Human Tutoring Improves the Impact of AI Tutor Use on Learning Outcomes

Ashish Gurung[1(✉)], Jionghao Lin[1,2], Jordan Gutterman[1], Danielle R Thomas[1], Alex Houk[1], Shivang Gupta[1], Emma Brunskill[3], Lee Branstetter[1], Vincent Aleven[1], and Kenneth Koedinger[1]

[1] Carnegie Mellon University, Pittsburgh, PA, USA
{agurung,jgutterm,dchine,ahouk,shivangg,branstet,va0e,kk1u}@andrew.cmu.edu
[2] The University of Hong Kong, Pok Fu Lam, Hong Kong
jionghao@hku.hk
[3] Stanford University, Stanford, CA, USA
ebrun@cs.stanford.edu

Abstract. High-impact (human) tutoring and computer-based AI tutors are widely recognized for their effectiveness in supporting learning. However, human tutoring is costly and difficult to scale, whereas AI tutors vary widely in their ability to adapt to students' academic and motivational needs. Our study presents a formative evaluation of a year-long implementation of virtual human-AI tutoring during the classroom use of AI tutors. Using year-long log data and standardized state tests, we examine the real-world impact of human-AI tutoring through measures of learning both within the AI tutor and on external standardized assessments. Through propensity score matching, we compare 356 seventh-grade students who received human-AI tutoring with 317 from the previous school year who received AI-only tutoring. The human-AI group demonstrated significantly higher growth and was 0.36 grade levels ahead by year's end. Although there was no overall difference in state test scores, we found a significant interaction between human-AI tutoring and time-on-task (i.e., AI tutor use). For each standard deviation (3.26 h) increase in AI tutor use, the human-AI group improved by 0.28 standard deviations on state tests, compared to 0.06 for the AI-only group. This finding suggests that human tutors enhance the benefits of AI tutors, with gains increasing with time-on-task. Our findings replicate prior studies on human-AI tutoring over a longer time scale, spanning an entire school year. An important insight of this work is that students' AI tutor usage data, particularly time-on-task, can serve as a valuable indicator of learning progress as well as a measure for identifying students in need of additional support to fully benefit from AI tutors.

Keywords: Human-AI Tutoring · EdTech · AIED · ITS · Program Implementation

A. I. Cristea et al. (Eds.): AIED 2025, LNAI 15880, pp. 393–407, 2025.
https://doi.org/10.1007/978-3-031-98459-4_28

1 Introduction

High-impact tutoring is well-documented for its effectiveness in improving student learning outcomes [14,28,29]. The National Student Support Accelerator (NSSA) defines high-impact tutoring as a structured form of instruction, delivered one-on-one or in small groups, that supplements students' classroom experiences and targets specific learning goals[1]. However, human tutoring presents significant challenges in terms of scalability and cost. As Bloom [7] observed, one-on-one tutoring is "too costly for most societies to bear," prompting researchers to explore scalable instructional methods that can approximate its benefits.

This challenge has driven ongoing efforts to develop scalable tutoring alternatives. One prominent approach has been the development of AI tutors (e.g., Intelligent Tutoring Systems (ITS), Computer Assisted Learning Platforms (CALP)), which aim to replicate the benefits of human tutoring in a more scalable and cost-effective manner. Notable examples include Cognitive Tutor [4], AutoTutor [18], and ASSISTments [20]. Like human tutoring, ITS has been shown to enhance student learning, with multiple meta-analyses confirming their effectiveness [8,25,35]. VanLehn's meta-analysis further suggests that ITS, in certain contexts, can be as effective as human tutoring (see [35]).

While Bloom's 2-sigma effect [7] over "traditional instruction" remains the aspirational benchmark, recent efforts have built on VanLehn's [35] findings by exploring hybrid tutoring approaches that combine the automation and scalability of AI tutors with the cognitive and motivational support of human tutors [1]. In human-AI tutoring, human tutors provide personalized instructional and motivational support, while ITS ensures consistent, adaptive feedback and hints. Studies comparing virtual human-AI tutoring with AI-only tutoring—already a strong baseline given the efficacy of ITS—have shown that incorporating human tutors further enhances student engagement and learning outcomes [12,34]. Notably, initial research has found that human-AI tutoring can double gains in math test scores [12] and significantly improve student engagement and performance within AI tutors [34]. The benefit of human-AI tutoring reported in [34] is particularly striking, as it was achieved through virtual tutoring by undergraduate tutors, demonstrating a potential solution for scaling human-AI tutoring.

