative. Yet, its effectiveness, and the role of how much and how well tutoring is delivered, remain under-characterized. We examine whether variation in the quantity and quality of remote human tutoring is associated with students’ in-platform learning progress and external assessment outcomes. Using tutoring records, platform logs, prior achievement, and end-of-year assessments from two U.S. middle-school implementations (IXL: N=203; MATHia: N=157), we examine tutoring dosage, sustained sessions (≥ 1 minute), and, for the transcript-available site, LLM-derived dialogue quality ratings. Regression models show that more frequent sustained tutoring sessions are positively associated with in-platform math progress after controlling for prior achievement, although the strength and robustness of this association vary across sites and model specifications. In the transcript-available site, LLM-derived tutoring rating explains additional variation in in-platform progress beyond dosage. However, neither dosage nor dialogue quality shows robust direct associations with external test scores after accounting for prior achievement and class-level heterogeneity, underscoring the need for cautious interpretation of observational dosage estimates.

Keywords: One-on-one human tutoring · Hybrid human-AI tutoring · Software learning progress · K-12 mathematics education.

1 Introduction

Personalized mathematics software is now used by millions of learners internationally, reshaping how instruction is delivered and studied at scale [8,12,30]. Prior work has found that students’ software-based progress is often positively associated with external performance measures. For example, large-scale observational analyses of the IXL software platform report that schools using IXL outperform non-IXL schools on state assessments, and students who master more skills per week tend to show greater test score gains [2]. Similarly, within MATHia, students who complete more practice workspaces tend to perform better in subsequent Algebra I courses, particularly among those with lower pre-test scores [34,28]. These findings are primarily correlational, and may reflect differences in student ability, engagement, or instructional context. Evidence from randomized controlled trials provides a more nuanced picture. For instance, evaluations of MATHia (Cognitive Tutor) have demonstrated positive but heterogeneous effects on student achievement [27], while a recent randomized study of IXL found significant improvements on diagnostic assessments but not on state accountability exams [10]. Recent large-scale analyses of computer-assisted learning in real-world settings similarly highlight that the effectiveness of such systems depends critically on usage intensity and implementation context [12]. Taken together, these results suggest that while in-software progress is often predictive of learning outcomes, it may not consistently translate into broader academic performance. This variability highlights the need to understand what additional supports or contextual factors help convert software practice into robust learning gains, motivating a growing body of work in this area [17,5,6].

One increasingly common instructional context is the integration of one-on-one (1:1) human tutoring alongside AI-driven mathematics software, often referred to as *hybrid human-AI tutoring* [31]. Hybrid human-AI tutoring provides the opportunity for rich, systematic data streams (e.g., tutor-student transcripts, interaction logs, and fine-grained timing information) that were largely unavailable in earlier intelligent tutoring system studies [17]. One-on-one tutoring with a human tutor is widely recognized as one of the most effective interventions in education [33,20]. Prior studies have found that 1:1 or small-group tutoring can dramatically improve student achievement. A meta-analysis of dozens of tutoring programs found an overall average positive effect on learning outcomes, indicating substantial gains [26]. These effects were stronger when tutors were professional teachers or paraprofessionals (as opposed to volunteers or parents) and when tutoring was delivered during the school day [26]. These impressive results underscore the potential of 1:1 tutoring to accelerate learning.

Despite the known benefits of 1:1 human tutoring on student learning, it is very costly to implement and difficult to scale [14,19]. Hybrid human-AI tutoring poses a solution to alleviating some of the cost [7,31] by offloading some support to an AI math software tutor. Such an approach also offers potential scalability given the ubiquitous use of math software in schools [15]. Relatively little is known about how human tutoring and AI tutor software-based learning may relate in such settings. Prior work establishes that making progress

in math software correlates with better test outcomes [2], and that tutoring on its own produces gains [26,29]. Prior hybrid tutoring models may alternate between scheduled human-led sessions and separate software-based instruction (e.g., Saga Education [3]). In contrast, we study a drop-in support model in which students work continuously in AI-based software while human tutors provide brief, opportunistic 1:1 assistance, a lower-intensity setting that may yield different relationships between human tutoring dosage and outcomes. This more closely resembles classroom “over-the-shoulder” support, as explored in prior work on teacher attention and orchestration tools such as Lumilo [1], where instructors dynamically allocate attention across students. Understanding how tutoring functions in this lower-intensity, opportunistic setting is critical, as it differs substantially from high-dosage or scheduled tutoring models.

Moreover, joint human tutoring and AI tutoring have shown positive effects on students’ AI tutor software progress. For example, Thomas et al. [31] found that the addition of human tutoring significantly improved both student engagement (as measured by time on task) and learning progress (e.g., skills mastered) across multiple implementation sites and AI platforms, compared to students receiving AI-only instruction. Gurung et al. [13] found that the benefits of human–AI tutoring were especially effective for lower-performing students in outcomes compared to similarly performing peers who received only AI tutoring.

