Journal of Information Technology Education: Research

An Official Publication of the Informing Science Institute
InformingScience.org

JITEResearch.org

Volume 24, 2025

CHATGPT-ASSISTED RETRIEVAL PRACTICE AND EXAM SCORES: DOES IT WORK?

Ibnatul Jalilah Yusof — Universiti Teknologi Malaysia, Skudai, Johor, Malaysia — ijalilah@utm.my

ABSTRACT

Aim/Purpose	This paper examines the potential of ChatGPT-assisted retrieval practice to enhance students' final exam performance. ChatGPT was utilized to generate questions and deliver timely feedback during retrieval practice, supporting learning in large class settings where providing personalized feedback is often challenging.
Background	Students often excel in continuous assessments yet face significant challenges in final exams, largely due to the demanding nature of these exams that require the recall and application of accumulated knowledge. This persistent issue highlights a gap in traditional study practices and underscores the need for innovative strategies to support long-term memory retention. This study explores how ChatGPT can bridge this gap by supporting retrieval practice, an evidence-based strategy known to improve long-term memory retention.
Methodology	This study adopts a retrospective cohort design, comparing final exam scores between previous cohorts who did not use ChatGPT for retrieval practice (control group) and current cohorts who did (experimental group). ChatGPT was used to generate objective questions for retrieval practice and provide immediate feedback to students. The primary sample consists of second-year education students enrolled in the *Measurement and Evaluation in Education* course, with 64 students randomly selected for each group. In addition to analyzing exam scores, the study incorporates complementary findings from students' feedback collected at the end of the semester to gain a deeper understanding of their experiences with ChatGPT-assisted retrieval practice.
Contribution	As higher education in Malaysia increasingly shifts towards alternative assessments, this study highlights a simple yet impactful retrieval practice as a learning strategy that integrates formative assessment with feedback. With the aid of generative AI such as ChatGPT, such strategies can be implemented

Accepting Editor Stamatis Papadakis | Received: December 31, 2024 | Revised: March 3, March 6, March 13, 2025 | Accepted: March 15, 2025.
Cite as: Yusof, I. J. (2025). ChatGPT-assisted retrieval practice and exam scores: Does it work?. *Journal of Information Technology Education: Research, 24,* Article 8. https://doi.org/10.28945/5474

	easily and effectively by offering a practical solution that enhances exam performance while reducing the lecturer's workload.
Findings	The results of the study indicate that students who engaged in AI-assisted retrieval practice using ChatGPT performed significantly better on their final exams compared to those who did not. A Welch's t-test revealed a significant difference in exam scores between the control group (students without ChatGPT-assisted practice) and the experimental group (students with ChatGPT-assisted practice). While the findings demonstrate the effectiveness of ChatGPT in enhancing academic performance, they also underscore the importance of complementing AI support with human feedback to address complex learning needs and provide deeper contextual understanding.
Recommendations for Practitioners	Practitioners should consider using AI tools like ChatGPT to support retrieval practice by helping students generate questions and receive immediate feedback, particularly in large classes. AI feedback should be complemented with human input to address complex questions and provide deeper understanding. Providing students with guidance on using AI tools effectively can maximize their benefits while balancing group activities with individual tasks ensures both collaboration and personalized engagement. Regularly gathering student feedback and tracking performance can help refine the integration of AI tools to better meet learning goals.
Recommendations for Researchers	Future studies should investigate the long-term impact of ChatGPT-assisted retrieval practice on memory retention and explore how different types of ChatGPT-generated feedback influence learning outcomes. Additionally, the research could examine how the integration of lecturer-generated or student-generated questions in ChatGPT-assisted retrieval practice affects learning and whether varying these approaches could further enhance memory retention and exam performance across different educational contexts and subjects.
Impact on Society	ChatGPT-assisted retrieval practice, when effectively implemented, transforms retrieval practice from being just an assessment tool into a powerful learning strategy that enhances memory retention. With ChatGPT providing timely feedback and supporting educators and students, it can reduce the burden on educators in large classrooms while empowering students to take greater control of their learning.
Future Research	Future research should explore the longer and more frequent use of ChatGPT-assisted retrieval practice to give students more opportunities for individual engagement. Studies could also focus on improving ChatGPT's feedback for more complex questions and finding the best ways to combine AI feedback with human input. Expanding research into different subjects and tracking long-term effects on learning and performance would help better understand how ChatGPT can be effectively used in education.

Keywords	AI-assisted learning, retrieval practice, ChatGPT, formative assessment, final exams

BACKGROUND

The Code of Practice for Programme Accreditation (COPPA) for Malaysia Higher Education (Malaysian Qualifications Agency, 2018) defines continuous assessment as "assessments conducted throughout the duration of a course/module for the purpose of determining student attainment."

These assessments, which may include presentations and projects, typically contribute 40-60% of the overall grade and provide ongoing feedback that helps students improve over the semester (Malaysian Qualifications Agency, 2018). In contrast, final exams are summative assessments: "assessment of learning which summarizes the student progress at a particular time and is used to assign the student a course grade" (Malaysian Qualifications Agency, 2018, p. iii). Occurring only once, final exams offer no opportunity for feedback, making the effective recall of accumulated knowledge critical for success (French et al., 2024).

In my five years of teaching measurement and evaluation in education to undergraduate students, I have consistently observed that students who perform well in continuous assessments often struggle with final exams. This concern is also reflected in the research, which indicates that while students tend to excel in coursework, their final exam scores are generally lower (Flom et al., 2023; Richardson, 2015). Factors contributing to lower exam scores can be categorized into those related to exam items, e.g., item difficulty and format (Abdullah et al., 2012; Maeda, 2021; Rudolph et al., 2019; Smolinsky et al., 2020), and those related to examinees, e.g., test anxiety, working memory capacity, and study habits (Dewi & Mangunsong, 2012; Jamieson et al., 2016; Maeda, 2021; McDougall & Gruneberg, 2002). Notably, poor performance is often linked to inadequate retrieval of learned information, as illustrated by Ebbinghaus' forgetting curve, which demonstrates memory decline without reinforcement (Donker et al., 2022; McDougall & Gruneberg, 2002).

Retrieval practice is a cornerstone of effective learning that involves actively recalling information from memory (Jaeger et al., 2024; Ritchie et al., 2013; Sana & Yan, 2022). This technique has significantly boosted long-term retention and deepened conceptual understanding, outperforming passive review methods (Carpenter et al., 2016; Deng et al., 2015; Jaeger et al., 2014; Ritchie et al., 2013). However, its traditional implementation faces several challenges. Effective retrieval practice hinges on the provision of timely and specific feedback that is tailored to individual student needs; yet, in large-class settings (e.g., typical educator-to-student ratios of 1:30 in Malaysian universities), providing such personalized feedback is a monumental task (Christiansen et al., 2024; Metu et al., 2024; Singh, 2019; Tan, 2020). Therefore, educators often resort to generic feedback due to time constraints, and delays in feedback can allow misconceptions to persist (Kubik et al., 2021).

Additionally, traditional retrieval practice methods such as low-stakes quizzes, multiple-choice questions, and short-answer questions require substantial preparation, administration, and grading effort (Sana & Yan, 2022). These labor-intensive processes limit the frequency and scalability of retrieval practice, preventing the optimally spaced repetitions that research suggests are necessary for robust learning (Ritchie et al., 2013; Sana & Yan, 2022). Recent studies have begun to explore the potential of ChatGPT to generate tailored retrieval questions automatically. For instance, Indran et al. (2023) and Zuckerman et al. (2023) demonstrated that ChatGPT can produce high-quality, contextually relevant questions that align with learning objectives, significantly reducing the burden on educators while enabling more frequent, adaptive retrieval practice.

These capabilities of ChatGPT make it a promising tool for implementing effective retrieval practices, which rely on the generation of questions and feedback to support student learning. Nevertheless, to date, only one study has specifically investigated the use of ChatGPT for retrieval practices (Meier & Löfqvist, 2024). While this study reported positive outcomes, such as improved learning retention and performance, its findings are based on a relatively small sample size (n=17). As a result, the conclusions of their study should be regarded as preliminary and interpreted with caution. They also suggested that further research is needed to validate these results, particularly in larger educational contexts, to fully understand the potential and limitations of ChatGPT in facilitating retrieval practice.

Recognizing this gap, I see an opportunity to utilize ChatGPT's capabilities to improve students' final exam performance. ChatGPT offers key benefits such as automated question generation, instant feedback, and adaptive learning support, reducing the burden of manual assessment tasks (Kasneci et

al., 2023; Steiss et al., 2024). For students, it enables personalized learning by delivering tailored questions and immediate feedback, reinforcing memory retention and conceptual understanding (Holmes et al., 2019; Luckin & Holmes, 2016)

Drawing on established theories of formative assessment and retrieval practice (Halamish & Bjork, 2011; Karpicke, 2017), this study investigates whether integrating ChatGPT into retrieval practice can effectively address traditional limitations such as delayed and generic feedback in large-class settings, ultimately improving final exam performance. This paper aims to contribute to the growing body of work on artificial intelligence (AI) applications in education by explicitly linking these challenges to current research gaps and engaging with key academic literature.

LITERATURE REVIEW

RETRIEVAL PRACTICE AS LEARNING STRATEGIES

Retrieval practice is grounded in the principle that actively recalling information from memory strengthens long-term retention (Ariel & Karpicke, 2018; Karpicke & Roediger, 2007). Traditionally, testing has been viewed primarily as a means to measure learning, often associated with rote memorization. However, retrieval practice goes beyond assessment – it is a powerful learning strategy that enhances memory, boosts motivation to study, and significantly improves long-term retention compared to methods like repeated studying or reading (Abel & Bäuml, 2020; Casselman, 2024; Karpicke, 2017; McDougall & Gruneberg, 2002). According to the retrieval-based learning approach (Jaeger et al., 2014; Karpicke, 2017; Karpicke & Roediger, 2007), each instance of retrieval strengthens the memory trace, increasing the likelihood of long-term retention.

Glover (1989) argues that repeated retrieval practice before a final assessment significantly enhances students' ability to recall and apply information during exams. Ariel and Karpicke (2018) further highlight that after students successfully recall information for the first time, continued retrieval practice is crucial, as multiple attempts lead to greater retention compared to a single recall. This iterative process reinforces memory, making it more durable and accessible when needed. Similarly, studies by Ma et al. (2020) and YeckehZaare et al. (2019) demonstrate that the effectiveness of retrieval practice increases with repeated attempts.

While retrieval practice is beneficial on its own, embedding it within formative assessment practice, such as providing feedback (e.g., providing correct answers), significantly enhances its effectiveness (Agarwal et al., 2016, 2021; Jaeger et al., 2024; Roediger & Butler, 2011), especially for students with lower memory capacity (Agarwal et al., 2016, 2021). Formative assessment, as conceptualized by Black and Wiliam (1998), emphasizes the role of feedback in guiding student learning by providing actionable insights that help learners refine their understanding. Feedback allows students to compare their responses to the correct ones, reinforcing accurate information and correcting misunderstandings. Sadler (1998) further argues that effective formative assessment requires students to recognize the gap between their current performance and the desired learning outcome, making feedback an essential component in this process.

Although formative assessment is often seamlessly integrated into teaching strategies (Ruiz-Primo, 2011; Trumbull & Lash, 2013), it also demands that students take a more focused approach to learning and put in greater effort to improve their own learning (Casselman, 2024; Trumbull & Lash, 2013). This aligns closely with the principles of retrieval practice, which similarly requires students to exert more mental effort (Agarwal et al., 2016; Hui et al., 2022). However, in traditional classroom settings, delays in feedback delivery often limit its effectiveness. Large student-to-teacher ratios further constrain educators' ability to provide timely and individualized responses, leaving many students without the guidance needed to improve their recall and application of knowledge (Ruiz-Primo, 2011; Trumbull & Lash, 2013).

Although retrieval practice has been widely validated across educational levels (Carpenter, 2023; Carpenter et al., 2022; Jaeger et al., 2014), its implementation in higher education remains inconsistent. Several studies demonstrate its effectiveness through different instructional approaches. Greving and Richter (2022) examined the testing effect in university lectures and found that short-answer questions significantly enhanced retention, whereas multiple-choice questions had little impact. Bego et al. (2024) investigated spaced retrieval practice in STEM courses through bi-weekly quizzes, finding substantial benefits for engineering and health science students. Broeren et al. (2021) explored self-regulated retrieval practice, showing that while instructional interventions improved students' use of retrieval strategies, they did not necessarily translate into higher test scores without additional support.

Notably, the studies discussing question formats in retrieval practice primarily examined their effectiveness during each retrieval attempt rather than their direct impact on final exam performance. While these studies provide strong evidence that different question types influence retention and recall in practice sessions, they do not explicitly explore whether these benefits carry over to high-stakes summative assessments. This distinction is critical, as final exams require students to retrieve accumulated knowledge under exam conditions that introduce additional cognitive demands, such as test anxiety and time constraints.

Research suggests that retrieval practice is most effective when it aligns with final exam formats. Morris et al. (1977) propose that memory performance improves when the method of learning matches the method of recall, indicating that study methods should mirror test conditions to enhance retrieval efficiency. Jensen et al. (2014) emphasize that the structure of objective questions such as multiple-choice, fill-in-the-blank, and short-answer questions used throughout the semester influences students' learning strategies, ultimately impacting their final exam performance. Cummings (2020) and Morano (2019) also support this view, arguing that practicing recall in a format similar to the final exam strengthens students' ability to retrieve and apply knowledge during high-stakes assessments. These findings highlight the importance of structuring retrieval practice to reflect final exam conditions, as it influences how students engage with the material, the effort they invest, and their study strategies outside of class (Agarwal et al., 2021; Casselman, 2024).

