Learning Behaviors Mediate the Effect of AI-powered Support for Metacognitive Calibration on Learning Outcomes

HaeJin Lee
haejin2@illinois.edu
School of Information Sciences, University of Illinois Urbana–Champaign
Champaign, IL, USA

Frank Stinar
fstinar2@illinois.edu
School of Information Sciences, University of Illinois Urbana–Champaign
Champaign, IL, USA

Ruohan Zong
rzong2@illinois.edu
School of Information Sciences, University of Illinois Urbana–Champaign
Champaign, IL, USA

Hannah Valdiviejas
hannahcohn1011@gmail.com
Society for Research in Child Development
Washington, District of Columbia
USA

Dong Wang
dwang24@illinois.edu
School of Information Sciences, University of Illinois Urbana–Champaign
Champaign, IL, USA

Nigel Bosch
pnb@illinois.edu
School of Information Sciences and Department of Educational Psychology, University of Illinois Urbana–Champaign
Champaign, IL, USA

ABSTRACT

Students struggle with accurately assessing their own performance, especially given little training to do so. We propose an AI-powered training tool to help students improve "metacognitive calibration," or the ability to accurately predict their own learning, potentially enhancing learning outcomes by enabling students' use of metacognition-informed learning behaviors. We present results from a randomized controlled trial (N = 133) assessing the effectiveness of the tool in a college-level computer-based learning environment. The AI-driven tool significantly improved learning gains compared to the control group by 8.9% (t = -2.384, p = .019), and this effect was significantly mediated by learning behaviors. Overconfident students who received the intervention showed significantly greater metacognitive calibration improvement than the control group by 4.1% (t = 2.001, p = .049). These insights highlight the value of AI-powered metacognitive calibration training and the importance of promoting specific metacognition-informed learning behaviors in computer-based learning.

CCS CONCEPTS

• **Human-centered computing** → **Human computer interaction (HCI)**; **Empirical studies in HCI**; • **Applied computing** → **E-learning**.

KEYWORDS

Explainable AI, Human-computer Interaction, Self-regulated Learning, Metacognitive Calibration, Computer-based Learning Environments

CHI '25, April 26-May 1, 2025, Yokohama, Japan

ACM ISBN 979-8-4007-1394-1/25/04
https://doi.org/10.1145/3706598.3713960

ACM Reference Format:
HaeJin Lee, Frank Stinar, Ruohan Zong, Hannah Valdiviejas, Dong Wang, and Nigel Bosch. 2025. Learning Behaviors Mediate the Effect of AI-powered Support for Metacognitive Calibration on Learning Outcomes. In *CHI Conference on Human Factors in Computing Systems (CHI '25), April 26-May 1, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3706598.3713960

1 INTRODUCTION

Metacognition, commonly referred to as thinking about one's thinking, is a high-order thinking skill that includes crucial cognitive activities such as planning, monitoring, reflecting, and evaluating one's learning strategies and knowledge [27, 35, 54]. Among the various metacognitive skills, metacognitive calibration—which refers to the accuracy of students' evaluation of their own knowledge or learning—is particularly crucial [3, 70, 99]. Metacognitive calibration enables students to allocate their time and effort more effectively because students can accurately judge how well they know a topic, allowing students to focus on learning challenging material while spending less time on topics they have already mastered [20]. Such strategic allocation of time and study resources likely contributes to the positive correlation observed between metacognitive calibration and learning outcomes [47, 53, 98]. However, students sometimes struggle to make accurate evaluations of their performance [8, 17]. When students make inaccurate judgments of their knowledge, it can lead to ineffective learning strategies and hinder academic progress and outcomes [49]. Students' challenges in accurately assessing their understanding are particularly pronounced in online learning environments, where limited external feedback may exacerbate the problem [11, 21]. For instance, students in computer-based learning environments may miss out on contextual cues and real-time interactions with instructors and peers, which could be crucial for evaluating their comprehension.

Although it is common for students to struggle to accurately assess their own knowledge, this metacognitive *mis*calibration is typically not an inadequacy on the part of students but rather a consequence of inadequate support and practice. With proper guidance

and assistance, students can enhance their metacognitive calibration abilities [1, 92]. Therefore, numerous interventions have been developed to support metacognitive calibration in both traditional and computer-based learning environments [8, 36, 40, 58, 63, 92]. In these metacognitive calibration interventions, a common approach employed is to provide feedback to students after their assessments. This feedback focuses on the discrepancy between how students perceive their performance and their actual results, intending to make students more aware of any inaccuracies in their self-evaluations [36, 40]. While such interventions can improve students' metacognitive calibration, these interventions are limited by the fact that feedback is only provided *after* the learning process, once students' performance scores are available. Feedback provided only after assessments may miss crucial opportunities to correct students' miscalibration and influence students' learning strategies and behaviors earlier, before it is too late to adapt for the assessment in question.

This limitation in the timing of interventions raises the central questions of our paper: What if we train students' metacognitive calibration during the learning process, rather than after? Could this earlier intervention lead to improved learning outcomes? If so, what underlying mechanisms would explain these improvements? For example, could these improvements be partly attributable to behavioral changes, as we propose? Or perhaps unobserved cognitive factors, such as increased attention to specific parts of a lecture based on improved calibration? Answering these questions is crucial for human–computer interaction (HCI) researchers and practitioners. Investigating whether early, real-time AI-driven metacognitive support enhances learning outcomes or calibration accuracy could provide valuable insights into the effectiveness of early interventions in supporting metacognitive skills and improving overall student learning. Additionally, if early intervention improves learning outcomes, understanding the causal mechanisms behind these effects could inform the design of learning software to target specific factors (e.g., learning behaviors relevant to self-regulated learning) that enhance student learning outcomes. By prioritizing these factors in the development of learning interfaces, researchers can develop proactive learning systems that provide timely and effective support, ultimately improving both student metacognitive calibration and learning outcomes.

In HCI research, various approaches have been proposed to measure and support facets of student cognitive behaviors [9, 13, 79, 96]. For instance, Cabales [13] developed Muse, a chatbot designed to foster student metacognitive reflection, and Wang et al. [96] introduced MindDot, a concept map-based learning system aimed at promoting the use of comparative strategies. Although these tools have proven effective in enhancing student metacognition, there remains a lack of empirical studies on developing tools specifically tailored to support metacognitive calibration. Since metacognition and self-regulated learning are difficult to observe and measure directly [5, 101], studies often miss the opportunities to closely observe behavioral changes resulting from interventions, making it difficult to understand the mechanisms linking the intervention to learning outcomes. Therefore, tools designed to support student metacognitive calibration typically operate on the theoretical premise that improving calibration will lead to better academic outcomes, but with few empirical specifics to inform the design of educational software. This challenge is evident in specific domains, such as computing education. As Prather et al. [74] highlighted, although researchers are interested in measuring and supporting student metacognition and self-regulated learning in programming [15, 55, 89], there is a lack of research grounding the measurements of self-regulated learning and metacognition in established literature, which limits the availability of empirical evidence necessary to develop effective learning software.

Motivated by these limitations, we develop a novel artificial intelligence (AI)-driven metacognitive calibration training tool that provides early, real-time feedback using the machine learning model-predicted end-of-learning performance scores to correct students' potential early miscalibration. We conducted a randomized controlled trial (N = 133) to examine whether the early AI-driven intervention leads to enhanced learning outcomes and metacognitive calibration improvement, including across various potential moderating factors such as race/ethnicity and gender. We measure metacognitive calibration based on students' performance estimations (i.e., predicted test grades), as this provides a direct and widely used approach in education to assess calibration relative to actual scores [32, 69]. While other facets, such as confidence and peer-relative placement [57], may capture additional aspects of calibration, our focus on performance estimations allows us to assess calibration based on the alignment between students' estimated test outcomes and objectively measured results, offering quantifiable measure that supports the study's objective. Further, we examined student learning behaviors to determine whether the intervention leads to different learning strategies, which in turn can explain the enhanced learning outcome, if observed. We present mediation analysis results that uncover the mechanisms underlying the effectiveness of AI-driven interventions on student learning outcomes, using two mediators that capture different categories of learning behavior.

The following are the research questions we aim to answer:

- **RQ1. Do students who receive an AI-powered metacognition-supporting intervention improve their metacognitive calibration more than their peers who do not receive the intervention?**
  Hypothesis: We anticipate that students receiving the intervention will improve metacognitive calibration more than the control group, consistent with previous studies [80, 92], as the real-time, AI-predicted post-learning test score will help correct miscalibration. However, we expect improvements to vary across subgroups based on confidence levels, gender, and race. We expect that both overconfident and underconfident students will show similar improvements, as the intervention will help both groups adjust their overestimation or underestimation of their knowledge by identifying gaps. However, we expect male students to show greater improvement than females, based on prior studies [37], and similar improvements among racial groups, as research has shown mixed results [61].

- **RQ2. Do students receiving an AI-powered metacognition-supporting intervention demonstrate better learning outcomes than the control group?**
  Hypothesis: We hypothesize that students receiving the AI-powered intervention will achieve greater learning gains than

those in the control group. The intervention is expected to correct miscalibration, allowing students to adjust their learning strategies and behaviors, ultimately leading to improved academic performance.

- **RQ3. Do students' learning behaviors mediate the impact of an AI-powered metacognitive calibration support tool on learning gains?**
  Hypothesis: We hypothesize that students' learning behaviors (i.e., what types of studying activities they choose and in what order) will mediate the effect of the intervention on learning gains. Specifically, we expect the intervention to lead to behaviors informed by improved metacognitive calibration, which will subsequently result in improved learning outcomes. By enabling students to actively manage and adjust their learning strategies, the intervention is expected to indirectly enhance learning gains through changes in learning behaviors.

We anticipate that our work will contribute the following:

(1) We developed an AI-driven intervention that provides early, real-time metacognitive calibration support by using AI-predicted students' end-of-learning scores, aiming to correct any early miscalibration students might have.

(2) Our study is the first to investigate the mechanisms of *how* and *why* an early AI-powered metacognitive support tool enhances student learning gains, specifically by using student learning behaviors as mediators.

(3) We advocate for a shift toward early interventions that take into account the nuances of student confidence levels, whether they are more prone to overconfidence or underconfidence, while encouraging engagement in targeted learning behaviors.

2 RELATED WORK

2.1 Metacognitive Calibration and Self-regulated Learning in Computer-based Learning Environments

Metacognition involves understanding and regulating one's cognitive processes, allowing students to continuously evaluate their knowledge and select effective learning strategies for different tasks accordingly [19, 27, 35]. Given the complex nature of metacognition, various theoretical models have been proposed to conceptualize and formalize different facets of metacognition, each offering a different perspective [12, 30, 35, 72, 81]. For instance, Efklides [30] highlights three key components of metacognition: metacognitive knowledge, experience, and skills. In Efklides's model, metacognitive knowledge refers to an individual's understanding of personal, task, and strategy-related factors, including self-awareness. Metacognitive experiences, on the other hand, manifest through real-time judgments, estimates, feelings, and task-specific knowledge [29]. Therefore, when students estimate their own understanding or performance (i.e., metacognitive judgments), they are engaging in metacognitive experiences. Moreover, this metacognitive judgment measured based on test performance can be further distinguished by their grain size [42]. For instance, global judgments relate to overall test performance (at the test level), while local judgments focus on each individual question or item within the test.

Metacognitive skills involve the deliberate application of strategies to regulate cognitive processes. These skills include a variety of strategies, such as planning, monitoring, regulating cognitive activities, and evaluating the outcomes of task execution. Metacognitive calibration refers to the degree to which students' self-assessment of their understanding aligns with external measures, such as performance scores [87]. When students are well-calibrated, their self-assessments closely match their actual performance, allowing them to make informed decisions about study strategies and areas needing improvement. While the exact categorization of metacognitive calibration within the scope of metacognition, and consequently its measurement, has not yet reached consensus [2], researchers agree that calibration is a key skill of metacognitive monitoring [95], which refers to an individual's awareness of their cognitive processes, such as comprehension or task performance [106].

