How Adding Metacognitive Requirements in Support of AI Feedback in Practice Exams Transforms Student Learning Behaviors

Mak Ahmad
University of California, Davis
Davis, CA, USA
shahmad@ucdavis.edu

Prerna Ravi
Massachusetts Institute of Technology
Cambridge, MA, USA
prernar@mit.edu

David Karger
Massachusetts Institute of Technology
Cambridge, MA, USA
karger@mit.edu

Marc Facciotti
University of California, Davis
Davis, CA, USA
mtfacciotti@ucdavis.edu

Abstract

Providing personalized, detailed feedback at scale in large undergraduate STEM courses remains a persistent challenge. We present an empirically evaluated practice exam system that integrates AI generated feedback with targeted textbook references, deployed in a large introductory biology course. Our system specifically aims to encourage *metacognitive behavior* by asking students to *explain their answers* and *declare their confidence*. It uses OpenAI's GPT-4o to generate personalized feedback based on this information, while directing them to relevant textbook sections. Through detailed interaction logs from consenting participants across three midterms (541, 342, and 413 students respectively), totaling 28,313 question-student interactions across 146 learning objectives, along with 279 post-exam surveys and 23 semi-structured interviews, we examined the system's impact on learning outcomes and student engagement. Analysis showed that across all midterms, the different feedback types showed no statistically significant differences in performance, though there were some trends suggesting potential benefits worth further investigation. The system's most substantial impact emerged through its required confidence ratings and explanations, which students reported transferring to their actual exam strategies. Approximately 40% of students engaged with textbook references when prompted by feedback—significantly higher than traditional reading compliance rates. Survey data revealed high student satisfaction (M=4.1/5), with 82.1% reporting increased confidence on midterm topics they had practiced, and 73.4% indicating they could recall and apply specific concepts from practice sessions. Our findings demonstrate how thoughtfully designed AI-enhanced systems can scale formative assessment while promoting sustainable study practices and self-regulated learning behaviors, suggesting that embedding structured reflection requirements may be more impactful than sophisticated feedback mechanisms.

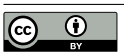
CCS Concepts

• **Human-centered computing** → **Empirical studies in HCI**; • **Computing methodologies** → *Artificial intelligence*; • **Applied computing** → **Interactive learning environments**; *Computer-assisted instruction*; • **Social and professional topics** → **Student assessment**.

Keywords

AI-enhanced feedback, Practice exams, Self-regulated learning, Metacognition, Learning at scale, Confidence ratings, Student explanations, Textbook engagement, Higher education, Biology education

ACM Reference Format:
Mak Ahmad, Prerna Ravi, David Karger, and Marc Facciotti. 2025. How Adding Metacognitive Requirements in Support of AI Feedback in Practice Exams Transforms Student Learning Behaviors. In *Proceedings of the Twelfth ACM Conference on Learning @ Scale (L@S '25), July 21–23, 2025, Palermo, Italy*. ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3698205.3729542



L@S '25, Palermo, Italy

ACM ISBN 979-8-4007-1291-3/2025/07
https://doi.org/10.1145/3698205.3729542

1 Introduction

It is a truism among faculty that students do not read the textbook [10, 16]. In contrast, our experience is that students are strongly motivated to tackle practice exams in the run-up to actual exams [6, 40]. This paper describes an experimental deployment, in a class of 1002 students, of a system we developed that aimed to turn practice exams into a teaching tool. While the AI features meant to be the system's centerpiece did not drive improvements in student exam performance, two other deliberate design choices aimed at shaping student engagement with the AI feedback—requiring students to 1) rate their confidence and 2) explain their reasoning while attempting every question—were unexpectedly impactful. These features had unanticipated positive effects on students' metacognition, internal motivation, and long-term exam preparation strategies.

To make multiple-choice practice exams more interactive we developed a tool that could give students *formative feedback* [19, 28] on their answer choices, rather than simple right and wrong responses. We used a Large Language Model (LLM) to either (i) guide students to an appropriate place in the textbook to acquire the knowledge they needed to answer correctly or (ii) *generate*

textual guidance about the material the student was getting wrong. To make this feedback more targeted, our design required that each student *explain* their answer when they submitted it; this explanation could help the LLM understand what the student did not understand, in order to steer them to the right textbook section or to generate suitable corrective guidance. Based on analysis of prior educational feedback literature, we also surmised that it would be beneficial to record students' *confidence* in their answer, and use that to further customize the feedback [19, 28].

We ran a randomized controlled trial in a 1002-student general biology class, comparing these two feedback approaches (guide-to-textbook and AI-generated feedback) to a control that only told students whether their answers were correct. Contrary to our expectations based on prior work on elaborated feedback, we found no significant difference between treatment conditions. Instead, the most substantial impact came from the metacognitive requirements we had initially included as supporting elements for the feedback system. This finding suggests that structured reflection may be more valuable for student learning than the specific content of feedback provided.

While the benefits of practice testing are well-documented [6, 40], providing effective, personalized feedback to large student populations remains a significant challenge [4, 56]. The emergence of sophisticated AI systems capable of generating contextual feedback offers promising new approaches to this challenge [50, 55], especially when considered within the framework of self-regulated learning and student success [19, 56], though careful integration is needed to balance AI assistance with student agency [1].

Traditional practice exam approaches face several key limitations. Static feedback and delayed grading often fail to provide students with the timely guidance they need [4, 29], while faculty time constraints make it difficult to offer detailed, personalized feedback to hundreds of students [45]. Furthermore, students frequently focus on their numerical grades rather than developing deeper understanding of the course material [53, 56]. This tendency is exacerbated by the perceived disconnect between exam performance and specific textbook content, making it challenging for students to effectively address their knowledge gaps [9, 26].

Because students are motivated by practice exams, we aimed to turn them into better teaching tools by providing customized feedback on student responses. The theory of effective feedback systems draws from several key education research areas. Self-regulated learning theory emphasizes the importance of helping students develop metacognitive strategies and self-assessment skills [19, 56]. Student confidence plays a crucial role in both learning and assessment, influencing how feedback is received and integrated [19, 28]. This work suggests that effective feedback systems should not only provide accurate information, but also support students in developing better study strategies and engaging more meaningfully with course materials.

Despite growing interest in AI-enhanced educational tools, significant gaps remain in our understanding of their effectiveness. Limited research exists on AI-generated feedback in authentic learning contexts, particularly regarding how different types of feedback compare in supporting student learning. While research has established a connection between student confidence and the effectiveness of feedback [19, 27], its application in the context of AI-powered feedback remains unexplored. While research has established ways to improve alignment between instruction and assessment [2, 9], the impact of directly connecting feedback to specific textbook content remains unexplored, particularly in the context of AI-enhanced learning systems at scale.

To address these gaps, this study investigates two areas:

RQ1: Does supplementation of traditional midterm exam study resources with an AI-powered practice exam system with textbook linking impact student exam performance and related study behaviors?

RQ2: How do design elements of the system accompanying AI feedback influence student attitudes, engagement, and self-regulated learning behaviors?

This study introduces a novel approach to practice exam feedback by combining AI-generated explanations with direct textbook references, implemented in a large enrollment undergraduate biology course. The use of required confidence ratings and student explanations as personalized inputs for the LLM also introduces a new dimension to the AI feedback literature. By examining both learning outcomes and student engagement patterns, we contribute to the understanding of how technology can enhance educational feedback at scale.

2 Related Work

2.1 Feedback in Educational Settings

Prior work has established foundational insights into how *timing* and *type* of feedback could influence the quality of learning.