These findings reinforce the need for structured retrieval practice that aligns with final exams and provides frequent feedback opportunities. Hence, this study posits that retrieval practice can improve students' final exam scores. To achieve this, learning activities should incorporate questions that mirror the format of the final exam, ensuring consistency between practice and assessment. Additionally, retrieval practice should be implemented multiple times, as repeated attempts significantly enhance memory retention, even though the optimal frequency remains unspecified. Finally, incorporating feedback, such as providing correct answers after each retrieval attempt, is essential for reinforcing accurate knowledge and addressing misunderstandings.

However, traditional retrieval practice methods such as quizzes, fill-in-the-blank, multiple-choice questions, and short-answer exercises require substantial time and effort for educators to prepare, administer, and grade (Meier & Löfqvist, 2024; Sana & Yan, 2022). The manual nature of these tasks limits the frequency of retrieval opportunities and prevents students from engaging in optimal spaced practice (Sana & Yan, 2022). To address these limitations, AI-driven tools such as ChatGPT offer a potential solution for question generation and feedback delivery.

Generative-Artificial Intelligence as a Tool in Enhancing Learning

Artificial intelligence (AI), particularly generative models like ChatGPT, has become widely used in education, enhancing teaching, learning, and assessment. Originally designed for text generation (Floridi & Chiriatti, 2020; Kikalishvili, 2023; Strasser, 2024), ChatGPT has since improved in contextual understanding, fluency, and reasoning (OpenAI, 2023; Yu, 2023). The release of GPT-4 in 2023 introduced multi-modal capabilities, while CustomGPT in 2024 allowed users to fine-tune AI responses for specific tasks (Almasre, 2024; OpenAI, 2023).

For educators, ChatGPT serves as a versatile tool that improves lesson planning, content creation, and student support. It also assists in grading and feedback generation, reducing workload while ensuring timely and consistent responses (Kiryakova & Angelova, 2023; Modran et al., 2024). Beyond administrative tasks, educators use ChatGPT to facilitate discussions, scaffold learning, and simulate real-world problem-solving, encouraging deeper student engagement (Dai et al., 2023; Kiryakova & Angelova, 2023; Lo, 2023).

ChatGPT acts as a virtual assistant for students, offering personalized learning support, improving motivation, and enhancing understanding of complex topics (Dai et al., 2023; Rasul et al., 2023; Sain et al., 2024). It helps generate ideas, provides constructive feedback, and increases engagement in learning activities (Rasyid et al., 2024; Sain et al., 2024). Studies highlight that students find ChatGPT to be an accessible and effective tool for improving learning outcomes (Rasyid et al., 2024; Sain et al., 2024).

ChatGPT as a feedback mechanism

Feedback is a critical component of effective retrieval practice, as it helps reinforce accurate knowledge, correct misconceptions, and improve long-term retention (Agarwal et al., 2016; Roediger & Butler, 2011). In traditional retrieval practice settings, feedback is often delivered manually by educators, posing significant challenges in timeliness, personalization, and scalability, particularly in large-class settings (Ruiz-Primo, 2011; Trumbull & Lash, 2013). The high student-to-teacher ratio in higher education (e.g., 1:30 or higher) further constrains instructors' ability to provide individualized feedback, leaving many students without the necessary guidance to refine their recall strategies (Christiansen et al., 2024; Metu et al., 2024).

In contrast, ChatGPT can provide immediate, personalized feedback, reducing the turnaround time for response evaluation and allowing students to correct errors in real-time (Kasneci et al., 2023; Steiss et al., 2024). Unlike traditional manual feedback, ChatGPT-generated feedback can be tailored to individual responses, offering explanations, hints, and contextualized corrections that cater to different levels of understanding (Holmes et al., 2019; Luckin & Holmes., 2016).

Studies have shown that AI-powered feedback mechanisms can enhance student engagement, retention, and metacognitive awareness. Meier and Löfqvist (2024) explored the effectiveness of ChatGPT as a real-time feedback provider in retrieval practice. They found that students who received ChatGPT-generated explanations demonstrated improved recall accuracy and conceptual understanding. Similarly, Cheung et al. (2023) reported that ChatGPT-assisted feedback improved self-regulated learning behaviors, as students could engage in iterative learning cycles by refining their responses based on ChatGPT-generated suggestions.

However, despite these advantages, ChatGPT-generated feedback has limitations. Concerns have been raised regarding potential biases in ChatGPT-generated explanations, inaccuracies in grading open-ended responses, and the inability to fully replicate human expertise in nuanced feedback (Özbay, 2024). Additionally, the effectiveness of ChatGPT as a feedback tool depends on how well students interpret and apply its generated responses, raising questions about whether AI feedback should complement, rather than replace, human guidance (Kiryakova & Angelova, 2023).

ChatGPT as a test questions generator

ChatGPT has demonstrated significant utility in providing feedback, generating complex assessment items, and addressing a critical challenge in education – the time and effort required to create high-quality test questions. Traditionally, instructors manually design test questions, which can be labor-intensive, especially in large-class settings where frequent assessments are needed (Meier & Löfqvist, 2024). Studies have highlighted ChatGPT's effectiveness in reducing academic workload while maintaining quality standards, making it a valuable tool for educators (Cheung et al., 2023).

One of ChatGPT's strengths is its ability to generate diverse question types tailored to different cognitive levels and assessment needs. Research has shown that ChatGPT can effectively create structured, well-balanced questions, ranging from basic recall questions to higher-order critical thinking tasks. For example, Fernández et al. (2024) and Meier and Löfqvist (2024) demonstrated ChatGPT's capability to generate short-answer questions, while Cheung et al. (2023), Özbay (2024), Rivera-Rosas et al. (2024) and Yusof and Ismail (2023) found that it can also produce multiple-choice questions (MCQs) that closely resemble human-crafted exam items. Cheung et al. (2023) further reported that expert evaluations rated ChatGPT-generated questions as comparable in quality to those created by experienced educators, emphasizing their potential to complement traditional question design methods.

However, while ChatGPT's capabilities in generation questions are promising, researchers emphasize the need for careful evaluation and refinement of its generated content. ChatGPT-generated questions may sometimes lack detailed phrasing, contain unintended biases, or fail to align perfectly with learning objectives. As Özbay (2024) and (Cheung et al., 2023) pointed out, while ChatGPT is a powerful assistant for educators, its outputs still require human oversight to ensure accuracy, fairness, and pedagogical relevance. Furthermore, studies have demonstrated that the quality of ChatGPT-generated questions depends heavily on the specificity of the prompt given (Kıyak et al., 2024; Kıyak & Emekli, 2024; Rivera-Rosas et al., 2024). Well-structured prompts that specify question type, difficulty level, content focus, and cognitive skill level yield more precise and pedagogically sound results.

RESEARCH METHODOLOGY

RESEARCH DESIGN

The study used a retrospective cohort design to investigate the impact of AI-assisted retrieval practice support on students' final exam scores. In this design, the use of AI-assisted retrieval practice (exposure) and the final exam scores (outcome) had already occurred when the study commenced. The research involved gathering existing information on students' engagement with AI-assisted retrieval practice from historical data and combining this with their final exam scores. Specifically, the study compared the final exam scores of a previous cohort of students who did not utilize AI-assisted retrieval practice with those of the current cohort who did.

At the time of implementation, the use of AI-assisted retrieval practice was focused solely on supporting students' learning and improving exam outcomes, particularly in response to underperformance in the previous cohort. The decision to write this paper and analyze the data as part of a research study was made later and students were not initially aware that their results would be used for research purposes.

ETHICAL CONSIDERATIONS

This study utilized secondary data originally collected for educational purposes for both the control and experimental groups. At the time of implementation, students participated in ChatGPT-assisted retrieval practice solely as a learning activity to enhance their understanding and exam preparation, with no prior intention of their scores being used for research. The decision to analyze their exam performance for this study was made retrospectively.

To ensure ethical integrity, institutional guidelines on secondary data use were followed, with all data anonymized and no additional risk posed to participants. In accordance with the Malaysian Code of Responsible Conduct in Research (National Science Council, 2020), researchers using secondary data must implement appropriate data management practices, including anonymization. Therefore, for this paper, the following measures were taken:

(i) The cohorts' names were adjusted to maintain anonymity.

(ii) Random sampling was employed to ensure that no individual data could be identified.
(iii) The dataset of students' final exam scores was recorded in a manner that prevented subject identification and privacy could only be accessed with a two-identification password, in compliance with Malaysian Personal Data Protection Act (Ministry of Digital, 2010) guidelines.

Population and Sampling

The participants in this study were second-year education students enrolled in the compulsory Measurement and Evaluation in Education course. They were divided into two groups – a control group and an experimental group. The control group consisted of a previous cohort that did not participate in AI-assisted retrieval practice, while the experimental group comprised students from a cohort that did. Both groups took final exams, which, although not identical, followed the same format, covering the same topics with 60 multiple-choice questions, and assessed across similar taxonomy levels, ranging from remembering to analyzing.

Table 1 provides a breakdown of the participants in the study across the control and experimental cohorts. The control group consisted of students from three cohorts between Sem 1 A, Sem 2 A, and Sem 2 B, with a total of 197 students. Specifically, the cohort sizes were 54, 85, and 58 students, respectively. The experimental group comprised students from two cohorts: Sem 1 C and Sem 2 C. The cohort sizes were 60 and 126, respectively, with a total of 186 students in the experimental group.

Table 1. Total students across semesters

Cohort (controlled)	No. of students	Cohort (experimental)	No. of students
Sem 1 A	54	Sem 1 C	60
Sem 2 A	85	Sem 2 C	126
Sem 2 B	58		
Total	**197**	**Total**	**186**

Since the primary objective was to compare the means between the control and experimental groups, the sample size was calculated using GPower, based on the criteria d = 0.5, α = .05, and (1 - β) = .80 for an independent t-test (Kang, 2021; Kim & Park, 2019), as shown in Figure 1. The sample size for each group was set at 64. An equal sample size in both groups, with a 1:1 ratio, maximizes the statistical power of the analysis while minimizing the risk of Type I and Type II errors (Kim & Park, 2019; Rusticus & Lovato, 2014).

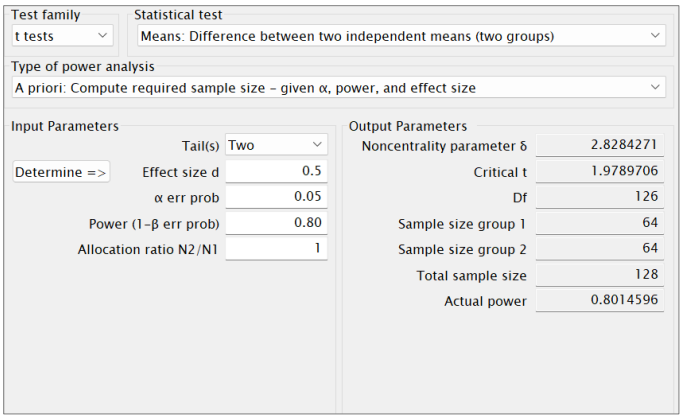

Figure 1. Required sample size for independent t-test

Additionally, equal sample sizes strengthen internal validity by reducing potential biases in variance estimation, ensuring compliance with the assumption of homogeneity of variance in t-test analysis (Rusticus & Lovato, 2014). In line with best practices in educational research, balanced group sizes enhance the comparability and interpretability of findings, particularly in controlled intervention studies (Lydersen, 2018). Finally, the minimum required sample size for each group in a t-test is 40 (Skaik, 2015), making 64 participants per group sufficient for this study

To ensure all the participants in each group were randomly assigned, the systematic sampling technique was applied. First, the 197 students were randomly arranged in a list using Microsoft Excel. Then, systematic random sampling was applied with a sampling interval of approximately 3 (197/64 ≈ 3). Every third student on the list was selected and assigned to the control group. To ensure that the control group included exactly 64 students, one additional student was randomly selected, as selecting every third student alone would have resulted in only 63 students. Similarly, for the experimental group, the 186 students were randomly arranged in Microsoft Excel, and using the same systematic random sampling method with a sampling interval of 3 (186/64 ≈ 3), every third student was selected and assigned to the experimental group.

BASELINE SCORE

In addition to randomization to minimize confounding factors, this study used coursework scores as baseline data. These scores were analyzed using an independent t-test, with the necessary assumptions such as normality and homogeneity of variance for the t-test examined beforehand.

The results of the normality tests for both the control and experimental groups are presented in Table 2. The Kolmogorov-Smirnov and Shapiro-Wilk analyses indicated that all p-values exceeded the 0.05 significance threshold. Therefore, the assumption of normality holds for both groups, suggesting that the data are normally distributed and suitable for parametric testing.