Metacognitive calibration has commonly been measured by calculating the difference between a student's perceived and actual performance on an assessment [32, 69]. Such measurement has enabled the identification of notable differences in metacognitive calibration between low- and high-achieving students [68]. While low-achievers are more prone to overconfidence—an issue related to but distinct from overestimation—high-achieving students typically demonstrate better calibration [41, 50]. However, high-achieving students may sometimes show underconfidence [38]. Underconfident students may waste effort over-studying material they already understand, while overconfident students risk neglecting areas that need more attention, potentially leading to poorer outcomes. This contrast between overconfidence in low-achievers and underconfidence in high-achievers emphasizes the complexity of supporting metacognitive calibration, highlighting the need for tailored interventions to effectively address these distinct challenges.

Moreover, positive associations are further demonstrated between metacognition and academic performance, as well as between metacognition and learning strategies [1, 43, 73, 97]. Sun et al. [88] identified positive relationships between students' metacognitive experience and academic performance for English as a foreign language students. Particularly in relation to metacognitive calibration, Zhou [108] found a positive association between university students' metacognitive calibration and their online information search performance scores. Zhao and Ye [106] explored the effect of metacognitive calibration accuracy in computer-based learning environments, and they found that learners with better calibration performed better on both exams and assignments. The significance of metacognition is also well-established in self-regulated learning (SRL) research, where metacognitive strategies (e.g., goal setting, self-reflection, and self-evaluation) are a central component of many SRL models [31, 66, 71, 100, 109, 110].

While it is known that both metacognitive calibration and the use of SRL strategies are associated with improved academic performance [44, 45, 103], the specific mechanisms through which these elements interact and lead to enhanced learning outcomes remain unclear. Although students with strong SRL skills achieve better learning outcomes [25, 94], it is still unknown whether the effectiveness of interventions aimed at improving metacognitive calibration directly translates into more effective use of SRL strategies and,

consequently, better academic performance. Understanding the underlying mechanisms, particularly the role of behavioral changes, is crucial for developing interventions that not only enhance students' calibration but also foster more meaningful engagement with SRL strategies. Our study seeks to fill this gap by investigating whether early interventions targeting metacognitive calibration influence students' engagement with SRL strategies and, if so, how these changes contribute to improved learning outcomes.

2.2 Metacognitive Calibration Supporting Tools

Given that students often struggle to accurately assess their own performance, numerous tools have been developed to enhance students' metacognitive calibration skills [8, 36, 92]. Studies employed numerous approaches to developing these tools. One common approach is presenting student-facing dashboards, which often provide students with performance information [36, 40, 65, 80, 92], usually after the assessments. For instance, Foster et al. [36] observed that students' calibration accuracy did not improve after receiving feedback on their perceived and actual test results. Saenz et al. [80] conducted a study comparing the effectiveness of five different interventions to improve students' calibration skills. These five approaches included: 1) salient feedback, which provided clear information on both performance and prediction accuracy, 2) a review session where students revisited test questions and their performance grade predictions, 3) an incentive-based approach where students had the opportunity to earn $50 for accurate predictions, 4) a motivational lecture focused on the importance of personal motivations and the role of academic information in making accurate predictions, and 5) reflective practices that required students to reflect on their predictions over an extended period without receiving any feedback. The results revealed that only the motivational lecture and salient feedback enhanced students' prediction accuracy. On the other hand, Callender et al. [14] found that both incentives and feedback had a positive impact on improving student metacognitive calibration, while Miller [56] discovered that feedback did not lead to a significant improvement in calibration performance, except for lower-performing students.

Urban and Urban [92] found that combining SRL training, peer evaluation, and calibration feedback led to improvements in students' calibration abilities. However, Emory and Luo [32] found no difference between the control and intervention group who received metacognitive monitoring interventions in terms of the relative and absolute metacognitive calibration accuracy of community college students in computer-based learning environments. This result highlights the need for more focused research on the effectiveness of calibration training tools in online settings, particularly for diverse student populations such as those in community colleges. In sum, while some approaches, such as salient feedback and motivational lectures, show promise, inconsistent results regarding the effectiveness of these tools underscore that calibration is difficult to change. These findings further highlight the need for more targeted interventions designed to proactively teach calibration skills and explore the potential impact of providing feedback before, rather than after, student assessments.

Although studies generally show that post-feedback metacognitive support positively impacts student learning, many existing tools, such as dashboards, have a significant limitation: these tools often provide feedback (e.g., performance discrepancies) without offering guidance on how to interpret and act upon that feedback. Providing feedback alone approach leaves students to navigate the information on their own, often failing to encourage deeper reflection or the examination of strategies that may have contributed to their performance gaps. While these tools may indirectly improve student metacognitive calibration, Dunlosky and Thiede [28] further emphasizes the limitations of such tools, noting that simply providing accurate performance feedback does not necessarily lead to deeper cognitive processing or improvements in learning strategies. Moreover, as students continue studying, their judgments of their own performance are likely to change, highlighting the limitation of these tools in failing to provide real-time interventions that could offer timely support.

This need for early, real-time proactive intervention is especially critical in computer-based learning environments, where students are often required to exercise greater levels of metacognitive skills and independence, placing considerable demands on their metacognitive abilities [51]. As Tankelevitch et al. [90] argue, AI-driven systems can impose high metacognitive demands on users. Such challenges highlight the need for more targeted research into the development of metacognitive calibration tools that can effectively enhance metacognitive skills, especially within online learning contexts [4]. AI-driven approaches present a promising avenue for addressing the limitations of current metacognitive support tools, particularly by focusing on early-stage metacognitive calibration.

HCI research has explored supporting metacognition and self-regulated learning across various domains [13, 26, 79, 96]. For instance, Desai and Chin [26] developed Health Buddy, a voice agent to support self-regulated learning, while Reza and Yoon [79] developed the Computer-Assisted Shadowing Trainer, to support self-regulation in foreign language listening practice. Moreover, numerous studies have leveraged AI-driven approaches to support learners in online education. Recent work has particularly explored the potential of AI, especially large language models, to assist learners in various contexts, such as coding [16, 59, 104], inspiring motivation [18], and teaching mathematical language [105].

Despite the potential of AI, there have been relatively few studies leveraging AI techniques to design tools specifically for supporting student metacognitive calibration. By using machine learning to predict students' real-time performance based on their trace data—which are the digital footprints students leave as they interact with learning interfaces, including clicks, submission, or time spent on tasks—AI-driven feedback can deliver more precise and timely feedback. This approach may addresses a key limitation of traditional tools that rely on post-assessment feedback and miss out on capturing the nuances of students' interactions with the online learning platforms. In our study, we developed an AI-driven intervention that offers early, real-time metacognitive calibration support, providing immediate feedback to teach students to improve calibration and enhance student learning outcomes.

3 RESEARCH CONTEXT AND DATA

3.1 Self-guided Online Learning System

We developed a self-guided online learning system that allowed students to study introductory statistics at their own pace. The learning system had four distinct subtopics, each visually represented by icons shown in Figure 1. Each subtopic-specific module contained four different learning activities: a reading, a quiz, a set of worked examples, and a summary. Each learning activity was designed with a specific learning objective and students had the freedom to choose the order of the learning activities, regardless of the subtopic, and to revisit activities as much as desired. The readings, which consisted of four to six pages, provided comprehensive information on the subject matter. The quizzes, consisting of approximately 10 questions each, allowed students to assess their understanding of the material. After completing a quiz, students were informed whether their answers were correct or incorrect, but correct answers for the incorrect responses were not revealed, encouraging self-guided learning. The examples not only included the correct answers but also walked students through the proper methods to approach and solve the problems. The summaries offered a brief recap of each module's key concepts, enabling students to quickly review the material for each subtopic. While students were not required to complete all the subtopics during the learning session, the system allowed them to revisit and repeat any activity as desired.

Participants began the study by completing a demographics survey and taking a pretest regarding the introductory statistics content which was covered in the learning session. The pretest was designed to assess students' prior knowledge of the material, and after completing the pretest, they were asked to estimate their performance without seeing their actual scores. We used the students' estimated and actual pretest scores to evaluate their initial metacognitive calibration, which is explained in detail in Section 4.2. Afterward, students engaged in a 60-minute self-paced learning session, with a timer that displayed the remaining time only during active interaction with the software, helping to maintain their focus. The study, lasting approximately 90 minutes, required students to interact with the system to learn foundational concepts in statistics. Participants took a posttest and were once again asked to estimate their scores. Similarly, we used students' estimated and actual posttest scores to measure their posttest metacognitive calibration.

3.2 Data Collection

The study included 134 college students recruited from the online platform Prolific [67]. One participant was excluded for not completing the introductory survey, resulting in a final sample of 133 college students from North America. The study was approved by institutional review board (IRB), and each student consented to the data being collected. Data were collected in January 2024. Participants reported demographics in open-ended text boxes, which we provide to contextualize the research and enable comparisons in future research (e.g., meta-analytic measurement of effect heterogeneity). For gender, 48.1% of participants identified as male, 46.6% as female, and 5.3% as non-binary. Regarding race, 33.8% of participants identified as White, 25.6% as Black, 18.0% as multiracial, 12.8% as Asian, 9.0% as Hispanic/Latinx, and one additional group not specified due to identifiability concerns with small group size. Each student was compensated $20 USD on completion.

The students were randomly assigned to either the control (n = 53) or experiment (n = 80) conditions by employing unequal randomization probabilities, which is a common method in randomized controlled study when there is a need to prioritize detailed analysis of the intervention group [91]. Therefore, we intentionally assigned more students to the intervention group to gather additional data about the intervention (e.g., usability perceptions) that we intend to analyze for future work on intervention improvements. Students in the control condition completed the online learning session as outlined above. Students in the experiment condition received interventions along with AI-predicted posttest grade three times (at 15, 30, and 45 minutes) during the hour-long learning session. We provide further elaboration on the AI-driven interventions students received in the following section.

3.3 AI-driven Metacognitive Calibration Intervention

For the experimental group, we designed an AI-powered intervention to train students' metacognitive calibration. The intervention provided students with an AI-predicted posttest grade and guidance on interpreting and using this prediction grade to improve their metacognitive calibration. The interventions encouraged students to reflect on their performance with prompts such as, "Take a moment to reflect on our prediction. How does it compare to the score you predicted for yourself?" These reflective prompts encouraged students to consider potential misalignments between their own estimations and the AI-predicted scores. Additionally, students were encouraged to evaluate their knowledge and focus on sections where they felt they were struggling. We leveraged an AI model to automatically predict students' potential learning outcomes based on both performance and interaction data collected during the learning session: 1) Performance data: This included a pretest score and quiz scores for each of the four subtopics, which served as effective indicators of students' comprehension of content for different topics during the learning session. 2) Interaction data: These data comprised the time spent on each topic and the number of times each topic was accessed, providing quantitative measures of the time and effort a student spent on each topic in the learning session.