Studies have shown that merely indicating correctness provides insufficient scaffolding for conceptual growth, whereas elaborative feedback explaining *why* an answer was correct or incorrect fostered better retention and application [28]. Researchers have devised models to assess effective feedback using three major questions asked by teachers and/or students: 1) *Where am I going?* (goal alignment), 2) *How am I going?* (progress evaluation), and 3) *Where to next?* (future guidance) [19]. This framework emphasizes moving beyond binary correctness indicators to address cognitive processes and self-regulation, aligning feedback with learners' developmental trajectories. This distinction has been corroborated by longitudinal cognitive diagnostic assessments, which found that cognitive diagnostic feedback tailored to learners' mastery of specific attributes significantly outperformed traditional correct-incorrect response feedback in challenging domains [50]. Similarly, studies have shown that high-information feedback (combining task-level correctness with self-regulation from monitoring attention, emotions, or motivation during learning) yielded better effects than simple forms of reinforcement or punishment at the task level [19, 34, 56]. Such findings underscore the importance of granular, attribute-level feedback in addressing knowledge gaps.

Timing further modulates feedback's utility [19]. Immediate feedback seems more effective for learning procedural skills and correcting errors in practice, while delayed feedback appears to better support long-term retention of conceptual knowledge [4, 29, 50]. Finally, learner confidence and self-assessment profoundly influence feedback efficacy, as overestimations or underestimations of competence can distort receptivity. Researchers found that students who were confident in their answers benefited most from

simple feedback about correctness, while students with low confidence needed more detailed guidance explaining underlying concepts and connecting them to foundational knowledge they might be missing [19, 27]. Furthermore, research has shown that self-regulated learners—those adept at monitoring and adjusting their strategies—derive more benefit from feedback than peers reliant on external guidance, emphasizing the interplay between feedback design and metacognitive skills [5]. This approach leverages metacognitive awareness, enabling learners to reconcile discrepancies between perceived and actual understanding.

Our study builds upon these foundational insights by integrating automated feedback with confidence-aligned strategies and textbook references, aiming to provide personalized, high-information feedback at scale while promoting self-regulated learning behaviors in a large undergraduate biology course.

2.2 Textbook Reading Practices in Higher Education

While academic reading fosters disciplinary discourse and improves writing skills [23, 33, 43], studies across disciplines reveal low compliance with reading assignments, with as few as 27% of students completing readings before class [10] and 72% admitting to rarely or never reading on schedule [11, 46]. Reasons for this include competing demands on time, lack of motivation, and misalignment between student and faculty expectations [39, 41, 47]. While motivated students with a high need for cognition or mastery-oriented goals may engage with textbooks more deeply [14, 47], others perceive reading as effortful and unnecessary if lectures or other resources provide sufficient information [37, 42, 46, 53]. Furthermore, students struggle when faculty do not explicitly integrate assigned readings into classroom discussions or assessments, reinforcing the perception that reading is optional [3, 16, 21, 35, 47].

Scholars have devised ways to improve student engagement with reading assignments. Prior work has found that introducing just-in-time quizzes during reading or in-class quizzes following a chapter's reading assignment can improve reading habits [18, 20, 22, 25, 46]. Beyond quizzes, question-based approaches foster self-assessment; research has also found that asking students to pose a question after reading the textbook enhances comprehension and leads to better exam performance [36, 54]. Critical scoped reading has also been emphasized through structured prompts [52]. Different reading formats have also been explored. While some studies suggested readers (a thematic compilation of excerpts from different authors grouped to explain a topic), over traditional textbook formats for increasing engagement and critical thinking [22], others found textbook format choice had little impact [15]. Other approaches, such as experiential learning models [48] and self-monitoring strategies [7], also aim to enhance reading compliance and academic performance.

Current practices also reveal fundamental gaps between course materials and assessments. Recent work highlights that assessment and instruction are often conceived as separate in both time and purpose: without explicit links between both, students struggle to recognize how instructional materials—including readings—contribute to building the skills needed for success on exams [26]. The constructive alignment framework [2] emphasizes the need for assessments that are directly tied to instructional content, ensuring coherence between teaching methods, evaluation, and learning objectives. Embedding instructional resources directly into assessments may also provide students with immediate access to relevant materials at critical learning moments [9], addressing the challenge of expecting students to independently seek out these resources. Strengthening the link between course materials, assessments, and feedback can support deeper engagement and knowledge retention.

Our study addresses these gaps by embedding links to relevant textbook sections within practice assessments, alongside AI-generated personalized feedback. We frame these references as optional, goal-oriented learning supports—designed not as required reading, but as resources students can actively use to deepen their understanding of course objectives and see clear, actionable connections to the exam questions they find most relevant.

2.3 AI Systems in Educational Assessment

Researchers found that generative AI offers the potential to deliver immediate and diverse feedback, along with opportunities for self-assessment, across a wide variety of learning contexts. This can encourage students to study independently and develop strong self-regulation, including goal setting, self-monitoring, self-assessment, and adaptive learning strategies [57]. For example, recent papers argue that AI-powered assessment tools can analyze student explanations, identifying concepts and scientific principles that may be missing or misapplied, while also making suggestions for how instructors can use these data to better guide student thinking [26]. LLM-powered reflection prompts have been shown to significantly improve student learning outcomes at scale [30]. Additionally, researchers reported that AI-based tools enhance real-time formative feedback by supporting personalized learning, adaptive test adjustments, and real-time classroom analysis, with students expressing strong support for these capabilities [55].

Despite these potential benefits, AI-powered feedback presents several challenges. AI systems may assess student responses using criteria different than course instructors', potentially leading to misaligned feedback without sufficient course-specific context [38, 57]. Students who use LLMs as personal tutors by conversing about the topic and asking for explanations benefit from usage, whereas learning is impaired for students who excessively rely on LLMs to solve practice exercises for them [31]. This suggests that AI-powered feedback is most effective when used as a supplement to student engagement rather than as a replacement for active learning. Excessive reliance on AI feedback can discourage students from developing important critical thinking skills through independent problem-solving [24]. A recent study also showed AI feedback deployed at scale that addresses several effective feedback components at the task and process level but still misses self-reflection feedback elements needed to motivate students [13].

Existing studies on AI-generated feedback in education also face limitations due to small sample sizes [17, 49], indicating the challenge of assessing the generalizability of AI-based interventions in different educational contexts. Several challenges complicate large-scale deployment of AI-powered assessment tools: instructors need time to learn and integrate new assessment types into their courses, technical infrastructure must be reliable and accessible to all students, and the systems must maintain consistent performance

throughout the term [45]. The need for instructors to verify the accuracy and pedagogical appropriateness of AI-generated feedback increases resource demands when implementing personalized feedback at scale [38].

In this paper, we present a novel practice exam tool for midterm preparation that combines AI-generated feedback with self-regulatory learning components. We propose strategies to balance automation with pedagogical rigor, ensuring that AI-driven feedback enhances learning through structured metacognitive elements like confidence ratings and reasoning explanations. Analyzing the effects of AI-generated feedback at scale, we build on previous controlled studies of automated feedback [32].

3 System Design and Implementation

Our system aimed to provide a deeper version of the typical interactive practice exam experience for a 1002-student undergraduate biology course. The designed workflow is as follows:

(1) We created a *question bank* of 400 multiple-choice practice exam questions (150 image-based) across the three midterms, each labeled with relevant *learning objectives*. We then prepared a small number of practice exams each drawing 25 questions from the question bank.
(2) Given a particular practice exam, in *test mode* (Figure 1) students answer each question. In addition to choosing a multiple choice answer, students must indicate a *confidence level* and provide an *explanation* for their choice
(3) In *review mode* (Figure 2) the system provides feedback on each answered question. In addition to reporting right-or-wrong, the system can provide *textbook links* to relevant content, and/or *custom feedback* generated by an LLM from the student's choice, confidence, and explanation.