Table 2. Normality test for baseline score

Tests of normality							
		Kolmogorov-Smirnova			Shapiro-Wilk		
		Statistic	df	Sig.	Statistic	df	Sig.
Coursework	Controlled	0.096	64	.200*	0.974	64	0.19
	Experimental	0.072	64	.200*	0.966	64	0.076
* This is a lower bound of the true significance							
a Lilliefors Significance Correction							

The results of the homogeneity of variance test, as shown in Table 3, indicate that Levene's statistics for all criteria (based on mean, median, median with adjusted degrees of freedom, and trimmed mean) were non-significant, with p-values greater than 0.05 ($p > 0.418$ for all). Therefore, the assumption of homogeneity of variances is satisfied, suggesting that the variances between the control and experimental groups are equal, and the data is suitable for parametric analysis using an independent t-test.

Table 3. Test of homogeneity of variance for baseline score

Test of homogeneity of variance					
		Levene Statistic	df1	df2	Sig.
Coursework	Based on mean	0.661	1	126	0.418
		Levene Statistic	df1	df2	Sig.
	Based on median	0.626	1	126	0.43

Test of homogeneity of variance					
Coursework	Based on the median and with adjusted df	0.626	1	120.359	0.43
	Based on trimmed mean	0.635	1	126	0.427

Table 4 shows the results of the independent samples t-test for coursework scores between the control and experimental groups. The t-test for equality of means revealed no statistically significant difference in coursework scores between the two groups, with a two-sided p-value = 0.05 (t(126) = -1.98). The mean difference between the control and experimental groups was -1.01 (SE = 0.51), with a 95% confidence interval ranging from -2.02 to -0.00074.

Table 4. Independent samples t-test for borderline score

Independent samples test									
		t-test for equality of means							
		t	df	Significance		Mean difference	Std. error difference	95% confidence interval of the difference	
				One-sided p	Two-sided p			Lower	Upper
Course work	Equal variances assumed	-1.98	126	0.025	0.05	-1.01094	0.51047	-2.0211	-0.00074
	Equal variances not assumed	-1.98	123.58	0.025	0.05	-1.01094	0.51047	-2.0213	-0.00055

Given that -0.00074 is extremely close to zero when rounded to two decimal places, the upper limit is 0.00, which indicates that the interval effectively includes zero. This supports the conclusion that there is no statistically significant difference in coursework scores between the two groups (Altman, 2005; Hoekstra et al., 2014).

Implementation of AI-Assisted Retrieval Practice

The implementation of this study is based on formative assessment practices (Bell & Cowie, 2001; Ruiz-Primo, 2011), which consist of three key components:

(i) gathering information,
(ii) interpreting information, and
(iii) using the information.

Additionally, it is crucial to explicitly explain and clarify learning goals at the beginning of each class to ensure focus and direction (Bell & Cowie, 2001; Ruiz-Primo, 2011).

Figure 2 illustrates the process of formative assessment, encompassing four key stages: clarifying learning goals, gathering information, interpreting the information, and using the information. Each stage is designed to continuously inform and improve student learning through ongoing feedback and instructional adjustments. These phases, except the first one, were integrated with ChatGPT to enhance the learning experience and align with the course learning outcome (CLO).

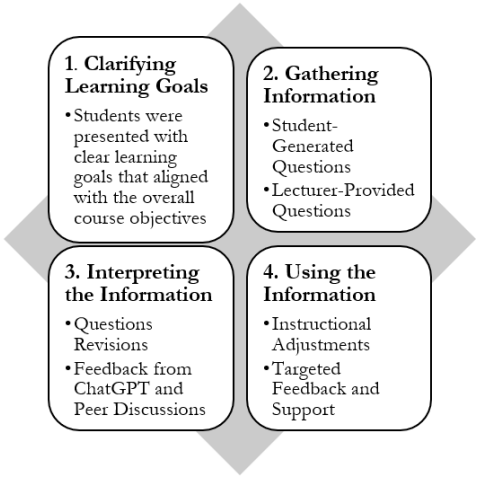

Figure 2. Process of formative assessment

Clarifying learning goals

The Measurement and Evaluation in Education course spans one semester, covering 12 topics over approximately 14 weeks. Each topic is addressed in a weekly two-hour class session. The course is structured around four distinct learning outcomes, which are presented to students during the first week of the semester. Each learning outcome is evaluated using a different assessment method. The final exam, designed to assess the first learning outcome, encompasses all the topics covered in the course. This exam measures students' abilities across the cognitive levels of remembering, understanding, applying, and analyzing and typically consists of 60 multiple-choice questions.

To ensure clarity and focus, the relevant learning outcomes are reiterated at the beginning of each class session, helping students understand the specific goals for that week's topic. During the course, students were also informed about the objectives of the AI-assisted retrieval practice, with a specific focus on improving exam performance. The need for improvement was highlighted by referencing the unsatisfactory exam scores from the previous year. Additionally, the role of AI in generating questions and providing feedback was explained.

Gathering information

Pavelea and Moldovan (2020) have studied a few factors that contribute to students' academic performance. They reported that some of the significant contributors are attention during courses and involvement in activities that influence final examination scores. Therefore, to implement retrieval practice, weekly tasks were designed to actively engage students in recalling and applying course content. These tasks involved both lecturer-generated and student-generated questions. The use of questions as a central task is supported by evidence indicating that test-like activities significantly enhance learning and improve long-term memory retention (Desy et al., 2017; Karpicke, 2017).

Validity in formative assessment depends on alignment with learning objectives, ensuring that assessments measure what they intend to assess (Divjak et al., 2024; Stobart, 2012). Developing clear guidelines based on teachers' knowledge and practices further strengthens assessment validity (Aglanovna et al., 2024). To ensure that both student- and lecturer-generated questions meet formative assessment standards:

(i) All questions were directly linked to weekly course topics and learning objectives (only up to the applying level), enhancing instructional coherence and engagement.
(ii) Students developed questions strictly based on assigned topics from the learning handbook, maintaining content relevance. Additionally, specific commands for ChatGPT to

generate multiple-choice questions (MCQs), as utilized in Yusof and Ismail (2023), were applied to guide question formulation systematically.

(iii) Before use, student-generated questions were reviewed and revised by the lecturer for clarity, appropriateness, and alignment with assessment criteria, preventing ambiguities and ensuring accuracy.

There were two rounds of retrieval practice: one during class and another during revision week. In the context of a two-hour class, students were prompted to answer the questions 10 to 15 minutes after the lecture without referring to any notes. Following this, feedback was provided, and a class discussion was held to deepen understanding. Students then submitted all tasks through the e-learning platform. As a final step in preparation for the examination, students were required to revisit and answer all the tasks again during their study week.

During revision week, students revisited the same questions from each week's practice, including both lecturer-generated and student-generated questions. Additionally, each topic was supplemented with extra questions voluntarily developed by students using ChatGPT. These questions were first verified by the lecturer before being shared with their peers. A notable challenge in this implementation was the low participation in creating and sharing additional questions, likely due to the voluntary nature of the task. However, it is worth noting that most students actively engaged by submitting answers to the questions developed by their peers.

The approach of prompting students to answer questions after a lecture and revisit these tasks during their study week is grounded in the principle that retrieval practice enhances long-term retention. As demonstrated by Roediger and Karpicke (2006), while repeated studying may be beneficial for immediate recall, retrieval practice proves more effective for delayed recall, such as weeks after the initial learning or during exams held day. This aligns with the design of this course, where students engage in retrieval practice shortly after the lecture and then again during their study week, effectively preparing them for the final exam.

Lecturer-generated questions

Students were provided with a variety of objective question types, including multiple-choice questions, true-false questions, word puzzles, and short-answer questions. Each task required students to complete the answers independently.

Table 5 outlines the weekly topics involved in retrieval practice and specifies the type of tasks assigned, including multiple-choice questions, short-answer questions, and word puzzles, along with the number of items and the duration of each practice session. These lecturer-generated questions are designed to reinforce students' understanding of the weekly content, with practice sessions lasting between 10 and 15 minutes.

Table 5. Lecturer-generated questions

Week	Topic	Type of task	No. of items	Practice duration (mins)
Week 1	Introduction to testing, measurement, evaluation, and assessment	Multiple-choice question, short answer	20	15
Week 2	Formative and summative assessment	Multiple-choice question, word puzzles	20	15
Week 3	Measurement scales			
Week 9	Traditional and alternative assessment	True/false	10	10
Week 10	Validity and reliability	Short answer	10	10

Week	Topic	Type of task	No. of items	Practice duration (mins)
Week 11	Basic statistics (mean, mode, median, variance, standard deviation)	Multiple-choice question	2 data sets	15
Week 12	Item analysis (difficulty and discrimination index)	Multiple-choice question	2 data set	15

Student-generated questions

In Week 5, during the topic of item development, students engaged in a group-based retrieval practice activity designed to reinforce their understanding. Working in groups, they developed three objective questions that targeted the remembering, understanding, and applying levels of the Revised Bloom's Taxonomy. These questions must be based on the topics covered in the course and were created with the assistance of ChatGPT.

Figure 3 presents a sample of three questions (in Malay language) developed by one of the student groups. These questions focus on the topic of Alternative and Traditional Assessment and include three types of objective questions: a fill-in-the-blank question, a true-false question, and a multiple-choice question. Once the questions were developed, each group shared their set with the rest of the class. For example, in a class with five groups, each would respond to the questions the other four groups created. This process included answering the questions and critically evaluating their peers' work to ensure that the multiple-choice stems were clearly written, the options were unambiguous, and the questions adhered to the correct format.

Topik	**Tahap Kognitif**			**Total**
	Mengingati	**Memahami**	**Mengaplikasi**	
Pentaksiran Tradisional dan Pentaksiran Alternatif	(1, Isi tempat kosong) Dalam pentaksiran _____, penilaian biasanya dilakukan melalui ujian objektif seperti peperiksaan pilihan berganda.	(1, Betul/Salah) Pentaksiran tradisional adalah secara tidak langsung manakala pentaksiran alternatif mentaksir kemahiran berfikir yang lebih tinggi.	(1, MCQ) Apakah perbezaan utama antara pentaksiran tradisional dan pentaksiran alternatif dalam konteks mengaplikasikan pengetahuan? A. Pentaksiran tradisional menekankan penguasaan fakta, manakala pentaksiran alternatif menekankan kemahiran berfikir kritis. B. Pentaksiran tradisional tidak melibatkan ujian bertulis, manakala pentaksiran alternatif hanya menguji hafalan. C. Pentaksiran tradisional hanya melibatkan ujian akhir, manakala pentaksiran alternatif melibatkan projek sahaja.	**3 soalan**

Figure 3. Sample of students' generated questions

Interpreting information

After each retrieval practice session, a discussion was held immediately. For lecturer-generated questions, the correct answers were reviewed and discussed with the students to ensure understanding. For student-generated questions, each group member presented their answers to the class. Any misconceptions that arose during these presentations were addressed with immediate feedback, allowing students to correct their understanding in real-time. Students were also encouraged to use ChatGPT to verify their answers.

During revision week, students primarily relied on ChatGPT for feedback. After answering all the questions, especially those they had generated themselves using ChatGPT, they used the tool to receive immediate feedback. Students only sought further clarification from the lecturer via WhatsApp in cases of confusion or ambiguity, as shown in Figure 4, where they asked for clarity on the questions they developed and confirmation of the feedback provided by ChatGPT.

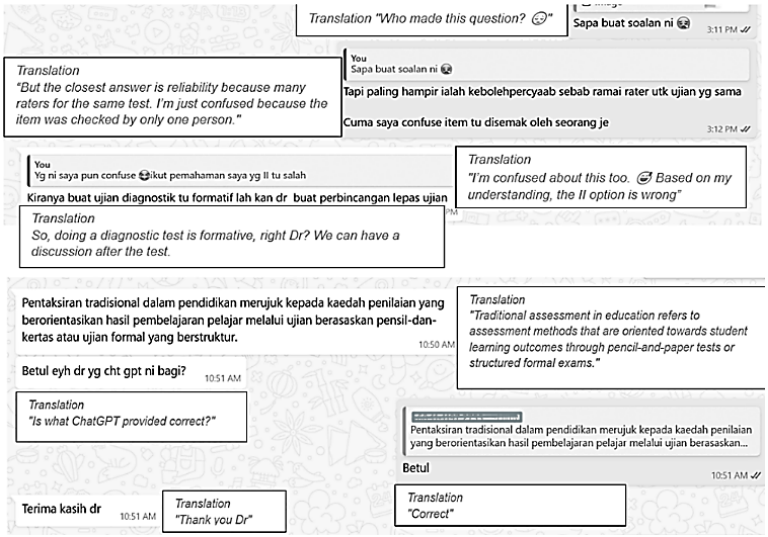

Figure 4. Feedback on student-developed questions and follow-up clarifications

Using information

After receiving feedback from the lecturer, ChatGPT, and peer discussions during each class practice, students were asked to write brief reflections on their learning experiences, outline their expectations for the next retrieval practice session, and prepare for the upcoming final exams.

For example, one area where students commonly expressed confusion during the first week of retrieval practice was the overlap between the concepts of testing, measurement, evaluation, and assessment. These terms are often used interchangeably, which can create confusion. One student reflected, "*I am confused about the differences between testing, measuring, evaluating, and assessing. They seem to overlap, and I find it difficult to clearly understand where one concept ends and the other begins*". Similarly, another student expressed, *"I still don't understand the key differences between these terms. All the answer options seem almost the same to me. Please pray I get an A for this course!"* Another student noted, "*Some books use the term evaluation, while some articles refer to assessment, and their definitions make it very hard for me to understand the difference. However, thank you, Dr., for the explanation; it has helped clarify things a bit.*"

To address this concern, a more detailed explanation with clear examples and scenarios was provided during the following week's class, as the two-hour session initially allocated did not allow enough time for an in-depth discussion. Therefore, any concerns that arose during the week were addressed in the subsequent class session to ensure students fully understood the distinctions between these concepts. Additionally, short supplementary learning videos were created using Loom, as shown in

Figure 5, specifically tailored for students needing extra support or clarification on key concepts. These videos provided students with the flexibility to access the content at their own pace and revisit explanations as needed.