To predict learning outcomes based on the student performance and interaction data, we employed a random forest model with an attention mechanism [93]. Training data were taken from a previous unrelated study in the same learning platform, which did not have AI-powered interventions. We used the random forest model [10] because of its effectiveness in handling tabular student data and its computational efficiency, allowing it to produce real-time predictions during interventions with minimal disruption to students' learning. The attention mechanism was included due to its success in various AI applications through learning to assign adaptive importance to different input features. This adaptability is crucial since students may demonstrate different behaviors during the learning process, leading to different factors contributing to their final outcomes. For instance, one student may achieve high

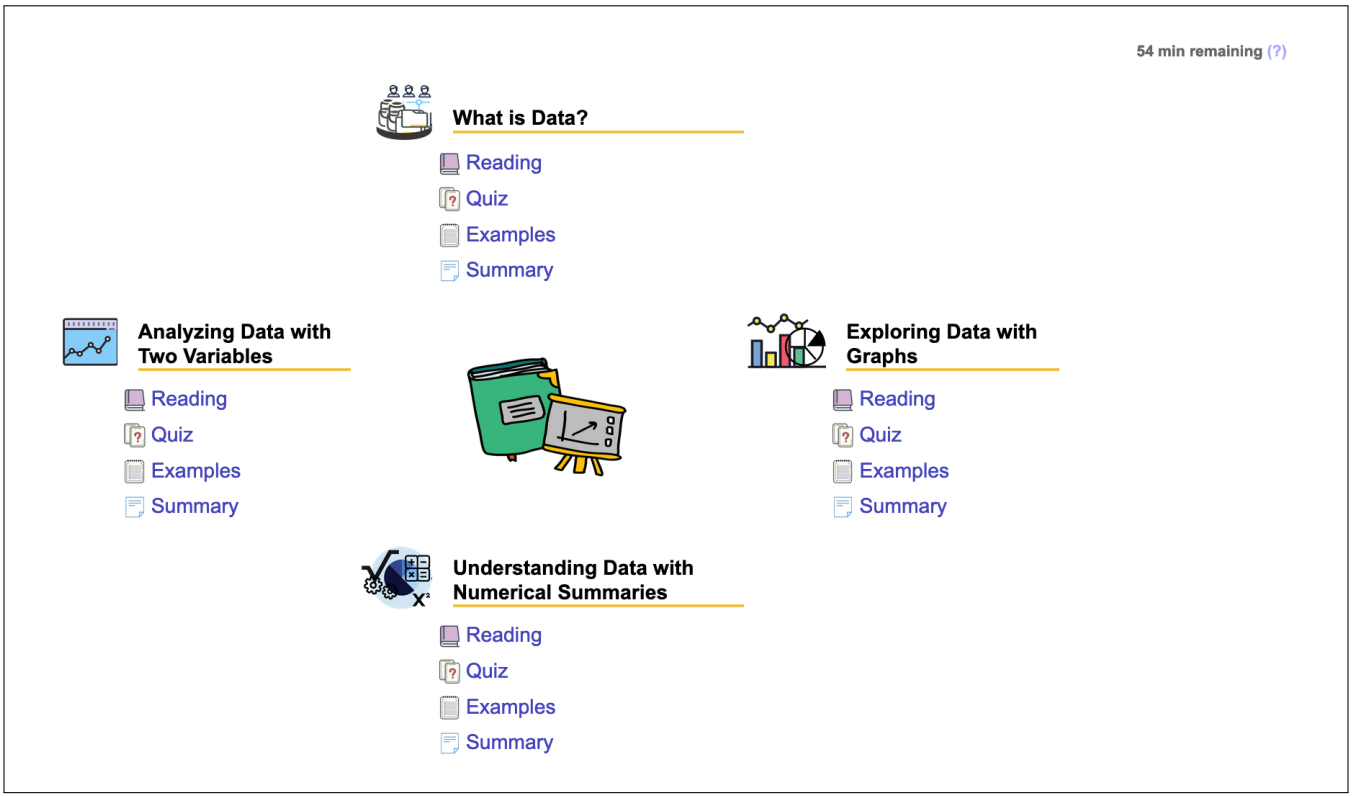


Figure 1: Screenshot of the main page menu of the self-guided online learning system

final grades by spending more time on challenging topics, while another may do so by keeping an effective learning pace. The attention mechanism captures these adaptive differences in feature importance, allowing our AI model to predict learning outcomes for each student more effectively.

The intervention presented students with the AI-predicted posttest score, along with corresponding guidance, a visualization of students' performance and interaction data compared to the average data of all students, and indications of how important each type of interaction data was for prediction as measured by the attention weights learned by our random forest model (example in Figure 2). To represent these weights, we used variations in color darkness in our input data graph: darker colors indicate higher importance weights, while lighter colors denote lower weights. We also randomly assigned students in the experiment condition to receive slight variations on this intervention, one without attention weights visualized by color darkness and one with no graph. However, the core functionality of the intervention remained the same, including the AI-predicted learning outcome. We grouped these three experiment condition variations together for analysis, given that they share the same purpose and core functionality, and plan to explore potential differences in usability and user perceptions in future work with larger samples that will afford statistical power for comparisons between variations.

4 MEASURES

4.1 Student Learning Behaviors

To assess students' learning behaviors, we used SRL-related learning measures established in another study, where trace data from the computer-based learning environment were analyzed [52]. Specifically, SRL-related strategies were conceptualized as sequences of learning activities—Read, Quiz, Example, and Summary—each representing the student's engagement in that learning activity. Using constrained sequential pattern discovery [46, 64, 102, 107], six frequent learning sequences were identified: Read → Quiz, Quiz → Read, Quiz → Quiz, Quiz → Example, Read → Example, and Quiz → Summary. For this study, we categorized the six frequent learning patterns into two broader categories of learning patterns: *seeking knowledge* and *seeking assessment* (Table 1). This grouping was based on the similarity of the learning strategies that each frequent learning strategy was associated with. Frequent learning patterns such as Quiz → Read, Quiz → Example, Read → Example, and Quiz → Summary are categorized under the seeking knowledge learning pattern category. For instance, the patterns Quiz → Read, Quiz → Example, and Quiz → Summary suggest that after taking a quiz, students attempt to review reading material, examples, or summaries. Such behaviors potentially indicate that students identified knowledge gaps through the quiz and sought additional information to fill those gaps or to further consolidate their understanding. Detailed descriptions of how each frequent

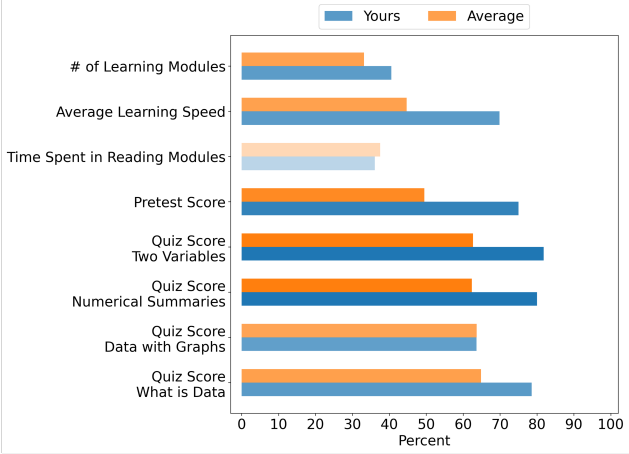
While you have been learning, we predicted what your posttest score might be based on your studying habits we noticed.

We acknowledge that our prediction may not be perfect, especially because our prediction is based on only what we have seen about your studying habits in this short session.

Our prediction is that you will score 72% on your posttest.

Take a moment to reflect on our prediction. How does it compare to the score you predicted about yourself?

The graph below shows your learning behaviors compared to others who have used this statistics learning software.



The darker bars shows that our artificial intelligence algorithm believes that behavior impacts positive performance posttest. The artificial intelligence algorithm believes the lighter bars are less important.

To improve your prediction by having it be a number closer to ours, you may consider evaluating your knowledge by reviewing quizzes and the summaries of the section you think you are struggling with.

We will show you a new prediction we made for your posttest score in 15 minutes.

Continue

Figure 2: Screenshot of the experiment condition intervention with an attention-weighted visualization of interaction and performance data alongside the AI-predicted posttest score. The attention weights are represented through variations in color darkness, with darker colors indicating higher importance.

learning pattern fits into the categories of learning patterns are provided in Table 1.

Seeking assessment reflects students' use of learning strategies where students evaluate their comprehension through assessments (e.g., taking a quiz) to evaluate how much of the material students have objectively understood. This category of learning patterns includes frequent learning patterns like Read → Quiz and Quiz → Quiz. These learning patterns, such as taking a quiz after reading or completing multiple quizzes, suggest that students are actively assessing their knowledge to identify areas needing improvement [82]. Using the two categories of learning patterns, we counted the frequency of corresponding learning patterns by analyzing students' trace data. We transformed the trace data into a comprehensive list of learning activities, capturing the entire sequence of actions students engaged in during the learning session. We then used this list to tally the occurrences of frequent learning patterns. To measure the seeking knowledge strategy, we aggregated the counts for Quiz → Read, Quiz → Example, Read → Example, and Quiz → Summary sequences. Read → Quiz and Quiz → Quiz sequences were used to assess the seeking assessment strategy.

4.2 Metacognitive Calibration

We asked students to estimate their pretest and posttest scores as percentages (i.e., "What do you think your grade will be on the test you just took? (0%–100%)") after taking their pretest and posttest. These estimations were used to assess students' initial and final levels of metacognitive calibration by comparing students' perceived pretest and posttest performance to actual scores.

Table 1: Categories of student learning patterns and their corresponding frequent patterns are presented. The "Associated Learning Strategies" column outlines the potential use of SRL-related strategies for each frequent pattern, as supported by the literature. The "Description" column provides further detail on how each frequent learning pattern fits within the two broader categories of learning patterns.

Categories of learning patterns	Frequent learning patterns	Associated learning strategies	Description
Seeking knowledge	Quiz→ Read	• *Keeping records and monitoring* [111] • *Seeking information* [111] • *Search* [83]	Students identify gaps in their knowledge from a quiz and respond by seeking additional information by reading, helping them fill those gaps and enhance their understanding.
	Quiz→ Examples	• *Keeping records and monitoring* [111] • *Seeking information* [111] • *Help-seeking* [23]	Seeking clarification by reviewing examples to understand how to approach or solve questions correctly, thereby enhancing or consolidating understanding of the material.
	Read→ Examples	• *Seeking information* [111] • *Elaboration* [83]	Seeking to further reinforce understanding from reading material to see how the material is applied in practice by going through worked-out examples.
	Quiz→ Summary	• *Keeping records and monitoring* [111] • *Seeking information* [111] • *Search* [83]	After taking the quiz, students seek to clarify or reinforce students' understanding by reviewing summarized concepts. This helps address any gaps or consolidate students' knowledge, leading to a more cohesive grasp of the material.
Seeking assessment	Read→ Quiz	• *Seeking evaluation* [111]	When students read material and take quiz, they are assessing their understanding, enabling students to evaluate their performance.
	Quiz→ Quiz	• *Rehearsing and memorizing* [111] • *Repeating* [83]	Active efforts to engage in self-assessment help students evaluate and reassess their understanding. Students continuously test their knowledge to monitor learning progress and identify areas needing improvement.

- **Initial Metacognitive calibration (Initial-MC)**: We calculated initial-MC by subtracting the actual pretest scores from the estimated scores, using this difference as a metric for initial calibration. Initial-MC allows us to assess the degree of students' metacognitive calibration (i.e., how accurately students are aware of their own performance) before students begin the learning session.
- **Post Metacognitive calibration (Post-MC)**: We subtracted the actual posttest scores from the students' predicted posttest grades for post-MC. Post-MC allows us to assess the extent of students' metacognitive calibration after the learning session (and thus after interventions if they were in the experiment condition).

Both negative and positive values in initial-MC and post-MC assessments reflect students' confidence levels regarding their performance. We adapted the approaches used in prior research [85, 86] to measure confidence. These approaches assess confidence in test-taking contexts by asking students, immediately after completing a test, how confident they are in their performance or how well they believe they performed. A negative value in initial-MC and post-MC suggests that students estimated a lower score than they actually achieved, indicating underconfidence. Conversely, a positive value in initial-MC and post-MC indicates overconfidence, meaning students anticipated a higher grade than they achieved. A value of zero in both the initial-MC and post-MC assessments suggests that students made minimal error in evaluating their performance, indicating neither underconfidence nor overconfidence.

- **Change in Metacognitive calibration (ΔMC)**: We assessed the change in students' metacognitive calibration by calculating the difference between their initial-MC and post-MC measures. We subtracted the initial-MC from the post-MC to determine how students' metacognitive calibration shifted after the learning session. This change provides insight into

how students' awareness of their performance evolved. For the intervention group, in particular, this shift is crucial as it allows us to assess the effectiveness of the intervention in enhancing students' ability to accurately evaluate their own performance.

A positive ΔMC indicates that students' post-MC is higher than their initial-MC, whereas a negative ΔMC means students' initial-MC was greater than their post-MC. A zero ΔMC suggests that the student's metacognitive calibration remained unchanged before and after the learning session. We refer readers to Table 2, which provides examples and interpretations of both positive and negative ΔMC cases, offering a clearer understanding of the potential shifts in metacognitive calibration. These changes in metacognitive calibration can occur regardless of whether students initially exhibited overconfidence (MC > 0), underconfidence (MC < 0), or had accurate calibration (MC = 0).

