Improving Tutor Discourse Practices via AI-Enhanced Coaching: A Piecewise Latent Growth Curve Modeling Approach

Sandra Sawaya[1][0009-0009-9680-249X], Jennifer Jacobs[1] [0000-0002-2300-4771], Robert Moulder[1] [0000-0001-7504-9560], Brent Milne[2][0009-0004-4891-1847], Tom Fischaber[2][0009-0007-3825-796X], Sidney K. D'Mello[1] [0000-0003-0347-2807]

[1] University of Colorado Boulder, Colorado, USA
sandra.sawaya@colorado.edu
[2] Saga Education, USA

Abstract. Instructional coaching – where coaches observe and provide feedback and guidance to improve practice – is a highly effective job-embedded professional learning approach but is difficult to scale. To address this, we developed a Hybrid Human-Agent Tutoring (HAT) platform which provides human coaches with AI feedback on the quality of discourse practices used by the human tutors assigned to them and guides their coaching sessions. We investigated whether HAT resulted in growth in tutors' use of discourse moves known to foster rich mathematical discussions (e.g., pressing for reasoning) in collaboration with a large provider of tutoring services to underrepresented youth. Using a piecewise latent growth modeling approach, we found significant improvements in tutors' use of four of six discourse moves, with negligible changes for the other two. Importantly, the introduction of HAT resulted in a reversal of decline in usage of key discourse moves. Coaches' usage patterns of HAT varied, though they mostly reported positive perceptions of the system. We discuss the implications of automated AI feedback tools such as HAT in scaling high-dosage tutoring programs effectively.

Keywords: Instructional Coaching, Discourse Analytics, Automated Feedback, High-Dosage Tutoring, Piecewise Latent Growth Modeling.

1 INTRODUCTION

Research has demonstrated that human tutoring improves student achievement. According to a recent meta-analysis of 89 randomized experiments, studies on pre-K to 12 tutoring programs reported statistically significant impacts on student learning outcomes, with an overall pooled effect size of 0.29 SD [1]. On average, effect sizes were larger for high-dosage tutoring (HDT) programs, which are small-group or one-on-one tutoring sessions, held during the school day, multiple times a week, with trained tutors. Research has generally shown that these programs can help students make significant progress on learning including in districts that support low-income students [2]. However, scaling HDT programs while retaining their effectiveness is complex and challenging [3], mainly due to costs involved in supporting a qualified tutoring workforce and lack of access to trained tutors in underserved communities [4].

One approach to scale is to increase the number of students per tutoring session by relaxing the gold standard of one-on-one tutoring to small-group tutoring [3]. Another approach is to hire paraprofessional tutors with college degrees, but no formal instructional training or background [1]. However, these paraprofessionals need ongoing professional learning (PL) to become effective, especially in the more complex context of small-group tutoring which involves managing social dynamics among students in addition to the learning content.

Instructional coaching, where coaches meet regularly with teachers and provide them with contextualized feedback, is a widely used PL strategy in the classroom context, with strong evidence supporting its effectiveness [5]. Indeed, the National Student Support Accelerator identifies ongoing coaching for paraprofessional tutors as a key criterion – and challenge – for the successful implementation of HDT programs [6]. However, instructional coaching itself can be costly [7], as it requires coaches to observe classrooms, identify instructional insights, and engage in an observation and feedback coaching cycle. These traditional methods rely on in-person observations, which are subjective, resource-intensive, and difficult to scale effectively. To address this challenge, the present study examines the use of AI-based approaches to support instructional coaching aimed at improving tutoring practice with an eye towards enhancing student learning outcomes.

Related Work. Researchers have explored the use of AI to provide automated feedback to scale *teacher* professional learning. Specifically, teachers upload recordings of their classroom sessions to the AI tools, which then apply cutting-edge advances in speech and discourse processing to provide teachers with automated, data-driven feedback on their classroom discourse [8], such as student-to-teacher talk ratio, teacher questions, student uptake of ideas, and so on. The focus on improving discourse is motivated by considerable evidence suggesting that dialogic, or discourse-based teaching environments can increase student engagement and student learning gains [9]. Examples include the *TalkMoves* application [10], *M-Powering Teachers* [11], the *Teacher Talk Tool* [12], and the commercial platform *TeachFX* (www.teachfx.com). AI feedback tools for teachers have been shown to significantly increase teacher and student talk moves in a correlational study [10], increase the uptake of student ideas, and improve overall student satisfaction in an online experimental study [11]. Researchers have also investigated how to integrate these tools into classroom PL models [13].

However, whereas teachers are experienced educators with degrees in education, effectively scaling HDT programs requires investments in PL technologies for novice, paraprofessional tutors. There are emerging efforts to develop automated feedback tools for tutors. For example, *Talk Meter* provides a feedback visualization on student-to-tutor talk ratio during a tutoring session. In a month-long experiment, researchers found that when *Talk Meter* was shown to tutors, tutor talk decreased by an average of 14% in comparison to a control group [14]. Another example is *Tutor CoPilot*, an AI tool that provides math tutors with real-time guidance during their tutoring session. In a two-month experiment, researchers saw improvements in students' mastery of topics as measured through passing an exit ticket [15]. Although these studies serve as a useful proof-of-concept of providing feedback to tutors, the short-duration of these interventions from a few weeks to a few months, raises questions about their sustained impact. These tools have also yet to be integrated into an instructional coaching model which has different affordances than providing feedback directly to teachers or tutors.