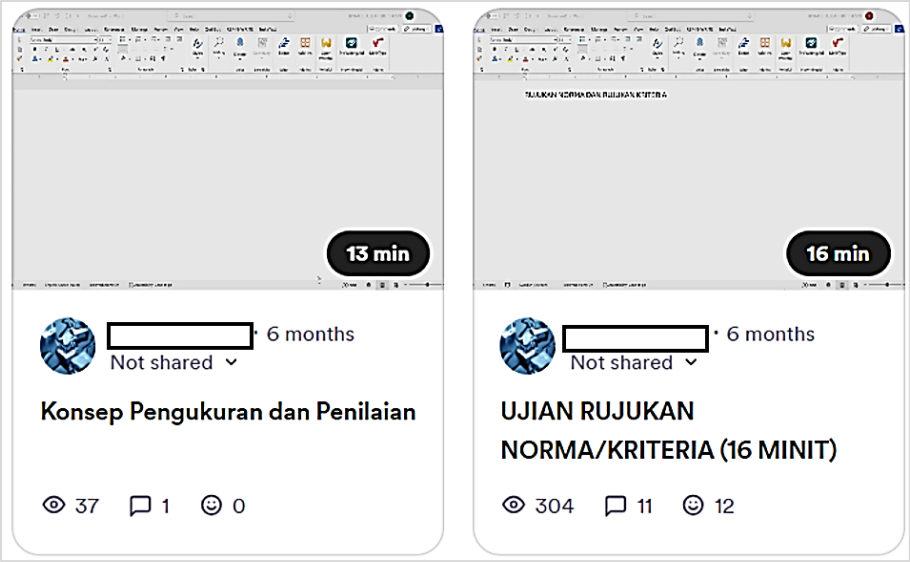

Figure 5. Examples of short learning videos on specific topics

Students were also encouraged to write their expectations for the next retrieval practice based on the feedback they received. Several students expressed gratitude for the feedback provided during retrieval practice, noting its impact on their confidence and performance expectations. One student reflected, *"Thank you, Dr., for the explanation. I feel more confident now, and I hope this retrieval practice will help me perform better in the final exams. The feedback was really useful, and I believe that by continuing with these practices, I can improve in the areas I'm struggling with."* Additionally, students highlighted the novelty and effectiveness of the retrieval practice. As another student shared, *"I've never done retrieval practice like this before, and I realize how effective it is for identifying my weak points. I hope you can provide more questions in the next session so I can better understand where I need to focus my studies."*

However, during the revision week, no formal reflections were requested, as formal classes had already concluded by that time. Instead, students were encouraged to independently engage with ChatGPT to review the material and clarify any lingering doubts. They were also advised to reach out to the lecturer for feedback or additional guidance as needed in the lead-up to the final exam. This approach allowed students to take ownership of their learning while still having flexible access to both ChatGPT-assisted support and lecturer feedback, ensuring they were well-prepared for the final assessment.

Lastly, at the end of the semester, a questionnaire was administered to gather feedback on students' experiences and perceptions regarding retrieval practice, question development, and feedback mechanisms. The questionnaire was self-developed and has been implemented over two semesters. The complete questionnaire is provided in the Appendix for reference. The questionnaire consisted of five sections, each containing four items, measured on a 4-point Likert scale ranging from 1 (strongly disagree) to 4 (strongly agree):

(i) *Perceptions of Retrieval Practice:* Focused on evaluating students' views on the effectiveness of retrieval practice in enhancing their learning and exam preparation.
(ii) *Preference for Student-Developed vs. Lecturer-Generated Questions:* Examined preferences and perceived benefits of questions developed by peers versus those generated by lecturers.

(iii) *Experience with ChatGPT in Developing Questions:* Explored how ChatGPT was used to assist students in developing questions and its perceived utility in reinforcing learning.
(iv) *Experience with ChatGPT Feedback:* Assessed the clarity, usefulness, and accessibility of feedback provided by ChatGPT during retrieval practice sessions.
(v) *Preference for Feedback from Lecturer vs. ChatGPT:* Investigated students' preferences for feedback sources, comparing lecturer feedback with ChatGPT-generated feedback.

A principal-components analysis of residuals (PCAR) was conducted to assess the unidimensionality of the instrument across two cohorts, as shown in Table 6. Item reliability was high for both cohorts, with Cohort 1 achieving a reliability of 0.85 and Cohort 2 achieving 0.91. The variance explained by measures was 25.0% for Cohort 1 and 16.7% for Cohort 2, aligning closely with expected values. However, eigenvalues for the 1st contrast exceeded the threshold of 2.0 in Cohort 1 (3.19), suggesting potential multidimensionality. In contrast, Cohort 2 exhibited lower eigenvalues (1.95), indicating stronger unidimensionality.

Table 6. Principal component analysis of residuals

Cohort	Total students	Total submission (n)	Item reliability (n=20)	Principal-components analysis of residuals			
				Raw variance explained by measures		Raw unexplained variance	
				Observed	Expected	Eigenvalue (1st contrast)	Eigenvalue (2nd contrast)
Sem B 1	60	29	0.85	25.0%	24.9%	3.19	2.42
Sem B 2	126	73	0.91	16.7%	16.6%	1.95	1.89

To investigate the higher eigenvalues observed in Cohort 1 further, an analysis of residual correlations was conducted to identify potential local dependence among items, as shown in Table 7. The analysis revealed that all standardized residual correlations were below the threshold of +0.7, indicating no evidence of highly locally dependent items (Linacre, 2025). The highest residual correlation was 0.45 between items A1 and D3, suggesting moderate shared variance that does not warrant item removal. These findings suggest that the observed higher eigenvalues in Cohort 1 may be attributed to its smaller sample size rather than significant multidimensionality or local dependence among the items. This interpretation is further supported by the lower eigenvalues observed in Cohort 2 (1.95 and 1.89), which demonstrate stronger unidimensionality, likely due to the larger sample size providing more stable variance estimates.

Table 7. Largest standardized residual correlations

Correlation	Entry number item 1	Entry number item 2
0.45	1 A1	15 D3
0.44	5 B1	11 C3
0.39	8 B4	15 D3
0.37	2 A2	19 E3
0.34	15 D3	19 E3
0.32	1 A1	8 B4
0.31	17 E1	20 E4
-0.61	1 A1	9 C1
-0.59	12 C4	19 E3
-0.55	14 D2	20 E4
-0.53	4 A4	20 E4

Correlation	Entry number item 1	Entry number item 2
-0.52	2 A2	12 C4
-0.37	11 C3	17 E1
-0.37	2 A2	13 D1
-0.36	4 A4	7 B3
-0.34	5 B1	15 D3
-0.33	3 A3	9 C1
-0.33	10 C2	11 C3
-0.32	9 C1	15 D3
-0.31	4 A4	6 B2

FINDINGS

COMPARISON OF EXAM SCORES BETWEEN CONTROL AND EXPERIMENTAL GROUPS

Prior to conducting the t-test, assumption analyses were performed to ensure the validity of the results. Specifically, the normality of the data and the homogeneity of variances were evaluated, as these assumptions are critical for the appropriate application of the t-test and the validity of its conclusions.

The normality of exam scores for both the control and experimental groups was assessed using the Kolmogorov-Smirnov and Shapiro-Wilk tests, as shown in Table 8. In the control group, the Kolmogorov-Smirnov test yielded a p-value of .200, while the Shapiro-Wilk test produced a p-value of .294. Similarly, in the experimental group, the p-values were .200 for the Kolmogorov-Smirnov test and .486 for the Shapiro-Wilk test. As all p-values exceeded the standard threshold of $p>.05$, these findings confirm that the exam scores in both groups are normally distributed, thereby satisfying the assumption of normality necessary for further analysis.

Table 8. Normality test for exam score

Tests of normality							
Group		**Kolmogorov-Smirnova**			**Shapiro-Wilk**		
		Statistic	**df**	**Sig.**	**Statistic**	**df**	**Sig.**
Exam-score	Controlled	0.089	64	.200*	0.978	64	0.294
	Experimental	0.051	64	.200*	0.982	64	0.486
*. This is a lower bound of the true significance.							
a. Lilliefors Significance Correction							

During the analysis of exam scores using Stem & Leaf plot as shown in Table 9, an outlier was identified in the control group with a score of 10, which is lower than the rest of the data. According to Aguinis et al. (2013), outliers should not be automatically removed unless there is clear evidence that they result from measurement error or do not represent the population of interest. The outlier was retained because it represents a legitimate data point within the natural variability of the student population.

Table 9. Stem-and-leaf plot

Examscore stem-and-leaf plot for group = controlled		Examscore stem-and-leaf plot for group = experimental	
Frequency	Stem & Leaf	Frequency	Stem & Leaf
1.00	1. 0	5.00	2. 00111
12.00	1. 555667888999	4.00	2. 2222
16.00	2. 0001222223344444	6.00	2. 444555
17.00	2. 55556666788899999	11.00	2. 66666667777
15.00	3. 000111122222244	12.00	2. 888888899999
3.00	3. 666	9.00	3. 000001111
Stem width:	10.00	7.00	3. 2222233
Each leaf:	1 case(s)	6.00	3. 444445
		4.00	3. 6667
		Stem width:	10.00
		Each leaf:	1 case(s)

Furthermore, Levene's test for homogeneity of variances was conducted to assess whether the variances of exam scores were equal between the control and experimental groups, as shown in Table 10. The results indicated that the assumption of equal variances was violated, as the test yielded significant p-values across all four methods: based on the mean (Levene statistic = 7.209, p = .008), based on the median (Levene statistic = 7.132, p = .009), based on the median with adjusted degrees of freedom (Levene statistic = 7.132, p = .009), and based on the trimmed mean (Levene statistic = 7.173, p = .008). Since all p-values were less than the threshold of $p<.05$, this indicates that the variances between the groups were not equal, and therefore, the assumption of homogeneity of variances was not met.

Table 10. Homogeneity of variance

Test of homogeneity of variance					
Test of homogeneity of variance		**Levene statistic**	**df1**	**df2**	**Sig.**
Test of homogeneity of variance	Based on Mean	7.209	1	126	0.008
	Based on Median	7.132	1	126	0.009
	Based on Median and with adjusted df	7.132	1	118.379	0.009
	Based on trimmed mean	7.173	1	126	0.008

Therefore, this study employed Welch's t-test instead of the standard independent t-test, as the data met the assumption of normality but not homogeneity of variances. Welch's t-test is recommended in such cases because it provides better control of Type I error rates when the assumption of homogeneity of variances is violated (Delacre et al., 2017).

Table 11 presents the descriptive statistics for the exam scores of the control and experimental groups. The control group (N = 64) had a mean exam score of 25.44 with a standard deviation of 5.95 and a standard error of the mean of 0.743. The experimental group (N = 64) had a higher mean exam score of 28.81 with a standard deviation of 4.37 and a standard error of the mean of 0.547. These descriptive statistics indicate that the experimental group, which utilized AI-assisted retrieval practice, performed better on average than the control group.

The results in Table 12 indicate a significant difference in final exam scores between the groups ($t(115.734) = -3.647$, $p < 0.001$). The mean difference in scores is -3.36453, with a standard error of 0.92258. The 95% confidence interval for the difference in means ranges from -5.19186 to -1.53719, indicating that the difference in scores is statistically significant and likely not due to random chance.

Table 11. Descriptive statistics for exam scores

Group statistics					
	ID	**N**	**Mean**	**Std. deviation**	**Std. error mean**
ExamScore	Controlled	64	25.4405	5.94552	0.74319
	Experimental	64	28.805	4.37319	0.54665

Table 12. Independent samples test for exam score

Independent samples test									
		t-test for equality of means							
		t	**df**	**Significance**		**Mean difference**	**Std. error difference**	**95% confidence interval of the difference**	
				One-sided p	**Two-sided p**			**Lower**	**Upper**
Exam-score	Equal variances assumed	-3.647	126	0.000	0.000	-3.36453	0.92258	-5.19029	-1.53877
	Equal variances not assumed	-3.647	115.734	0.000	0.000	-3.36453	0.92258	-5.19186	-1.53719

STUDENT FEEDBACK ON RETRIEVAL PRACTICE AND CHATGPT-ASSISTED

At the end of the semester, following the completion of the final exam, students were invited to complete a questionnaire. The primary purpose of collecting this feedback was to identify areas for improvement in subsequent semesters. Of the 186 students in the experimental group (across two semesters), a total of 102 students responded to the questionnaire. Since the questionnaire consists of Likert-scale items, which are considered ordinal data (Sullivan & Artino, 2013; Yusof, 2024), descriptive analyses, such as percentages, median, and mode, were utilized to summarize the responses.

Perceptions of retrieval practice

Perceptions of students toward retrieval practice were measured based on:

(i) Perception A1: Retrieval practice helped me retain information better for the final exam.
(ii) Perception A2: Retrieval practice improved my understanding of key course concepts.
(iii) Perception A3: Retrieval practice sessions were a valuable addition to my overall learning process.
(iv) Perception A4: I feel more confident about answering exam questions after engaging in retrieval practice.