We measured *metacognitive calibration improvement* by calculating the average of the absolute value of ΔMC for instances where students showed enhanced metacognitive calibration following the learning session. By "enhancement in metacognitive calibration," we refer to cases where students showed a shift in the direction of perfect calibration after the learning session. This includes students who were initially underconfident and exhibited a positive ΔMC, whether they became less underconfident, achieved perfect calibration, or shifted to overconfidence (an over-correction possibility that merits exploration in future work with intervention design changes). Similarly, for students who were initially overconfident, a negative ΔMC indicates a shift in the expected direction away from overconfidence. The extent of improvement in metacognitive calibration provides valuable insights into how much students enhanced their ability to accurately assess their performance.

- **Metacognitive calibration improvement |ΔMC|**: We assessed the magnitude of students' improvement in metacognitive calibration by calculating the absolute values of ΔMC for instances where students demonstrated enhanced calibration following the learning session. |ΔMC| allows us to quantify the extent of calibration improvement, providing a clear measure of how much students adjusted their self-assessment accuracy after engaging in the learning session.

A larger ΔMC reflects greater improvement in self-assessment accuracy. For example, if Student A was initially underconfident with an initial-MC of -8 and became less underconfident with a post-MC of -5, their improvement in metacognitive calibration is moderate, with a change of 3 points. In contrast, if Student B also started with an initial-MC of -5 but achieved perfect calibration with a post-MC of 0, their improvement is much greater, with a change of 5 points.

5 METHODS

5.1 RQ1. Impact of AI-powered intervention on student metacognitive calibration

We assessed metacognitive calibration improvements across three subgroups—initial confidence levels, gender, and race/ethnicity—by comparing control and intervention groups to evaluate the effectiveness of the AI-driven tool in enhancing metacognitive calibration. For initial confidence, we compared improvements between underconfident (i.e., students with initial-MC < 0) and overconfident (i.e., students with initial-MC > 0) students in both groups, investigating whether the tool's impact varied based on the confidence level. For demographics, we analyzed metacognitive calibration improvement between female and male students and across race/ethnicity groups (White, Black, Asian, and Multiracial). To maximize the use of available data in gender- and race-related analyses, we did not further subdivide these groups (e.g., by initial confidence level) as subgroup sample sizes were already small.

5.2 RQ2. Effectiveness of AI-powered intervention on student learning outcomes

We ran a linear regression analysis to examine how learning gains were influenced by the intervention and demographic variables (i.e., race and gender). The dependent variable, student learning gain, measured as the difference between pretest and posttest grades, is a continuous variable. The independent variables included group (a binary variable indicating whether a student was in the control or intervention group), race (categorized as White, Black, Asian, Hispanic/Latinx, and Multiracial), gender (categorized as male and female), the interaction between group and race, and the interaction between group and gender. Participants who identified as a member of very small race/ethnicity or gender groups ($n \leq 7$) were excluded due to anonymity concerns and the small sample sizes. This regression model enabled us to assess the main effects of group, race, and gender, as well as any interaction effects between these variables on learning gain. We checked the assumptions of linear regression, ensuring that linearity, independence, homoscedasticity, and normality were met.

5.3 RQ3. Mediating role of learning behaviors in AI intervention effects on learning gains

To explore whether students' learning behaviors mediated the effects of our AI-powered intervention on learning gains, we conducted a multiple mediator analysis [34]. We hypothesized that the intervention would influence learning gains through two specific mediators: *seeking knowledge* and *seeking assessment* (see Figure 3). We further discuss how our preliminary analysis results on mediator and outcome variables informed the development of the proposed mediation path diagram in section 6.3.1. In our mediation model, the independent variable was *Intervention*, a binary variable coded as 0 for the control group and 1 for the intervention group. The two mediator variables were *seeking knowledge* and *seeking assessment*, which were measured as count variables representing the frequency of each behavior. To approximate normal distributions, we applied square-root transformations to these mediator variables. The dependent variable was *learning gain*, a continuous variable calculated as the difference between pretest and posttest grades, which follows a normal distribution. We analyzed mediation within a path analysis framework using the structural equation modeling (SEM) software package, Mplus version 8 [62]. Mediation effects were estimated using the product of coefficients approach, consistent with the methods proposed by Preacher and Hayes [75, 76]. We employed a maximum likelihood estimator with standard errors

Table 2: Examples of changes in metacognitive calibration (ΔMC) are presented for two cases, one where the value is positive and one where it is negative. The "Cases" column outlines potential scenarios for initial and post-calibration conditions. The "Example" column provides specific values to illustrate each case, while the "Interpretation" column explains how to interpret the corresponding scenario.

ΔMC	Cases		Example	Interpretation
Positive (ΔMC>0)	Initial overconfidence (Initial-MC>0)	Post-MC > 0	Initial-MC=10, Post-MC=15, ΔMC = 5	Initially overconfident and became more overconfident
	Initial underconfidence (Initial-MC<0)	Post-MC<0	Initial-MC=-2, Post-MC=-1, ΔMC=1	Initially underconfident and became less underconfident
		Post-MC=0	Initial-MC=-1, Post-MC=0, ΔMC=0	Initially underconfident and achieved perfect calibration
		Post-MC>0	Initial-MC=-1, Post-MC=1 ΔMC=2	Initally underconfident and became overconfident
	Initial perfect calibration (Initial-MC=0)	Post-MC>0	Initial-MC=0, Post-MC=1 ΔMC=1	Initial perfect calibration and became overconfident
Negative (ΔMC<0)	Initial underconfidence (Initial-MC<0)	Post-MC < 0	Initial-MC=-3, Post-MC=-5, ΔMC = -2	Initially underconfident and became more underconfident
	Initial overconfidence (Initial-MC<0)	Post-MC<0	Initial-MC=1, Post-MC=-1, ΔMC=-2	Initially overconfident and became underconfident
		Post-MC=0	Initial-MC=3, Post-MC=0, ΔMC=-3	Initially overconfident and achieved perfect calibration
		Post-MC>0	Initial-MC=15, Post-MC=10 ΔMC=-5	Initally overconfident and became less overconfident
	Initial perfect calibration (Initial-MC=0)	Post-MC<0	Initial-MC=-1, Post-MC=-1 ΔMC=-1	Initial perfect calibration and became underconfident

robust to violations of normality to account for any deviations from normality in the data.

6 RESULTS

6.1 RQ1. Impact of AI-powered intervention on student metacognitive calibration

In **RQ1**, we examined whether students who received the AI-driven intervention showed enhanced metacognitive calibration, aiming to evaluate the effectiveness of the early intervention in improving metacognitive calibration. We anticipated similar improvements for overconfident and underconfident students, as well as across racial groups, with greater improvement expected for male students compared to females. Results partially supported our hypothesis. We observed varying improvements in metacognitive calibration across student groups, including those categorized by initial confidence level, gender, and race, which may suggest differential impacts of AI-driven support on student metacognitive improvement (Appendix Table 3). Students with initial underconfidence who received the intervention exhibited an average improvement in metacognitive calibration that was not significantly better than the control group: |ΔMC| (absolute value of change in metacognitive calibration) in the control group was 12.7% versus 15.0% in the intervention group, t = 0.293, p = .771. For students who had initial overconfidence, students in the intervention group showed 4.1% better metacognitive calibration enhancement (control: 12.2%, intervention: 16.3%, t = 2.001, p = .049). This significantly greater improvement in metacognitive calibration for overconfident students implies that the AI-driven tool could be especially beneficial in supporting overconfident students in enhancing their metacognitive calibration. We observed no significant per-group effects of the intervention on metacognitive calibration across gender or race/ethnicity (Table 3), indicating no evidence of moderation, though subsample sizes were especially small for some groups so future work will be needed to establish a tight confidence interval.

6.2 RQ2. Effectiveness of AI-powered intervention on student learning outcomes

Building on the findings from **RQ1**, where we explored the effectiveness of the intervention in improving metacognitive calibration, we investigate in **RQ2** whether this real-time AI-powered intervention enhances learning outcomes. We hypothesized that students receiving the AI-powered intervention would achieve greater learning gains than those in the control group. Our results support this hypothesis. On average, students in the intervention group achieved higher learning gains, with a mean of 16.3% (SD = 22.0%, n = 80), compared to a mean gain of 7.4% (SD = 20.0%, n = 53) in the control group. The difference was statistically significant (t = -2.384, p =

.019), indicating the effectiveness of the metacognitive calibration intervention in improving learning outcomes. When accounting for additional factors such as race/ethnicity and gender in a linear regression model, the effect of the intervention was no longer statistically significant (p = .248), possibly because of the loss of statistical power due to inclusion of demographic variables (Appendix Table 4). The regression analysis revealed no significant main or interaction effects of demographic variables on learning gains, indicating no evidence of differential effects on learning overall nor due to the intervention itself, though like the main intervention effect in this model, more data may be needed to improve statistical power for such fine-grained effects to emerge. However, it remains promising that no significantly inequitable results emerged from these data.

6.3 RQ3. Mediating role of learning behaviors in AI intervention effects on learning gains

6.3.1 Preliminary analysis results. Our preliminary analyses on mediators and outcome variables informed the development of the proposed mediation path diagram (Figure 3). A univariate analysis of the two categories of learning patterns across the entire student sample (N = 133) revealed that students engaged more frequently in seeking knowledge than in seeking assessment. On average, students employed the seeking knowledge 8.917 times (SD = 6.879), while students engaged in seeking assessment 5.910 times (SD = 4.626) during a self-paced learning session. However, we observed that students in the intervention group (n = 80) engaged more frequently in both seeking knowledge and seeking assessment than those in the control group (n = 53). The control group showed a mean of 7.585 times (SD = 6.576) for seeking knowledge and a mean of 5.019 times (SD = 4.992) for seeking assessment. In comparison, the intervention group demonstrated higher means of 9.800 times (SD = 6.973) for seeking knowledge and 6.063 times (SD = 3.623) for seeking assessment. Given our hypothesis that the intervention would lead to better learning outcomes, we further examined the associations between these two categories of learning patterns and learning gains to determine whether these behaviors could potentially mediate the anticipated effects of the intervention. The Spearman *rho* analysis [84] revealed a significant association between seeking knowledge and learning gains (ρ = .597, p <.001), whereas the association between seeking assessment and learning gains was not statistically significant (ρ = .113, p = .194). These preliminary findings suggest that both seeking knowledge and seeking assessment may serve as mediators in explaining the relationship between the AI-driven intervention and student learning gains, with seeking knowledge potentially emerging as the stronger mediator.

6.3.2 Mediation analysis results. From **RQ2**, we observed that students who received the AI-powered intervention showed increased learning gains. In **RQ3**, we examine through mediation analysis whether this improvement was partly due to the intervention prompting students to adapt their behaviors. That is, does the intervention improve learning in part because it causes students to adapt their behaviors? Unstandardized parameter estimates for the mediation model are presented in Figure 3, with corresponding standard error (SE) estimates provided in parentheses. We also provide bias-corrected and accelerated (BCa) 95% confidence intervals (CIs) calculated through 10,000 bootstrap samples. Regressing seeking knowledge on the intervention showed that the intervention significantly increased engagement in seeking knowledge (a_1 = .460 (.213), p = .031, BCa 95% CI = [.110, .807]). This result indicates that students who received the intervention engaged in seeking knowledge behavior 0.460 times more than those who did not receive the intervention. Conversely, there was no significant effect of the intervention on seeking assessment (a_2 = .170 (.171), p = .318, BCa 95% CI = [-.118, .437]).