Hybrid Human-Agent Tutoring (HAT) Platform. To address the challenge of providing effective PL for tutors through instructional coaching at scale, we combine the scalability of automated feedback tools with the effectiveness of human instructional coaching. Specifically, we developed an automated, data-driven, AI tool for tutor discourse analytics called HAT that analyzes recordings of small-group tutoring sessions for evidence of tutors' use of high-impact discourse practices (based on the framework of *Academically Productive Talk* (APT) [16]) and provides coaches with feedback visualizations on these practices. According to our theory of change (Fig. 1), coaches review HAT-generated tutor discourse analytics, engage in sense-making of the information to provide feedback and guidance to their tutors. Tutors who act on this feedback should improve their tutoring practices, leading to enhanced student achievement outcomes. In this paper, we test the first component of this hypothesis that *an automated tutor discourse analytics tool embedded in an instructional coaching model is associated with improvements in tutor practice.*

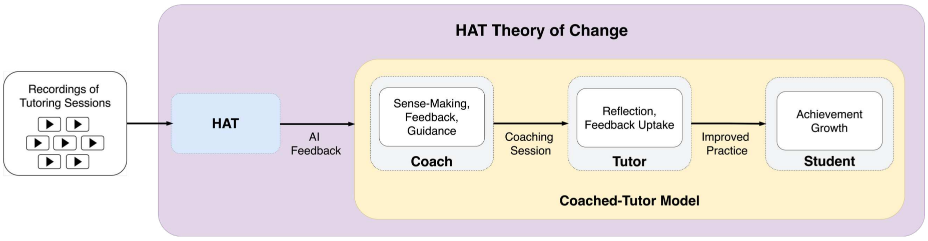


Fig. 1. HAT Theory of Change. Adapted from [17]

Current Study. To test our hypothesis, we collaborated with Saga Education (Saga), a national tutoring service that provides high-dosage tutoring (HDT) to Title I schools in the U.S. (i.e., public schools with predominantly low-income students). As part of Saga's program, students receive HDT from a paraprofessional tutor three to five times a week. Tutoring sessions are recorded with their consent and recordings are processed through HAT. Coaches then use the automated tutor discourse analytics from HAT in their ongoing biweekly coaching sessions with their tutors as part of Saga's professional learning (PL) model.

Because our work is conducted as part of a research-practice partnership with an authentic tutoring provider and HAT was rolled out to all coaches (i.e., no control group), it was not feasible to conduct a randomized controlled trial (RCT). Accordingly, we used a piecewise latent growth-curve modeling (LGCM) approach: a powerful quasi-experimental design method related to interrupted-time series analyses. Specifically, we used LGCM to assess changes in tutoring practices after the launch of

HAT (post-intervention) compared to changes that would be expected without the intervention (i.e., counterfactual). This design has an advantage over traditional pre-post designs in that it provides a counterfactual comparison of change in the absence of the intervention rather than simply contrasting pre and post values (Fig. 2). LGCM allows for the estimation of individual trajectories of change over time within a structural equation modeling framework [18]. It enables us to analyze how the introduction of HAT affected the rate of change in tutor discourse practices (changes in slopes), and to determine any immediate effects of transitioning to HAT (changes in intercepts).

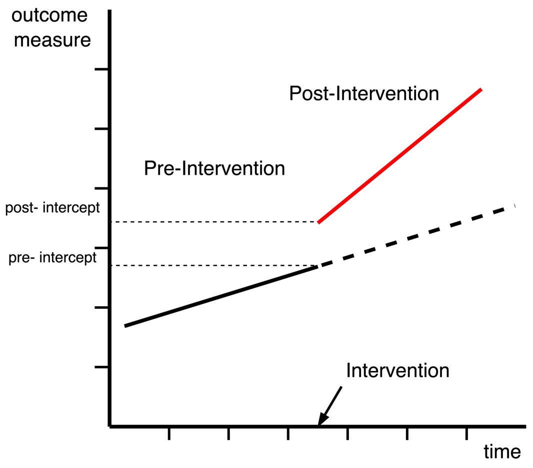


Fig. 2. Representation of a piecewise latent growth-curve model (adapted from [19]). Solid lines represent observed rates of change in outcome (slopes) before (black) and after the intervention (red), while the dashed line is the projected trend had the intervention not occurred (counterfactual). The change in intercept at the point of intervention reflects its immediate effect, whereas the difference in pre- and post-intervention slopes demonstrates changes in growth.

Our work is **novel** in that whereas most existing work focuses on direct-to-teacher AI feedback tools or preliminary work on direct-to-tutor feedback, our work focused on developing AI tools to support the professional learning (PL) of novice tutors embedded in a coached-tutor framework. Our use of piecewise LGCM as a quasi-experimental approach is also novel in this context and can inspire other research efforts that aim for rigorous evaluation methods when RCTs are not feasible. Our specific research questions are: **RQ1:** What are coaches' patterns of use and perceptions of HAT? **RQ2:** To what extent do tutor discourse practices change as a result of coaches' access to HAT?