Table 13 shows the median and mode for all four items were consistently 3, indicating that most students agreed with the statements.

Table 13. Perceptions of retrieval practice

Statistics	**Perception_A1**	**Perception_A2**	**Perception_A3**	**Perception_A4**
Median	3	3	3	3
Mode	3	3	3	3

Figure 6 shows a significant proportion of students expressed positive perceptions, with 71.64% agreeing and 28.40% strongly agreeing that retrieval practice helped them retain information better

for the final exam (A1). Similarly, 53.90% agreed, and 46.10% strongly agreed that it improved their understanding of key concepts (A2). Retrieval practice was also seen as a valuable addition to the learning process, with 56.90% agreeing and 43.10% strongly agreeing (A3). However, confidence in answering exam questions (A4) was slightly lower, with 54.90% agreeing, 26.50% strongly agreeing, and 18.60% disagreeing. These results indicate an overall positive perception of retrieval practice, though confidence in exam performance shows room for improvement.

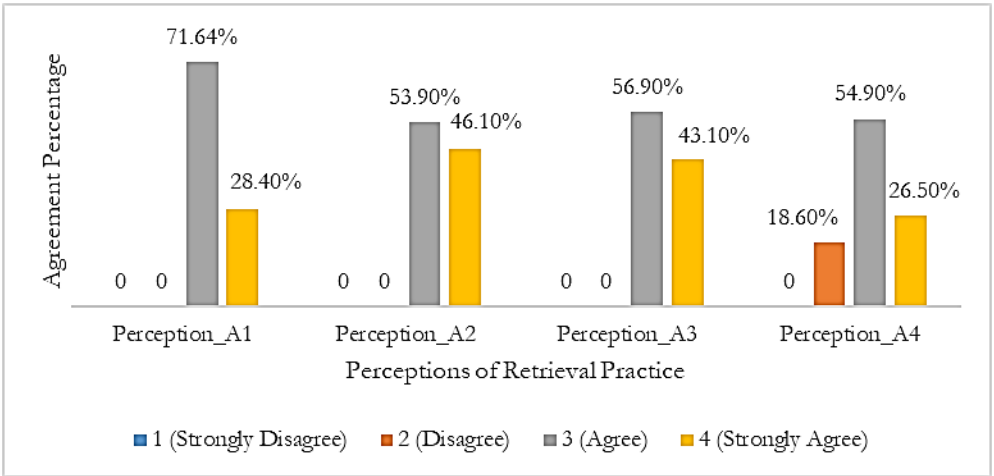

Figure 6. Percentage of agreement (perceptions of retrieval practice)

Preference for student-developed vs. lecturer-developed questions

Students' preferences for student-developed versus lecturer-developed questions were measured based on:

(i) Preference B1: I preferred lecturer-generated questions because they provided a clearer structure for retrieval practice.
(ii) Preference B2: I found student-developed questions more relatable and better aligned with my learning needs.
(iii) Preference B3: Lecturer-generated questions were well-structured and closely matched the format of the final exam.
(iv) Preference B4: I found it easier to learn from retrieval practice sessions that used lecturer-generated questions.

Table 14 illustrates students' preferences for lecturer-generated versus student-developed questions during retrieval practice. The median and mode for all four statements were consistently 3, indicating a general agreement across the responses.

Table 14. Preference for student-developed vs. lecturer-generated questions

Statistics	Preference_B1	PreferenceQS_B2	PreferenceQS_B3	PreferenceQS_B4
Median	3	3	3	3
Mode	3	3	3	3

Figure 7 shows the percentage of students who agreed with each statement regarding their preference for student-developed versus lecturer-developed questions. For B1, which assessed the preference for lecturer-generated questions due to their clearer structure, 41.2% of students agreed, and 30.4% strongly agreed, reflecting a positive inclination toward structured questions provided by the lecturer. Similarly, for B3, 56.9% agreed, and 43.1% strongly agreed that lecturer-generated questions were well-structured and closely aligned with the final exam format. In contrast, B2, which explored the

relatability of student-developed questions, showed more varied responses, with 34.3% agreeing, 22.5% strongly agreeing, and 25.5% strongly disagreeing. Lastly, for B4, which evaluated the ease of learning from lecturer-generated questions, 54.9% agreed, and 26.5% strongly agreed, further emphasizing the preference for lecturer-generated questions. Overall, while students acknowledged the relatability of student-developed questions, the structured and exam-oriented nature of lecturer-generated questions was clearly favored.

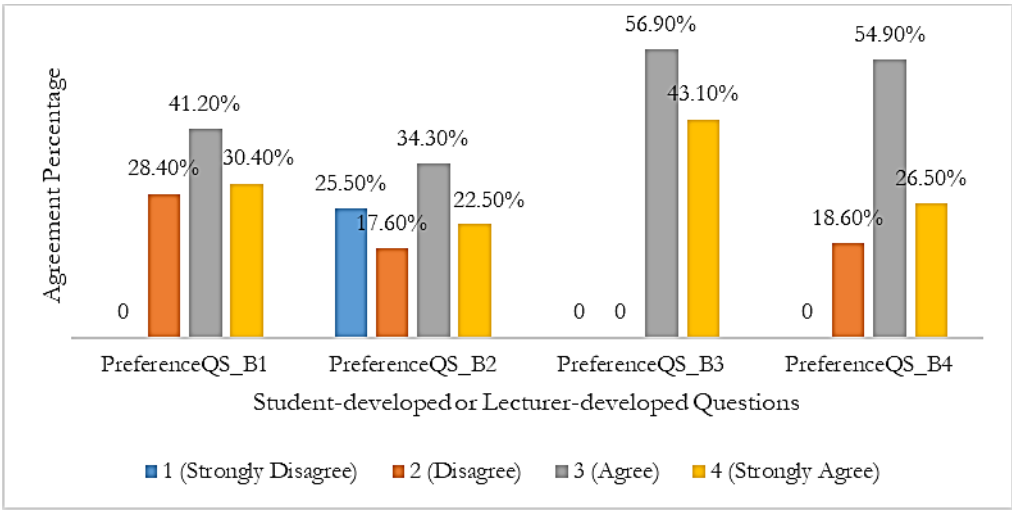

Figure 7. Preference for student-developed vs. lecturer-developed questions

Students' experience with ChatGPT in developing questions

As students were tasked with generating questions using ChatGPT, the following statements assess their experiences with the tool:

(i) ChatGPTQS_C1: ChatGPT helped me create better questions for retrieval practice.
(ii) ChatGPTQS_C2: ChatGPT was useful in helping me structure questions that aligned with the learning objectives.
(iii) ChatGPTQS_C3: I felt more confident in the quality of the questions I developed with ChatGPT's assistance.
(iv) ChatGPTQS_C4: ChatGPT made it easier to generate diverse types of questions for retrieval practice.

Table 15 summarizes students' experiences with ChatGPT in developing questions for retrieval practice. The median for all items was consistently 3, indicating general agreement among respondents, while the mode varied, with C2 having a mode of 4, reflecting a stronger positive response for this item.

Table 15. Students' experience with ChatGPT in developing questions

Statistics	ChatGPTQS C1	ChatGPTQS C2	ChatGPTQS C3	ChatGPTQS C4
Median	3	3	3	3
Mode	3	4	3	3

Figure 8 shows agreement for each statement. For C1, which explored whether ChatGPT helped students create better questions, 33.3% agreed, and 25.5% strongly agreed, highlighting a moderately positive experience. For C2, which assessed ChatGPT's usefulness in structuring questions aligned

with learning objectives, 34.3% agreed, and 22.5% strongly agreed, showing relatively stronger agreement compared to the other items. Similarly, for C3, which measured students' confidence in the quality of questions developed with ChatGPT, 30.4% agreed, and 21.6% strongly agreed, indicating moderate confidence levels. C4, which evaluated whether ChatGPT made it easier to generate diverse question types, had the highest proportion of strong agreement, with 36.3% strongly agreeing and 15.7% agreeing. This indicates that ChatGPT was particularly appreciated for its ability to assist in creating varied question formats. Overall, while students found ChatGPT helpful in various aspects of question development, its role in generating diverse questions stood out as the most impactful.

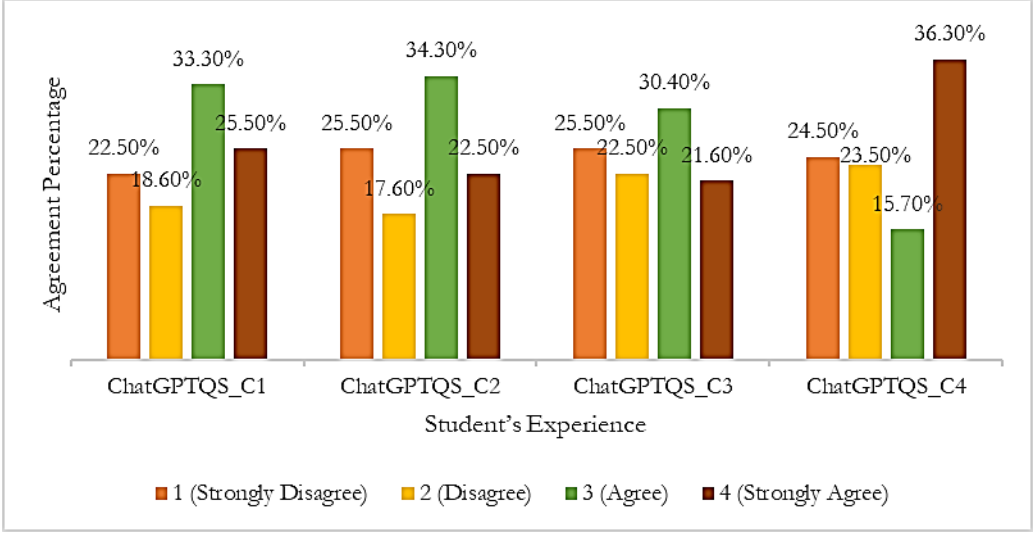

Figure 8. Students' experience with ChatGPT in developing questions

Students' experience with ChatGPT as a feedback tool

To assess students' experience using ChatGPT for feedback, the following statements measure their perceptions:

(i) ChatGPTFeed_D1: ChatGPT provided sufficient feedback for me to understand my mistakes.
(ii) ChatGPTFeed_D2: I feel less confident relying solely on ChatGPT feedback.
(iii) ChatGPTFeed_D3: ChatGPT's feedback was accessible whenever I needed support for reviewing my answers.
(iv) ChatGPTFeed_D4: ChatGPT's feedback sometimes lacked the depth needed for complex questions.

Table 16 summarizes students' experiences with ChatGPT as a feedback provider. The median for most items was 3, indicating a neutral to moderately positive perception of ChatGPT's feedback. However, D3 had a median of 4, suggesting stronger agreement that ChatGPT was accessible for reviewing answers. The mode followed a similar pattern, with D4 having the lowest rating (mode = 2), indicating that students found ChatGPT's feedback lacking in depth for complex questions.

Table 16. Experience with ChatGPT feedback

Statistics	ChatGPTFeed D1	ChatGPTFeed D2	ChatGPTFeed D3	ChatGPFeed D4
Median	3	3	4	2
Mode	3	3	4	2

Figure 9 illustrates that students generally perceived ChatGPT's feedback as both sufficient and accessible. For D1 (ChatGPT provided sufficient feedback for me to understand my mistakes), 29.4% agreed, and 22.5% strongly agreed, indicating that many students found ChatGPT's feedback helpful. However, 25.5% strongly disagreed, suggesting that a notable portion of students felt the feedback did not adequately address their mistakes. For D3 (ChatGPT's feedback was accessible whenever I needed support for reviewing my answers), responses were overwhelmingly positive, with 55.9% strongly agreeing and 21.6% agreeing, highlighting that students widely valued ChatGPT's availability during retrieval practice.

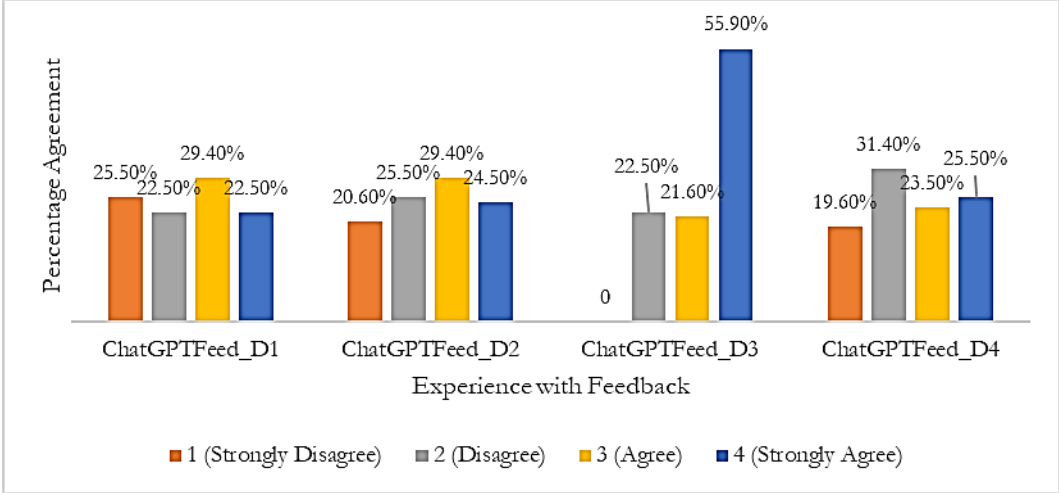

Figure 9. Students' experience with ChatGPT as a feedback tool

Conversely, for D2 (I feel less confident relying solely on ChatGPT feedback), the responses reflected some reservations, with 29.4% agreeing and 24.5% strongly agreeing, showing that a significant number of students were hesitant to depend entirely on ChatGPT's feedback. However, 20.6% strongly disagreed, indicating that some students felt confident using ChatGPT independently. For D4 (ChatGPT's feedback sometimes lacked the depth needed for complex questions), 31.4% disagreed, while 25.5% strongly agreed, revealing a divide in perceptions regarding ChatGPT's effectiveness in handling complex topics. While some students appreciated its feedback, others found it lacking the depth necessary for more advanced inquiries.