However, it remains unclear under what conditions combining 1:1 human tutoring with AI tutoring yields consistent improvements, and how to optimize such interventions. In particular, stakeholders are interested in understanding how variation in 1:1 human tutoring dosage is associated with subsequent software-based learning progress and performance on formal assessments in real-world implementations. Because tutoring was allocated in part based on student need rather than randomly assigned, we do not estimate causal effects. Instead, we use observational learning analytics models to characterize robust patterns of association across two implementation contexts, while controlling for prior achievement and class-level heterogeneity where possible. This design allows us to examine not only whether students who receive more tutoring make more progress, but also whether the quality of tutor–student dialogue, as measured from full-session transcripts using an LLM-based rubric, explains additional variation beyond dosage alone.

In this paper we address this gap by analyzing a hybrid human-AI tutoring intervention to examine how variation in tutoring time, session frequency, and dialogue quality are associated with learning. We operationalize each student’s human tutoring dosage as both the total number of 1:1 human tutoring sessions and the total minutes of 1:1 human tutoring. We also measure each student’s AI tutor math progress (skills proficient or workspaces mastered in the tutoring software) and their external assessment scores. Studying human tutoring dosage in real-world settings presents several challenges: tutoring is not randomly assigned, students with greater need often receive more support, and interactions vary widely in duration and purpose. We therefore adopt an observational approach and focus on characterizing associations under these constraints. Using

a series of regression models, we aim to answer the following research questions (RQs):

- RQ1.** How is tutoring dosage, measured by total tutoring time and session frequency, associated with students’ in-software math learning progress?
- RQ2.** Does LLM-derived dialogue quality explain additional variation in in-software progress beyond tutoring dosage and prior achievement?
- RQ3.** Are tutoring dosage and LLM-derived dialogue quality associated with external standardized test performance after accounting for prior achievement and class-level heterogeneity?

2 Related Work

Human Tutoring and Math Edtech Synergy There is extensive evidence that individualized tutoring is effective. A systematic review of K–12 tutoring trials found consistent positive impacts on learning [26]. High-quality programs (often using certified teachers or well-trained tutors) yield larger gains than volunteer programs or after-school efforts [14,26]. For example, evaluations of Saga Education’s in-school tutoring found tutored students (where 2 students work with 1 tutor) improved their reading percentile rank and scores [18]. However, nearly all prior work treats tutoring as a fixed “dose” and does not extensively explore how variation in the amount of tutor time matters.

To leverage technology for scale, recent initiatives have paired human tutors with math software. Saga’s high-dosage model alternated small-group tutoring with computer-assisted learning (CAL); in a large randomized trial, students in this model gained about 0.23 SD on a standardized math test, almost as much as in Saga’s traditional 2:1 tutoring, while reducing per-pupil cost by 30% [4]. Complementing this, recent learning analytics work has begun to examine how human support interacts with student engagement in digital environments. For example, Borchers et al. [5] analyze fine-grained interaction data and show that teacher visits are temporally associated with reduced disengagement in AI-supported math learning. Other research has compared blended to software-only tutoring: a quasi-experimental study of middle-school students found the hybrid human-AI group outperformed a group using only AI-tutoring [13]. In that study, the addition of modest human support (~weekly 1:1 Zoom sessions) led to significantly higher learning growth (approximately 0.36 grade levels) relative to software-only students. Thomas et al. [31] similarly report that adding human scaffolding to an AI tutor increased students’ software usage and proficiency. Together, these studies show that tutoring supplemented with technology can support learning, but prior work has not jointly examined tutoring dosage and dialogue quality within hybrid human–AI tutoring.

Tutor Attention and Spatial Analytics A separate line of work has examined how teachers allocate attention in classrooms with AI tutors [22,25,24,23,16,17]. For example, Karumbaiah et al. performed a spatiotemporal analysis of teacher

behavior in an AI-enabled math classroom [17]. In that study, researchers defined teacher visits as episodes when a teacher stays physically close to a student while the student uses an AI tutor. They found that teachers are highly engaged and these visits do provide immediate benefits (*e.g.*, students reduce off-task behaviors). However, teachers inevitably could not reach every struggling student, suggesting the limits of in-person attention. In a related study using a smart-glasses “awareness” tool, Karumbaiah et al. [16] measured three multimodal attention signals: gaze fixations, use of an analytics “deep dive” feature, and physical visits. They found that teachers’ use of the deep-dive analytics tool was most strongly associated with students’ learning gains (after accounting for student ability and tutor logs), whereas mere proximity visits had weaker or even mixed effects. These findings illuminate how teachers notice student needs in real time, but they do not quantify the overall amount of tutoring attention each student receives nor its impact on outcomes. This work characterizes attention at fine granularity, but does not link total human tutoring dosage with student outcomes.