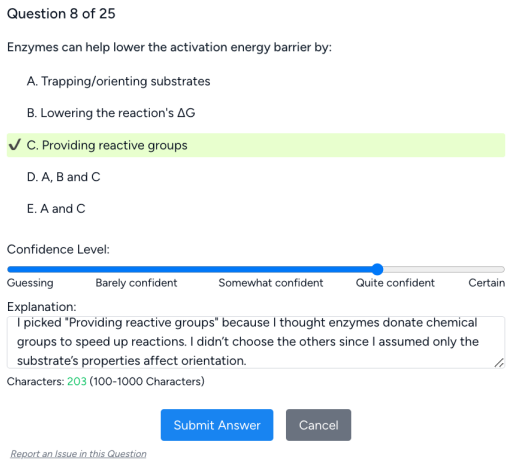


Figure 1: Test mode interface where students provide their answer, confidence rating, and reasoning.

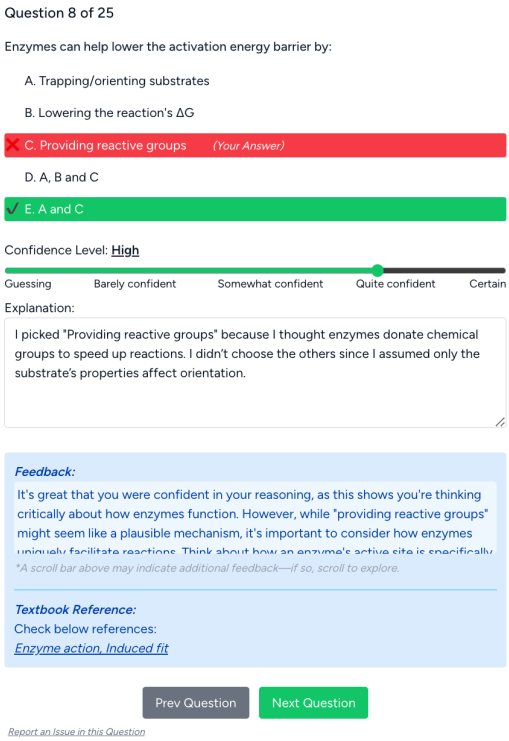


Figure 2: Review mode interface showing condition-specific feedback and textbook references in a scrollable panel.

3.1 Question Bank Development and Learning Objectives

We worked with the course instructor to validate learning objective mappings for approximately 400 multiple-choice questions from previous exams, including 150 image-based questions. To ensure reliable AI feedback generation, we identified relevant textbook sections and generated comprehensive summaries for each learning objective using Claude 3.5 Sonnet, which were then validated by the instructor. These summaries and textbook mappings were included in feedback prompts to constrain GPT-4o's responses to validated biological content rather than its general knowledge, significantly reducing the risk of hallucination. The entire data preparation process took over 30 hours.

3.2 Practice Exam Implementation and Deployment

The practice exam structure mirrored the actual midterm format while incorporating experimental elements. Each practice exam contained 25 questions, carefully selected to ensure coverage of all relevant learning objectives while maintaining the integrity of the condition assignment system. Students were required to provide both a confidence rating and written explanation for each answer before proceeding. Motivated by the related work on effective feedback timing, we implemented a two-phase approach. In the initial attempt phase, students answered all questions without immediate feedback. Upon completing the entire exam, students

entered a review mode that presented their responses alongside condition-specific feedback. This design choice was informed by research suggesting benefits of delayed feedback in certain learning contexts.

3.3 Feedback Generation

The AI feedback component used OpenAI's GPT-4o API. We selected GPT-4o as it is currently one of the most capable publicly available language models, with documented performance on complex reasoning tasks and educational content generation. The feedback generation pipeline begins by assembling a comprehensive context that includes the question content, correct answer, student's selected answer and explanation, relevant learning objective summaries, and the student's reported confidence level. This context feeds into a carefully engineered prompting system that constrains responses to 5-7 sentences while ensuring pedagogically appropriate tone and relevant textbook references.

The base prompt established guidelines for encouraging, constructive feedback that addressed specific elements of student reasoning without directly stating correct answers for incorrect responses. Feedback varied based on both correctness and confidence level - for instance, high-confidence incorrect answers received feedback that gently highlighted inconsistencies while guiding students to identify specific logical flaws. For low-confidence correct answers, the system focused on building confidence by connecting correct intuitions to fundamental principles. All prompts underwent multiple iterations of testing with course instructors to ensure alignment with course objectives.

To maintain reliability, the system implements robust error handling and fallback mechanisms. When API responses fail or take too long, the system gracefully degrades to simpler feedback modes (changing the experimental condition) rather than leaving students without guidance. The presentation layer implements a scrollable interface with careful attention to user experience, including visual cues encouraging students to engage with the full feedback content.

3.4 Interaction Tracking and Analytics

The system implements a comprehensive interaction tracking to understand how students engage with different types of feedback. Beyond basic metrics like time spent per question, we designed specific mechanisms to capture detailed interaction patterns. The feedback interface uses a deliberately constrained window height, enabling us to track whether students initiate any scrolling and whether they reach the end of the feedback content. For each question attempt, the system records temporal data between initial question load and final submission, as well as any revised answers and explanation changes during review mode. To understand varying study contexts, the system tracks device types and browsers used in both test and review modes, allowing analysis of different access patterns across these phases. All interaction data, including mouse hover patterns and textbook reference engagement, is stored in a structured format with timestamps and session identifiers. This data model enables correlation between interaction patterns and learning outcomes while controlling for variables like confidence levels and prior performance, all while maintaining reasonable storage requirements for our large student population.

3.5 Infrastructure and Technical Stack

The application (shown in figure 3) was built using Laravel 11.34 and Vue.js, hosted on AWS infrastructure. The data model centered on four key entities: a question bank containing 400 multiple-choice questions with associated metadata, learning objectives mapped to textbook sections, student profiles managing authentication and experimental conditions, and comprehensive interaction logs. Question images were stored in a GitHub repository for version control and reliable delivery. The backend uses PHP 8.3.14, MySQL 8.0.40, and Nginx for web serving. The system used Google OAuth integration configured for university email addresses, ensuring secure access while minimizing friction in the student experience. This approach maintained consistent experimental conditions across sessions while restricting access to enrolled students.

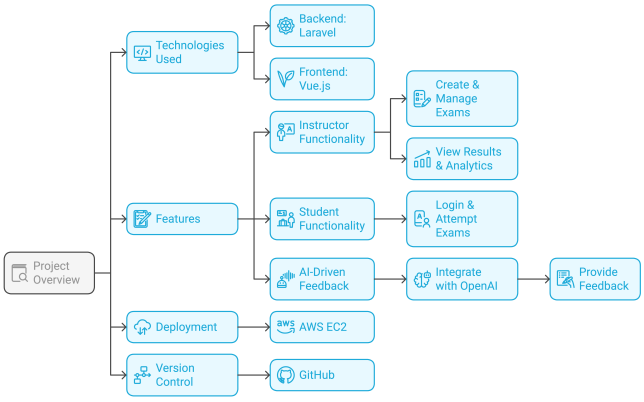


Figure 3: Technologies used to build our system.

4 Experimental Methodology

4.1 Study Context and Participants

This study was conducted in a first-year undergraduate general biology course offered at UC Davis. The course has 3 midterm exams. Use of the practice tool was assigned for credit as preparation for midterms 1 and 3 and made optional for midterm 2.

We studied 1002 students enrolled in a single offering of the course. The practice exam tool was deployed one week before each midterm, allowing students sufficient time to engage with the platform while maintaining proximity to the actual assessment. Prior to implementation, the IRB of the hosting institution deemed the project exempt from full review (IRB 1456274-2). All participating students consented to their data being used for research purposes.