2 HAT PLATFORM

HAT aims to address the challenge of scaling effective tutor PL through instructional coaching by automatically analyzing tutoring sessions and providing coaches with data-driven feedback on the quality of tutoring discourse based on the framework of *Academically Productive Talk* [16]. In partnership with Saga, HAT was integrated into Saga's tutoring platform and deployed as part of ongoing two-week observation cycles. Coaches use HAT feedback to deliver AI-enhanced coaching, helping tutors improve their practice.

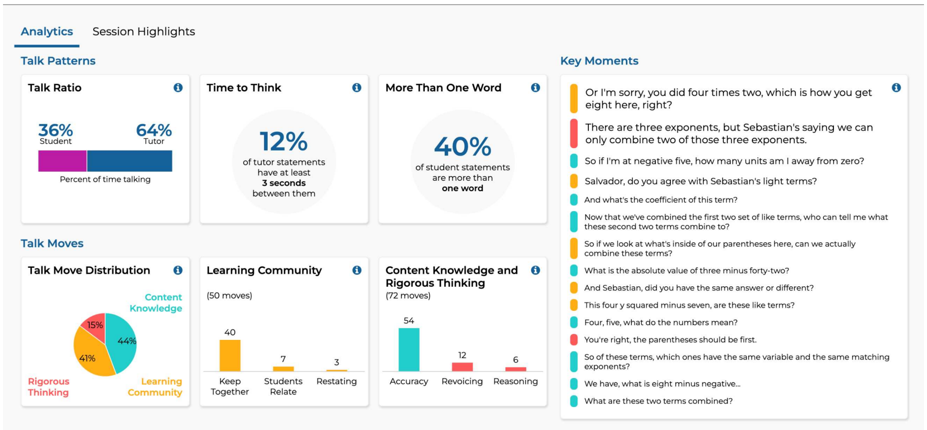
Focal Discourse Practices. *Academically Productive Talk* identifies teaching practices that foster rich and rigorous student-led discussions. These practices, known as *talk moves*, are discourse acts that encourage productive academic dialogue and promote equitable conversations between students and teachers or tutors [16]. While both student and teacher or tutor talk moves exist, our present work focused on processing tutoring sessions for tutor talk moves. There are several types and categories of *talk moves*: those that support the *learning community*, express *content knowledge*, and encourage *rigorous thinking* (see Table 1). Research has shown that these talk moves can help increase the level of academic rigor in the classroom [20] promote equitable participation among students [21], increase student learning and engagement [9].

Table 1. Tutor talk moves categories, types, and examples (adapted from www.talkmoves.com)

Category	Talk Move Type	Examples
Learning Community	**Keeping everyone together**	Tutor (T): What did Eliza just say?
	Getting students to **Relate** to others' ideas	T: Do you agree with her that the answer is 7/10?
	Getting students to **Restate** others' ideas	Student (S): Add 2 here. T: Add 2 here.
Content Knowledge	Pressing for **Accuracy**	T: Can you give an example of an ordered pair?
Rigorous Thinking	**Revoicing**	S: Add 2 here. T: Julie thinks we should add 2 to this part of the equation.
	Pressing for **Reasoning**	T: Why can I argue the slope should be increasing?

Technical Components. HAT includes several technical components, such as automatic speech recognition (ASR), discourse classification models, and LLM-generated summaries of a tutoring session (not discussed here). For ASR, we used the publicly available OpenAI Whisper Medium model to process the audio recording of tutoring sessions into transcripts. Then, we run these transcripts through a RoBERTa model that was first fine-tuned on the TalkMoves dataset of 560+ transcripts from recordings of K-12 math classrooms, then further fine-tuned on a dataset of 94 human-annotated tutoring session transcripts from Saga [22]. The model also trained on pairs of previous student utterances and current tutor utterances for additional context. Using recording-level 10-fold cross-validation, the model achieved an overall macro F1 score of .77 on human-annotated transcripts and .58 on ASR transcripts. We deemed this level of accuracy adequate for the present purposes since the classifications are aggregated across different levels of granularity which increases reliability [23].

HAT Interfaces. Over the course of four months, research team members, Saga coaches, coach supervisors, R&D staff members, and a designer met and co-designed the HAT user interfaces. For more information on how the interfaces were developed and how coaches use them, see [24]. The main HAT interface is an *Overview page* that provides a one-page overview of a tutor's use of discourse moves in a tutoring session (Fig. 3). In addition, two other interfaces (not shown here due to their delayed rollout, see *Procedure* section) were developed: a *Deep Dive* page that includes the session transcript, video recording, and a talk moves annotated video timeline, and a *Trends* page with a filter to select specific talk moves and a scatterplot that helps coaches pick a session to observe and view high-level trends to track change in tutors' use of discourse moves over time.



Or I'm sorry, you did four times two, which is how you get eight here, right?

There are three exponents, but Sebastian's saying we can only combine two of those three exponents.

So if I'm at negative five, how many units am I away from zero?

Salvador, do you agree with Sebastian's light terms?

And what's the coefficient of this term?

Now that we've combined the first two set of like terms, who can tell me what these second two terms combine to?

So if we look at what's inside of our parentheses here, can we actually combine these terms?

What is the absolute value of three minus forty-two?

And Sebastian, did you have the same answer or different?

This four y squared minus seven, are these like terms?