LLM-Based Instructional Quality Assessment Large language models (LLMs) have demonstrated increasing capability for analyzing instructional interactions, from classroom discourse to tutoring dialogues. For example, Long et al. showed GPT-4 achieved 90%+ inter-coder agreement with human experts on middle school mathematics classroom transcripts [21], while Demszky and Liu demonstrated that NLP-based feedback on tutoring quality improved instructor uptake by 10–13% across 1,136 teachers [11]. However, applications to tutoring quality assessment remain under-investigated: Tran et al. achieved Quadratic Weighted Kappa of 0.63–0.77 on the Instructional Quality Assessment framework using few-shot prompting at the utterance level [32], while Whitehill and LoCasale-Crouch approached human inter-rater reliability by combining LLMs with bag-of-words features [35]. Most directly relevant, Thomas et al. examined hybrid MATHia-human tutoring with 585 middle school students but did not assess tutor quality variation [31]; their quasi-experimental study found significant positive effects on in-platform proficiency, with lower-achieving students benefiting most and time-on-task increasing from 24 to 33 minutes weekly. Existing LLM research focuses on validating automated coding against human raters or demonstrating that feedback improves practice, not on linking LLM-derived quality scores to learning outcomes.

3 Design Context & Methods

3.1 Data

We analyzed two datasets from U.S. middle-grade schools implementing hybrid human-AI tutoring in regular mathematics instruction. All procedures were IRB-approved; data were collected under standard school consent protocols, de-identified prior to analysis, and included opt-out provisions where applicable. Tutoring sessions were conducted in a shared-attention format, where each tutor

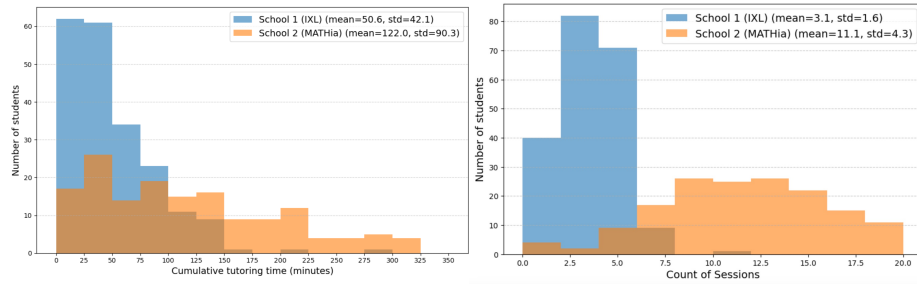


Fig. 1. Distributions of cumulative tutoring time and session counts per student across the two sites.

was assigned to multiple students (typically 4–6, varied by contexts). Tutors rotated across breakout rooms, providing brief, individualized assistance as needed. Students could also request help (e.g., via hand-raising), which influenced tutor prioritization. As a result, recorded tutoring “session time” reflects time in which support was available, rather than continuous one-on-one interaction.

At School 1, 203 students across grades 6–8 participated in weekly online math classes integrating IXL⁵ usage with remote 1:1 human tutoring delivered via Zoom breakout rooms. The program launched in late September 2024 and operated on a rotating schedule, such that students attended hybrid tutoring on different days each week. The dataset includes weekly IXL activity logs (e.g., dates, skills attempted, proficiency status), detailed 1:1 tutoring durations, and standardized assessments: a Spring 2024 state test used as a pre-diagnostic, and an end-of-year (EOY) state test in Spring 2025. In-software learning progress was assessed as the count of math skills proficient during using the software.

At School 2, 157 students participated in a parallel weekly hybrid tutoring program beginning in September 2024, combining MATHia⁶ usage with remote 1:1 human tutoring during 40-minute class sessions. MATHia logs captured fine-grained student interactions with workspaces, including timestamps and mastery indicators, while separate logs recorded the duration of each 1:1 tutoring session. Available assessments included a Spring 2024 state test as a pre-diagnostic, and an EOY state test in Spring 2025. In-software learning progress was assessed as the count of workspace mastered during using the software.

Quantifying Tutoring Scores of Full-Session Tutor-Student Dialog in School 2

We include an LLM-derived session-level tutoring quality score as a covariate to account for variation in instructional quality across human tutoring sessions. This score was derived from a separate, under-review study using a large language model (Gemini-2.5-pro [9]) to evaluate full-session tutoring transcripts. The assessment pipeline employs a two-stage prompting process: an *opportunity prompt* identifies pedagogically relevant moments (e.g., student errors or correct attempts, expressions of confusion, student questions, and productive struggle),

⁵ <https://www.ixl.com/>

⁶ <https://www.carnegielearning.com/solutions/math/mathia>

and an *evaluation prompt* assesses the tutor’s response to these moments against evidence-based best practices. The prompting rubrics were derived from human-coded guidelines and validated in a separate study, where LLM-derived transcript scores were associated with independent tutor-training performance and expected improvements in pedagogical behaviors over time. Each identified instance receives a binary quality score (0=ineffective, 1=effective). Session-level scores are computed by aggregating across up to six identified opportunities per session, yielding an overall score in the range [0, 6]. Transcripts were generated from Zoom recordings using Whisper and merged with synchronized chat logs. Tutoring transcripts were only available for School 2, and therefore LLM-derived rating measures are included only for that site.