Regressing seeking knowledge and seeking assessment, the two potential mediating paths, on learning gain showed that two mediators had different directions of effect on learning gain. Seeking knowledge significantly increased the learning gain (b_1 = 13.745 (1.545), p < .001, BCa 95% CI = [11.211, 16.289]), while seeking assessment significantly decreased the learning gain (b_2 = -7.596 (1.736), p < .001, BCa 95% CI = [-10.438, -4.781]). An insignificant direct effect of the intervention was observed after the mediators were included in the regression c' = 3.929 (2.972), p = .186, BCa 95% CI = [-1.102, 8.722]). The insignificant direct effect of the intervention implies that after including the two mediators in the regression model, the effect of the intervention on learning gains became insignificant, indicating that the relationship between the intervention and learning gains may be fully mediated by the mediator variables. However, we found that only seeking knowledge significantly mediated the effects of the intervention on learning gain (a_1b_1 = 6.327 (3.068), p = .039, BCa 95% CI = [1.564, 11.692]; a_2b_2 = -1.293 (1.397), p = .355, BCa 95% CI = [-3.882, .673]). In other words, the effect of the intervention on student learning gain is accounted for by seeking knowledge learning pattern category rather than having a direct effect on learning outcomes. This finding indicates that the intervention enhances students' learning gains by fostering greater engagement in seeking knowledge, which in turn drives the observed improvements in learning outcomes. We report the effect size for the indirect effect, a_1b_1, by computing the product of the standardized regression coefficients for a_1 and b_1, consistent with the completely standardized indirect effect proposed by Preacher and Kelley [77]. We obtained βa_1 = .186 and βb_1 = .680, resulting in $\beta a_1 b_1$ = .126 for the completely standardized indirect effect. According to Cohen [22], $\beta a_1 b_1$ = .126 indicates a medium mediation effect.

7 DISCUSSION

We discuss the implications of our findings for each research question and offer insights relevant to HCI. We examined the effectiveness of the real-time, AI-driven metacognitive calibration support tool by addressing three research questions, using data collected from 133 college students recruited from Prolific in an online learning environment. Our key findings are as follows:

- **RQ1.** Overconfident students who received the intervention showed significantly greater metacognitive calibration improvement than the control group by 4.1% (t = 2.001, p = .049). We did not observe differences in the improvement in metacognitive calibration across gender and race/ethnicity groups.

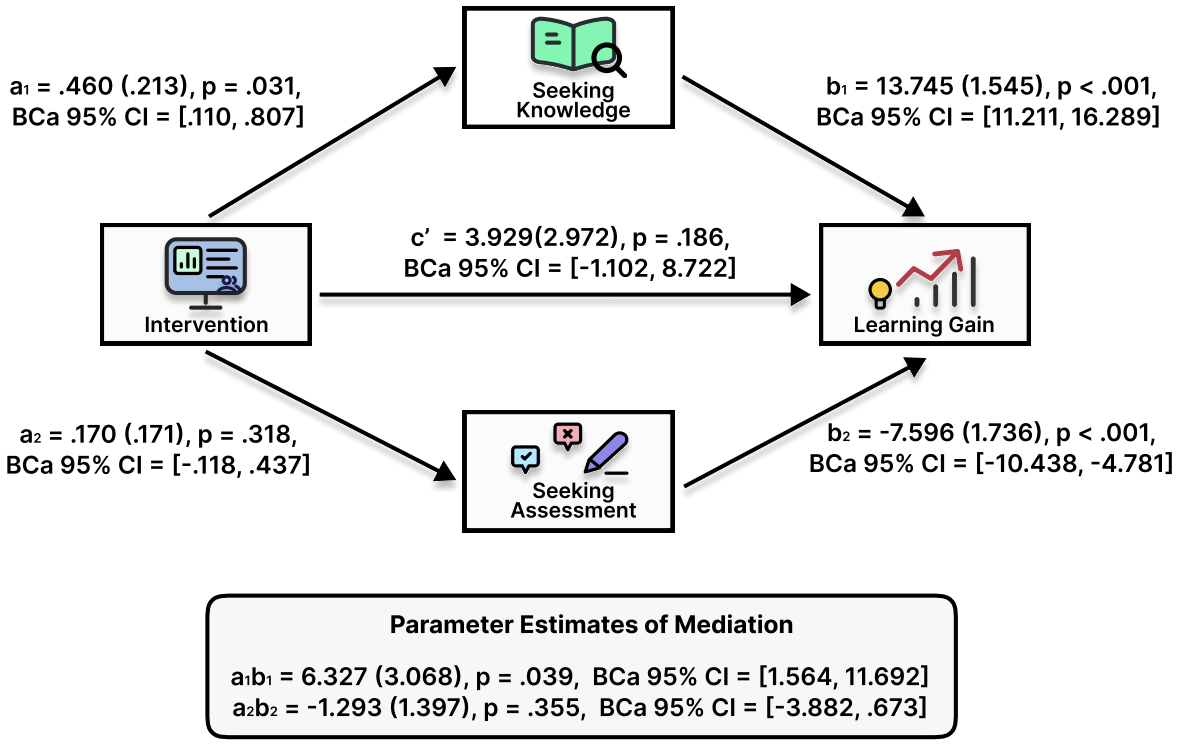


Figure 3: Path diagram showing the results of a multiple mediator analysis. The intervention is a binary variable indicating whether a student was in the control or intervention group. The two mediators, seeking knowledge and seeking assessment, represent distinct categories of student learning patterns. The outcome variable is learning gain. Unstandardized parameter estimates for the mediation model are presented with standard errors (SE) in parentheses, along with bias-corrected and accelerated (BCa) 95% confidence intervals (CIs) from 10,000 bootstrap samples.

- **RQ2.** We found a significant difference in learning gains between the control and intervention groups, with students who received the intervention achieving a mean learning gain of 16.3% compared to the control group. Regression analysis showed no significant effects of race, gender, or their interactions on the intervention's effectiveness.
- **RQ3.** Seeking knowledge, rather than seeking assessment, mediates the effect of the AI-driven intervention on learning gains. This finding highlights that students who received the intervention engaged more in seeking knowledge behaviors, which, in turn, led to increased learning gains.

7.1 RQ1. Impact of AI-powered intervention on student metacognitive calibration

We observed greater improvements in metacognitive calibration in the intervention group compared to the control group for students who were initially overconfident among various subgroups (i.e., initial confidence levels, gender, and race/ethnicity). This finding aligns with prior studies demonstrating the benefits of metacognitive support, often provided ad hoc, in enhancing metacognitive calibration [14, 65, 92]. However, while previous research has generally highlighted the effectiveness of interventions in improving metacognitive calibration, the literature also reflects inconsistencies, with some studies reporting no effects [32, 56]. While these inconsistencies pose challenges for researchers in identifying effective approaches to supporting metacognition, our novel approach to providing early feedback contributes to this growing body of research by showing that improvements can be achieved through early, real-time AI-driven interventions. To build on this work, comparative studies are needed to evaluate different approaches and determine whether they result in varying levels of improvement.

We observed overall improvements in metacognitive calibration in the control group as well, although the extent of this improvement was greater for the intervention group. This overall improvement across groups was expected, as control group students also had opportunities to assess their knowledge through quizzes, reviewing incorrect answers, and gaining insights into their understanding. Because most participants were unfamiliar with statistics, spending an hour engaging with the material likely aided all students in improving their ability to estimate their knowledge more accurately. Furthermore, unlike previous studies that lacked various

subgroup-specific analyses (e.g., [14, 80]), our work examined the effectiveness of real-time interventions across diverse subgroups of students. This approach addresses a gap in the literature regarding whether metacognitive support benefits all students equally, an essential step toward designing personalized interventions. Notably, the significant improvement among initially overconfident students who received the intervention, compared to those in the control group, suggests that early AI-driven support is particularly effective in helping these students recalibrate their self-assessments. However, while we observed significant improvements for overconfident students, no significant differences were found for underconfident students, highlighting the complexity of supporting metacognitive calibration based on student confidence levels. Our findings suggest that interventions aimed at improving metacognitive calibration should be tailored to address the specific needs of students, taking into account whether they are more likely to exhibit overconfidence or underconfidence. To this end, future research should explore how underconfident and overconfident students differ in their use of learning strategies upon receiving interventions. These insights would enable the development of personalized early AI-driven interventions that could target each group's specific needs and challenges, ensuring more effective support for recalibrating self-assessments and enhancing overall learning outcomes.

7.2 RQ2. Effectiveness of AI-powered intervention on student learning outcomes

The significant difference in learning gains, with the intervention group achieving a mean gain of 16.3% compared to 7.4% in the control group, highlights the effectiveness of early AI-driven support in improving learning outcomes in computer-based learning environments. This finding extends previous research on metacognitive calibration interventions, which has predominantly focused on improving calibration accuracy while often overlooking their impact on learning outcomes [63, 92] or reporting non-significant impacts on performance [8, 80]. While we observed improved learning outcomes among students who received the intervention, we encourage future studies to explore how students perceive early, real-time AI-driven interventions and their views on how the intervention impacts learning outcomes and metacognitive calibration. Gathering qualitative insights into which aspects of early interventions students find helpful, discouraging, or potentially disruptive will be essential for refining the specifics of these interventions—such as the optimal timing and frequency of delivery—and ensuring the effectiveness of early metacognitive calibration tools. For example, certain groups, such as students with Attention Deficit Hyperactivity Disorder (ADHD), might perceive these early interventions as disruptive, since these students might adopt different learning strategies and demonstrate behavioral differences [24, 78] when interacting with learning interfaces in computer-based learning environments. Understanding these diverse perspectives will allow for the refinement of early interventions and the personalization of these tools to better meet the needs of various student groups.

We did not observe significant main effects of gender or race/ethnicity and gender on learning gains, suggesting that the intervention did not result in significant differential effects across these demographic factors. However, the intervention leverages machine learning to predict student learning outcomes, which may raise questions regarding the issues of fairness. These variations may reflect inherent biases in the model, as research has shown that machine learning models can exhibit predictive discrepancies [6, 7, 48], especially when certain groups are underrepresented in the training data. This highlights the need for further investigation with larger samples to determine if there are indeed no differences between groups.

7.3 RQ3. Mediating role of learning behaviors in AI intervention effects on learning gains

Seeking knowledge, rather than seeking assessment, was the only significant mediator in the effect of the AI-driven intervention on learning gains. The intervention significantly increased student engagement in seeking knowledge behaviors, while no significant increase was observed in seeking assessment behaviors. This result suggests that the intervention encouraged students to adopt seeking knowledge strategies, which, in turn, led to improved learning outcomes. Most importantly, our finding contributes to the line of research on supporting student metacognitive calibration, specifically by highlighting how supporting calibration leads to learning outcomes—a relationship that has not been thoroughly explored in previous research [8, 63, 80].

One possible explanation for why seeking knowledge emerged as the only significant mediator could be attributed to the nature of the intervention feedback students received during learning. Since the metacognitive calibration intervention was intended to help students better understand their own level of knowledge, we might expect this to result in more behaviors acting upon that understanding—i.e., seeking knowledge to address identified knowledge gaps. In contrast, receiving the AI-predicted posttest grade along with their quiz performance for each subtopic during the intervention may not do much to motivate even more (potentially redundant) assessment, and could even reduce students' reliance on seeking assessment behaviors such as the frequent learning patterns of Quiz → Quiz and Read → Quiz.

Our contribution to understanding the mechanisms by which the early AI-driven intervention enhanced learning outcomes through student behavioral patterns has important practical implications, which we discuss further in the next section. Specifically, interventions can be designed to be more effective by encouraging students to engage in targeted learning behaviors, such as seeking knowledge, rather than solely providing feedback on discrepancies between their estimated and actual performance. This approach could lead to even greater improvements in learning outcomes.

7.4 HCI Implications

The results of this research offer practical implications for designing learning interfaces that enhance students' metacognitive calibration in computer-based learning environments. A key contribution of this work lies in uncovering the mechanisms through which real-time AI-driven intervention significantly increases student behaviors and, subsequently, learning outcomes. Specifically, our findings on **RQ3** highlight that the intervention led to increased engagement in *seeking knowledge* behaviors, which mediated the improvement in learning gains. Understanding this mechanism is crucial for HCI researchers and practitioners as it is important

to design learning interfaces that account for how interventions impact students' interactions (which we measure as self-regulated learning-relevant in this study).