In our study, we build on prior evaluations of human-AI tutoring conducted over a few months [12,34] by examining its feasibility across an entire school year. We conduct a formative evaluation of an at-scale implementation to examine the real-world impact of virtual human-AI tutoring. Specifically, we assess how virtual tutoring by undergraduates influences measures of learning within the AI tutors. Additionally, we also evaluate whether these effects extend beyond the AI tutors to improvements in state test performance. To achieve this, we employ a quasi-experimental approach with propensity score matching, comparing students who received human-AI tutoring with those from the previous

[1] NSSA Report: https://studentsupportaccelerator.org/sites/default/files/Higg_Impact_Tutoring_Definition.pdf.

academic year at the same school who received AI-only tutoring. This study aims to address the following research questions:

RQ 1. What are the impacts of human-AI tutoring on within-tutor measures of learning, compared to AI-only tutoring?

RQ 2. Do the effects of human-AI tutoring transfer to learning outcomes, as measured by standardized tests?

2 Background

2.1 Human Tutoring

Adaptive and personalized instructions from human tutors have consistently led to better learning outcomes than conventional classroom instruction [14,19,19]. The effectiveness of human tutors is often attributed to their ability to provide cognitive, motivational, and adaptive instructional support [2,35].

VanLehn's meta-analysis (see [35]), comparing ITS and human tutors, suggests that the success of human tutors stems from monitoring students' learning processes, identifying misunderstandings, and providing immediate feedback and scaffolding. The immediate feedback helps tutors address uncertainty in the student response, correct errors, and provide clarification [16,17] whereas scaffolding allows tutors to offer structured support that strengthens reasoning and grasp complex concepts [11,13].

Despite the well-established efficacy of high-impact human tutoring, the widespread adoption is hindered by scalability and cost constraints. Kraft and Falken [24] estimate the annual cost of high-impact tutoring in the United States at approximately $1,002 per K-12 student. Scaling this approach would require nearly 15 million tutors to support ∼49 million K-12 students, highlighting the feasibility challenges. Field experiments suggest even higher costs, ranging from $3,500 to $4,300 per student annually, depending on factors such as tutor expertise, group size, and session length [15,19].

2.2 Intelligent Tutoring Systems

From early on, researchers have tried to automate the benefits of human tutoring using AI-powered tutors. ITSs, in particular, use the estimation of students' mastery to tailor personalize adaptive instructional support and learning experience [8,25]. VanLehn's [35] meta-analysis suggests that ITSs are capable of performing at the same level as human tutors, offering a promising solution to the scalability challenges associated with human tutoring.

ITSs have been deployed across diverse educational domains, including mathematics [2,4,20], physics [23,33], writing [18,26], and programming [27]. Over time, advancements in ITSs have increasingly focused on automating adaptive feedback and scaffolding to narrow the gap between ITSs and human tutors. These advancements have been leveraged by commercial platforms, such as MATHia [4] and ALEKS [9], to deliver tutoring at scale.

Similarly, large-scale adaptive tutoring platforms, such as IXL[2] and i-Ready[3], have integrated select ITS features to provide instructional support to a larger student population. For instance, in 2022–2023, i-Ready (the system used in our study) served approximately 11 million K-8 students in the U.S. at an annual cost of $30 per student per subject [5]—a fraction of the cost of human tutoring.

2.3 Human-AI Tutoring

Human-AI tutoring offers a balanced approach by integrating the cognitive and motivational support of human tutors with the scalable, data-driven instruction of ITSs [1]. Based on individual student needs, human tutors provide personalized feedback and scaffolding, while ITSs ensures continuous, adaptive instructional support. This hybrid model seeks to combine the strengths of both approaches while maintaining scalability and cost-effectiveness.

Prior implementations have demonstrated the effectiveness of human-AI tutoring in improving student learning outcomes on tests, compared to propensity-matched students who did not receive human-AI tutoring [12]. Additionally, prior studies have found that human-AI tutoring enhances student outcomes within AI tutors [34]. These findings attest to the effectiveness of human-AI tutoring in both in-person [12] and virtual [34] settings. With costs ranging from $397 to $700 per student [12,34], human-AI tutoring offers a more feasible alternative to the cost of high-impact tutoring ($3,500 to $4,300) [19]. As such, the effectiveness, lower costs, and flexibility make human-AI tutoring a promising and sustainable instructional model.

Thomas et al. [34] suggest that human-AI tutoring enhances student learning by increasing opportunities, leading to higher lesson completion and mastery. This effect results from a combination of adaptive AI tutor-driven instruction and personalized human tutor support [1]. As students receive targeted assistance and real-time feedback, they solve more problems and engage with progressively challenging content. The conceptual framework proposed by Thomas et al. [34] is illustrated in Fig. 1.

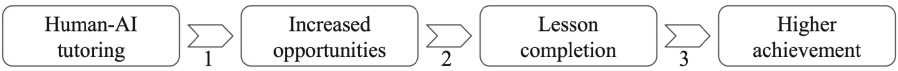

Fig. 1. Access to human-AI tutoring influences student engagement, which in turn improves student performance and leads to higher achievement [34].