3.2 Implementation Context and Potential Confounds

The data come from real-world hybrid human–AI tutoring implementations, providing natural variation in tutoring dosage but no random assignment. Students receiving more tutoring may differ systematically from others; we mitigate this by controlling for prior achievement and class fixed effects, but results remain associational. Need-based allocation is partly by design — students with lower pre-test scores were often prioritized for proactive tutoring — and naïve associations may therefore underestimate or even reverse true dosage effects. Tutoring dosage varied substantially across students: at School 1, cumulative tutoring time averaged 50.6 minutes (SD = 42.1) with mean 3.1 sessions, reflecting a 6-day rotating schedule with week-to-week variability; at School 2, averages were 122.0 minutes (SD = 90.3) and 11.1 sessions. These structural differences motivate separate site-level analyses throughout.

3.3 Analytical Framework

Let i index students, Y_i denote sum of students’ in-software math learning progress, measured via cumulative proficiency (School 1) or mastery (School 2) metrics. Let Z_i denote EOY standardized external assessment performance. We define student-level tutoring exposure variables aggregated across the study period: (i) (T_i): total 1:1 tutoring duration, (ii) (S_i): number of tutoring sessions, and (iii) (Q_i): average LLM-derived tutoring rating.

All external test outcomes are standardized (z -scores on grade level) to facilitate interpretation in terms of standard deviations. Math progress in software, dosage-related covariates are log transformed. Tutor quality score is z -normalized in multivariate/adjusted models. All analyses are conducted separately by site using one row per student. Tutoring dosage, in-software progress, prior achievement, tutoring quality, and external assessment outcomes are aggregated to student-level features before modeling. Extended models include class fixed effects to account for shared classroom, schedule, and implementation differences. We fit linear regressions with cluster-robust standard errors grouped by class ID. We report all three dosage operationalizations rather than selecting a single post

hoc “best” predictor, and emphasize patterns that are stable across model specifications. Model diagnostics did not indicate severe violations of linear-model assumptions; nevertheless, estimates remain associational because tutoring was not randomly assigned.

Dosage–Outcome Associations (RQ1) We use correlational linear regression models that relate summary-level 1:1 tutoring exposure to in-software learning progress. Let Y_{iw} denote in-software learning progress for student i in workspace w . We estimate:

$$Y_i = \beta_0 + \beta_1 D_i + \beta_2 P_i + \varepsilon_i, \quad (1)$$

where P_i is a pre-test (baseline achievement) measure. Let D_i denote a summary-level measure of tutoring dosage for student i . The key dosage regressor D_i is substituted across specifications:

1. **Tutoring duration:** $D_i = T_i$ (total minutes of 1:1 tutoring),
2. **Session count (all sessions):** $D_i = S_i$ (number of 1:1 tutoring sessions),
3. **Session count (at least 1 minute):** $D_i = S_i^{\geq 1\text{min}}$, where $S_i^{\geq 1\text{min}}$ counts only sessions with duration ≥ 1 minute. This specification is included to exclude extremely short or potentially not meaningful interactions (e.g., brief check-ins or logging artifacts) and better capture sessions with substantive instructional engagement⁷.

We report models *adjusted for prior achievement* by adding P_i .

Incremental Role of LLM-Derived Tutoring Rating (RQ2) To evaluate what extent LLM-derived tutoring rating explains variation in learning progress beyond tutoring dosage alone, we extend the baseline model:

$$Y_i = \beta_0 + \beta_1 \{T_i, S_i, S_i^{\geq 1\text{min}}\}_{\text{best}} + \beta_2 \bar{Q}_i + \beta_3 P_i + \varepsilon_i \quad (2)$$

where we report the most powerful covariate among $T_i, S_i, S_i^{\geq 1\text{min}}$ in the model. Improvements in model fit quantify the additional explanatory contribution of tutoring quality.

Relationships to External Academic Outcomes (RQ3) Finally, we relate cumulative tutoring experiences across workspaces to external standardized test outcomes:

$$Z_i = \beta_0 + \beta_1 \{T_i, S_i, S_i^{\geq 1\text{min}}\}_{\text{best}} + \beta_2 \bar{Q}_i + \beta_3 P_i + \varepsilon_i \quad (3)$$

⁷ Sensitivity analyses using 30-second, 2-minute, and 5-minute thresholds confirm that the main findings are robust to this choice for both in-software progress (School 1: $\beta = 0.28\text{--}0.45$, all $p \leq 0.044$; School 2: $\beta = 0.19\text{--}0.53$, all $p > 0.10$) and external exam scores (all $p > 0.40$ in School 1; all $p > 0.43$ in School 2 with the exception of an isolated significant negative estimate at the 2-minute threshold, which we attribute to subsample composition)

4 Results

4.1 Dosage–Outcome Associations (RQ1)

For regressions adjusting for pre-test scores, in School 1, associations between human tutoring dosage and progress remain robust after adjustment, including total tutoring duration ($\beta = 0.14$, $p < 0.001$) and session count measures (all sessions: $\beta = 0.37$, $p < 0.001$; sessions ≥ 1 minute: $\beta = 0.37$, $p < 0.001$). Pre-test scores are also positively associated with progress but do not attenuate dosage effects. In School 2, pre-test scores explain a substantial share of variance in progress ($R^2 \approx 0.43$ – 0.46). After adjustment, total tutoring duration and overall session count are not significant, while session frequency restricted to sessions lasting at least one minute remains positively associated with progress ($\beta = 0.68$, $p = 0.004$).