By gaining insight into the specific behaviors that contribute to enhanced learning outcomes, researchers and practitioners can create interfaces that not only provide feedback but also actively encourage and facilitate these impactful behaviors (e.g., [33, 60]). For instance, functions could be designed to promote seeking knowledge behaviors (spending time on materials that students have not mastered or find challenging), which could be supported by learning interfaces. Tankelevitch et al. [90] noted that incorporating metacognitive support strategies into generative AI systems could reduce the metacognitive demands on users. Likewise, researchers could consider incorporating features that facilitate these behaviors, through diverse ways such as adaptive content recommendations.

Incorporating features designed to promote students' engagement in seeking knowledge into learning interfaces may lead to greater engagement in seeking knowledge behaviors, ultimately resulting in even greater learning gains. Future studies could explore whether incorporating these features, specifically designed to target seeking knowledge behaviors, indeed enhances metacognitive calibration as well as learning outcomes. Moreover, as suggested by our results on **RQ1** and **RQ2**, there may be potential variations across subgroups of students, which could be another avenue for future research.

Aligned with the findings from **RQ1** and **RQ2**, our early, real-time AI-driven approach presents a promising alternative to post hoc feedback for supporting student metacognitive calibration in computer-based learning environments. By integrating AI-driven methods capable of real-time analysis of student interactions, HCI researchers and practitioners can design learning systems that include features to promote seeking knowledge behaviors, empowering students to adjust their self-assessments and learning strategies during the learning process. In sum, understanding how interventions affect students' interactions with learning interfaces is vital for designing systems that support actions leading to improved learning. By focusing on promoting beneficial behaviors like *seeking knowledge*, HCI practitioners can develop interfaces that not only support metacognitive calibration but also enhance overall learning experiences.

8 LIMITATIONS

Although we observed few significant differences between groups in terms of the intervention's effect on metacognitive calibration (**RQ1**) and learning outcomes (**RQ2**), these findings may have been influenced by insufficient sample size. Future studies should aim to replicate these results with larger, more diverse populations to uncover what heterogeneous treatment effects there may be. Additionally, our online learning system may differ from other platforms, such as massive open online courses and semester-long online college courses. These platforms typically offer longer learning periods, a wider variety of learning activities, and distinct platform features. Future research should investigate whether our findings can be generalized to other online learning environments. We grouped three slight variations of the AI-driven interventions (i.e., one without attention weights visualized by color darkness and one with no graph). While the core functionality of the interventions, including AI-predicted posttest grades with reflective prompts, remained the same, the small sample size limited our ability to thoroughly analyze differences across the three variations. Future work will explore these differences in greater depth with larger samples, enabling more robust comparisons and providing insights into whether specific graphical details lead to differential impacts.

We assessed metacognitive calibration based on students' estimations of their own performance (i.e., predicted pretest and posttest grades). Measuring metacognitive calibration using students' estimation of their performance compared to their actual grade has been a common approach in literature [32, 39, 69]. However, metacognitive calibration comprises multiple facets [29, 30], including confidence and placement [57], which may not have been fully captured in this study. Future studies could examine whether our early, real-time intervention approach can be applied to these additional facets of metacognitive calibration or explore how different facets might require distinct forms of support. Such studies could provide a deeper and more holistic understanding of how to enhance student metacognitive calibration effectively.

9 CONCLUSION

Supporting students in making accurate judgments about their performance is critical for enhancing learning outcomes, particularly in computer-based learning environments. The effectiveness of our early AI-driven intervention in improving metacognitive calibration for overconfident students, along with the observed increase in learning gains, highlights the value of early interventions for metacognitive calibration. Additionally, this result paves the way for applying advanced AI approaches to develop even more sophisticated early interventions that use the same approach for a variety of upcoming assessments (e.g., quizzes, exams) where students might benefit from learning how to self-evaluate accurately in advance of the assessment and adjust their learning accordingly. Our work in identifying the mechanisms behind the intervention's impact on student learning outcomes also highlights the importance of promoting targeted learning behaviors in computer-based environments. In particular, the findings of **RQ3** could be beneficial in designing and developing learning interfaces that facilitate the types of knowledge-seeking activities students are likely to benefit from after a metacognitive calibration intervention.

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Award No. IIS-2202481.

REFERENCES

[1] Rasha M. Abdelrahman. 2020. Metacognitive awareness and academic motivation and their impact on academic achievement of Ajman University students. *Heliyon* 6, 9 (Sept. 2020), e04192. https://doi.org/10.1016/j.heliyon.2020.e04192

[2] Ahmet Oguz Akturk and Ismail Sahin. 2011. Literature review on metacognition and its measurement. *Procedia - Social and Behavioral Sciences* 15 (Jan. 2011), 3731–3736. https://doi.org/10.1016/j.sbspro.2011.04.364

[3] Patricia A. Alexander. 2013. Calibration: What is it and why it matters? An introduction to the special issue on calibrating calibration. *Learning and Instruction* 24 (April 2013), 1–3. https://doi.org/10.1016/j.learninstruc.2012.10.003

[4] Roger Azevedo, François Bouchet, Melissa Duffy, Jason Harley, Michelle Taub, Gregory Trevors, Elizabeth Cloude, Daryn Dever, Megan Wiedbusch, Franz Wortha, and Rebeca Cerezo. 2022. Lessons learned and future directions of MetaTutor: Leveraging multichannel data to scaffold self-regulated learning

with an intelligent tutoring system. *Frontiers in Psychology* 13 (June 2022), 813632. https://doi.org/10.3389/fpsyg.2022.813632

[5] Roger Azevedo, Daniel C. Moos, Amy M. Johnson, and Amber D. Chauncey. 2010. Measuring cognitive and metacognitive regulatory processes during hypermedia learning: Issues and challenges. *Educational Psychologist* 45, 4 (Oct. 2010), 210–223. https://doi.org/10.1080/00461520.2010.515934

[6] Ryan S. Baker and Aaron Hawn. 2022. Algorithmic bias in education. *International Journal of Artificial Intelligence in Education* 32, 4 (Dec. 2022), 1052–1092. https://doi.org/10.1007/s40593-021-00285-9

[7] Clara Belitz, HaeJin Lee, Nidhi Nasiar, Stephen E. Fancsali, Steve Ritter, Husni Almoubayyed, Ryan S. Baker, Jaclyn Ocumpaugh, and Nigel Bosch. 2024. Hierarchical dependencies in classroom settings influence algorithmic bias metrics. In *Proceedings of the 14th Learning Analytics and Knowledge Conference*. ACM, New York, NY, 210–218. https://doi.org/10.1145/3636555.3636869

[8] Linda Bol and Douglas J. Hacker. 2001. A comparison of the effects of practice tests and traditional review on performance and calibration. *The Journal of Experimental Education* 69, 2 (Jan. 2001), 133–151. https://doi.org/10.1080/00220970109600653

[9] Nigel Bosch, Yingbin Zhang, Luc Paquette, Ryan S. Baker, Jaclyn Ocumpaugh, and Gautam Biswas. 2021. Students' verbalized metacognition during computerized learning. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, Vol. 680. ACM, New York, NY, 1–12. https://doi.org/10.1145/3411764.3445809

[10] Leo Breiman. 2001. Random forests. *Machine Learning* 45, 1 (2001), 5–32. https://doi.org/10.1023/A:1010933404324

[11] Rex Bringula, Jon Jester Reguyal, Don Dominic Tan, and Saida Ulfa. 2021. Mathematics self-concept and challenges of learners in an online learning environment during COVID-19 pandemic. *Smart Learning Environments* 8, 1 (Oct. 2021), 22. https://doi.org/10.1186/s40561-021-00168-5

[12] A. L. Brown. 1987. Metacognition, executive control, self-regulation, and other more mysterious mechanisms. In *Metacognition, Motivation, and Understanding*. Erlbaum, Hillsdale, NJ, 65–116.

[13] Victoria Cabales. 2019. Muse: Scaffolding metacognitive reflection in design-based research. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 1–6. https://doi.org/10.1145/3290607.3308450

[14] Aimee A. Callender, Ana M. Franco-Watkins, and Andrew S. Roberts. 2016. Improving metacognition in the classroom through instruction, training, and feedback. *Metacognition and Learning* 11, 2 (Aug. 2016), 215–235. https://doi.org/10.1007/s11409-015-9142-6

[15] Jennifer Campbell, Diane Horton, and Michelle Craig. 2016. Factors for success in online CS1. In *Proceedings of the 2016 ACM Conference on Innovation and Technology in Computer Science Education*. ACM, New York, NY, 320–325. https://doi.org/10.1145/2899415.2899457

[16] John Chen, Xi Lu, Yuzhou Du, Michael Rejtig, Ruth Bagley, Mike Horn, and Uri Wilensky. 2024. Learning agent-based modeling with LLM companions: Experiences of novices and experts using ChatGPT & NetLogo Chat. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 1–18. https://doi.org/10.1145/3613904.3642377

[17] Peggy P. Chen. 2003. Exploring the accuracy and predictability of the self-efficacy beliefs of seventh-grade mathematics students. *Learning and Individual Differences* 14, 1 (2003), 77–90. https://doi.org/10.1016/j.lindif.2003.08.003

[18] Alan Y. Cheng, Meng Guo, Melissa Ran, Arpit Ranasaria, Arjun Sharma, Anthony Xie, Khuyen N. Le, Bala Vinaithirthan, Shihe (Tracy) Luan, David Thomas Henry Wright, Andrea Cuadra, Roy Pea, and James A. Landay. 2024. Scientific and fantastical: Creating immersive, culturally relevant learning experiences with augmented reality and large language models. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 1–23. https://doi.org/10.1145/3613904.3642041

[19] Eric C. K. Cheng and Joanna K. M. Chan. 2021. Metacognition and metacognitive learning. In *Developing Metacognitive Teaching Strategies Through Lesson Study*, Eric C. K. Cheng and Joanna K. M. Chan (Eds.). Springer, Singapore, 11–24. https://doi.org/10.1007/978-981-16-5569-2_2

[20] Chih-Yueh Chou, K. Robert Lai, Po-Yao Chao, Chung Hsien Lan, and Tsung-Hsin Chen. 2015. Negotiation based adaptive learning sequences: Combining adaptivity and adaptability. *Computers & Education* 88 (Oct. 2015), 215–226. https://doi.org/10.1016/j.compedu.2015.05.007

[21] Chih-Yueh Chou and Nian-Bao Zou. 2020. An analysis of internal and external feedback in self-regulated learning activities mediated by self-regulated learning tools and open learner models. *International Journal of Educational Technology in Higher Education* 17, 1 (Dec. 2020), 55. https://doi.org/10.1186/s41239-020-00233-y

[22] Jacob Cohen. 1988. *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates, Hillsdale, NJ. https://doi.org/10.4324/9780203771587

[23] Linda Corrin, Paula G. De Barba, and Aneesha Bakharia. 2017. Using learning analytics to explore help-seeking learner profiles in MOOCs. In *Proceedings of the 7th International Learning Analytics & Knowledge*. ACM, New York, NY, 424–428. https://doi.org/10.1145/3027385.3027448

[24] D. Daley and J. Birchwood. 2010. ADHD and academic performance: why does ADHD impact on academic performance and what can be done to support ADHD children in the classroom? *Child: Care, Health and Development* 36, 4 (July 2010), 455–464. https://doi.org/10.1111/j.1365-2214.2009.01046.x

[25] Amy L. Dent and Alison C. Koenka. 2016. The relation between self-regulated learning and academic achievement across childhood and adolescence: A meta-analysis. *Educational Psychology Review* 28, 3 (Sept. 2016), 425–474. https://doi.org/10.1007/s10648-015-9320-8

[26] Smit Desai and Jessie Chin. 2023. Ok Google, let's learn: Using voice user interfaces for informal self-regulated learning of health topics among younger and older adults. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 1–21. https://doi.org/10.1145/3544548.3581507

[27] John Dunlosky and Janet Metcalfe. 2009. *Metacognition*. Sage Publications, Thousand Oaks, Calif.