Thomas et al. [34] found that students receiving human-AI tutoring had higher time on tasks (engagement), completed more problems (progress), and mastered more skills (achievement) than their peers receiving AI-only tutoring.

[2] https://www.ixl.com/.
[3] https://www.curriculumassociates.com/programs/i-ready.

This effect was observed across three different sites using different AI tutors, i.e., MATHia, i-Ready, and IXL.

In our study, we extend prior research demonstrating the feasibility of human-AI tutoring for improving student learning within ITS [34] and on standardized tests [12]. While these prior studies report on the short-term benefits of human-AI tutoring, the current study evaluates its effectiveness on a year-long implementation. Specifically, we examine its impact within-tutor and out-of-tutor measures of learning.

3 Methods

We use a quasi-experimental design with propensity score matching to compare human-AI tutoring with AI-only tutoring. For **RQ1**, we examine the effect of human-AI tutoring on within-tutor measures of learning in i-Ready (Fig. 1). Next, we evaluate whether these improvements in i-Ready translate to better performance on standardized state tests (**RQ2**).

3.1 Dataset

The dataset contains iReady[4] data for Grade 7 students from a school in the U.S., collected during the 2022–2023 and 2023–2024 school years in the U.S. There were 402 students in the first year and 446 in the second. The dataset includes i-Ready diagnostic scores, state test scores for Grade 6 (pre) and Grade 7 (post), and demographic information. The sample was 50.6% female and 49.4% male, with 65% Hispanic, 17% Asian, 9.5% African American, 5.7% White, and 2.64% other races. Additionally, 13.5% received Special-Ed services, 87.5% were socioeconomically disadvantaged, and 18% were English Language Learners.

The i-Ready data includes detailed logs at the student, lesson, problem, and attempt levels. In both academic years, students used i-Ready as part of their regular math curriculum. The only major instructional change in 2023–2024 was the introduction of virtual tutors, who provided virtual tutoring support on Mondays. All data collection procedures followed IRB-approved protocols, ensuring ethical data handling and student privacy (Table 1).

Table 1. Student data from two consecutive academic years where students received AI-only or human-AI tutoring.

Academic Year	Tutoring Type	Teachers (N)	Classes (N)	Students (N)
2022–2023	ITS (AI only tutoring)	3	12	402
2023–2024	Human-AI tutoring	4	14	446

[4] https://i-readycentral.com/articles/middle-school-3/.

3.2 Personalization in I-Ready

In i-Ready, a student's math proficiency is estimated using a computer-adaptive diagnostic test [36] administered at the start of the school year. This test helps ensure students are assigned content appropriate to their skill level. Lessons typically conclude with a *Quiz Module* that assesses mastery of the topic. These quizzes are developed using Item Response Theory to provide precise mastery estimates. Students who pass (with a score above 64%) progress to the next lesson, while those who do not may be required to redo the lesson or complete additional prerequisite content. As students demonstrate mastery and complete more lessons, they are assigned increasingly advanced material. Thus, an increase in the grade level of assigned lessons reflects growth in mathematical ability. Learning in i-Ready is therefore measured through student performance (i.e., passing) on lesson quizzes and the grade level of the assigned lessons.

3.3 Human-AI Tutoring

The tutors were undergraduate students from a U.S. university, trained on tutoring strategies to identify and address student needs, such as social-emotional learning, advocacy, and relationship building [10]. While students could access i-Ready at any time, both groups used it during a designated period, on Mondays, during which the human-AI group received virtual tutoring via Zoom.

Students on Zoom were randomly assigned to individual breakout rooms, whereas tutors were assigned a set of rooms, maintaining a 1:4 tutor-to-student ratio. Tutors rotated among rooms in a round-robin fashion to monitor progress and provide support. Although the class period was 40 min, tutoring sessions lasted approximately 28–32 min, depending on the school's bell schedule and classroom context. Given this rotation model, each student typically received about 5–8 min of one-on-one human tutoring per session.

3.4 Study Design

Propensity Score Matching: We used *full* PSM [31] to estimate the effects of human-AI tutoring compared to AI-only tutoring. This approach improves the validity of the comparison by balancing covariates, reducing selection bias, and enhancing effect estimates [22,31]. We applied *logit* distance with calipers set at 0.1 standard deviations (SD) to ensure strong matches. We use student-level covariates, including gender, ethnicity, SES, ELL status, special education status, and i-Ready diagnostic score. The i-Ready diagnostic scores were used over prior state test scores to account for potential summer learning loss.[5].