Summary across contexts Overall, associations between human tutoring dosage and in-software progress vary by instructional context. In School 1, both tutoring time and frequency are consistently associated with greater progress, whereas in School 2, where prior achievement dominates, only sustained tutoring interactions (sessions ≥ 1 minute) show a reliable association. These results suggest that effective operationalizations of human tutoring dosage, and their relationship to learning, may differ across platforms and student populations.

Model Fit and Robustness (RQ1). The baseline models account for human tutoring dosage and prior achievement only, yielding $R^2 = 0.09$ (School 1) and $R^2 = 0.46$ (School 2). Extending these models with class fixed effects, grade level, tutoring quality rating (School 2 only), and weeks of platform usage (Table 1) substantially improves fit in both settings ($R^2 \approx 0.39$ – 0.40 for School 1; $R^2 \approx 0.70$ – 0.73 for School 2). In School 1, the dosage association is positive and significant across all three operationalizations: session count (≥ 1 min; $\beta = 0.42$, $p = 0.006$), unrestricted session count ($\beta = 0.41$, $p = 0.025$), and log instruction duration ($\beta = 0.10$, $p = 0.033$). Grade proficiency emerges as a strong positive predictor across all models ($\beta \approx 0.09$ – 0.11 , $p < 0.001$), while pre-test score is consistently non-significant ($p > 0.60$), suggesting that classroom-level curriculum difficulty accounts for much of the variance previously attributed to prior achievement. In School 2, none of the three dosage operationalizations reach significance in the extended models (session count ≥ 1 min: $\beta = 0.53$, $p = 0.259$; unrestricted session count: $\beta = 0.53$, $p = 0.259$; log duration: $\beta = 0.01$, $p = 0.955$). Pre-test score remains a strong positive predictor across all specifications ($\beta \approx 0.21$ – 0.22 , $p < 0.001$).

4.2 Incremental Role of LLM-Derived Tutoring Rating (RQ2)

We examine to what extent LLM-derived tutoring rating explains additional variation in students’ in-software math progress beyond human tutoring dosage and prior achievement. This analysis is conducted in School 2 (MATHia). The

Table 1. Regression of **in-software math progress** on human tutoring dosage and covariates.

Predictor	School 1 (IXL)			School 2 (MATHia)		
	<i>S1</i>	<i>S2</i>	<i>Dur.</i>	<i>S1</i>	<i>S2</i>	<i>Dur.</i>
<i>Human tutoring dosage</i>						
Sess. count (≥ 1 min)	0.42** (0.154)	—	—	0.53 (0.470)	—	—
Sess. count (all)	—	0.41* (0.182)	—	—	0.53 (0.470)	—
Duration	—	—	0.10* (0.045)	—	—	0.01 (0.141)
<i>Prior achievement</i>						
Pre-test score	0.03 (0.064)	0.02 (0.066)	0.03 (0.070)	0.22*** (0.042)	0.22*** (0.042)	0.21*** (0.036)
<i>Other controls</i>						
Tutoring quality rating	—	—	—	0.16*** (0.043)	0.16*** (0.043)	0.15*** (0.035)
Grade level	0.11*** (0.011)	0.11*** (0.011)	0.09*** (0.013)	0.32** (0.093)	0.32** (0.093)	0.41*** (0.057)
Year-week count	−0.03 (0.051)	−0.03 (0.058)	0.03 (0.037)	−0.03 (0.034)	−0.03 (0.034)	0.02 (0.023)
Class FEs	Yes	Yes	Yes	Yes	Yes	Yes
R^2	.397	.393	.384	.710	.710	.704
N	199	199	199	87	87	87

S1 = session count (≥ 1 min); *S2* = session count (all); *Dur.* = log instruction duration. Cluster-robust SEs (class level) in parentheses. [†] $p < 0.10$, $*p < 0.05$, $**p < 0.01$, $***p < 0.001$. ^aBase models (dosage only, no class FEs or extended controls) yielded $R^2 = 0.09$ (School 1) and 0.46 (School 2).

inclusion of LLM-derived tutoring quality rating as a covariate in School 2 (Table 1) reveals a consistent and statistically significant positive association with in-software math progress across all three dosage operationalizations ($\beta \approx 0.15$ – 0.16 , $p < 0.001$), independent of session frequency, duration, prior achievement, grade level, and class fixed effects. This pattern suggests that the quality of tutoring interactions carries incremental explanatory power beyond dosage alone.

4.3 Relationships to External Academic Outcomes (RQ3)

We examine whether tutoring-related measures are associated with external assessment performance after controlling for prior achievement. In base models, prior achievement is the dominant predictor in both schools. Session count is positively associated with external scores in School 1 ($\beta = 0.25$, $p = 0.019$) but negatively associated in School 2 ($\beta = -0.41$, $p = 0.036$), likely reflecting need-based allocation. LLM-derived tutoring quality shows a small positive base-model association in School 2 ($\beta = 0.10$, $p = 0.044$), but this association is not robust to the extended specifications below.