In total, 836 students (541 consenting to research) used the system for midterm 1, 760 students (342 consented) voluntarily used it for midterm 2, and 877 students (413 consented) for midterm 3, generating 28,313 question-student interactions. The high ongoing participation demonstrated sustained engagement with the tool.

4.2 Experimental Design and Condition Assignment

We ran a randomized controlled trial over four feedback conditions. Condition 1 only indicates whether the student's answer was right

or wrong. Condition 2 augments the responses in the review phase with links to specific relevant (online) textbook sections. Condition 3 in the review phase delivers AI-generated personalized feedback using GPT-4o, factoring the student's explanation and confidence level to provide targeted guidance. Condition 4 provides both the textbook links of Condition 2 and the personalized feedback of Condition 3. Questions were carefully distributed across conditions to ensure each student experienced all feedback types while maintaining learning objective independence.

The system maintains consistent assignment of experimental conditions. Instead of randomly assigning conditions each time a student interacts with a question, the system implements a deterministic assignment mechanism based on a hash function that considers the IDs of student, question, and practice exam. This ensures that students consistently experience the same experimental conditions for specific questions across multiple sessions, while maintaining an even distribution of conditions across students.

4.3 Quantitative Analysis Framework

Our analysis examined feedback condition impacts on student performance through both ANOVA and linear regression. The primary unit of analysis was student performance on individual midterm questions, mapped to practice questions through learning objectives. We employed one-way ANOVA with Tukey's HSD test ($\alpha = 0.05$) to identify differences between conditions, supplemented by a linear regression model accounting for practice exposure. The model used question-level scores as the dependent variable, with condition dummy variables and an exposure measure (number of relevant practice questions attempted per learning objective) as predictors.

To analyze how practice exam performance predicted real exam success, we mapped questions through learning objectives, with each question testing multiple objectives (average 2.3 per question). We weighted objectives equally within questions and calculated overlap weights between practice and real exam questions based on shared objectives. For example, if a practice question tested three objectives (GC.31, GC.32, GC.33) and a real exam question tested two of these (GC.31, GC.32), the overlap weight would be 0.67, representing the proportion of shared objectives.

4.4 Survey Assessment

A post-midterm survey (n=279, 25.8% response rate) collected data across four categories: study habits and tool usage, feedback effectiveness, learning behaviors, and tool experience. Questions included multiple-choice, 5-point Likert scales, and open-ended responses to assess both quantitative impact and qualitative experiences with the system.

4.5 Qualitative Interview Study

To understand how students integrated the practice exam system into their study processes, the two lead authors conducted 23 semi-structured remote interviews with 19 participants via Zoom, each lasting 30-45 minutes. Four participants were interviewed twice– during and after the course– to examine whether and how their experiences with the tool influenced future study practices. We recruited these students via course announcements and incentivised participation with a $10 gift card. Students in our interview sample had varied backgrounds in their motivations for taking the course, including fulfilling major requirements, preparing for medical school, or exploring biology as a field. Their grade expectations ranged from simply passing with a C to aiming for a high A. Additionally, they differed in their preferred exam preparation strategies, the resources they prioritized, and the timing of their practice exam usage within their study schedules. The interview protocol explored five areas: study strategies and resource usage for exam preparation, tool usage and experience when attempting the exam, feedback and learning, perceived impact on midterm performance, and feature suggestions for future integration. We investigated how students engaged with different feedback types, their experience with confidence ratings and explanation requirements, and instances where the system influenced their understanding. Follow-up interviews with the four returning participants allowed us to investigate the persistence of study behaviors across courses and the tool's long-term influence on metacognitive study habits.

Following the interviews, the two authors conducted an inductive thematic analysis of the transcripts, allowing themes to emerge organically from the data [8, 12]. The process began with an initial review of the transcripts, during which notes were taken. A second reading facilitated the development of preliminary codes refined through discussion and iteration between the authors. After coding, related quotes were organized into overarching themes and categories. These themes and their definitions were further reviewed and refined until consensus and saturation were achieved. This qualitative analysis provided valuable context for interpreting the quantitative findings from the survey and system interaction logs.

5 Results

5.1 Model Diagnostics

Diagnostic tests indicated minimal autocorrelation (Durbin-Watson = 1.764) and acceptable multicollinearity (condition number = 10.8). While the Jarque-Bera test (JB = 2846.646, p < 0.001) suggested non-normality in the residuals, this doesn't invalidate our regression results given our large sample size (n = 10,820). With samples this large, the sampling distribution of regression coefficients approaches normality regardless of residual distribution, allowing for reliable statistical inference despite the normality violation.

5.2 RQ1: Impact on Performance

Analysis of variance across all three midterms showed no consistently significant differences between feedback conditions ($F = 2.139$, $p = 0.093$). While pairwise comparisons using Tukey's HSD test suggested some differences between conditions, these did not reach statistical significance when considering the complete dataset.

The linear regression model using Condition 1 as baseline indicated a positive trend for Condition 2 (textbook references) compared to basic feedback, equivalent to approximately 0.7% improvement in performance, though this difference was not statistically significant ($\beta = 0.0073$, $p = 0.328$). Similarly, other conditions showed trends but no statistically significant differences. Mean scores across all midterms by condition were:

- Condition 1 (basic feedback): $M = 0.730$, $SD = 0.444$, $n = 7698$

Themes	Definition
A: Background Knowledge	
Course Importance & Motivation	How students perceive the importance of [Anonymous course] in their academic and career trajectory, including their motivation for taking the class and desired grade outcomes.
Study Strategies & Preparation Methods	The approaches students take to prepare for the midterm, including study materials, resources, and scheduling (including when they used the practice exam).
Identifying Knowledge Gaps & Study Challenges	The ways students assess their understanding and the difficulties they encounter while studying.
General Role/Importance of Practice Exams	How students use practice exams (both PDF and online tool) to test their knowledge and adjust their study strategies.
PDF vs. Online Tool Preferences	The perceived advantages and disadvantages of the PDF practice test compared to the online tool.
B: Perceptions and Self-regulation during ATTEMPT/TEST Mode	
Expectations & First Impressions of the Online Tool	Students' initial reactions to using the online practice tool and how it compared to their expectations.
Confidence Ratings Perception during ATTEMPT mode	The impact of rating confidence levels (e.g. on self-awareness) when attempting the question
Explanation Requirement Response	Details on perception towards writing down thought process and its usefulness
Confidence Ratings Use during REVIEW mode	The impact of confidence levels on error correction, feedback attention, perceived knowledge mastery, etc. when reviewing feedback
C: Engagement with Feedback and Learning Resources during Review Mode	
Perceived effectiveness of AI generated feedback	Student engagement with AI-generated explanations and whether they found them useful for understanding concepts. Including anecdotes
Feedback Awareness Gap	Student understanding (or lack thereof) about feedback availability being condition-dependent; includes misconceptions about technical issues vs. experimental conditions
Engagement with Textbook Links	How and when students used the textbook links provided within the tool. Including anecdotes
Engagement with External Resources	Instances when students tried to seek additional explanations/resources outside of the tool
D: Perceived Effects on Midterm and Knowledge Transfer	
Perceived Midterm Performance	The extent to which students felt that the practice exams prepared them for the actual midterm.
Learning Transfer Evidence	Specific instances where tool usage directly aided exam performance; includes concept recognition and application
E: Future Directions	
Usability & Technical Challenges with the Tool	Details on tool's functionality and ease of use
Future Study Approaches	How using the tool influenced students' preparation for future exams and courses.
Student-Driven Suggestions for Tool Improvement	The changes students feel would make the tool more effective.