Post-matching, we assess balance using the absolute Standardized Mean Difference (|SMD|) [6,32]. SMD measures the standardized difference between group means, providing a scale-independent metric for comparing covariate distributions. An |SMD| below 0.1 indicates minimal imbalance, and 0.1–0.2 suggests moderate imbalance requiring review, risking biased estimates [6].

[5] The student's Grade 6 state scores and their performance on the fall i-Ready diagnostic tests were strongly correlated ($Pearson'sr = 0.81, p < 0.001$).

3.5 Longitudinal Evaluation of Human-AI Tutoring

For (**RQ1**), we adopt the conceptual framework proposed by Thomas et al. [34] (Figure 1) to evaluate the impact of human-AI tutoring on engagement and within-tutor measure of learning (pass percentage and lesson grade level).

1. **Engagement**: Monthly student time on task in i-Ready, including both time spent with and without support from human tutors in the human-AI group.
2. **Performance**: Lesson pass percentage represents the proportion of lessons students pass each month.
3. **Achievement**: Grade level indicates the average grade level of lessons students worked on each month and serves as a proxy for student proficiency.

For student i in month j, we model student outcomes (Y_{ij}), as a function of human-AI tutoring (X_{ij}^{tutor}), the month of the school year (X_{ij}^{month}), their interaction, and student-level variance (u_i) as shown in Eq. 1.

$$Y_{ij} = \beta_0 + \beta_1 X_{ij}^{\text{tutor}} + \beta_2 X_{ij}^{\text{month}} + \beta_3 (X_{ij}^{\text{tutor}} \cdot X_{ij}^{\text{month}}) + u_j + \epsilon_{ij} \tag{1}$$

3.6 Impact of Human-AI Tutoring on State Test Performance

For **RQ2**, we first assess the impact of human-AI tutoring on student performance in standardized state tests (Eq. 2). We then extend the analysis to examine whether its effects vary by prior performance and time on task (Eqs. 3 and 4). Since motivational and instructional support from human tutors can augment the benefits of student engagement, we investigate whether the effectiveness of human-AI tutoring varies with students' level of engagement.

For student i, we model their state test score (Y_i), as a function of human-AI tutoring (X_i^{tutor}), prior state test score (X_i^{prior}), time on task (X_i^{time}), and their interactions as shown in Eqs. 2, 3 and 4.

$$Y_i = \beta_0 + \beta_1 X_i^{\text{tutor}} + \epsilon_i \tag{2}$$

$$Y_i = \beta_0 + \beta_1 X_i^{\text{tutor}} + \beta_2 X_i^{\text{prior}} + \beta_3 (X_i^{\text{tutor}} \cdot X_i^{\text{prior}}) + \epsilon_i \tag{3}$$

$$\begin{aligned} Y_i = {} & \beta_0 + \beta_1 X_i^{\text{tutor}} + \beta_2 X_i^{\text{prior}} + \beta_3 X_i^{\text{time}} \\ & + \beta_4 (X_i^{\text{tutor}} \cdot X_i^{\text{prior}}) + \beta_5 (X_i^{\text{tutor}} \cdot X_i^{\text{time}}) + \epsilon_i \end{aligned} \tag{4}$$

We estimate the Average Treatment Effect on the Treated (ATT) by using *estimatr* package in R to fit a weighted linear regression. The weights are derived from PSM, where treated units receive uniform weights, and control units are weighted to achieve covariate balance. Since the control group comes from the prior year, ATT is the appropriate measure, as it estimates the treatment effect on the current cohort using a counterfactual based on the prior year's data.

4 Results

4.1 Propensity Score Matching

The results of PSM, as outlined in Sect. 3.4, are presented in Table 2. The dataset was filtered to only include students who took the Grade 7 state test. A z-score threshold of (–3, 3) was used to remove outliers based on prior test scores, time on task, and lessons passed. Full PSM resulted in 319 students from the AI-only group being matched with 359 students from the human-AI group.

As described in Sect. 3.4, we use absolute SMD, a standardized measure of covariate balance, to assess group comparability. Before matching, the imbalance was minimal ($|SMD| < 0.1$) for most factors, with a moderate imbalance in ELL ($|SMD| = 0.13$), gender ($|SMD| = 0.11$), and SES ($|SMD| = 0.10$). Post-matching, the imbalance was minimal across all factors ($|SMD| \leq 0.05$).

Table 2. Results of PSM between AI-only and Human-AI tutoring.

	AI-only	Human-AI	Total (N)
Original Dataset	402	433	835
After pre-processing	319	366	685
Matched	319	359	678
Unmatched	0	7	7

4.2 RQ1: Longitudinal Evaluation of Human-AI Tutoring

To address **RQ1**, we conduct a longitudinal analysis of the impact of human-AI tutoring on student engagement and within-tutor measures of learning. While the start and end months differed, each school year lasted 36 weeks. For consistency, we define a "month" as a four-week interval from the start of the school year rather than aligning it with the actual calendar months.