Four, five, what do the numbers mean?

You're right, the parentheses should be first.

So of these terms, which ones have the same variable and the same matching exponents?

We have, what is eight minus negative...

What are these two terms combined?

Fig. 3. HAT *Overview* Page: Top row displays metrics related to talk patterns such as the percent of student-to-tutor talk ratio (*left*), percent of tutor statements with appropriate wait time (*middle*), and percent of student statements > one word (*right*). Bottom row displays metrics related to talk moves such as the percent of content knowledge, rigorous thinking, and learning community (*left*), frequency of occurrence of each learning community moves (*middle*), and those of content knowledge and rigorous thinking moves (*right*). The right panel displays selected examples of each tutor talk move as *Key Moments.* Information icons in each box provide additional information about each metric; these vary by metric and including definitions, suggestions for use in practice, and what research demonstrates about these metrics.

3 METHOD

We partnered with Saga to conduct a study on HAT in an authentic HDT context during the 2023-24 school year.

Participants. Participants were 37 coaches (27% White, 53% Black, 7% Hispanic, 11% Asian, and 2% identifying as two or more races) from four Title I public school districts and their assigned 172 tutors (28% White, 30% Black, 19% Hispanic, 8% Asian, 6% identifying as two or more races, <1% American Indian/Alaska Native, and 9% who did not disclose).

Procedure. Tutors provided remote tutoring to small groups of in-person students during the school day, conducting online sessions through Saga's tutoring platform with approximately 4,300 ninth-grade students. These remote tutoring sessions were recorded with participant consent in accordance with school district policies. The primary purpose of these recordings is to ensure the safety of students and tutors. Additionally, they are used to support non-evaluative efforts to improve the quality of tutoring. The recordings were processed by Saga and anonymized transcripts were processed through HAT to produce tutor discourse analytics for their use during their coaching sessions.

HAT analytics were available to Saga coaches on a rolling basis throughout the school year beginning in December 2023 in one school district and January 2024 in the rest of the school districts (see Fig. 4). Specifically, the main *Overview page* launched in December 2023 or January 2024 (depending on the school district), followed by the *Deep Dive* and *Trends pages* in March 2024. Due to the delayed rollout, these latter pages did not receive as much usage as the main *Overview* (per our HAT clickstream analysis), so we consider the *Overview page* to be the main HAT intervention.

Measures. To track **coach usage** of HAT, we tracked each time a coach logged in and what they clicked on from when HAT was first implemented to the end-of-the-school year (see Fig. 4). We collected over 4,000 clicks and determined the total time each coach spent using HAT from the logs. To measure **coach perceptions of HAT**, we administered an informal, researcher-developed, end-of-year survey to determine how useful the HAT feedback was for coaches over the course of the year and the extent to which they incorporated its feedback into their coaching conversations with tutors. The survey included 14 questions across the following topics *use of the AI application* (e.g., “Did you look at feedback for each of your tutors?”), *perceptions of utility* of the tool (e.g., “Overall, how useful is the feedback for you as a coach?”), and *incorporating the AI tool into coaching* (e.g., “To what extent has the feedback informed how you worked with tutors?”). In our survey analysis, we focus on analyzing a subset of questions related to coaches’ perceptions of HAT utility and how they incorporated its feedback into their coaching. Twenty-nine coaches responded to the survey administered between May and June 2024 (76% response rate).

Lastly, to examine **changes in discourse practices**, we analyzed tutorial session data from HAT’s discourse models during a period of October 2023 to April 2024. This period excludes start-of-school sessions what were focused on relationship building and end-of-school sessions focused on review of content for standardized testing. Initial filtering also excluded sessions that were too short or long and sessions with invalid tutor identifiers (e.g., tutors from districts outside of our study). This resulted in 43,731 sessions from 158 tutors being used for analysis. This number of tutors accounts for tutors who left Saga during the study period. To facilitate longitudinal analysis, session dates were transformed into a numeric month variable relative to the earliest recorded session. Since HAT was rolled out on a different time frame to each school district, a key transformation involved classifying each session as either pre- or post-HAT period, with sessions occurring before coded as 0 and all others as 1. Monthly averages of each tutor talk move (and talk ratio) were computed for each tutor and normalized by number of sessions within a month, which comprised the time series used in the analyses.

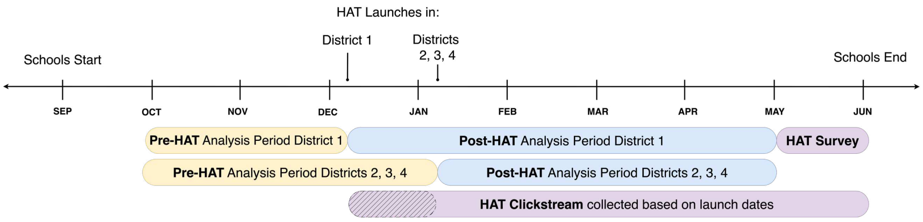

HAT Launches in:
District 1
Districts 2, 3, 4
Schools Start
Schools End
SEP
OCT
NOV
DEC
JAN
FEB
MAR
APR
MAY
JUN
Pre-HAT Analysis Period District 1
Post-HAT Analysis Period District 1
HAT Survey
Pre-HAT Analysis Period Districts 2, 3, 4
Post-HAT Analysis Period Districts 2, 3, 4
HAT Clickstream collected based on launch dates

Fig. 4. Timeline of HAT launched, analysis periods, and survey administration.