Table 1: Final set of interview categories (A-E), themes under each, and their definitions

- Condition 2 (textbook references): $M = 0.739$, $SD = 0.439$, $n = 6917$
- Condition 3 (AI feedback): $M = 0.745$, $SD = 0.436$, $n = 6723$
- Condition 4 (AI + textbook): $M = 0.726$, $SD = 0.446$, $n = 6975$

The most consistent finding across all midterms was that student confidence ratings strongly predicted performance ($\beta = 0.063$, $p < 0.001$), with higher confidence associated with approximately 6.3% better performance on exam questions. This robust relationship reinforces the value of metacognitive elements in the system.

A separate analysis of lower-performing students (those scoring in the bottom 20% on initial practice exams) showed a potentially different pattern. While the overall model approached but did not reach statistical significance ($F = 2.48$, $p = 0.059$), there was a trend suggesting that Condition 4 (AI + textbook) might provide additional benefit compared to baseline ($\beta = 0.049$, $p = 0.067$). This represents a potential performance improvement equivalent to approximately one additional correct question on a 20-question exam, and could indicate that combined feedback might be more beneficial for students who would benefit from additional learning support—a hypothesis warranting further investigation.

5.2.1 Survey Results. Based on 279 survey responses, students rated different feedback types on a 5-point scale. Correctness feedback received the highest average rating ($M = 3.82$, $SD = 0.94$), followed by combined AI and textbook feedback ($M = 3.52$, $SD = 1.13$), AI-generated explanations alone ($M = 3.44$, $SD = 1.08$), and textbook references ($M = 3.41$, $SD = 1.12$). A strong majority (82.1%) of survey respondents reported increased confidence on the midterm for topics they had practiced using the system. When asked about concept recall, 73.4% of survey respondents indicated they could specifically remember and apply concepts from their practice sessions during the actual midterm.

5.3 RQ2: Impact on Engagement and Learning Behaviors

5.3.1 Survey Results. Following midterm 1, we conducted a post-exam survey that received 279 responses (25.8% response rate). These survey results reflect student experiences with the system during preparation for the first midterm examination only.

Of the total 1002 enrolled students, 804 (80.2%) used the online practice exam tool for midterm 1, with 541 consenting to have their data included in the research. Among the survey respondents, 65.9% reported using both the online tool and a separate PDF practice exam, while 26.9% reported using only the online tool. The PDF practice exam contained different questions than the online version, offering complementary practice opportunities.

Regarding tool satisfaction after midterm 1, students rated the system's ease of use favorably ($M = 4.1/5$, $SD = 0.89$). 76.3% of survey respondents indicated they would use it again for future exams - a prediction validated by the 76% voluntary adoption rate when the tool became optional for midterm 2. The most frequently requested improvement was enhanced AI explanations (68.4% of respondents), particularly for complex topics and incorrect answers.

Student survey responses indicated changes in study behaviors after using the online tool for midterm 1. Among those who used the tool, 73.4% reported adjusting their study approaches based on the feedback received. Analysis of midterm 1 interaction data showed that approximately 28% of students clicked on provided textbook links, with similar rates between those who received only textbook references and those who received both AI feedback and textbook references. By midterm 3, this engagement with textbook references increased to approximately 39%.

5.3.2 Qualitative insights from interviews. For this paper's RQs, we focus on themes from categories B, C, and D (See Table 1) below.

Student perceptions and self-regulation during test mode: The integration of confidence ratings and explanation requirements

in the online practice tool played a significant role in shaping students' self-assessment, engagement with the material, and overall learning experience. Their effectiveness however varied based on individual study habits, highlighting both strengths and challenges of incorporating structured reflection into an assessment tool.

For many of the interviewees, the explanation requirement functioned as a tool for deeper thinking and self-reflection. P6 found that it forced them to justify their choices rather than rely on intuition, stating, *"Sometimes I'd think, I'm pretty sure it's this,' but then I'd realize, I don't actually have a reason to think that."* Similarly, others noted that the process of writing out explanations made them question their assumptions and reconsider their answers, and distinguish between educated guesses and actual understanding, a level of engagement they did not typically experience when using traditional PDF practice exams. Participants even stated that writing their reasoning sometimes increased their confidence, as they could clearly see their logical progression. Beyond reinforcing content knowledge, these features also helped students develop metacognitive skills. P13 compared the explanation requirement to their preferred study method of teaching concepts to friends, noting that *"it was almost like I was trying to convince myself why I thought this answer was the best."* P16 reported a notable shift in their study approach beginning with Midterm 2, where they started crafting more elaborate explanations, even for questions they felt uncertain about. They began associating the depth of their explanations with a clearer understanding of concepts—demonstrating increased metacognitive awareness and internal motivation. This change was also reinforced by a noticeable improvement in the quality of AI feedback P16 received: when their explanations were more detailed, the AI responses became more substantive and helpful.

Despite the benefits, the reflection components were not universally appreciated. Some students viewed them as a waste of their time and admitted to writing filler responses, especially when they had absolutely no idea what the answer was. P2 acknowledged the psychological benefits of reflection but found it unnecessary for questions where they were "100% certain." P17, who reported having ADHD, found that the requirement was mentally exhausting, particularly when applied to every question. They suggested that it should be optional for straightforward problems, allowing students to focus their energy on more complex ones. The confidence feature also had its detractors. Some students struggled with rating their confidence accurately, often defaulting to a neutral rating on the scale. They rarely selected the highest confidence level, noting that they generally did not feel fully confident in their answers until they received confirmation. P9, who described themselves as an overthinker, found the confidence slider challenging because they could always identify both strengths and weaknesses in their reasoning, leading them to repeatedly select "somewhat confident." P19 appreciated that the confidence slider forced them to *"be honest with myself"* rather than instinctively selecting an answer, but they also noted that it made the practice exam take significantly longer to complete. In contrast, P7 found an unexpected benefit in the system—when they marked *"just purely guessing"* in the confidence slider, they received more detailed feedback, which ultimately improved their confidence when retrying those questions. P16, who initially felt stressed about articulating their reasoning, adapted by writing about what confused them, which sometimes led to realizations that they knew more than they had initially thought.

Student engagement with feedback and learning resources during review mode: Students' interaction with AI-generated feedback and their study habits were heavily shaped by the confidence ratings they assigned during their initial attempts. Many students reported that their confidence level determined how thoroughly they engaged with the feedback and whether they consulted additional resources. Student engagement with feedback varied significantly based on the intersection of confidence and correctness. As reported in our survey, students were most likely to thoroughly engage with feedback when their confidence level mismatched the outcome—specifically when high-confidence answers were incorrect or low-confidence answers were correct. P8 noted that when they got a low-confidence question correct, they reviewed the feedback to ensure that their reasoning aligned with the correct answer, rather than assuming they had simply guessed correctly. In contrast, if they were highly confident and answered correctly, they often skimmed the feedback for confirmation or disregarded it entirely. Finally, moments where students felt completely certain about an answer but discovered they were wrong prompted the strongest engagement—these instances led them to scrutinize the feedback, revisit textbook readings, and seek further clarification from external sources to resolve their misconceptions. However, not all students found the confidence ratings beneficial during review, particularly those who struggled with accurately assessing their own confidence from the outset. A smaller subset of students largely disregarded their confidence ratings altogether, prioritizing whether their answers were correct or incorrect instead. Notably, these were often the same students who had found the confidence rating requirement cumbersome during the initial attempt phase.