Model Fit and Robustness (RQ3). Extended models with class fixed effects and grade-level controls (Table 2) raise explained variance from $R^2 = 0.53$ to ≈ 0.65 in School 1 and from 0.66 to ≈ 0.79 in School 2. Across all six specifications, no tutoring dosage measure reaches significance in either school, and tutoring quality rating remains non-significant in School 2 ($\beta \approx 0.04$, $p > 0.37$) once class-level heterogeneity is absorbed. Prior achievement emerges as the sole robust predictor (School 1: $\beta \approx 0.58$, $p < 0.001$; School 2: $\beta \approx 0.41$, $p < 0.01$), suggesting that neither the quantity nor the LLM-based rating of human tutoring interactions independently predicts external standardized test performance beyond what students already knew at the start of the year.

4.4 Descriptive Pathways from Tutoring to External Outcomes

To explore whether in-software progress provides a plausible pathway linking tutoring dosage to external performance, we estimated chained associational models using session count (≥ 1 minute). Because tutoring was not randomly assigned, these analyses are descriptive rather than causal mediation tests. In School 1, session count predicted in-software progress ($\beta_1 = 0.029$, $p = 0.001$), while progress was positively but not significantly associated with external scores ($\beta_2 = 0.417$, $p = 0.105$); the implied indirect association was not significant (Sobel $z = 1.23$, $p = 0.219$). In School 2, session count predicted in-software progress ($\beta_1 = 0.026$, $p < 0.001$), and progress predicted external performance ($\beta_2 = 0.928$, $p < 0.001$), yielding a significant indirect association (Sobel $z = 2.79$, $p = 0.005$). However, the direct association between session count and external scores was negative ($\beta_3 = -0.042$, $p = 0.019$), consistent with need-based confounding.

5 Discussion

This study examined how 1:1 human tutoring dosage and instructional quality relate to in-software learning and external assessments across two hybrid human-AI tutoring implementations, contributing to a growing body of work on how human support complements AI-driven math instruction [5,13,31]. Rather than comparing human and AI tutoring, we jointly analyze dosage, sustained interaction thresholds, and LLM-derived dialogue quality across two platforms.

Our first key finding is that repeated substantive tutoring sessions (≥ 1 minute) are positively associated with in-software math progress after controlling for prior achievement, consistent with the broader tutoring literature showing meaningful positive effects of 1:1 human support on learning [26,33]. This association held across multiple dosage operationalizations in School 1, while in School 2 effects attenuated once class fixed effects were absorbed, possibly reflecting the higher baseline tutoring intensity at that site and diminishing returns at higher dosage levels [19]. This context-dependence echoes findings from [17], who showed that teacher visits in AI-enabled classrooms provide immediate benefits whose downstream impact is moderated by implementation context.

Table 2. Regression of **external exam scores** on human tutoring dosage and covariates.

Predictor	School 1 (IXL)			School 2 (MATHia)		
	<i>S1</i>	<i>S2</i>	<i>Dur.</i>	<i>S1</i>	<i>S2</i>	<i>Dur.</i>
<i>Human tutoring dosage</i>						
Sess. count (≥ 1 min)	0.10 (0.146)	—	—	−0.14 (0.311)	—	—
Sess. count (all)	—	0.12 (0.150)	—	—	−0.14 (0.311)	—
Duration	—	—	0.02 (0.041)	—	—	−0.08 (0.108)
<i>Prior achievement</i>						
Pre-test score	0.58*** (0.089)	0.57*** (0.088)	0.58*** (0.090)	0.41** (0.144)	0.41** (0.144)	0.41** (0.150)
<i>Other controls</i>						
Tutoring quality rating	—	—	—	0.04 (0.047)	0.04 (0.047)	0.04 (0.045)
Grade level	−0.16*** (0.016)	−0.16*** (0.016)	−0.16*** (0.020)	0.06 (0.063)	0.06 (0.063)	0.07 [†] (0.037)
Year-week count	0.04 (0.059)	0.03 (0.056)	0.05 (0.046)	−0.02 (0.025)	−0.02 (0.025)	−0.02 (0.023)
Class FEs	Yes	Yes	Yes	Yes	Yes	Yes
R^2	.647	.647	.646	.790	.790	.791
N	199	199	199	87	87	87

S1 = session count (≥ 1 min); *S2* = session count (all); *Dur.* = log instruction duration. Cluster-robust SEs (class level) in parentheses. [†] $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. ^aSchool 2 models restricted to students with available transcript data for tutoring quality scoring.

Second, in School 2, LLM-derived tutoring quality explains additional variation in in-software progress beyond dosage and prior achievement. This extends work validating automated instructional quality assessment [21,11,32,35] by linking such scores to student learning outcomes, and aligns with evidence that tutoring effectiveness depends on instructional practice as well as time allocated [26,33].