4 RESULTS

4.1 Patterns of Usage and Coach Perceptions (RQ1)

Patterns of Usage. All 37 coaches used HAT and accessed its analytics for an average of four of their tutors (coaches support between four to six tutors at a time). Individually, coaches logged in an average of 1.9 times a month (ranging from < 1 to 6 times a month), spent an average of 3 mins and 18 secs each time (ranging from < 1 to 16 mins), and used HAT for a total of 45 mins and 15 secs on average (ranging from less than a minute to a little over 3 hours). This wide range suggests that there was considerable variation in how HAT was used. As expected, the HAT *Overview* page (available from December to June 2024) was used the most (76%) in comparison to the *Deep Dive* (13%) and *Trends* page (11%) (both released in March).

Coach Survey. In our analysis, we focus on select questions to illustrate coaches' perceptions of HAT utility and how they incorporated its feedback into their coaching. Related to perceptions of utility, most coaches found the platform very or extremely useful (72%) and were very or extremely confident they could interpret the information (76%). Related to incorporating HAT into their work with tutors, more than half of the coaches reported that HAT was informing their work with tutors (55%), and that they were using the information to develop coaching goals (59%). Most coaches reported talking about HAT feedback with their tutors (83%) and showing some of its feedback to them (77%). In general, coaches were positive about HAT and made comments such as: "We used the analytics to start conversations about sessions and to push tutors to think critically on their lesson. The [*Overview* page] is a great way for (coaches) to see a snapshot of the lesson" and "It helps me, in a shorter amount of the time, to give specific and directed feedback that is tangible for [tutors] to work on and see growth in."

4.2 Changes in Tutor Discourse Practice (RQ2)

Having demonstrated that coaches used HAT and had generally positive perceptions of it, we turned to testing our main hypothesis that coaches using HAT would result in downstream changes in tutoring practices.

Analytic Approach. To investigate the changes in tutor discourse before we introduced HAT (pre-HAT) and then after (post-HAT), we used a piecewise latent growth curve model (LGCM) [18] (Sec 1. Fig 2). These two phases were modeled with separate intercepts and slopes, which reflect the initial status and rate of change during each phase. The inclusion of random effects and slopes allows us to account for individual variability in tutor behavior. We used the following equation for the model:

$$Y_{it} = \alpha_{i1} + \lambda_{t1}\beta_{i1} + \alpha_{i2} + \lambda_{t2}\beta_{i2} + \epsilon_{it}$$

$$\alpha_{i1}, \beta_{i1}, \alpha_{i2}, \beta_{i2} \sim N(G, \Sigma)$$

Yit is the observed outcome (e.g., a tutor discourse variable) for individual i at time t, $\alpha i1$ is the intercept (initial status) for individual i during the first phase (before HAT or pre-HAT), $\beta i1$ represents the slope (rate of change) for individual i during the pre-HAT phase, $\lambda t1$ are the time loadings for the pre-HAT phase, capturing the time points before HAT, $\alpha i2$ is the intercept for individual i during the post-HAT phase (after HAT implemented), $\beta i2$ represents the slope (rate of change) for individual i during the post-HAT phase, $\lambda t2$ are the time loadings for the post-HAT phase, capturing the HAT period or the period when HAT was implemented, G is a vector of means of each growth parameter, Σ is a covariance matrix of each growth parameter, and ϵit is the residual error term. This equation models the overall trajectory across both phases, with separate parameters for the intercept and slope in each phase, allowing us to compare the rate of change in tutor practices before and after HAT (slope change) and immediate effects of the HAT intervention (intercept change).

We ran separate models for the six talk moves and an additional model for the ratio of tutor-to-student talk (talk ratio) to investigate if any effects might simply be attributable to changes in the amount (vs. nature) of talk. Data were analyzed using the lavaan package in R and missing data were handled with full-information maximum likelihood (FIML), which allows for unbiased parameter estimation under the assumption that data are "missing at random".

Model Effects. The piecewise LGCM results reveal two main patterns (and two sub patterns) in tutors' use of talk moves pre-HAT and post-HAT (see Table 2). **Pattern 1:** Pre-HAT, tutors' overall *talk ratio*, and use of *keep together* and *get to relate* task moves were all flat (as measured by non-significant slopes), indicating stability and consistency in the rate of using these discourse moves, and the overall student-to-tutor talk ratio. Post-HAT, the intercept changes of these discourse practices were not significant, and their slopes also remained flat, indicating that HAT did not have any effect (positive or negative) on changing these tutor practices. Whereas both pre- and post-slopes for *reasoning* were flat, suggesting no increase nor decrease in the use of this talk move over time, there was a significant intercept change suggesting that the introduction of HAT led to an increase in the use of this talk move (Pattern 1a), while the flat post-HAT slope indicates no change in the rate of use of this discourse move. **Pattern 2:** Pre-HAT, tutors' use of *restating*, *pressing for accuracy*, and *revoicing* moves were all decreasing (as measured by significant, negative slopes), indicating a decline in the rate of using these discourse moves prior to HAT. Post-HAT, the slopes of these discourse practices show a significant and positive change, indicating that HAT had a positive effect on the trajectory of use of these moves. Of these three talk moves, there was a significant difference in the intercept for *restating*, suggesting that the intervention had a more immediate effect on this talk move as well (Pattern 2b). Fig. 5 provides a graphical representation of these main patterns in the data.