Students had mixed perceptions of the AI-generated feedback, with some finding it highly useful for correcting misconceptions and others criticizing its limitations. P1 explained that the AI feedback was most helpful when they already had some understanding of the concept because it provided additional details that reinforced their knowledge. For some students, the AI feedback served as a scaffold to refine their reasoning. P5 appreciated that the feedback did not simply confirm whether an answer was right or wrong but also analyzed their explanation, pointing out gaps in their reasoning. They found this especially valuable when the feedback suggested alternative ways of thinking about a problem. P10 recalled an instance where they received AI feedback that critiqued their explanation despite selecting the correct answer, which helped them realize that their reasoning was flawed and could lead to mistakes in future assessments. P12 found it particularly helpful in making interdisciplinary connections, as they were simultaneously taking a chemistry course and appreciated how the AI linked chemistry concepts to biological applications. Some students described how the AI's structured explanations helped them *"get to the answer faster rather than thinking about it in an abstract way."* However, others expressed frustration when the feedback lacked depth. They argued that a direct three-sentence summary explaining why an answer was correct or incorrect would have been most useful. P7 similarly noted that while they valued guiding (socratic-style) questions in feedback, they sometimes found them too vague, leading them to

search for additional clarification in other resources. Students reported the AI-generated text was sometimes too dense and difficult to read in the small feedback window, making it less engaging.

Students had varied opinions of the AI's qualities in the feedback. P15 expressed skepticism toward AI-generated feedback, preferring clear disclosure that the responses were AI-generated rather than potentially mistaken for instructor-written explanations. Their distrust stemmed from the *"robotic"* tone, occasional inaccuracies, and an overall uncertainty about the reliability of the feedback. There was a stark contrast in other students' receptivity, with others appreciating the tone and quality of the AI-generated feedback. w One recurring suggestion among students was the inclusion of human-verified explanations. P9 proposed incorporating short instructor- or TA-led videos to accompany the AI feedback, arguing that a concise one- to three-minute explanation would be more digestible than lengthy text-based responses. They also pointed out that students sometimes turn to external sources like YouTube because they prefer human explanations over AI-generated ones.

Students exhibited varying engagement levels with the textbook links accompanying the feedback. Some, like P3, frequently used the links as a reference, while others, like P6, admitted that they rarely clicked on them despite recognizing their potential utility. P1 recommended that textbook links should lead directly to specific subsections rather than entire chapters. P7 echoed this, explaining that they often found themselves scrolling through lengthy readings to locate relevant information. Several students reported turning to external resources when the provided textbook links or AI feedback were insufficient. Some students corroborated the AI feedback with at least two sources (one of which was the textbook links), ensuring that they fully understood the concept before moving on.

Perceived tool effects on midterm performance knowledge transfer: Students overwhelmingly found the tool to be instrumental in their preparation for the midterm. However, perspectives varied regarding the extent and nature of their effectiveness.

Students discussed how they instinctively recalled and reused strategies (they had picked up from the tool and its feedback) when attempting the actual midterm. Multiple students reported encountering identical or highly similar questions in the midterm, which allowed them to leverage their prior mistakes and apply the correct reasoning. For P9, pattern recognition played a significant role in rectifying misconceptions. They remembered a question they had answered incorrectly twice on the practice exam and used the feedback to solidify the necessary procedural steps. When encountering a similar question on the midterm, they systematically applied the same analytical approach, demonstrating a transfer of conceptual understanding rather than mere memorization. P12 shared that they often struggled with confusion around really complex concepts. The AI feedback played a crucial role in prompting them to double-check their understanding, ensuring they reviewed the right information instead of reinforcing misconceptions. This extra review not only clarified tricky details but gave them greater confidence when answering similar questions on the actual midterm.

Finally, students reported various study and test-taking strategies that emerged from using the tool. P16 adopted a structured approach by answering the questions they were most confident in first before revisiting the more uncertain ones, attributing this strategy to the tool's emphasis on confidence reflection. P15, who had extended exam time accommodations due to stress-related challenges, found that practicing with the tool helped with pace:

> *"I start stressing out if I notice that the exam time is really short...But I noticed myself slowing down because of the tool...I have a hard time slowing down on exams because of the time limit, so having the [tool] practice exam beforehand for thinking through my answers, really helped...It helps you learn, like, 'Okay, this is a bad habit of mine. Let's slow it down and unlearn that.'"*

P19 adopted the explanation strategies from the practice tool by deliberately writing out their thought process for each question on the midterm. However, they worried that if they initially rationalized an incorrect answer during practice, it could inadvertently reinforce a misconception and lead them to recall the wrong response later. Despite these nuances, students overwhelmingly agreed that engaging with the practice tool significantly contributed to their increased confidence for the midterm.

These behaviors carried over into future courses, even after students no longer had access to our tool. P16 spoke about continuing the practice of gauging their confidence on low-certainty questions and writing out their reasoning in detail as a way to reflect on their learning. They described this as a valuable method for tracking their progress in their new courses. Similarly, P4 shared that they began using ChatGPT in their new classes to replicate the feedback loop provided by our tool. They tackled new practice problems by writing out full justifications for their answers, and then submitting those to GPT for critique. This allowed them to reframe complex concepts in their own words and receive targeted feedback on the precision and completeness of their understanding. These examples illustrate not only the cross-contextual transfer and persistence of reflective learning strategies, but also the adaptive repurposing of available AI tools to meet similar cognitive and reflective goals.

6 Discussion

6.1 RQ1: Impact on Performance

Our findings present interesting contrasts with prior work in educational feedback and student engagement. While Kulhavy et al. [27] argued that simple correctness feedback was insufficient, our results showed a more nuanced picture. Across all midterms, we found no statistically significant differences between the various feedback conditions, suggesting that in the presence of structured metacognitive activities, the specific type of feedback may be less important than previously thought. The most robust finding was the strong relationship between student confidence ratings and performance, highlighting the importance of metacognitive elements in the process. This suggests that requiring students to engage in structured self-explanation and confidence assessment may itself create sufficient cognitive engagement to enhance learning, potentially diminishing the relative impact of feedback types.

Furthermore, our analysis of students who initially scored in the bottom 20% on their practice exams revealed a particularly interesting trend: while the overall impact of combined AI and textbook feedback (Condition 4) was modest, these struggling students showed marginally significant improvements with this condition ($\beta = 0.049$, $p = 0.067$). This suggests that more comprehensive

feedback approaches may be especially beneficial for supporting students needing additional assistance, showing the potential value of differentiated feedback based on student performance levels.

6.2 RQ2: Impact on Engagement and Learning Behaviors

Our findings reinforce existing literature on students' motivation in academic reading and assessment engagement. Prior research has shown that students often prioritize exam performance over deep engagement with course text materials [22, 42]. This study further highlights that students' primary motivation for using the practice tool was to improve their midterm performance, aligning with previous findings that students are driven more by assessment outcomes than intrinsic engagement with course content [16, 53].

Traditional practice exams can only tell students whether their answers are correct or not, which means students are only incentivized to provide answers. But an AI feedback tool that adapts its responses to students' explanations and confidence can incentivize students to invest in structured reflection about explanations and confidence, creating an assessment-driven learning experience that encourages deeper cognitive engagement. The positive reception is evidenced by the 76% voluntary adoption rate for midterm 2, showing students found value in the tool beyond course requirements.

The structured self-explanation and confidence rating requirements in the tool reflect established pedagogical approaches that promote metacognitive awareness [14, 52]. Many students reported that articulating their reasoning forced them to reconsider assumptions, distinguish between educated guesses and true understanding, and refine their problem-solving strategies. This supports previous research demonstrating that self-explanation and question-generation enhance comprehension and retention [36, 44, 54]. However, our findings also reveal potential drawbacks, particularly for students who found the requirement mentally exhausting or redundant when their confidence was on the extreme ends of the spectrum (*"very confident"* or *"basically guessing"*). Thus, for students with minimal confidence, providing an option to bypass explanation requirements could be beneficial. Many students instinctively turned to textbook links before engaging with AI feedback to build foundational knowledge in such cases, aligning with existing research on confidence and feedback receptivity [19, 27].