Moreover, for external standardized assessments, prior achievement dominated across all specifications, and neither dosage nor tutoring quality showed independent direct effects once class-level heterogeneity was absorbed. This aligns with prior evaluations of MATHia and IXL reporting stronger effects on curriculum-aligned measures than on state accountability exams [27,10], and with large-scale analyses of computer-assisted learning highlighting the sensitivity of effects to outcome measurement [12]. Descriptive pathway analyses further suggest a plausible pattern in which tutoring dosage is more directly associated with software mastery gains, which are in turn associated with external performance, consistent with the chained learning hypothesis advanced by [28,34]. However, because tutoring was not randomly assigned, these pathways should not be interpreted as causal mediation. The competitive descriptive pathway pattern in School 2 — where a significant negative direct association coexists with a positive indirect

association — likely reflects need-based confounding, a well-documented challenge in observational tutoring research whereby students receiving more support are those struggling most [17].

Limitations and Future Work. This study is observational, so results should be interpreted as associations rather than causal effects. Unmeasured factors such as student motivation, attendance, and need-based tutor allocation may bias dosage estimates. LLM-derived quality scores were available only for School 2 and reflect session-level aggregates rather than moment-to-moment instructional decisions. We also lack participant feedback or qualitative implementation data from students and tutors, limiting interpretation of why particular interactions were more or less productive. Future work should use stronger causal designs, richer multimodal interaction data, and longer-term follow-up.

6 Conclusions

This study provides an observational learning analytics account of hybrid human–AI tutoring across two middle-school mathematics implementations. Repeated substantive tutoring sessions were generally associated with greater in-software progress, though this pattern varied by site and model specification. In the transcript-available site, LLM-derived dialogue quality explained additional variation beyond dosage and prior achievement. However, neither dosage nor dialogue quality showed robust direct associations with external standardized test scores once prior achievement and class-level heterogeneity were included. Future work should use stronger causal designs and richer qualitative or multimodal evidence to examine when platform-proximal progress translates into broader assessment gains.

Acknowledgments

This work was made possible with the support of the Learning Engineering Virtual Institute. The opinions, findings, and conclusions are those of the authors.

References

1. Aleven, V., Blankestijn, J., Lawrence, L., Nagashima, T., Taatgen, N.: A dashboard to support teachers during students’ self-paced ai-supported problem-solving practice. In: European Conference on Technology Enhanced Learning. pp. 16–30. Springer (2022)
2. An, X.: Associations between ixl usage and ixl impact in math and ela: Analysis of data from seven states (2021)
3. Ander, R., Guryan, J., Ludwig, J.: Improving academic outcomes for disadvantaged students: Scaling up individualized tutorials. The Hamilton Project–Brookings (2016)

4. Bhatt, M.P., Guryan, J., Khan, S.A., LaForest-Tucker, M., Mishra, B.: Can technology facilitate scale? evidence from a randomized evaluation of high dosage tutoring. Tech. rep., National Bureau of Economic Research (2024)
5. Borchers, C., Wang, Y., Karumbaiah, S., Ashiq, M., Shaffer, D.W., Aleven, V.: Revealing networks: Understanding effective teacher practices in ai-supported classrooms using transmodal ordered network analysis. In: Proceedings of the 14th Learning Analytics and Knowledge Conference. pp. 371–381 (2024)
6. Cassar, S., Neal, M., Schofield, C.: Edtech, social relationships and learning: A study of relationships in the tech-mediated blended classroom. In: INTED2022 Proceedings. pp. 4174–4182. IATED (2022)
7. Chine, D.R., Brentley, C., Thomas-Browne, C., Richey, J.E., Gul, A., Carvalho, P.F., Branstetter, L., Koedinger, K.R.: Educational equity through combined human-ai personalization: A propensity matching evaluation. In: International conference on artificial intelligence in education. pp. 366–377. Springer (2022)
8. Clark, A.E., Nong, H., Zhu, H., Zhu, R.: Compensating for academic loss: Online learning and student performance during the covid-19 pandemic. *China Economic Review* **68**, 101629 (2021)
9. Comanici, G., Bieber, E., Schaeckermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al.: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint arXiv:2507.06261 (2025)
10. Copeland, S., Cook, M.A., Grant, A.A., Ross, S.M.: Randomized-control efficacy study of ixl math in holland public schools. Center for Research and Reform in Education (2023)
11. Demszky, D., Liu, J.: M-powering teachers: Natural language processing powered feedback improves 1: 1 instruction and student outcomes. In: Proceedings of the Tenth ACM Conference on Learning@ Scale. pp. 59–69 (2023)
12. Eames, T., Brunskill, E., Yamkovenko, B., Weatherholtz, K., Oreopoulos, P.: Computer-assisted learning in the real world: How khan academy influences student math learning. *Proceedings of the National Academy of Sciences* **123**(1), e2507708123 (2026)
13. Gurung, A., Lin, J., Gutterman, J., Thomas, D.R., Houk, A., Gupta, S., Brunskill, E., Branstetter, L., Aleven, V., Koedinger, K.: Human tutoring improves the impact of ai tutor use on learning outcomes. In: International Conference on Artificial Intelligence in Education. pp. 393–407. Springer (2025)
14. 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.: Not too late: Improving academic outcomes among adolescents. *American Economic Review* **113**(3), 738–765 (2023)
15. Hillmayr, D., Ziernwald, L., Reinhold, F., Hofer, S.I., Reiss, K.M.: The potential of digital tools to enhance mathematics and science learning in secondary schools: A context-specific meta-analysis. *Computers & Education* **153**, 103897 (2020)
16. Karumbaiah, S., Borchers, C., Falhs, A.C., Holstein, K., Rummel, N., Aleven, V.: Teacher noticing and student learning in human-ai partnered classrooms: a multimodal analysis. In: Proceedings of the 17th International Conference of the Learning Sciences-ICLS 2023, pp. 1042–1045. International Society of the Learning Sciences (2023)
17. Karumbaiah, S., Borchers, C., Shou, T., Falhs, A.C., Liu, P., Nagashima, T., Rummel, N., Aleven, V.: A spatiotemporal analysis of teacher practices in supporting student learning and engagement in an ai-enabled classroom. In: International conference on artificial intelligence in education. pp. 450–462. Springer (2023)