Trend Analysis: The monthly trends in student time on task, pass percentage, and grade level are illustrated in Fig. 2. Initially, the human-AI group exhibited higher engagement (Fig. 2a), but it declined over time, while engagement in the AI-only group remained stable. Despite this decline, the human-AI group performed better (Fig. 2b) and achieved more (Fig. 2c), during the later months of the school year.

Impact of Human-AI Tutoring: We use linear mixed-effects models, as outlined in Sect. 3.5, to assess the impact of human-AI tutoring. The results are presented in Table 3.

Access to human-AI tutoring significantly increased engagement in the initial months ($\beta = 0.31$, $p < 0.001$), with students averaging 1.68 h per month in i-Ready, compared to 1.37 h in the AI-only group (intercept). However, engagement declined significantly over time in the human-AI group ($\beta = -0.08$, $p < 0.001$), while no such decline was observed in the AI-only group.

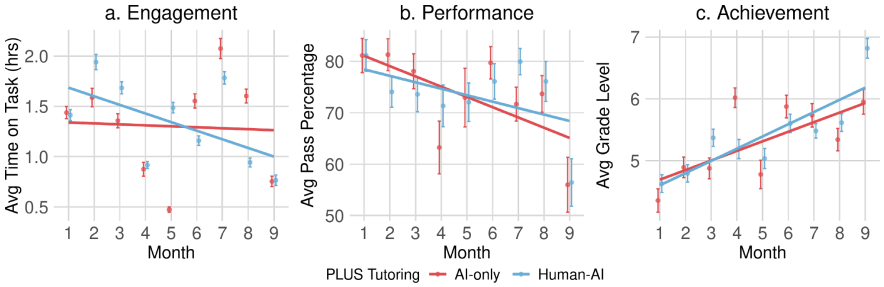

Fig. 2. Visualizing trends in engagement, and within-tutor measures of learning on a monthly basis (error bars represent 95% confidence intervals).

Initially, the AI-only group had a higher pass percentage than the human-AI group ($\beta = -0.05$, $p = 0.014$). However, pass rates declined over time, likely due to increasing lesson difficulty. This decline was steeper for the AI-only group ($\beta = -0.02$, $p < 0.001$), whereas access to human tutors helped mitigate it in the human-AI group ($\beta = 0.01$, $p = 0.004$). By mid-year, pass percentages were higher in the human-AI group than in the AI-only group.

There was no significant initial difference in grade-level between groups ($\beta = -0.12$, $p = 0.186$). While both groups demonstrated significant learning over time, the AI-only group progressed at a rate of 0.16 grade levels per month ($\beta = 0.16$, $p < 0.001$). In contrast, access to human tutors ($\beta = 0.04$, $p < 0.001$) increased the progress of the human-AI group to 0.20 grade levels per month, i.e., by the 9th month the human-AI group was ahead by 0.36 grade levels.

Table 3. Comparing the impact of human-AI tutoring with AI-only tutoring.

Predictors	time on task		pass percentage		grade level	
	β	SE	β	SE	β	SE
(Intercept)	1.37^{***}	0.03	0.82^{***}	0.01	4.84^{***}	0.07
human-AI Tutoring	0.31^{***}	0.04	-0.05^{*}	0.02	–0.12	0.09
Month	–0.01	0.01	-0.02^{***}	0.00	0.16^{***}	0.01
human-AI Tutoring × Month	-0.08^{***}	0.01	0.01^{**}	0.00	0.04^{***}	0.01

$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

4.3 RQ2: Evaluation of Human-AI Tutoring on Learning Outcomes

For **RQ2**, we examine whether improvements in i-Ready translate to better performance on standardized state tests. Following the methodology in Sect. 3.6, we estimate the ATT for human-AI tutoring. All features are standardized, allowing estimates to be interpreted in standard deviation (SD). The main effect of

human-AI tutoring on learning outcomes is presented in Table 4 (model 1). There was no significant difference in state test scores ($\beta = -0.04, p = 0.66$) between the human-AI and AI-only groups, despite the human-AI group showing significantly higher achievement in **RQ1**.

As part of our formative evaluation, we extend our analysis to investigate the differential impacts of human-AI tutoring by interacting it with students' prior state test scores and time on task in i-Ready. When controlling for prior performance and time on task (Model 3 in Table 4), we observed significant interactions between human-AI tutoring and both total time on task ($\beta = 0.19, p < 0.01$) and prior performance ($\beta = -0.15, p = 0.03$), indicating that tutoring effects vary by student engagement and baseline ability. In contrast, time on task had no significant effect in the AI-only group ($\beta = 0.05, p = 0.53$).