The AI-generated feedback adapted to students' confidence levels and explanations and played a critical role in steering students towards the textbook. Students reported verifying AI feedback against human-authored content, reflecting AI skepticism [51], but also enabling clarification of misconceptions and supplementing less directive AI feedback. This suggests future systems should support layered scaffolding that balances open-ended prompting with direct instructional resources, and facilitates verification behaviors to promote critical AI literacy. The observed transfer of reasoning and pacing strategies from the practice tool to the midterm underscores the potential of structured AI interactions to influence students' metacognitive behaviors. This aligns with prior work on value of practice testing as a learning strategy [6, 40]. Some students, however, noted concerns that articulating reasoning early in the process might sometimes lead to reinforcement of incorrect thinking. Future systems could address this by including iterative feedback loops that prompt students to revisit and refine their reasoning over time. Designing AI feedback supporting both immediate understanding and sustained, reflective learning would help.

While some students found the integration of AI and textbook references helpful for mapping instructional materials to exam questions [2, 9], others expressed frustration when referred to dense textbook sections. This aligns with prior findings that students often prefer digestible explanations over extensive references [3, 22]. Future iterations could enhance usability through more targeted textbook linking, AI-generated summaries of relevant sections, or incorporating human-verified content like short instructor videos. The student explanations collected can inform refined AI prompts, enabling more personalized feedback aligned with student reasoning patterns. Students found answering identical questions on multiple attempts ineffective; varying question formats on reattempts would better support knowledge transfer and retention [40].

More fundamentally, our findings point to a possible redesign opportunity for educational platforms. Rather than positioning metacognitive components as supporting features for content delivery, future systems might invert this relationship—making metacognitive development the explicit design goal while AI feedback serves as scaffolding. Such systems could analyze confidence patterns over time, provide targeted feedback on explanation quality, and gradually reduce structured support as students develop independent reflection. This would leverage technology to cultivate transferable critical thinking skills that persist beyond specific course contexts.

7 Limitations

Our study focused on a single undergraduate biology course, potentially limiting generalizability across STEM disciplines. While GPT-4o proved effective for feedback generation, other AI approaches might offer different benefits. Self-reported survey data has inherent biases, and the lack of baseline biology knowledge assessment made it difficult to control for varying levels of prior subject expertise.

8 Conclusion

Our work demonstrates how AI-enhanced practice systems can support learning at scale through careful integration of technology and pedagogical principles. While textbook references alone showed modest performance gains, the system's greater impact emerged in transforming students' learning behaviors and metacognitive strategies. The high textbook engagement rate and successful adoption of self-assessment practices suggest that contextual, just-in-time support can effectively motivate student engagement with course materials. As institutions explore AI integration in education, our results emphasize designing systems that enhance rather than replace traditional learning resources, while supporting the development of sustainable study practices.

9 Acknowledgments

We thank Mason Johnstone for contributing practice exam question drafts from the course for consideration.

References

[1] Mak Ahmad and Kwan-Liu Ma. 2024. More Than Chatting: Conversational LLMs for Enhancing Data Visualization Competencies. In *EuroVis 2024 - Education*

Papers, Elif E. Firat, Robert S. Laramee, and Nicklas Sindelv Andersen (Eds.). The Eurographics Association. doi:10.2312/eved.20241056

[2] John Biggs and Catherine Tang. 2011. Train-the-trainers: Implementing outcomes-based teaching and learning in Malaysian higher education. *Malaysian Journal of Learning and Instruction* 8 (2011), 1–19.

[3] Brian Brost and Karen Bradley. 2006. Student compliance with assigned reading: A case study. *Journal of the Scholarship of Teaching and Learning* (2006), 101–111.

[4] Andrew C Butler, Jeffrey D Karpicke, and Henry L Roediger III. 2007. The effect of type and timing of feedback on learning from multiple-choice tests. *Journal of Experimental Psychology: Applied* 13, 4 (2007), 273.

[5] Deborah L Butler and Philip H Winne. 1995. Feedback and self-regulated learning: A theoretical synthesis. *Review of educational research* 65, 3 (1995), 245–281.

[6] Shana K Carpenter. 2012. Testing enhances the transfer of learning. *Current directions in psychological science* 21, 5 (2012), 279–283.

[7] Mei-Mei Chang. 2010. Effects of self-monitoring on web-based language learner's performance and motivation. *Calico Journal* 27, 2 (2010), 298–310.

[8] Kathy Charmaz. 2008. Grounded theory as an emergent method. *Handbook of emergent methods* 155 (2008), 172.

[9] Grace Chipperfield, Lauren Butterworth, and Pablo Munguia. 2022. Embedding resources into digital assessment rubrics: Bringing academic support directly to students. *Journal of Academic Language and Learning* 16, 1 (2022), C1–C11.

[10] Michael A Clump, Heather Bauer, and Catherine Bradley. 2004. The extent to which psychology students read textbooks: a multiple class analysis of reading across the psychology curriculum. *Journal of Instructional Psychology* 31, 3 (2004), 227–233.

[11] Patricia A Connor-Greene. 2000. Assessing and promoting student learning: Blurring the line between teaching and testing. *Teaching of Psychology* 27, 2 (2000), 84–88.

[12] Juliet Corbin et al. 1990. Basics of qualitative research grounded theory procedures and techniques. (1990).

[13] Wei Dai, Yi-Shan Tsai, Jionghao Lin, Ahmad Aldino, Hua Jin, Tongguang Li, Dragan Gašević, and Guanliang Chen. 2024. Assessing the proficiency of large language models in automatic feedback generation: An evaluation study. *Computers and Education: Artificial Intelligence* 7 (2024), 100299.

[14] W Pitt Derryberry and Steven R Wininger. 2008. Relationships among Textbook Usage and Cognitive-Motivational Constructs. *Teaching Educational Psychology* 3, 2 (2008), n2.

[15] Cheryl Cisero Durwin and William M Sherman. 2008. Does choice of college textbook make a difference in students' comprehension? *College teaching* 56, 1 (2008), 28–34.

[16] Michelle French, Franco Taverna, Melody Neumann, Lena Paulo Kushnir, Jason Harlow, David Harrison, and Ruxandra Serbanescu. 2015. Textbook use in the sciences and its relation to course performance. *College Teaching* 63, 4 (2015), 171–177.

[17] Wayne Geerling, G Dirk Mateer, Jadrian Wooten, and Nikhil Damodaran. 2023. Is ChatGPT smarter than a student in principles of economics. *Available at SSRN* 4356034 (2023).

[18] Sarah J Hatteberg and Kody Steffy. 2013. Increasing reading compliance of undergraduates: An evaluation of compliance methods. *Teaching Sociology* 41, 4 (2013), 346–352.

[19] John Hattie and Helen Timperley. 2007. The power of feedback. *Review of educational research* 77, 1 (2007), 81–112.

[20] Cynthia E Heiner, Amanda I Banet, and Carl Wieman. 2014. Preparing students for class: How to get 80% of students reading the textbook before class. *American Journal of Physics* 82, 10 (2014), 989–996.

[21] Mary E Hoeft. 2012. Why university students don't read: What professors can do to increase compliance. *International journal for the scholarship of teaching and learning* 6, 2 (2012), 12.

[22] Jay R Howard. 2004. Just-in-time teaching in sociology or how I convinced my students to actually read the assignment. *Teaching Sociology* 32, 4 (2004), 385–390.

[23] Pamela J Howard, Meg Gorzycki, Geoffrey Desa, and Diane D Allen. 2018. Academic reading: Comparing students' and faculty perceptions of its value, practice, and pedagogy. *Journal of College Reading and Learning* 48, 3 (2018), 189–209.