Table 2. Piecewise growth curve model results pre- and post-HAT

	Intercepts			Slopes		
Outcome	**Pre**	**Post**	**Change**	**Pre**	**Post**	**Change**
Tutor Ratio	0.75[a]	0.76[a]	0.01 (.388)	0.00 (.802)	0.00 (.071)	-0.01 (.248)
Keep Together	17.4[a]	12.9[a]	-4.48 (.283)	1.10 (.165)	0.32 (.359)	-0.78 (.399)
Get to Relate	0.37[a]	0.41[a]	0.04 (.355)	-0.04 (.156)	0.00 (.755)	0.03 (.217)
Reasoning	1.85[a]	2.10[a]	**0.25***	-0.04 (.431)	-0.03 (.558)	0.01 (.899)
Restating	0.51[a]	0.60[a]	**0.09***	**-0.06****	**0.03***	**0.10*****
Accuracy	30.1[a]	32.0[a]	1.73 (.148)	**-1.38***	**0.84***	**2.22****
Revoicing	1.85[a]	2.87[a]	0.12 (.330)	**-0.14 (.027)**	**0.15*****	**0.29*****

Significant at $p < .05$*, $p < .01$**, $p < .001$***; [a]All intercept terms are significant at $p < .001$; non-significant p values in parentheses

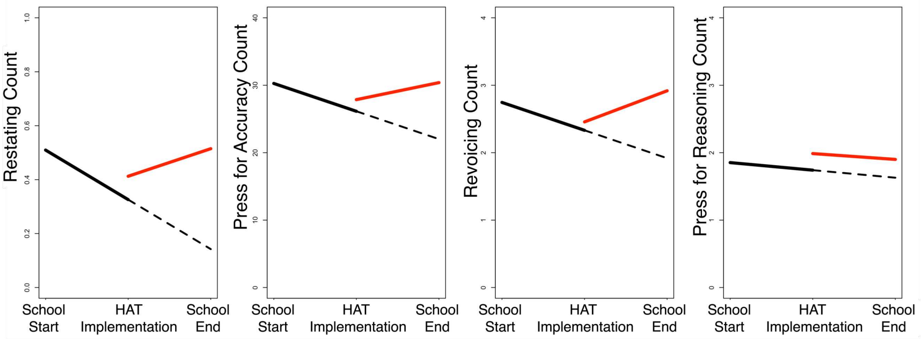


Fig. 5. Significant Changes in Tutor Discourse Pre- and Post-HAT Implementation. From left to right changes in tutor *restating*, *pressing for accuracy*, *revoicing*, and *reasoning*. The first solid line in black (pre-HAT) demonstrates the actual rate of change in discourse moves before the HAT implementation. The dotted line indicates the expected trajectory without HAT. The second solid line in red (post-HAT) demonstrates the actual rate of change in discourse after the HAT implementation.

5 DISCUSSION

We developed an automated, data-driven, AI feedback *tutor discourse analytics* tool embedded within a coached-tutor framework where novice tutors receive ongoing professional learning (PL) from experienced coaches. HAT provided coaches with analytics on their tutors' use of discourse practices, specifically talk moves. We tested part of our theory of change that by providing coaches with automated tutoring analytics in a coached-tutor framework, we can improve tutor practices.

Main Findings and Implications. Analysis of the clickstream data revealed that all coaches used HAT (albeit at varying rates), and survey results indicated that the majority found it useful and integrated its analytics into their coaching sessions with tutors. These results suggest that coaches find value in using an automated tutor discourse feedback tool such as HAT for AI-enhanced coaching. This finding was also confirmed by [24] through a think-aloud study which revealed that coaches have integrated HAT in their workflows either as a central tool that guides their observations and feedback, or as supplementary tool to augment feedback notes they take during a live observation.

We also found that use of HAT led to long-term statistically significant improvements in three of the six measured tutor talk moves: *restating*, *pressing for accuracy*, and *revoicing* (Pattern 2) with a more immediate effect on *restating* (Pattern 2a). We also observed an increase in tutors' use of the *reasoning* talk move after HAT, but no increase in rate of use over time (Pattern 1a). These findings, consistent with those of a similar previous study [17], have broad implications for instructional quality and student outcomes. Specifically, in a study looking at the relationship between talk moves and math quality of instruction (MQI), researchers found that teacher *restating* was significantly positively correlated with several MQI domains (e.g., Overall Richness) and that teacher *revoicing* and *reasoning* were significantly positively correlated with and predictive of overall MQI [25]. Other studies have shown that tutors' use of *revoicing* and *reasoning* can predict student math achievement as measured on practice problems on an Intelligent Tutoring System, and led to students' deeper understanding of complex material and higher-level engagement with other's ideas [26]. Thus, it is highly encouraging that HAT had positive effects on these key discourse variables.