To contextualize these results (Model 3 in Table 4), we conducted a two-sample t-test comparing time on task between the human-AI and AI-only groups. While the difference was statistically significant (0.77 hours, $t = 3.04$, $p = 0.03$), it only amounted to 1.28 additional minutes of i-Ready usage per week by the human-AI group. This minimal difference suggests that the observed interaction effect between human-AI tutoring and time on task is not simply a function of greater i-Ready dosage and likely reflects qualitative differences in engagement due to the cognitive and motivational support provided by human tutors.

Table 4. Evaluating the impact of human-AI tutoring on state test scores.

	(model 1)		(model 2)		(model 3)	
	post test score		post test score		post test score	
Predictors	β	SE	β	SE	β	SE
(Intercept)	0.02	0.08	0.01	0.06	0.02	0.06
human-AI tutoring	–0.04	0.09	–0.03	0.07	–0.06	0.07
prior test score	–	–	0.77^{***}	0.06	0.77^{***}	0.06
human-AI tutoring×prior test score	–	–	–0.11	0.07	-0.15^{*}	0.07
time on task	–	–	–	–	0.05	0.05
human-AI tutoring×time on task	–	–	–	–	0.19^{**}	0.07

$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

Figure 3 illustrates the relationship between human-AI tutoring and time on task (Model 3 in Table 4). The actual time values are used for interpretability, and students are categorized as low or high performers using median prior test scores. Scatter plots display observed scores, while regression lines represent predicted scores. Figure 3 helps illustrate how the benefits of human-AI tutoring on state test performance are conditioned on students' use of i-Ready. For low-performing students, the crossover point—where human-AI tutoring begins to outperform AI-only tutoring—occurs at approximately 11 h of total time on task, compared to around 14 h for high-performing students.

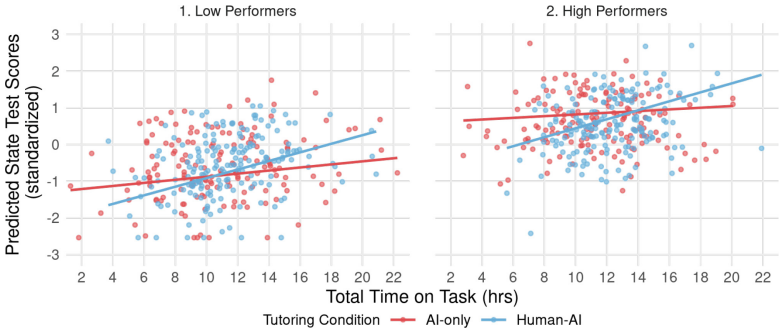

Fig. 3. Relationship between time on task and state test scores. (Regression lines show predicted scores whereas scatter plots display observed scores.)

5 Discussion and Conclusion

Human-AI tutoring Improves Student Learning Outcomes in i-Ready. From our exploration of **RQ1**, we find that access to human-AI tutoring significantly enhances student learning outcomes within i-Ready. On average, students in the human-AI group progressed at a significantly higher rate of 0.2 *grade levels* per month, compared to 0.16 *grade levels* per month for students in the AI-only group (Table 3). By the end of the 9-month school year, the human-AI group was ahead by 0.36 grade levels. Similar to Thomas et al. [34], we find that access to human-AI tutoring improves student learning outcomes within ITS.

Despite the observed improvements in within-tutor learning, student engagement gradually declined over time (Fig. 2a). This may be due to the novelty wearing off, increasing lesson difficulty, or students learning to disengage while tutors rotated between breakout rooms. The decline was more pronounced in the human-AI group, suggesting missed opportunities to sustain engagement and support continued progress. Given that engagement (time on task) is a key factor influencing student learning, monitoring engagement and providing timely support are essential for maximizing the benefits of human-AI tutoring. Prior research on teaching augmentation tools has demonstrated their potential to enhance teachers' ability to identify students and provide timely support to students [3,21]. Our findings highlight opportunities to integrate similar tools into human-AI tutoring environments to improve tutors' ability to provide timely motivational and instructional support to students.

Higher AI Tutor Use in the Human-AI Group is Associated with Better Performance on the State Test. While we did not observe a significant average treatment effect on the treated (ATT) for human-AI tutoring (**RQ2**), further analysis revealed a significant interaction between tutoring condition and time on task. One standard deviation increase in time on task (3.26 h) was associated with a 0.28 SD (24.27-point) gain in state test scores for the human-AI

group, compared to only 0.06 SD (5.2 points) in the AI-only group. This suggests that the benefits of human-AI tutoring are conditioned upon students' continued use of the AI tutor.