[24] Lasse X Jensen, Alexandra Buhl, Anjali Sharma, and Margaret Bearman. 2024. Generative AI and higher education: a review of claims from the first months of ChatGPT. *Higher Education* (2024), 1–17.

[25] Bethany C Johnson and Marc T Kiviniemi. 2009. The effect of online chapter quizzes on exam performance in an undergraduate social psychology course. *Teaching of Psychology* 36, 1 (2009), 33–37.

[26] Michael Klymkowsky and Melanie M Cooper. 2024. The end of multiple choice tests: using AI to enhance assessment. *arXiv preprint arXiv:2406.07481* (2024).

[27] Raymond W Kulhavy and William A Stock. 1989. Feedback in written instruction: The place of response certitude. *Educational psychology review* 1 (1989), 279–308.

[28] Raymond W Kulhavy, Mary T White, Bruce W Topp, Ann L Chan, and James Adams. 1985. Feedback complexity and corrective efficiency. *Contemporary educational psychology* 10, 3 (1985), 285–291.

[29] James A Kulik and Chen-Lin C Kulik. 1988. Timing of feedback and verbal learning. *Review of educational research* 58, 1 (1988), 79–97.

[30] Harsh Kumar, Ruiwei Xiao, Benjamin Lawson, Ilya Musabirov, Jiakai Shi, Xinyuan Wang, Huayin Luo, Joseph Jay Williams, Anna N Rafferty, John Stamper, et al. 2024. Supporting self-reflection at scale with large language models: Insights from randomized field experiments in classrooms. In *Proceedings of the eleventh ACM conference on learning@ scale*. 86–97.

[31] Matthias Lehmann, Philipp B Cornelius, and Fabian J Sting. 2024. AI Meets the Classroom: When Does ChatGPT Harm Learning? *arXiv preprint arXiv:2409.09047* (2024).

[32] Abe Leite and Saúl A Blanco. 2020. Effects of human vs. automatic feedback on students' understanding of AI concepts and programming style. In *Proceedings of the 51st ACM Technical Symposium on Computer Science Education*. 44–50.

[33] Tara Lockhart and Mary Soliday. 2016. The critical place of reading in writing transfer (and beyond): A report of student experiences. *Pedagogy* 16, 1 (2016), 23–37.

[34] Richard S Lysakowski and Herbert J Walberg. 1981. Classroom reinforcement and learning: A quantitative synthesis. *The Journal of Educational Research* 75, 2 (1981), 69–77.

[35] JaneMaree Maher and Jennifer Mitchell. 2010. I'm not sure what to do! Learning experiences in the humanities and social sciences. *Issues in Educational Research* 20, 2 (2010), 137.

[36] Gili Marbach-Ad and Phillip G Sokolove. 2000. Can undergraduate biology students learn to ask higher level questions? *Journal of Research in Science Teaching: The Official Journal of the National Association for Research in Science Teaching* 37, 8 (2000), 854–870.

[37] Teresa Murden and Cindy S Gillespie. 1997. The role of textbooks and reading in content area classrooms: What are teachers and students saying. *Exploring literacy* (1997), 87–96.

[38] Sasha Nikolic, Scott Daniel, Rezwanul Haque, Marina Belkina, Ghulam M Hassan, Sarah Grundy, Sarah Lyden, Peter Neal, and Caz Sandison. 2023. ChatGPT versus engineering education assessment: a multidisciplinary and multi-institutional benchmarking and analysis of this generative artificial intelligence tool to investigate assessment integrity. *European Journal of Engineering Education* 48, 4 (2023), 559–614.

[39] Susan Bobbitt Nolen. 1996. Why study? How reasons for learning influence strategy selection. *Educational Psychology Review* 8 (1996), 335–355.

[40] Henry L Roediger and Andrew C Butler. 2011. The critical role of retrieval practice in long-term retention. *Trends in cognitive sciences* 15, 1 (2011), 20–27.

[41] Ernst Z Rothkopf. 1988. Perspectives on study skills training in a realistic instructional economy. In *Learning and study strategies*. Elsevier, 275–286.

[42] John Sappington, Kimberly Kinsey, and Kirk Munsayac. 2002. Two studies of reading compliance among college students. *Teaching of psychology* 29, 4 (2002), 272–274.

[43] Sima Sengupta. 2002. Developing academic reading at tertiary level: A longitudinal study tracing conceptual change. *The reading matrix* 2, 1 (2002).

[44] Betty Lou Smith, William G Holliday, and Homer W Austin. 2010. Students' comprehension of science textbooks using a question-based reading strategy. *Journal of Research in Science Teaching: The Official Journal of the National Association for Research in Science Teaching* 47, 4 (2010), 363–379.

[45] Adele Smolansky, Andrew Cram, Corina Raduescu, Sandris Zeivots, Elaine Huber, and Rene F Kizilcec. 2023. Educator and student perspectives on the impact of generative AI on assessments in higher education. In *Proceedings of the tenth ACM conference on Learning@ Scale*. 378–382.

[46] Helen St Clair-Thompson, Alison Graham, and Sara Marsham. 2018. Exploring the reading practices of undergraduate students. *Education Inquiry* 9, 3 (2018), 284–298.

[47] Keith Starcher and Dennis Proffitt. 2011. Encouraging Students to Read: What Professors Are (and Aren't) Doing About It. *International Journal of Teaching and Learning in Higher Education* 23, 3 (2011), 396–407.

[48] Stephanie Stokes-Eley. 2007. Using Kolb's experiential learning cycle in chapter presentations. *Communication Teacher* 21, 1 (2007), 26–29.

[49] Petra Stutz, Maximilian Elixhauser, Judith Grubinger-Preiner, Vivienne Linner, Eva Reibersdorfer-Adelsberger, Christoph Traun, Gudrun Wallentin, Katharina Wöhs, and Thomas Zuberbühler. 2023. Ch (e) atGPT? An anecdotal approach addressing the impact of ChatGPT on teaching and learning GIScience. (2023).

[50] Fang Tang and Peida Zhan. 2021. Does diagnostic feedback promote learning? Evidence from a longitudinal cognitive diagnostic assessment. *AERA Open* 7 (2021), 23328584211060804.

[51] Andrea Tick. 2024. Exploring ChatGPT's Potential and Concerns in Higher Education. In *2024 IEEE 22nd Jubilee international symposium on intelligent systems and informatics (SISY)*. IEEE, 000447–000454.

[52] Terry Tomasek. 2009. Critical reading: Using reading prompts to promote active engagement with text. *International journal of teaching and learning in higher education* 21, 1 (2009), 127–132.

[53] Mario Vafeas. 2013. Attitudes toward, and use of, textbooks among marketing undergraduates: an exploratory study. *Journal of Marketing Education* 35, 3 (2013), 245–258.

[54] Dianna L Van Blerkom, Malcolm L Van Blerkom, and Sharon Bertsch. 2006. Study strategies and generative learning: What works? *Journal of College Reading and Learning* 37, 1 (2006), 7–18.

[55] Ben Ward, Deepshikha Bhati, Fnu Neha, and Angela Guercio. 2024. Analyzing the Impact of AI Tools on Student Study Habits and Academic Performance. *arXiv preprint arXiv:2412.02166* (2024).

[56] Benedikt Wisniewski, Klaus Zierer, and John Hattie. 2020. The power of feedback revisited: A meta-analysis of educational feedback research. *Frontiers in psychology* 10 (2020), 487662.

[57] Qi Xia, Xiaojing Weng, Fan Ouyang, Tzung Jin Lin, and Thomas KF Chiu. 2024. A scoping review on how generative artificial intelligence transforms assessment in higher education. *International Journal of Educational Technology in Higher Education* 21, 1 (2024), 40.