We did not observe significant changes in tutors' use of *keeping everyone together* moves and *getting students to relate to others' ideas* (Pattern 1), which are both *learning community* moves (see Table 1). These findings may be attributed to the inherent challenges with tutoring small groups. In a study where researchers interviewed 37 small-group tutors, they found that all tutors reported interpersonal conflicts between at least two students and that while students were from the same classroom, they had differing needs [27]. Furthermore, prior research suggests that novice teachers often struggle to address students' contributions in group settings when compared to experienced teachers, and therefore focus more on disseminating content knowledge [28]. In fact, in our results, the largest growth occurred in tutors' use of *pressing for accuracy*, a *content knowledge* move (see Table 1). The complex context of small-group tutoring, coupled with the novice nature of Saga's tutors, make attending to the learning community a challenge which may explain these non-significant results. Furthermore, we did not see significant changes in the *tutor-to-student talk ratio* ostensibly because HAT aimed to improve the *quality* of tutor discourse rather than the *quantity* of talk.

Overall, these results have broad implications for maintaining quality while scaling high-dosage tutoring (HDT) programs. Tools like HAT, which provide automated AI feedback, can facilitate the scaling of ongoing instructional coaching, a key strategy for tutor professional learning (PL) in HDT programs [3, 6]. By automating the delivery of data-driven AI feedback, these tools can streamline coaching workflows, making them more efficient and impactful. These measures can enable coaches to support more tutors without compromising the quality of their feedback.

Limitations and Future Work. HAT was a new automated tutoring analytics tool rolled out to Saga coaches. This presented multiple challenges. First, rolling out a new technology to actual users in the real world is a complex activity that in this case, did not align with the needs for an experimental research design. However, the use of piecewise latent growth curve modeling, which examined changes in behavior immediately during and after launching HAT, provides some support for causal effects [29]. Second, technology adoption takes time: time to embed in daily work practices and time to build trust with an AI-enabled system. This consideration would have impacted how coaches used the platform and the extent of its impact on tutor discourse changes. Furthermore, HAT was embedded in Saga's tutoring platform, and some user experience and interface issues with that technology may have limited coaches' use of HAT's discourse analytics. And, because HAT was integrated into broader Saga's practices, the research team could not ensure a high-fidelity implementation of our tool.

In terms of future work, we aim to test the second part of our theory of change (Fig. 1) by demonstrating a relationship between changes in tutor practice supported by use of HAT and student achievement outcomes. We also will investigate whether coaches' use of HAT, both of the interface itself and of its feedback in the coaching sessions, moderates changes in tutors' practices and associated student learning outcomes, thereby connecting all components of our theory of change. Finally, our work to date has focused on discourse analytics for virtual tutoring sessions where novice tutors receive instructional support from an experienced coach. Expanding to in-person tutoring or models with no or limited coaching support will require advancements in speech diarization, automatic speech recognition, and direct-to-tutor feedback tools.

6 CONCLUSION

High-dosage tutoring (HDT) plays an important role in improving student achievement outcomes. While this model has demonstrated success, achieving impact at scale continues to present challenges. Technology, however, provides promising solutions [1]. HAT aims to address the challenge of scaling effective tutor professional learning (PL) by automatically analyzing tutoring sessions and providing tutor coaches with data-driven feedback to guide their coaching. It enables a tutoring model where paraprofessional, novice tutors can receive AI-enhanced PL and support through a coached-tutor framework. Our overall findings indicated that HAT, when used by coaches to provide targeted feedback, positively impacted tutors' discourse practices and mitigated decline in tutor discourse quality over time. Leveraging advanced AI models for tutoring session analysis and feedback has the potential to keep support costs down, broaden access to tutor PL, and maintain consistency and quality in tutoring services at scale.

Acknowledgments. The authors would like to thank Peter Foltz, Scott Dooley, Chelsea Chandler, Chelsea Brown, and Tamara Sumner (CU Boulder) and Devren Hobbs, Denise Parham-Gantt, and Quionna Rivers (Saga Education) for their efforts in making this work possible. This study was funded by Learning Engineering Virtual Institute (LEVI). The opinions expressed in this paper are those of the authors only and do not reflect the views of the LEVI program.