Feedback preference

In addition to receiving feedback from ChatGPT, students also receive feedback from lecturers during class or through WhatsApp. Their feedback preferences were measured based on the following statements:

(i) PreferenceFeed_E1: I found lecturer feedback to be more comprehensive than ChatGPT's feedback.
(ii) PreferenceFeed _E2: ChatGPT's feedback was quicker and more convenient than receiving feedback from the lecturer.
(iii) PreferenceFeed _E3: I trust lecturer feedback more than ChatGPT feedback for ensuring accurate understanding.
(iv) PreferenceFeed _E4: I would like a combination of feedback from both ChatGPT and the lecturer in future learning sessions.

Table 17 summarizes students' feedback preferences regarding ChatGPT and lecturer-generated feedback. The median and mode for E2 and E3 was 3, indicating general agreement that ChatGPT's feedback was quick and convenient, but lecturer feedback was more comprehensive and trustworthy. Meanwhile, E1 and E4 had a mode and median of 4, suggesting stronger agreement that lecturer

feedback provided more depth and that students preferred a combination of both ChatGPT and lecturer feedback in future learning sessions.

Table 17. Feedback preference

Statistics	PreferenceFeed E1	PreferenceFeed E2	PreferenceFeed E3	PreferenceFeed E4
Median	4	3	3	4
Mode	4	3	3	4

Figure 10 shows students' preferences regarding feedback sources, showing distinct patterns of agreement and disagreement across the items. For E1 (I found lecturer feedback to be more comprehensive than ChatGPT's feedback), responses were highly positive, with 54.9% strongly agreeing and 26.5% agreeing, indicating a strong consensus on the depth and comprehensiveness of lecturer feedback. Similarly, for E3 (I trust lecturer feedback more than ChatGPT feedback for ensuring accurate understanding), 52.0% agreed, and 48.0% strongly agreed, reinforcing students' trust in lecturer-provided feedback. For E2 (ChatGPT's feedback was quicker and more convenient than receiving feedback from the lecturer), 49.0% agreed, and 24.5% strongly agreed, highlighting ChatGPT's accessibility. However, 26.5% disagreed, suggesting that while ChatGPT was valued for its speed, it did not fully meet all students' expectations.

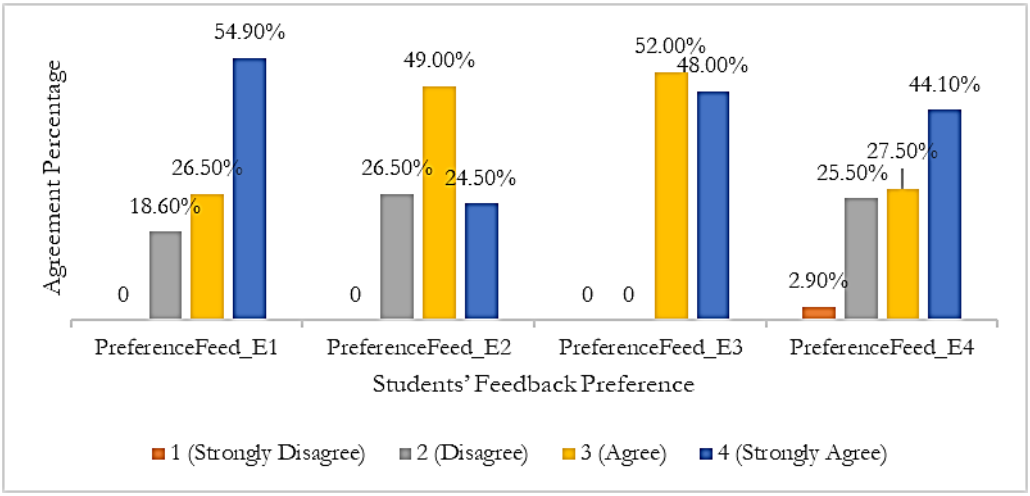

Figure 10. Students' feedback preferences

For E4 (I would like a combination of feedback from both ChatGPT and the lecturer in future learning sessions), 44.1% strongly agreed, and 27.5% agreed, indicating a clear preference for a blended feedback approach, combining the efficiency of ChatGPT with the depth and reliability of lecturer feedback.

DISCUSSION

The implementation of retrieval practices during the course was motivated by the observation that students' final exam scores were generally lower compared to their coursework marks, even among those who performed well in coursework. Retrieval practices were introduced due to their proven effectiveness in sustaining memory over time (Bishop, 2022; Jaeger et al., 2014; Karpicke, 2017; Moreira et al., 2019), ensuring that knowledge is retained until the final exams. Research has shown

that retrieval practices become even more effective when they are similar format to final exams, repeated and accompanied by formative feedback (Agarwal et al., 2016, 2021; Casselman, 2024; Roediger & Butler, 2011).

Building on this evidence, the current implementation utilized ChatGPT as a tool to facilitate student engagement by encouraging its use for both creating questions and obtaining immediate feedback on practice responses. This dual functionality of ChatGPT was designed to enhance the effectiveness of retrieval practice by integrating active learning with timely feedback, thereby better-supporting students in their exam preparation.

Impact of Retrieval Practice on Exam Performance

The implementation of retrieval practices during the course demonstrated a positive impact on students' final exam performance, as evidenced by the improved scores in the experimental group compared to the control group. This improvement underscores the effectiveness of retrieval practices in enhancing long-term retention and application of knowledge. The observed progress aligns with research emphasizing the role of retrieval practice in strengthening memory and fostering deeper learning (Agarwal et al., 2021; Casselman, 2024; Karpicke, 2017; Moreira et al., 2019).

The retrieval practices in this study were conducted twice before the final exam: first during weekly class sessions, where individual topics were addressed, and again during the study week, which reviewed all topics covered throughout the course. This two-stage approach allowed students to revisit the material at different intervals, utilizing the spacing effect, which is known to enhance memory retention over time. Prior research highlights that retrieval practice, particularly when spaced out over intervals such as days or weeks, significantly improves long-term retention (Ariel & Karpicke, 2018; Ma et al., 2020; YeckehZaare et al., 2019). The results of this study align with these findings, demonstrating that retrieval practice is an effective tool for supporting academic performance.

A potential confounding factor in this study was the possibility that students independently engaged in additional retrieval practices outside the structured sessions. However, the random selection of participants for both the experimental and control groups minimized this risk (Shadish et al., 2011), ensuring that external influences, such as independent retrieval efforts, were evenly distributed across the groups. This approach strengthens the validity of the findings, providing robust evidence that the structured retrieval practice sessions were a significant factor in the observed improvement in exam performance. These results are consistent with YeckehZaare et al. (2019), who demonstrated that retrieval tools effectively enhance learning outcomes, even when controlling for external factors such as prior knowledge or study habits, further underscoring the reliability and utility of retrieval practices in improving academic achievement.

Furthermore, students' perceptions of the retrieval practice, collected after the final exam, indicated overwhelmingly positive feedback. Most students agreed that retrieval practice helped them retain information better for the final exam, demonstrating its value in enhancing memory retention. Additionally, many students reported that retrieval practice improved their understanding of key course concepts, reinforcing its role in fostering deeper learning. The sessions were widely regarded as a valuable addition to their overall learning process, with students noting that the practice bolstered their confidence in answering exam questions.

During the implementation of retrieval practice, the primary focus was on improving final exam performance, treating retrieval practice as a formative activity to support learning rather than an opportunity for detailed quantitative analysis. Hence, feedback was provided immediately after each session to address misconceptions and reinforce learning, consistent with research demonstrating that feedback enhances the effectiveness of retrieval practice (Agarwal et al., 2021; Casselman, 2024; Jaeger et al., 2024). However, the absence of recorded individual performance data during these sessions presented limitations. Specifically, it prevented the identification of students with lower memory capac-

ity, who are known to benefit significantly from retrieval practice when paired with immediate feedback and targeted support (Agarwal et al., 2016; Racsmány et al., 2020; Roediger & Butler, 2011). This limitation also restricted a more detailed analysis of how each retrieval session contributed to the progress of students with varying memory capabilities.

Despite these limitations, the findings underscore the importance of embedding retrieval practice within formative assessments and incorporating immediate feedback to maximize its benefits. By providing opportunities for repeated recall and addressing misunderstandings in real-time, retrieval practice proved to be a valuable tool for enhancing long-term retention and supporting students' academic performance (Casselman, 2024; Hui et al., 2022). The findings corroborate those of Meier and Löfqvist (2024), affirming ChatGPT's potential as a valuable tool for enhancing retrieval practices. Its demonstrated ability to generate targeted questions and provide effective feedback highlights its capacity to support and enrich learning processes.

ChatGPT-Assisted in Retrieval Practice

This study clarifies the role of ChatGPT in supporting retrieval practice by demonstrating its potential to enhance both student performance and learning experiences. While research on ChatGPT's role in retrieval practice is still emerging, scholars such as Meier and Löfqvist (2024) suggest that it can be a powerful tool for generating questions and providing personalized feedback, thereby reducing educators' workload and enhancing student engagement (Kiryakova & Angelova, 2023; Modran et al., 2024). This aligns with the findings of this study, which demonstrated that students who engaged in ChatGPT-assisted retrieval practices achieved higher final exam scores compared to those who did not.

ChatGPT's role in question development

One of the key ways ChatGPT supported retrieval practice was by helping students create effective questions (Meier & Löfqvist, 2024). It was particularly useful in aligning these questions with the course's learning objectives and encouraging students to think critically about the material. Additionally, ChatGPT's ability to produce a variety of question types (Indran et al., 2023; Zuckerman et al., 2023) can make the process more engaging and help students explore the course content from multiple perspectives.

In this study, while most students agreed that ChatGPT helped them develop questions aligned with the learning outcomes, they ultimately preferred lecturer-generated questions, viewing them as more structured and better suited to assessment expectations. This preference may also highlight a limitation in the implementation of ChatGPT-assisted question development. Students were instructed to create their own questions during only one class session, limiting their exposure to this practice. Additionally, due to the constraints of the two-hour class, which included developing, answering, and providing feedback on questions, the activity was conducted collaboratively in groups rather than individually.

Although this group-based approach was efficient, it may have restricted opportunities for personalized engagement with the material (Silseth & Furberg, 2024; Theobald et al., 2017). In the context of this study, it also constrained students' ability to fully explore and leverage ChatGPT's potential for tailoring questions to their unique learning needs. While ChatGPT proved useful in supporting question development, the implementation lacked the depth and repetition necessary for students to build confidence and proficiency in independently crafting high-quality, outcome-aligned questions. Only a few students voluntarily developed their own questions during the study week, suggesting that students may require more structured opportunities to independently explore ChatGPT's potential.

Another key challenge observed was related to the complexity of crafting multiple-choice questions (MCQs). Students were provided with general prompts to develop MCQs, but these prompts did not specify key elements such as answer keys, options, and distractors. As a result, many students made

errors in structuring their answer choices and distractors, which are essential components of well-designed MCQs. This issue suggests that while ChatGPT assisted in question generation, it did not provide enough guidance on constructing high-quality answer choices and plausible distractors. Studies have demonstrated that the quality of ChatGPT-generated questions depends heavily on the specificity of the prompt given (Kıyak et al., 2024; Kıyak & Emekli, 2024; Rivera-Rosas et al., 2024).

The survey's results further reinforced this challenge, as students expressed a preference for lecturer-developed questions. One possible explanation is that lecturer-generated MCQs tend to be better structured and refined, ensuring alignment with course learning outcomes and assessment expectations. Additionally, students may not have had enough practice in designing effective MCQs, leading them to rely more on questions created by instructors. Future implementations should consider providing more structured guidance on MCQ development, such as step-by-step prompts or AI-assisted refinement tools, to help students generate better-quality questions.

Another limitation of the current implementation was that ChatGPT was not utilized to generate higher-order questions that require analyzing, evaluating, or creating. In this study, students primarily developed lower-order MCQs focused on remembering, understanding, and applying, as these levels were directly aligned with the course's learning objectives. However, Bloom's Taxonomy suggests that higher-order thinking skills, such as analyzing, evaluating, and creating, are essential for deeper learning and critical thinking development (Anderson, 2005; Krathwohl, 2002)

The absence of higher-level question generation may have limited students' ability to engage in deeper cognitive processing. Although developing such questions were not explicitly required in this course, future studies could explore the potential of ChatGPT in scaffolding students' ability to generate more complex and cognitively demanding questions. This could involve using customized prompts to encourage ChatGPT to generate higher-order questions, allowing students to develop a broader range of question types beyond basic recall-based MCQs.