Aleven et al. [1] posit that cognitive and motivational support from human tutors—such as providing encouragement, clarifying misconceptions, and connecting new tasks to prior topics—can enhance the learning benefits of AI tutors. Our findings build on this claim by offering nuanced insight into how human tutoring contributes to student learning. The significant interaction between time on task and human-AI tutoring (Model 3 in Table 4) suggests that the benefits of human tutoring are contingent upon students' continued use of the AI tutor. As shown in Fig. 3, students in the human-AI group with higher time on task tend to outperform their AI-only peers. In contrast, the absence of a significant relationship in the AI-only group further underscores that comparable levels of AI tutor use, without human support, are not as beneficial to learning.

State Tests May not Fully Capture Learning Gains. One possible explanation for the lack of a main effect on state test performance is that most students began the year significantly below grade level (i.e., two grade levels behind). Because standardized tests are designed to assess on-grade-level proficiency, they may fail to detect progress among students working on out-of-level content [30,37]. Based on i-Ready grade-level estimates (Table 3), both the human-AI (6.52) and AI-only (6.16) groups were working on early to mid Grade 6 content prior to the state test. However, the end-of-year Grade 7 assessment may not fully reflect this difference due to the limited number of items targeting Grade 6 content, which can reduce the precision of performance estimates for students working well below grade level.

Our formative evaluation of human-AI tutoring, delivered at approximately $700 per student, demonstrates that relatively small doses of continuous human support can meaningfully improve learning outcomes over time. The significant grade-level gains of the human-AI group (0.36) are especially compelling given the well-established effectiveness of AI tutors [25,35]. Unlike the NSSA's recommendation for high-impact tutoring—which emphasizes continuous, in-person instruction in one-on-one or small-group settings—we implemented an easier-to-scale virtual tutoring with a round-robin model. Despite these structural differences, students in the human-AI group learned more within the AI tutor and also demonstrated stronger transfer to the state test—though this transfer was observed only among students with higher AI tutor use (time on task).

Our findings present a promising solution to the long-standing tension in AIED between personalization and scalability. The current approach, though lightweight, demonstrates the feasibility of combining the individualized support of human tutoring with the continuous assistance of AI tutors. These findings lay a strong foundation for rethinking how human-AI collaboration can enhance both the learning experience for students and the overall effectiveness of AI tutors. As this model continues to evolve, it offers a path toward more engaging, responsive, and effective learning environments. These opportunities can be further amplified

by harnessing the full potential of AIED—through behavioral modeling, teaching augmentation, improved tutor training, and refined virtual tutoring norms.

Acknowledgments. The authors would like to thank Curriculum Associates, LLC, for their partnership and collaborative efforts in this research. This work was supported by the Learning Engineering Virtual Institute. The opinions, findings, and conclusions expressed in this material are those of the authors and do not necessarily reflect the views of the Institute or Curriculum Associates.

References

1. Aleven, V., et al.: Towards the future of AI-augmented human tutoring in math learning. In: AIED, pp. 26–31. Springer (2023)
2. Aleven, V., Mclaren, B., Roll, I., Koedinger, K.: Toward meta-cognitive tutoring: a model of help seeking with a cognitive tutor. Int. J. Artif. Intell. Educ. **16**(2), 101–128 (2006)
3. An, P., Holstein, K., d'Anjou, B., Eggen, B., Bakker, S.: The ta framework: Designing real-time teaching augmentation for k-12 classrooms. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, pp. 1–17 (2020)
4. Anderson, J.R., Corbett, A.T., Koedinger, K.R., Pelletier, R.: Cognitive tutors: lessons learned. J. Learn. Sci. **4**(2), 167–207 (1995)
5. Associates, C.: State of student learning in 2023: academic recovery remains slow in most grades. https://cdn.bfldr.com/LS6J0F7/at/35hn9w6qrwqbqqhp69g34jv/ca-state-of-student-learning-executive-summary-2023.pdf (2023), Accessed 22 Sep 2024
6. Austin, P.C.: An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivar. Behav. Res. **46**(3), 399–424 (2011)
7. Bloom, B.S.: The 2 sigma problem: the search for methods of group instruction as effective as one-to-one tutoring. Educ. Res. **13**(6), 4–16 (1984)
8. du Boulay, B.: Recent meta-reviews and meta-analyses of aied systems. IJAIED **26**(1), 536–537 (2016)
9. Canfield, W.: Aleks: a web-based intelligent tutoring system. Math. Comput. Educ. **35**(2), 152 (2001)
10. Chhabra, P., Chine, D., Adeniran, A., Gupta, S., Koedinger, K.: An evaluation of perceptions regarding mentor competencies for technology-based personalized learning. In: Society for Information Technology & Teacher Education International Conference, pp. 1812–1817. Association for the Advancement of Computing in Education (AACE) (2022)
11. Chi, M.T., Siler, S.A., Jeong, H., Yamauchi, T., Hausmann, R.G.: Learning from human tutoring. Cogn. Sci. **25**(4), 471–533 (2001)
12. Chine, D.R., et al.: Educational equity through combined human-AI personalization: a propensity matching evaluation. In: International Conference on Artificial Intelligence in Education, pp. 366–377. Springer (2022)
13. Core, M.G., Moore, J., Zinn, C.: The role of initiative in tutorial dialogue. In: 10th Conference of the European Chapter of the Association for Computational Linguistics, pp. 67–74. ACL (2003)