[28] John Dunlosky and Keith W. Thiede. 2013. Four cornerstones of calibration research: Why understanding students' judgments can improve their achievement. *Learning and Instruction* 24 (April 2013), 58–61. https://doi.org/10.1016/j.learninstruc.2012.05.002

[29] Anastasia Efklides. 2006. Metacognitive experiences: The missing link in the self-regulated learning process. *Educational Psychology Review* 18, 3 (Sept. 2006), 287–291. https://doi.org/10.1007/s10648-006-9021-4

[30] Anastasia Efklides. 2008. Metacognition: Defining its facets and levels of functioning in relation to self-regulation and co-regulation. *European Psychologist* 13, 4 (Jan. 2008), 277–287. https://doi.org/10.1027/1016-9040.13.4.277

[31] Anastasia Efklides. 2011. Interactions of metacognition with motivation and affect in self-regulated learning: The MASRL model. *Educational Psychologist* 46, 1 (Jan. 2011), 6–25. https://doi.org/10.1080/00461520.2011.538645

[32] Bethany Emory and Tian Luo. 2022. Metacognitive training and online community college students' learning calibration and performance. *Community College Journal of Research and Practice* 46, 4 (April 2022), 240–256. https://doi.org/10.1080/10668926.2020.1841042

[33] Katharina Engelmann, Maria Bannert, and Nadine Melzner. 2021. Do self-created metacognitive prompts promote short- and long-term effects in computer-based learning environments? *Research and Practice in Technology Enhanced Learning* 16 (Feb. 2021), 3. https://doi.org/10.1186/s41039-021-00148-w

[34] Amanda J. Fairchild and Heather L. McDaniel. 2017. Best (but oft-forgotten) practices: Mediation analysis. *The American Journal of Clinical Nutrition* 105, 6 (June 2017), 1259–1271. https://doi.org/10.3945/ajcn.117.152546

[35] John H. Flavell. 1979. Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American Psychologist* 34, 10 (Oct. 1979), 906–911. https://doi.org/10.1037/0003-066X.34.10.906

[36] Nathaniel L. Foster, Christopher A. Was, John Dunlosky, and Randall M. Isaacson. 2017. Even after thirteen class exams, students are still overconfident: The role of memory for past exam performance in student predictions. *Metacognition and Learning* 12, 1 (April 2017), 1–19. https://doi.org/10.1007/s11409-016-9158-6

[37] Antonio P. Gutierrez and Addison F. Price. 2017. Calibration between undergraduate students' prediction of and actual performance: The role of gender and performance attributions. *The Journal of Experimental Education* 85, 3 (July 2017), 486–500. https://doi.org/10.1080/00220973.2016.1180278

[38] Douglas J. Hacker and Linda Bol. 2004. Metacognitive theory: Considering the social–cognitive influences. In *Big Theories Revisited: Vol. 4: Research on Sociocultural Influences on Motivation and Learning*. Information Age, Greenwich, CT, 275–297.

[39] Douglas J. Hacker, Linda Bol, and Kamilla Bahbahani. 2008. Explaining calibration accuracy in classroom contexts: The effects of incentives, reflection, and explanatory style. *Metacognition and Learning* 3, 2 (Aug. 2008), 101–121. https://doi.org/10.1007/s11409-008-9021-5

[40] Douglas J. Hacker, Linda Bol, Dianne D. Horgan, and Ernest A. Rakow. 2000. Test prediction and performance in a classroom context. *Journal of Educational Psychology* 92, 1 (March 2000), 160–170. https://doi.org/10.1037/0022-0663.92.1.160

[41] Karen R. Harris, Steve Graham, Mary Brindle, and Karin Sandmel. 2009. Metacognition and children's writing. In *Handbook of Metacognition in Education*. Routledge/Taylor & Francis Group, New York, NY, US, 131–153.

[42] Marion Händel, Anique B. H. De Bruin, and Markus Dresel. 2020. Individual differences in local and global metacognitive judgments. *Metacognition and Learning* 15, 1 (April 2020), 51–75. https://doi.org/10.1007/s11409-020-09220-0

[43] Akbar Jalili, Masoud Hejazi, Gholamhossein Entesar Foumani, and Zekrollah Morovati. 2018. The relationship between meta-cognition and academic performance with mediation role of problem solving. *Quarterly Journal of Child Mental Health* 5, 1 (June 2018), 80–91. http://childmentalhealth.ir/article-1-379-en.html Publisher: Quarterly Journal of Child Mental Health.

[44] Sung-Hee Jin, Kowoon Im, Mina Yoo, Ido Roll, and Kyoungwon Seo. 2023. Supporting students' self-regulated learning in online learning using artificial intelligence applications. *International Journal of Educational Technology in Higher Education* 20 (June 2023), 37. https://doi.org/10.1186/s41239-023-00406-5

[45] Amy M. Johnson, Roger Azevedo, and Sidney K. D'Mello. 2011. The temporal and dynamic nature of self-regulatory processes during independent and externally assisted hypermedia learning. *Cognition and Instruction* 29, 4 (Oct. 2011), 471–504. https://doi.org/10.1080/07370008.2011.610244

[46] Jina Kang, Min Liu, and Wen Qu. 2017. Using gameplay data to examine learning behavior patterns in a serious game. *Computers in Human Behavior* 72 (July 2017), 757–770. https://doi.org/10.1016/j.chb.2016.09.062

[47] William L. Kelemen, Robert G. Winningham, and Charles A. Weaver. 2007. Repeated testing sessions and scholastic aptitude in college students' metacognitive accuracy. *European Journal of Cognitive Psychology* 19, 4-5 (July 2007), 689–717. https://doi.org/10.1080/09541440701326170

[48] René F. Kizilcec and Hansol Lee. 2022. Algorithmic fairness in education. In *The Ethics of Artificial Intelligence in Education*. Routledge, New York, NY, 174–202.

[49] Rob Klassen. 2002. A question of calibration: A review of the self-efficacy beliefs of students with learning disabilities. *Learning Disability Quarterly* 25, 2 (May 2002), 88–102. https://doi.org/10.2307/1511276

[50] Justin Kruger and David Dunning. 1999. Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology* 77, 6 (1999), 1121–1134. https://doi.org/10.1037/0022-3514.77.6.1121

[51] Susanne P. Lajoie and Roger Azevedo. 2006. Teaching and learning in technology-rich environments. In *Handbook of Educational Psychology* (2nd ed.). Lawrence Erlbaum Associates Publishers, Mahwah, NJ, US, 803–821.

[52] HaeJin Lee and Nigel Bosch. 2024. Subtopic-specific heterogeneity in computer-based learning behaviors. *International Journal of STEM Education* 11, 1 (2024), 61. https://doi.org/10.1186/s40594-024-00519-x

[53] Lin-Miao Lin and Karen M. Zabrucky. 1998. Calibration of comprehension: Research and implications for education and instruction. *Contemporary Educational Psychology* 23, 4 (Oct. 1998), 345–391. https://doi.org/10.1006/ceps.1998.0972

[54] Jennifer A. Livingston. 2003. *Metacognition: An overview*. Technical Report ED474273. ERIC. https://eric.ed.gov/?id=ED474273

[55] Lauri Malmi, Judy Sheard, Päivi Kinnunen, Simon, and Jane Sinclair. 2019. Computing education theories: What are they and how are they used?. In *Proceedings of the 2019 ACM Conference on International Computing Education Research*. ACM, New York, NY, 187–197. https://doi.org/10.1145/3291279.3339409

[56] Dionne A. Miller. 2015. Learning how students learn: An exploration of self-regulation strategies in a two-year college general chemistry class. *Journal of College Science Teaching* 44, 3 (2015), 11–16.

[57] Don A. Moore and Paul J. Healy. 2008. The trouble with overconfidence. *Psychological Review* 115, 2 (April 2008), 502–517. https://doi.org/10.1037/0033-295X.115.2.502

[58] Jason W. Morphew. 2021. Changes in metacognitive monitoring accuracy in an introductory physics course. *Metacognition and Learning* 16, 1 (April 2021), 89–111. https://doi.org/10.1007/s11409-020-09239-3

[59] Hussein Mozannar, Gagan Bansal, Adam Fourney, and Eric Horvitz. 2024. Reading between the lines: Modeling user behavior and costs in AI-assisted programming. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 1–16. https://doi.org/10.1145/3613904.3641936

[60] Anabil Munshi, Gautam Biswas, Ryan Baker, Jaclyn Ocumpaugh, Stephen Hutt, and Luc Paquette. 2023. Analysing adaptive scaffolds that help students develop self-regulated learning behaviours. *Journal of Computer Assisted Learning* 39, 2 (April 2023), 351–368. https://doi.org/10.1111/jcal.12761

[61] Caroline Z. Muteti, Brooke I. Jacob, and Jacinta M. Mutambuki. 2023. Metacognition instruction enhances equity in effective study strategies across demographic groups in the general chemistry I course. *Chemistry Education Research and Practice* 24, 4 (2023), 1204–1218. https://doi.org/10.1039/D3RP00103B

[62] Bengt Muthén and Linda Muthén. 2017. Mplus. In *Handbook of Item Response Theory* (1st ed.). Chapman and Hall/CRC, Boca Raton, FL, 507–518.

[63] Marloes L. Nederhand, Huib K. Tabbers, and Remy M.J.P. Rikers. 2019. Learning to calibrate: Providing standards to improve calibration accuracy for different performance levels. *Applied Cognitive Psychology* 33, 6 (Nov. 2019), 1068–1079. https://doi.org/10.1002/acp.3548

[64] Jeremy T. D. Ng, Yiming Liu, Didier S. Y. Chui, Jack C. H. Man, and Xiao Hu. 2023. Leveraging LMS logs to analyze self-regulated learning behaviors in a maker-based course. In *Proceedings of the 13th International Conference on Learning Analytics and Knowledge Conference*. ACM, New York, NY, 670–676. https://doi.org/10.1145/3576050.3576111

[65] John L. Nietfeld, Li Cao, and Jason W. Osborne. 2006. The effect of distributed monitoring exercises and feedback on performance, monitoring accuracy, and self-efficacy. *Metacognition and Learning* 1, 2 (Aug. 2006), 159–179. https://doi.org/10.1007/s10409-006-9595-6

[66] Ernesto Panadero. 2017. A review of self-regulated learning: Six models and four directions for research. *Frontiers in Psychology* 8 (April 2017), 422. https://doi.org/10.3389/fpsyg.2017.00422

[67] Eyal Peer, David Rothschild, Andrew Gordon, Zak Evernden, and Ekaterina Damer. 2021. Data quality of platforms and panels for online behavioral research. *Behavior Research Methods* 54, 4 (Sept. 2021), 1643–1662. https://doi.org/10.3758/s13428-021-01694-3

[68] Gordon Pennycook, Robert M. Ross, Derek J. Koehler, and Jonathan A. Fugelsang. 2017. Dunning–Kruger effects in reasoning: Theoretical implications of the failure to recognize incompetence. *Psychonomic Bulletin & Review* 24, 6 (Dec. 2017), 1774–1784. https://doi.org/10.3758/s13423-017-1242-7

[69] O. Pesout and J. L. Nietfeld. 2021. How creative am I?: Examining judgments and predictors of creative performance. *Thinking Skills and Creativity* 40 (June 2021), 100836. https://doi.org/10.1016/j.tsc.2021.100836

[70] Stephanie Pieschl. 2009. Metacognitive calibration—an extended conceptualization and potential applications. *Metacognition and Learning* 4, 1 (April 2009), 3–31. https://doi.org/10.1007/s11409-008-9030-4

[71] Paul R. Pintrich. 2000. The role of goal orientation in self-regulated learning. In *Handbook of Self-Regulation*. Academic Press, San Diego, 451–502. https://doi.org/10.1016/B978-012109890-2/50043-3

[72] Paul R. Pintrich, Christopher A. Wolters, and Gail P. Baxter. 2000. Assessing metacognition and self-regulated learning. In *Issues in the Measurement of Metacognition*. Buros Institute of Mental Measurements, Lincoln, NE.