ChatGPT's role in feedback for retrieval practice

Effective feedback is a crucial component of retrieval practice, as it helps students reinforce correct knowledge, identify misunderstandings, and refine their learning strategies. In the context of retrieval practice, timely and constructive feedback allows students to strengthen memory recall, adjust incorrect responses, and deepen their understanding of key concepts (Agarwal et al., 2016; Jaeger et al., 2024; Racsmány et al., 2020). This study examined students' experiences with ChatGPT-generated feedback as an aid in retrieval practice and their preferences between ChatGPT and lecturer-provided feedback. The survey results revealed both the strengths and limitations of ChatGPT as a feedback tool, particularly in its role in assisting students in retrieval practice by offering immediate responses and supporting independent learning.

One of ChatGPT's most notable strengths was its immediacy and accessibility, which students widely appreciated. Unlike lecturer feedback, which is often constrained by time and availability, ChatGPT allows students to receive feedback instantly, particularly outside classroom hours. Many students agreed or strongly agreed that ChatGPT was a useful tool for checking their answers, reviewing mistakes, and clarifying doubts in real-time, making it a valuable asset for self-regulated learning in retrieval practice. Furthermore, the ability to access feedback at any time was also beneficial during study week, when students required immediate responses to reinforce learning. This aligns with research emphasizing that timely feedback is critical in retrieval practice, as it prevents students from reinforcing incorrect knowledge and helps them adjust their understanding before misconceptions take hold (Agarwal et al., 2016; Roediger & Butler, 2011).

Despite ChatGPT's convenience, students overwhelmingly preferred lecturer feedback, particularly for gaining a thorough understanding of complex concepts. Many students felt less confident relying solely on ChatGPT, preferring to verify its feedback with lecturer explanations. A majority of students strongly agreed or agreed that lecturer feedback was more comprehensive, providing greater depth, clarity, and reassurance than ChatGPT's responses. One possible reason for this preference

was the inconsistencies some students encountered in ChatGPT's responses, particularly for difficult concepts. Similar findings have been reported that ChatGPT also struggles to provide reliable responses when addressing sophisticated or multifaceted issues (Naik et al., 2024; Tyson, 2023). Notably, ChatGPT occasionally provided incorrect answers to its own generated questions when students led the process without structured prompts. Since ChatGPT generates responses based on the phrasing and specificity of user input, the quality of its feedback depends heavily on how students frame their queries (Steiss et al., 2024). Furthermore, all students used the free version of ChatGPT, which may have limitations compared to premium versions, potentially affecting response accuracy, depth, and consistency.

Expectedly, students expressed a strong preference for a blended feedback model. A significant number of students strongly agreed or agreed that they would like to receive a combination of feedback from both ChatGPT and lecturers in future learning sessions. The findings suggest that ChatGPT and lecturer feedback serve distinct but complementary roles in retrieval practice. ChatGPT excels in delivering immediate, scalable feedback, making it ideal for quick self-assessment, factual verification, and independent study. However, lecturers remain essential for deeper discussions, personalized clarification, and conceptual accuracy. This finding echoes Otaki and Lindwall's (2024) findings, asserting that the distinction goes far beyond mere interaction. It underscores the critical role of face-to-face engagement in fostering meaningful interpersonal relationships rooted in temporal depth and emotional resonance. These connections transform students and teachers into active participants within shared communities of practice, promoting a deeper sense of collaboration, understanding, and shared purpose.

Research and Practical Implication

This study explores the practical and research implications of integrating ChatGPT into retrieval practices, emphasizing its transformative potential in educational settings. Beyond the immediate findings, the results point to actionable strategies for educators and institutions while identifying critical areas for future research.

Theoretical implications

The findings of this study contribute to the growing body of research on retrieval practice and its role in enhancing long-term memory retention and academic performance. By demonstrating that structured retrieval practice, facilitated through digital tools such as ChatGPT, leads to significant improvements in exam performance, this study strengthens the theoretical foundation of retrieval-based learning within cognitive psychology and educational assessment. Specifically, the results align with the retrieval practice effect (Roediger & Butler, 2011) and the spacing effect (Ariel & Karpicke, 2018), reinforcing the importance of repeated recall over distributed intervals to enhance learning outcomes.

Moreover, this study highlights the value of ChatGPT-assisted formative assessment, particularly through the integration of ChatGPT as a tool for retrieval practice. The ability of ChatGPT to provide immediate feedback aligns with research emphasizing that formative assessment is most effective when feedback is timely, targeted, and actionable. By facilitating frequent, low-stakes retrieval tasks, ChatGPT-supported retrieval practice offers a scalable approach to formative assessment that reinforces memory and understanding over time. The results support the argument that formative retrieval activities, particularly when enhanced with immediate feedback, contribute to deeper learning and academic success (Agarwal et al., 2016, 2021; McDougall & Gruneberg, 2002; Roediger & Butler, 2011).

While this study did not explicitly adopt the Self-Regulated Learning (SRL) theory (Zimmerman & Schunk, 2011) as a theoretical lens, the findings suggest potential intersections with this perspective. The use of ChatGPT enabled students to take greater control over their learning process by generating questions, evaluating their own answers, and engaging in independent study. This aligns with the

SRL model, which emphasizes metacognitive regulation, strategic learning, and self-motivation in academic achievement. Future research could further explore how ChatGPT-assisted retrieval practice fosters self-regulated learning behaviors, particularly in enhancing metacognitive awareness and strategic study habits.

Similarly, while scaffolding theory (Wood et al., 1976) was not a central framework in this study, the findings suggest that ChatGPT-generated feedback could serve as a form of scaffolding in retrieval-based learning. ChatGPT provided students with immediate explanations and guidance, allowing them to adjust their understanding without direct instructor intervention. This suggests that AI can act as a temporary support structure, gradually transferring learning responsibility to students who are a key principle in scaffolding. However, since students in this study still preferred lecturer-generated questions and feedback, future research should explore how AI-driven formative assessment can be structured to better support student autonomy while maintaining instructional quality.

Practical implications

The findings of this study provide several practical implications for the integration of ChatGPT-assisted retrieval practice in educational settings. ChatGPT was used to generate questions and provide immediate feedback. Educators can enhance existing retrieval practice strategies to improve student engagement and learning outcomes. However, structured implementation is required to maximize its effectiveness while addressing the identified limitations.

This study demonstrates that ChatGPT can serve as a valuable tool in retrieval practice, particularly in automating question development and providing instant formative feedback. The ability to generate diverse question types aligned with course objectives allows for more frequent and structured retrieval opportunities. However, to ensure the quality of ChatGPT-generated questions, human oversight is necessary, especially in designing multiple-choice questions (MCQs) that require well-balanced answer choices and plausible distractors. This was also highlighted in studies such as Cheung et al. (2023), Özbay (2024), and Yusof and Ismail (2023). Therefore, educators should develop structured prompting strategies to guide ChatGPT's output and ensure alignment with assessment criteria.

Future implementations of ChatGPT-assisted retrieval practice should also explore its role in fostering higher-order cognitive skill levels such as analyzing, evaluating, and creating. Educators can design ChatGPT-assisted retrieval activities that require students to engage in complex problem-solving rather than simple factual recall. By using customized prompts, ChatGPT can generate case-based or scenario-driven questions that prompt students to justify their reasoning and apply knowledge in new contexts (Mariano et al., 2024). However, the design of these activities should align with the intended learning objectives, as the effectiveness of formative assessment depends on its ability to accurately measure and support student learning progress (Aglanovna et al., 2024; Divjak et al., 2024; Stobart, 2012). Without clear alignment with course goals, ChatGPT-generated retrieval questions or feedback may become disconnected from broader learning objectives, reducing their impact on conceptual understanding.

While students recognized the usefulness of ChatGPT-generated questions, they expressed a preference for lecturer-developed questions, which they perceived as better structured and more reflective of assessment expectations. This suggests that ChatGPT-assisted retrieval practice should be implemented with scaffolding techniques, allowing students to transition from ChatGPT-assisted question development to independent question creation. Structured exercises could involve guided question refinement activities, where students first generate questions with ChatGPT and then critically evaluate and improve them under lecturer supervision.

The findings also highlight that ChatGPT's immediacy in providing feedback was particularly beneficial for independent learning and self-regulated study practices. However, concerns regarding the depth and accuracy of ChatGPT-generated feedback suggest that it should be integrated as part of a hybrid feedback model (Jacobsen & Weber, 2023; Nazaretsky et al., 2024). This approach would strengthen ChatGPT's ability to provide instant clarification and low-stakes formative feedback while

lecturers provide in-depth explanations and targeted guidance for higher-order thinking skills. Institutions may consider designing tiered feedback mechanisms, where ChatGPT-generated feedback is supplemented with lecturer-led discussions to deepen conceptual understanding.

Most importantly, to maximize the benefits of ChatGPT-assisted retrieval practice, both educators and students must be equipped with the necessary skills to effectively utilize ChatGPT in learning and assessment (Elbanna & Armstrong, 2024; Javaid et al., 2023). Professional development programs for educators should focus on designing effective ChatGPT-assisted retrieval activities, particularly in developing structured prompts that guide ChatGPT's question generation to align with learning objectives. Additionally, given the limitations of ChatGPT, professional development should also address potential biases, inconsistencies, and ethical concerns, helping educators develop strategies to critically assess and refine ChatGPT-generated content.

Similarly, students also require targeted training on how to effectively interact with ChatGPT for retrieval practice. Training should emphasize the importance of crafting effective prompts to generate well-structured questions, allowing students to experiment with different question formats such as multiple-choice, short-answer, and case-based problems. Furthermore, students must learn how to critically evaluate ChatGPT-generated responses, distinguishing between accurate and potentially misleading feedback (Dai et al., 2023; Sain et al., 2024). To encourage self-regulated learning, students should be guided on how to use ChatGPT-generated feedback as an initial source of clarification while seeking further guidance from lecturers for deeper understanding. Additionally, discussions on ethical use are essential to prevent over-reliance on ChatGPT and promote academic integrity.

Future research directions

The findings of this study provide several important implications for the integration of ChatGPT-assisted retrieval practice across diverse educational contexts. While this study demonstrated the effectiveness of ChatGPT-assisted retrieval practice in improving final exam performance, further research is needed to examine its long-term effects on memory retention. Retrieval practice is known to strengthen memory over time, but it remains unclear whether ChatGPT-assisted retrieval practice leads to sustained improvements in recall across extended periods. Longitudinal studies could explore whether repeated use of ChatGPT-generated retrieval practice enhances knowledge retention across semesters.

Another key area for future research is the impact of ChatGPT-assisted retrieval practice on higher-order cognitive skills. Traditional retrieval practice primarily strengthens factual recall, but research is needed to investigate whether ChatGPT-generated questions and feedback can also facilitate deeper learning, such as conceptual application, critical thinking, and problem-solving. Since this study primarily focused on recall-based objective questions, future studies could explore how ChatGPT can generate and support retrieval practice for higher-order questions, such as those requiring analysis, evaluation, and synthesis.

Additionally, the nature of feedback provided by ChatGPT, its immediacy, and the frequency of student engagement with the feedback may also influence higher-order cognitive development. Immediate feedback allows students to identify and correct misconceptions in real-time, preventing the reinforcement of incorrect knowledge. However, it remains unclear whether frequent use and exposure to ChatGPT-generated feedback fosters deeper metacognitive awareness and self-regulation skills. Future studies could examine whether students who regularly engage with ChatGPT-assisted retrieval practice and receive continuous feedback develop better problem-solving abilities, improved critical thinking skills, or enhanced conceptual understanding compared to those who rely solely on delayed human feedback.

Furthermore, future research should explore how ChatGPT's feedback can be refined to better support higher-order cognition. While ChatGPT can efficiently assess factual accuracy, its ability to provide detailed feedback on complex, open-ended responses remains limited. Investigating ways to en-

hance ChatGPT-driven formative feedback, such as integrating explanations, prompting deeper reflection, or incorporating scaffolding techniques, could significantly improve its effectiveness in fostering higher-order learning outcomes. Additionally, future studies should consider developing hybrid retrieval practice models that combine ChatGPT-generated feedback with human evaluation to optimize learning quality and ensure pedagogical rigor.

Additionally, future research should investigate students' perceptions of ChatGPT-generated questions and feedback further, focusing on factors such as trust, usability, engagement, and confidence in its reliability and effectiveness. This study found that while students appreciated the convenience and accessibility of ChatGPT-generated questions and feedback, they preferred lecturer-developed questions due to their structured nature and perceived alignment with assessment expectations. Hence, future studies could explore how student confidence in ChatGPT-assisted retrieval practice evolves over time and whether structured guidance on ChatGPT-generated content can enhance trust in its accuracy and educational value. Another important avenue is investigating whether training students to refine prompts and critically evaluate ChatGPT-generated responses can improve their confidence in using ChatGPT for retrieval practice.

CONCLUSION

This study demonstrated the significant potential of ChatGPT-assisted retrieval practice in enhancing student learning outcomes, particularly in improving final exam performance. Integrating ChatGPT into question development and formative feedback processes provided an accessible and scalable solution to support retrieval practice, especially in large-class settings where individualized feedback is often constrained. The findings affirm the effectiveness of retrieval practice as a robust learning strategy while also highlighting the role of ChatGPT in facilitating structured, frequent, and personalized engagement with course content.