14. Dietrichson, J., Bøg, M., Filges, T., Klint Jørgensen, A.M.: Academic interventions for elementary and middle school students with low socioeconomic status: a systematic review and meta-analysis. Rev. Educ. Res. **87**(2), 243–282 (2017)
15. Education Lab, T.U.O.C.: Realizing the promise of high dosage tutoring at scale. https://educationlab.uchicago.edu/wp-content/uploads/sites/3/2024/03/UChicago-Education-Lab-PLI-Technical-Report-03.2024.pdf (2024), Accessed 22 Sep 2024
16. Forbes-Riley, K., Litman, D.J.: Analyzing dependencies between student certainness states and tutor responses in a spoken dialogue corpus. Recent Trends Disc. Dial. 275–304 (2008)
17. Fox, B.A.: The Human Tutorial Dialogue Project: Issues in the Design of Instructional Systems. CRC Press (2020)
18. Graesser, A.C., Wiemer-Hastings, K., Wiemer-Hastings, P., Kreuz, R., Group, T.R., et al.: Autotutor: a simulation of a human tutor. Cogn. Syst. Res. **1**(1), 35–51 (1999)
19. Guryan, J., et al.: Not too late: improving academic outcomes among adolescents. Am. Econ. Rev. **113**(3), 738–765 (2023)
20. Heffernan, N.T., Heffernan, C.L.: The assistments ecosystem: building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching. IJAIED **24**, 470–497 (2014)
21. Holstein, K., Aleven, V.: Designing for human-AI complementarity in k-12 education. AI Mag. **43**(2), 239–248 (2022)
22. Hong, G., Raudenbush, S.W.: Effects of kindergarten retention policy on children's cognitive growth in reading and mathematics. Educ. Eval. Policy Anal. **27**(3), 205–224 (2005)
23. Katz, S., Jordan, P., Litman, D.: Rimac: a natural-language dialogue system that engages students in deep reasoning dialogues about physics. Soc. Res. Educ. Effect. (2011)
24. Kraft, M.A., Falken, G.T.: A blueprint for scaling tutoring across public schools. edworkingpaper no. 20-335. Annenberg Institute for School Reform at Brown University (2020)
25. Kulik, J.A., Fletcher, J.D.: Effectiveness of intelligent tutoring systems: a meta-analytic review. Rev. Educ. Res. **86**(1), 42–78 (2016)
26. McNamara, D.S., Levinstein, I.B., Boonthum, C.: istart: interactive strategy training for active reading and thinking. Behav. Res. Methods Instrum. Comput. **36**(2), 222–233 (2004)
27. Mitrovic, A.: An intelligent sql tutor on the web. Int. J. Artif. Intell. Educ. **13**(2–4), 173–197 (2003)
28. Muldner, K., Lam, R., Chi, M.T.: Comparing learning from observing and from human tutoring. J. Educ. Psychol. **106**(1), 69 (2014)
29. Nickow, A., Oreopoulos, P., Quan, V.: The impressive effects of tutoring on prek-12 learning: a systematic review and meta-analysis of the experimental evidence (2020)
30. Resch, A., Isenberg, E.: How do test scores at the floor and ceiling affect value-added estimates? Technical Report, Mathematica Policy Research (2014)
31. Rosenbaum, P.R.: Dropping out of high school in the united states: an observational study. J. Educ. Stat. **11**(3) (1986)
32. Rubin, D.B.: Using propensity scores to help design observational studies: application to the tobacco litigation. Health Serv. Outcomes Res. Method. **2**, 169–188 (2001)

33. Shute, V.J., et al.: The design, development, and testing of learning supports for the physics playground game. Int. J. Artif. Intell. Educ. **31**, 357–379 (2021)
34. Thomas, D.R., 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)
35. VanLehn, K.: The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. Educ. Psychol. **46**(4), 197–221 (2011)
36. Wainer, H., Dorans, N.J., Flaugher, R., Green, B.F., Mislevy, R.J.: Computerized Adaptive Testing: A Primer. Routledge (2000)
37. Zhu, L., Gonzalez, J.: Modeling floor effects in standardized vocabulary test scores in a sample of low ses hispanic preschool children under the multilevel structural equation modeling framework. Front. Psychol. **8**, 2146 (2017)