18. Khan, S., Bhatt, M., Guryan, J., Kim, J.: Can high-dosage literacy tutoring put first-graders on the road to success?. Society for Research on Educational Effectiveness (2024)
19. Kraft, M.A., Falken, G.T.: A blueprint for scaling tutoring and mentoring across public schools. *Aera Open* **7**, 23328584211042858 (2021)
20. Kulik, J.A., Fletcher, J.D.: Effectiveness of intelligent tutoring systems: a meta-analytic review. *Review of educational research* **86**(1), 42–78 (2016)
21. Long, Y., Luo, H., Zhang, Y.: Evaluating large language models in analysing classroom dialogue. *npj Science of Learning* **9**(1), 60 (2024)
22. Martinez-Maldonado, R., Echeverria, V., Mangaroska, K., Shibani, A., Fernandez-Nieto, G., Schulte, J., Buckingham Shum, S.: Moodoo the tracker: Spatial classroom analytics for characterising teachers’ pedagogical approaches. *International Journal of Artificial Intelligence in Education* **32**(4), 1025–1051 (2022)
23. Martinez-Maldonado, R., Kay, J., Yacef, K., Edbauer, M.T., Dimitriadis, Y.: Mt-classroom and mtdashboard: supporting analysis of teacher attention in an orchestrated multi-tabletop classroom (2013)
24. Martinez-Maldonado, R., Schulte, J., Echeverria, V., Gopalan, Y., Shum, S.B.: Where is the teacher? digital analytics for classroom proxemics. *Journal of Computer Assisted Learning* **36**(5), 741–762 (2020)
25. Martínez-Maldonado, R., Yan, L., Deppeler, J., Phillips, M., Gašević, D.: Classroom analytics: Telling stories about learning spaces using sensor data. In: *Hybrid learning spaces*, pp. 185–203. Springer (2022)
26. Nickow, A., Oreopoulos, P., Quan, V.: The impressive effects of tutoring on pre-k-12 learning: A systematic review and meta-analysis of the experimental evidence (2020)
27. Pane, J.F., Griffin, B.A., McCaffrey, D.F., Karam, R.: Effectiveness of cognitive tutor algebra i at scale. *Educational Evaluation and Policy Analysis* **36**(2), 127–144 (2014)
28. Ritter, S., Joshi, A., Fancsali, S., Nixon, T.: Predicting standardized test scores from cognitive tutor interactions. In: *Educational Data Mining 2013* (2013)
29. Robinson, C.D., Kraft, M.A., Loeb, S., Schueler, B.E.: Accelerating student learning with high-dosage tutoring. *edresearch for recovery design principles series*. EdResearch for recovery project (2021)
30. Steenbergen-Hu, S., Cooper, H.: A meta-analysis of the effectiveness of intelligent tutoring systems on k–12 students’ mathematical learning. *Journal of educational psychology* **105**(4), 970 (2013)
31. Thomas, D.R., Lin, J., Gatz, E., Gurung, A., Gupta, S., Norberg, K., Fancsali, S.E., Alevan, V., Branstetter, L., Brunskill, E., et al.: Improving student learning with hybrid human-ai tutoring: A three-study quasi-experimental investigation. In: *Proceedings of the 14th Learning Analytics and Knowledge Conference*. pp. 404–415 (2024)
32. Tran, N., Pierce, B., Litman, D., Correnti, R., Matsumura, L.C.: Analyzing large language models for classroom discussion assessment. In: *Proceedings of the 17th International Conference on Educational Data Mining*. pp. 500–510 (2024)
33. VanLehn, K.: The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational psychologist* **46**(4), 197–221 (2011)
34. Weeks, J., Baron, P.: Exploring math education relations by analyzing large data sets ii. research memorandum. ets rm-21-02. Educational Testing Service (2021)
35. Whitehill, J., LoCasale-Crouch, J.: Automated evaluation of classroom instructional support with llms and bows: Connecting global predictions to specific feedback. *Journal of Educational Data Mining* (2024)