References

1. Nickow, A., Oreopoulos, P., Quan, V.: The promise of tutoring for PreK–12 learning: A systematic review and meta-analysis of the experimental evidence. American Educational Research Journal. 61, 74–107 (2024).
2. Dietrichson, J., Bøg, M., Filges, T., Jørgensen, A.M.: Academic interventions for elementary and middle school students with low socioeconomic status: A systematic review and meta-analysis. Review of Educational Research. 87, 243–282 (2017).
3. Kraft, M.A., Falken, G.T.: A blueprint for scaling tutoring and mentoring across public schools. Aera Open. 7, 23328584211042858 (2021).
4. Aleven, V., Popescu, O., Koedinger, K.: Towards tutorial dialog to support self-explanation: Adding natural language understanding to a cognitive tutor. Artificial Intelligence in Education (2001).
5. Kraft, M.A., Blazar, D., Hogan, D.: The effect of teacher coaching on instruction and achievement: A meta-analysis of the causal evidence. Review of Educational Research. 88, 547–588 (2018).
6. Paraprofessionals as High-Impact Tutors: Opportunity and Guidance | National Student Support Accelerator, https://studentsupportaccelerator.org/briefs/paraprofessionals-high-impact-tutors-opportunity-and-guidance, last accessed 2025/02/10.
7. Knight, D.S.: Assessing the cost of instructional coaching. Journal of Education Finance. 52–80 (2012).
8. Wang, D., Tao, Y., Chen, G.: Artificial intelligence in classroom discourse: A systematic review of the past decade. International Journal of Educational Research. 123, 102275 (2024).
9. Böheim, R., Schnitzler, K., Gröschner, A., Weil, M., Knogler, M., Schindler, A.K., Alles, M., Seidel, T.: How changes in teachers' dialogic discourse practice relate to changes in students' activation, motivation and cognitive engagement. Learning, Culture and Social Interaction. 28, 100450 (2021).
10. Jacobs, J.., Suresh, A., Booth, B., Sumner, T., Bush, J., Brown, C., D'Mello, S.: Automating Feedback from Recorded Instructional Observations: Using AI to Detect and Support Dialogic Teaching. In: Kelly, S. (ed.) Research Handbook on Classroom Observation. Edward Elgar Publishing.
11. Demszky, D., Liu, J., Hill, H.C., Jurafsky, D., Piech, C.: Can automated feedback improve teachers' uptake of student ideas? Evidence from a randomized controlled trial in a large-scale online course. Educational Evaluation and Policy Analysis. (2023).
12. Kelly, S., Guner, G., Hunkins, N., D'Mello, S.K.: High school english teachers reflect on their talk: A atudy of response to automated feedback with the teacher talk tool. International Journal of Artificial Intelligence in Education. 1–35 (2024).
13. Jacobs, J., Scornavacco, K., Harty, C., Suresh, A., Lai, V., Sumner, T.: Promoting rich discussions in mathematics classrooms: Using personalized, automated feedback to support reflection and instructional change. Teaching and Teacher Education. 112, 103631 (2022).

14. Demszky, D., Wang, R., Geraghty, S., Yu, C.: Does feedback on talk time increase student engagement? Evidence from a randomized controlled trial on a math tutoring Platform. In: Proceedings of the 14th Learning Analytics and Knowledge Conference. pp. 632–644. ACM, Kyoto Japan (2024).
15. Wang, R.E., Ribeiro, A.T., Robinson, C.D., Loeb, S., Demszky, D.: Tutor CoPilot: A human-AI approach for scaling real-time expertise, http://arxiv.org/abs/2410.03017, (2025).
16. Michaels, S., O'Connor, C.: Conceptualizing talk moves as tools: Professional development approaches for academically productive discussion. Socializing Intelligence Through Talk and Dialogue. 347–362 (2015).
17. Booth, B., Jacobs, J., Bush, J., Milne, B., Fischaber, T., D'Mello, S.K.: Human-tutor coaching technology (HTCT): Automated discourse analytics in a coached tutoring model. Learning Analytics and Knowledge Conference. pp. 725–735 (2024).
18. Kohli, N., Harring, J.R.: Modeling growth in latent variables using a piecewise function. Multivariate Behavioral Research. 48, 370–397 (2013).
19. Harring, J.R., Strazzeri, M.M., Blozis, S.A.: Piecewise latent growth models: Beyond modeling linear-linear processes. Behav Res. 53, 593–608 (2021).
20. Wolf, M.K., Crosson, A.C., Resnick, L.B.: Accountable talk in reading comprehension instruction (2005).
21. Michaels, S., O'Connor, C., Resnick, L.B.: Deliberative discourse idealized and realized: Accountable talk in the classroom and in civic life. Stud Philos Educ. 27, 283–297 (2008).
22. Cao, J., Suresh, A., Jacobs, J., Clevenger, C., Howard, A., Brown, C., Milne, B., Fischaber, T., Sumner, T., Martin, J.H.: Enhancing talk moves analysis in mathematics tutoring through classroom teaching discourse, http://arxiv.org/abs/2412.13395, (2024).
23. Rushton, J.P., Brainerd, C.J., Pressley, M.: Behavioral development and construct validity: The principle of aggregation. Psychological Bulletin. 94, 18–38 (1983).
24. Sawaya, S., D'Mello, S.K.: Sense-Making with an AI-enhanced coaching tool: A think-aloud study (accepted). Artificial Intelligence in Education (2025).
25. Jacobs, J., Thomas, K., Engel, M., Bush, J.: The relationship between academically productive talk and instructional quality in mathematics lessons. Paper presented to the American Educational Research Association (2025).
26. Abdelshiheed, M., Jacobs, J., D'Mello, S.K.: Aligning tutor discourse supporting rigorous thinking with tutee content mastery for predicting math achievement. Artificial Intelligence in Education. pp. 150–164 (2024).
27. MacDonald, R.B.: Group tutoring techniques: From research to practice. Journal of Developmental Education. 17, 12–16 (1993).
28. Stahnke, R., Blömeke, S.: Novice and expert teachers' situation-specific skills regarding classroom management: What do they perceive, interpret and suggest? Teaching and Teacher Education. 98, 103243 (2021).
29. Pakpahan, E., Hoffmann, R., Kröger, H.: Statistical methods for causal analysis in life course research: an illustration of a cross-lagged structural equation model, a latent growth model, and an autoregressive latent trajectories model. International Journal of Social Research Methodology. 20, 1–19 (2017).