[73] Sradhanjali Pradhan and Parismita Das. 2021. Influence of metacognition on academic achievement and learning style of undergraduate students in Tezpur University. *European Journal of Educational Research* 10, 1 (Jan. 2021), 381–391. https://doi.org/10.12973/eu-jer.10.1.381

[74] James Prather, Brett A. Becker, Michelle Craig, Paul Denny, Dastyni Loksa, and Lauren Margulieux. 2020. What do we think we think we are doing? Metacognition and self-regulation in programming. In *Proceedings of the 2020 ACM Conference on International Computing Education Research (ICER '20)*. Association for Computing Machinery, New York, NY, USA, 2–13. https://doi.org/10.1145/3372782.3406263

[75] Kristopher J. Preacher and Andrew F. Hayes. 2004. SPSS and SAS procedures for estimating indirect effects in simple mediation models. *Behavior Research Methods, Instruments, & Computers* 36, 4 (Nov. 2004), 717–731. https://doi.org/10.3758/BF03206553

[76] Kristopher J. Preacher and Andrew F. Hayes. 2008. Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behavior Research Methods* 40, 3 (Aug. 2008), 879–891. https://doi.org/10.3758/BRM.40.3.879

[77] Kristopher J. Preacher and Ken Kelley. 2011. Effect size measures for mediation models: Quantitative strategies for communicating indirect effects. *Psychological Methods* 16, 2 (2011), 93–115. https://doi.org/10.1037/a0022658

[78] Abigail Reaser, Frances Prevatt, Yaacov Petscher, and Briley Proctor. 2007. The learning and study strategies of college students with ADHD. *Psychology in the Schools* 44, 6 (2007), 627–638.

[79] Mohi Reza and Dongwook Yoon. 2021. Designing CAST: A computer-assisted shadowing trainer for self-regulated foreign language listening practice. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 1–13. https://doi.org/10.1145/3411764.3445190

[80] Gabriel D. Saenz, Lisa Geraci, and Robert Tirso. 2019. Improving metacognition: A comparison of interventions. *Applied Cognitive Psychology* 33, 5 (2019), 918–929. https://doi.org/10.1002/acp.3556

[81] Gregory Schraw and Rayne Sperling Dennison. 1994. Assessing metacognitive awareness. *Contemporary Educational Psychology* 19, 4 (Oct. 1994), 460–475. https://doi.org/10.1006/ceps.1994.1033

[82] James R. Segedy, John S. Kinnebrew, and Gautam Biswas. 2015. Using coherence analysis to characterize self-regulated learning behaviours in open-ended learning environments. *Journal of Learning Analytics* 2 (May 2015), 1. https://doi.org/10.18608/jla.2015.21.3

[83] Christoph Sonnenberg and Maria Bannert. 2015. Discovering the effects of metacognitive prompts on the sequential structure of SRL-processes using process mining techniques. *Journal of Learning Analytics* 2, 1 (May 2015), 72–100. https://doi.org/10.18608/jla.2015.21.5

[84] C. Spearman. 1987. The proof and measurement of association between two things. *The American Journal of Psychology* 100, 3/4 (1987), 441–471. https://doi.org/10.2307/1422689

[85] Lazar Stankov and John D. Crawford. 1996. Confidence judgments in studies of individual differences. *Personality and Individual Differences* 21, 6 (Dec. 1996), 971–986. https://doi.org/10.1016/S0191-8869(96)00130-4

[86] Lazar Stankov and Jihyun Lee. 2008. Confidence and cognitive test performance. *Journal of Educational Psychology* 100, 4 (Nov. 2008), 961–976. https://doi.org/10.1037/a0012546

[87] Nancy J. Stone. 2000. Exploring the relationship between calibration and self-regulated learning. *Educational Psychology Review* 12, 4 (Dec. 2000), 437–475. https://doi.org/10.1023/A:1009084430926

[88] Qiyu Sun, Lawrence Jun Zhang, and Susan Carter. 2021. Investigating students' metacognitive experiences: Insights from the English as a Foreign Language Learners' Writing Metacognitive Experiences Questionnaire (EFLLWMEQ). *Frontiers in Psychology* 12 (2021), 744842. https://doi.org/10.3389/fpsyg.2021.744842

[89] Claudia Szabo, Nickolas Falkner, Andrew Petersen, Heather Bort, Kathryn Cunningham, Peter Donaldson, Arto Hellas, James Robinson, and Judy Sheard. 2019. Review and use of learning theories within computer science education research:

Primer for researchers and practitioners. In *Proceedings of the Working Group Reports on Innovation and Technology in Computer Science Education*. ACM, Aberdeen, Scotland, 89–109. https://doi.org/10.1145/3344429.3372504

[90] Lev Tankelevitch, Viktor Kewenig, Auste Simkute, Ava Elizabeth Scott, Advait Sarkar, Abigail Sellen, and Sean Rintel. 2024. The metacognitive demands and opportunities of generative AI. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 1–24. https://doi.org/10.1145/3613904.3642902

[91] David J. Torgerson and Carole J. Torgerson. 2008. *Designing randomised trials in health, education and the social sciences*. Palgrave Macmillan UK, London. https://doi.org/10.1057/9780230583993

[92] Kamila Urban and Marek Urban. 2019. Improving the accuracy of the self-evaluation during on-screen self-regulated learning through calibration feedback. In *Proceedings of the 13th International Technology, Education and Development Conference*. IATED, Valencia, Spain, 9002–9007. https://doi.org/10.21125/inted.2019.2239

[93] Lev V. Utkin and Andrei V. Konstantinov. 2022. Attention-based random forest and contamination model. *Neural Networks* 154 (Oct. 2022), 346–359. https://doi.org/10.1016/j.neunet.2022.07.029

[94] David C. D. van Alten, Chris Phielix, Jeroen Janssen, and Liesbeth Kester. 2020. Self-regulated learning support in flipped learning videos enhances learning outcomes. *Computers & Education* 158 (Dec. 2020), 104000. https://doi.org/10.1016/j.compedu.2020.104000

[95] Marcel V. J. Veenman, Bernadette H. A. M. Van Hout-Wolters, and Peter Afflerbach. 2006. Metacognition and learning: Conceptual and methodological considerations. *Metacognition and Learning* 1, 1 (April 2006), 3–14. https://doi.org/10.1007/s11409-006-6893-0

[96] Shang Wang, Deniz Sonmez Unal, and Erin Walker. 2019. MindDot: Supporting effective cognitive behaviors in concept map-based learning environments. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 1–14. https://doi.org/10.1145/3290605.3300258

[97] Richard T. Ward and Darrell L. Butler. 2019. An investigation of metacognitive awareness and academic performance in college freshmen. *Education* 139, 3 (March 2019), 120–126.

[98] Jennifer Wiley, Thomas D. Griffin, and Keith W. Thiede. 2005. Putting the comprehension in metacomprehension. *The Journal of General Psychology* 132, 4 (Oct. 2005), 408–428. https://doi.org/10.3200/GENP.132.4.408-428

[99] Philip H. Winne and Roger Azevedo. 2014. Metacognition. In *The Cambridge handbook of the learning sciences, 2nd ed.* Cambridge University Press, New York, NY, US, 63–87. https://doi.org/10.1017/CBO9781139519526.006

[100] Philip H. Winne and Allyson F. Hadwin. 1998. Studying as self-regulated learning. In *Metacognition in educational theory and practice*. Lawrence Erlbaum Associates Publishers, Mahwah, NJ, US, 277–304.

[101] Philip H. Winne and Nancy E. Perry. 2000. Measuring self-regulated learning. In *Handbook of Self-Regulation*. Elsevier, San Diego, CA, USA, 531–566. https://doi.org/10.1016/B978-012109890-2/50045-7

[102] Jacqueline Wong, Mohammad Khalil, Martine Baars, Björn B. de Koning, and Fred Paas. 2019. Exploring sequences of learner activities in relation to self-regulated learning in a massive open online course. *Computers & Education* 140 (Oct. 2019), 103595. https://doi.org/10.1016/j.compedu.2019.103595

[103] Zhihong Xu, Yingying Zhao, Jeffrey Liew, Xuan Zhou, and Ashlynn Kogut. 2023. Synthesizing research evidence on self-regulated learning and academic achievement in online and blended learning environments: A scoping review. *Educational Research Review* 39 (May 2023), 100510. https://doi.org/10.1016/j.edurev.2023.100510

[104] Litao Yan, Alyssa Hwang, Zhiyuan Wu, and Andrew Head. 2024. Ivie: Lightweight anchored explanations of just-generated code. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 1–15. https://doi.org/10.1145/3613904.3642239

[105] Chao Zhang, Xuechen Liu, Katherine Ziska, Soobin Jeon, Chi-Lin Yu, and Ying Xu. 2024. Mathemyths: Leveraging large language models to teach mathematical language through child-AI co-creative storytelling. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 1–23. https://doi.org/10.1145/3613904.3642647

[106] Lin Zhao and Chen Ye. 2020. Time and performance in online learning: Applying the theoretical perspective of metacognition. *Decision Sciences Journal of Innovative Education* 18, 3 (2020), 435–455. https://doi.org/10.1111/dsji.12216

[107] Zhichun Liu and Jewoong Moon. 2023. A framework for applying sequential data analytics to design personalized digital game-based learning for computing education. *Educational Technology & Society* 26, 2 (April 2023), 181–197. https://doi.org/10.30191/ETS.202304_26(2).0013

[108] Mingming Zhou. 2023. Students' metacognitive judgments in online search: A calibration study. *Education and Information Technologies* 28, 3 (March 2023), 2619–2638. https://doi.org/10.1007/s10639-022-11217-y

[109] Barry J. Zimmerman. 1989. A social cognitive view of self-regulated academic learning. *Journal of Educational Psychology* 81, 3 (1989), 329–339. https://doi.org/10.1037/0022-0663.81.3.329

[110] Barry J. Zimmerman and Adam R. Moylan. 2009. Self-regulation: Where metacognition and motivation intersect. In *Handbook of Metacognition in Education*. Routledge/Taylor & Francis Group, New York, NY, US, 299–315.

[111] Barry J. Zimmerman and Manuel M. Pons. 1986. Development of a structured interview for assessing student use of self-regulated learning strategies. *American Educational Research Journal* 23, 4 (1986), 614–628. https://doi.org/10.2307/1163093 Place: US Publisher: American Educational Research Assn.

A DETAILED REGRESSION ANALYSIS RESULTS

A.1 RQ1 Regression results

A.2 RQ2 Regression results

Table 3: The *t*-test results on metacognitive calibration improvement between control and intervention conditions, analyzed across student initial confidence levels, gender, and race/ethnicity.

	Group	Number of Students	Metacognitive calibration improvement	Statistics (*t*, *p*)
Initial confidence	Initial underconfidence	Control (n = 19)	12.7%	t = 0.293
		Intervention (n = 35)	15.0%	p = .771
	Initial overconfidence	Control (n = 31)	12.2%	t = 2.001
		Intervention (n = 42)	16.3%	p = .049
Gender	Female	Control (n = 28)	16.0%	t = 0.316
		Intervention (n = 34)	16.8%	p = .753
	Male	Control (n = 22)	14.3%	t = 1.513
		Intervention (n = 42)	19.1%	p = .135
Race/ethnicity	White	Control (n=19)	16.0%	t = 0.704
		Intervention (n = 25)	19.1%	p = .486
	Black	Control (n = 13)	20.3%	t = 0.396
		Intervention (n = 21)	15.0%	p = .695
	Asian	Control (n = 6)	11.7%	t = 1.716
		Intervention (n = 8)	20.1%	p = .107
	Multiracial	Control (n = 7)	6.5%	t = 1.999
		Intervention (n = 17)	17.8%	p = .058

Table 4: Table of regression results for RQ1. In the race/ethnicity category, White serves as the reference group, while Male is the reference group for the gender category.

Dependent Variable	Independent Variable	Coefficient	*p*-value
Learning Gain	Intercept	6.114	.334
	Group	9.283	.248
	Black	6.545	.426
	Asian	6.269	.487
	Hispanic/Latinx	-11.865	.289
	Multiracial	9.927	.343
	Female	-2.290	.725
	Group × Black	-4.663	.659
	Group × Asian	0.205	.987
	Group × Hispanic/Latinx	7.656	.600
	Group × Multiracial	-4.837	.699
	Group × Female	0.630	.939