Beyond its practical application, this study contributes to the broader field of AI-driven assessment methodologies by demonstrating how generative AI can be utilized to enhance formative learning experiences. The ability of ChatGPT to generate diverse question types and provide immediate feedback presents new opportunities for integrating AI into assessment design, reducing educator workload, and fostering self-regulated learning. Additionally, this study highlights the potential for scaling AI-assisted retrieval practice across various educational contexts, particularly in disciplines that rely on knowledge retention and conceptual application. Future research should explore how ChatGPT can be adapted to different subject areas, educational levels, and assessment formats, as well as its effectiveness in fostering higher-order cognitive skills such as critical thinking, problem-solving, and synthesis.

While the findings were promising, this study also identified several limitations. The limited exposure to ChatGPT-assisted question development, as it was integrated into only one session, may have restricted students' ability to fully utilize AI-generated retrieval practice. Additionally, the group-based nature of the implementation may have limited individual engagement with AI-generated questions, potentially affecting students' confidence in using ChatGPT independently. Furthermore, while ChatGPT provided immediate feedback, concerns about the depth and accuracy of AI-generated responses indicate that human oversight remains essential. These findings suggest that AI tools should be integrated as a complement rather than a substitute for educator-led instruction and formative assessment.

While the results were promising, certain limitations in the implementation were noted. Students had limited exposure to ChatGPT-assisted question development, as it was integrated into only one session. Additionally, the group-based format restricted individualized engagement, potentially limiting the depth of learning and personalized application of ChatGPT's capabilities. Feedback from students highlighted the accessibility and immediacy of ChatGPT's responses, yet its depth was sometimes inadequate for addressing more complex queries. These findings indicate that while ChatGPT

serves as a valuable complement to traditional teaching methods, it is not a replacement for the detailed feedback and contextual understanding provided by educators.

To enhance the effectiveness of ChatGPT-assisted retrieval practice, future implementations should focus on providing more structured guidance in AI-assisted question generation, promoting sustained individual engagement with AI tools, and refining hybrid feedback models that blend AI-generated and lecturer-driven responses. Longitudinal studies could further examine the long-term impact of AI-assisted retrieval practice on memory retention and higher-order cognitive skills, as well as the scalability of this approach in diverse learning environments.

This study highlights the importance of combining technological tools with human guidance to create an effective and well-rounded learning environment. While tools like ChatGPT provide quick access to information, generate questions, and offer immediate feedback, they do not fully replace the depth of understanding and personalized support that educators provide. Moving forward, efforts should focus on increasing students' engagement with these tools by incorporating structured activities that encourage independent practice, guided refinement of generated content, and support for higher-order thinking skills.

Additionally, improving feedback mechanisms by blending automated responses with teacher-led discussions can help address gaps in understanding. By thoughtfully integrating technology with traditional teaching methods, educational institutions can develop a more inclusive and adaptable learning framework that benefits students with different learning styles, cognitive abilities, and academic needs.

REFERENCES

Abdullah, H., Arsad, N., Hashim, F. H., Aziz, N. A., Amin, N., & Ali, S. H. (2012). Evaluation of students' achievement in the final exam questions for microelectronic (KKKL3054) using the Rasch model. *Procedia-Social and Behavioral Sciences*, *60*, 119-123. https://doi.org/10.1016/j.sbspro.2012.09.356

Abel, M., & Bäuml, K.-H. T. (2020). Would you like to learn more? Retrieval practice plus feedback can increase motivation to keep on studying. *Cognition*, *201*, 104316. https://doi.org/10.1016/j.cognition.2020.104316

Agarwal, P. K., Finley, J. R., Rose, N. S., & Roediger, H. L. (2016). Benefits from retrieval practice are greater for students with lower working memory capacity. *Memory*, *25*(6), 764-771. https://doi.org/10.1080/09658211.2016.1220579

Agarwal, P. K., Nunes, L. D., & Blunt, J. R. (2021). Retrieval practice consistently benefits student learning: A systematic review of applied research in schools and classrooms. *Educational Psychology Review*, *33*(4), 1409-1453. https://doi.org/10.1007/s10648-021-09595-9

Aglanovna, T. M., Muckanaevich, K. G., Toleuovna, D. A., Gizatovna, Z. S., & Shynbolatovna, T. A. (2024). Monitoring the validity of formative assessment: Tools and methods. *Evolutionary Studies in Imaginative Culture, 8.1*(S1), 955-968. https://doi.org/10.70082/esiculture.vi.929

Aguinis, H., Gottfredson, R. K., & Joo, H. (2013). Best-practice recommendations for defining, identifying, and handling outliers. *Organizational Research Methods*, *16*(2), 270-301. https://doi.org/10.1177/1094428112470848

Almasre, M. (2024). Development and evaluation of a custom GPT for the assessment of students' designs in a typography course. *Education Sciences*, *14*(2), 148. https://doi.org/10.3390/educsci14020148

Altman, D. G. (2005). Why we need confidence intervals. *World Journal of Surgery*, *29*(5), 554-556. https://doi.org/10.1007/s00268-005-7911-0

Anderson, L. W. (2005). Objectives, evaluation, and the improvement of education. *Studies in Educational Evaluation*, *31*(2-3), 102-113. https://doi.org/10.1016/j.stueduc.2005.05.004

Ariel, R., & Karpicke, J. D. (2018). Improving self-regulated learning with a retrieval practice intervention. *Journal of Experimental Psychology: Applied, 24*(1), 43–56. https://doi.org/10.1037/xap0000133

Bego, C. R., Lyle, K. B., Ralston, P. A., Immekus, J. C., Chastain, R. J., Haynes, L. D., Hoyt, L. K., Pigg, R. M., Rabin, S. D., Scobee, M. W., & Starr, T. L. (2024). Single-paper meta-analyses of the effects of spaced retrieval practice in nine introductory STEM courses: Is the glass half full or half empty? *International Journal of STEM Education*, *11*, Article 9. https://doi.org/10.1186/s40594-024-00468-5

Bell, B., & Cowie, B. (2001). The characteristics of formative assessment in science education. *Science Education*, *85*, 536-553. https://doi.org/10.1002/sce.1022

Bishop, D. (2022). *Supporting A-level sciences students to develop their revision strategies and exam technique* [Master's dissertation, University of Oxford]. https://ora.ox.ac.uk/objects/uuid:b6ac95d8-e2c7-46be-aa7b-2205307e8de8

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice*, *5*(1), 7-74. https://doi.org/10.1080/0969595980050102

Broeren, M., Heijltjes, A., Verkoeijen, P., Smeets, G., & Arends, L. (2021). Supporting the self-regulated use of retrieval practice: A higher education classroom experiment. *Contemporary Educational Psychology*, *64*, 101939. https://doi.org/10.1016/j.cedpsych.2020.101939

Carpenter, S. K. (2023). Encouraging students to use retrieval practice: A review of emerging research from five types of interventions. *Educational Psychology Review, 35*(4). https://doi.org/10.1007/s10648-023-09811-8

Carpenter, S. K., Lund, T. J., Coffman, C. R., Armstrong, P. I., Lamm, M. H., & Reason, R. D. (2016). A classroom study on the relationship between student achievement and retrieval-enhanced learning. *Educational Psychology Review*, *28*, 353-375. https://doi.org/10.1007/s10648-015-9311-9

Carpenter, S. K., Pan, S. C., & Butler, A. C. (2022). The science of effective learning with spacing and retrieval practice. *Nature Reviews Psychology, 1*, 496–511 https://doi.org/10.1038/s44159-022-00089-1

Casselman, C. (2024). An investigation into the impact of vocabulary retrieval practice as a method of formative assessment in a Latin AS-level unseen translation context. *Journal of Classics Teaching*, *25*(50), 123-128. https://doi.org/10.1017/S2058631024000692

Cheung, B. H. H., Lau, G. K. K., Wong, G. T. C., Lee, E. Y. P., Kulkarni, D., Seow, C. S., Wong, R., & Co, M. T. H. (2023). ChatGPT versus human in generating medical graduate exam questions – An international prospective study. *medRxiv*. https://doi.org/10.1101/2023.05.13.23289943

Christiansen, C., Calvert, C., & Morris, C. (2024). Factors affecting students' likelihood to access feedback. *Educational Researcher*, *53*(8), 478-480. https://doi.org/10.3102/0013189X241285416

Cummings, A. T. (2020). Correlation of student participation in practice exams and actual exam performance. *ASEE North Midwest Section Annual Conference*. 18. https://openprairie.sdstate.edu/asee_nmws_2020_pubs/18

Dai, W., Lin, J., Jin, H., Li, T., Tsai, Y. S., Gašević, D., & Chen, G. (2023, July). Can large language models provide feedback to students? A case study on ChatGPT. *Proceedings of the IEEE International Conference on Advanced Learning Technologies, Orem, UT, USA,* 323-325. https://doi.org/10.1109/ICALT58122.2023.00100

Delacre, M., Lakens, D., & Leys, C. (2017). Why psychologists should by default use Welch's t-test instead of student's t-test. *International Review of Social Psychology*, *30*(1), 92-101. https://doi.org/10.5334/irsp.82

Deng, F., Gluckstein, J., & Larsen, D. (2015). Student-directed retrieval practice is a predictor of medical licensing examination performance. *Perspectives on Medical Education, 4*(6, 308-313. https://doi.org/10.1007/s40037-015-0220-x

Desy, J., Busche, K., Cusano, R., Veale, P., Coderre, S., & McLaughlin, K. (2017). How teachers can help learners build storage and retrieval strength. *Medical Teacher*, *40*(4), 407-413. https://doi.org/10.1080/0142159X.2017.1408900

Dewi, N., & Mangunsong, F. (2012). Contribution of student's perception toward teacher's goal orientation and student's goal orientation as a mediator in test anxiety on elementary's final exams. *Procedia – Social and Behavioral Sciences*, *69*, 509-517. https://doi.org/10.1016/j.sbspro.2012.11.440

Divjak, B., Svetec, B., & Horvat, D. (2024). How can valid and reliable automatic formative assessment predict the acquisition of learning outcomes? *Journal of Computer Assisted Learning*, *40*(6), 2616-2632. https://doi.org/10.1111/jcal.12953

Donker, S. C., Vorstenbosch, M. A., Gerhardus, M. J., & Thijssen, D. H. (2022). Retrieval practice and spaced learning: Preventing loss of knowledge in Dutch medical sciences students in an ecologically valid setting. *BMC Medical Education*, *22*, Article 65. https://doi.org/10.1186/s12909-021-03075-y

Elbanna, S., & Armstrong, L. (2024). Exploring the integration of ChatGPT in education: Adapting for the future. *Management & Sustainability: An Arab Review*, *3*(1), 16-29. https://doi.org/10.1108/MSAR-03-2023-0016

Fernández, A. A., López-Torres, M., Fernández, J. J., & Vázquez-García, D. (2024). ChatGPT as an instructor's assistant for generating and scoring exams. *Journal of Chemical Education*, *101*(9), 3780-3788. https://doi.org/10.1021/acs.jchemed.4c00231

Flom, J., Green, K., & Wallace, S. (2023). To cheat or not to cheat? An investigation into the ethical behaviors of generation Z. *Active Learning in Higher Education*, *24*(2), 155-168. https://doi.org/10.1177/14697874211016147

Floridi, L., & Chiriatti, M. (2020). GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines*, *30*, 681-694. https://doi.org/10.1007/s11023-020-09548-1

French, S., Dickerson, A., & Mulder, R. A. (2024). A review of the benefits and drawbacks of high-stakes final examinations in higher education. *Higher Education, 88*, 893-918. https://doi.org/10.1007/s10734-023-01148-z

Glover, J. A. (1989). The "testing" phenomenon: Not gone but nearly forgotten. *Journal of Educational Psychology*, *81*(3), 392-399. https://doi.org/10.1037/0022-0663.81.3.392

Greving, S., & Richter, T. (2022). Practicing retrieval in university teaching: Short-answer questions are beneficial, whereas multiple-choice questions are not. *Journal of Cognitive Psychology*, *34*(5), 657-674. https://doi.org/10.1080/20445911.2022.2085281

Halamish, V., & Bjork, R. A. (2011). When does testing enhance retention? A distribution-based interpretation of retrieval as a memory modifier. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(4), 801-812. https://doi.org/10.1037/a0023219

Hoekstra, R., Morey, R. D., Rouder, J. N., & Wagenmakers, E.-J. (2014). Robust misinterpretation of confidence intervals. *Psychonomic Bulletin & Review*, *21*, 1157-1164. https://doi.org/10.3758/s13423-013-0572-3

Holmes, W., Bialik, M., & Fadel, C. (2019). *Artificial intelligence in education promises and implications for teaching and learning*. Center for Curriculum Redesign, Boston, MA, USA. https://discovery.ucl.ac.uk/id/eprint/10139722/

Hui, L., de Bruin, A. B. H., Donkers, J., & van Merriënboer, J. J. G. (2022). Why students do (or do not) choose retrieval practice: Their perceptions of mental effort during task performance matter. *Applied Cognitive Psychology*, *36*(2), 433-444. https://doi.org/10.1002/acp.3933

Indran, I. R., Paranthaman, P., Gupta, N., & Mustafa, N. (2023). Twelve tips to leverage AI for efficient and effective medical question generation: A guide for educators using Chat GPT. *Medical Teacher*, *46*(8), 1021-1026. https://doi.org/10.1080/0142159X.2023.2294703

Jacobsen, L. J., & Weber, K. E. (2023, September 29). *The promises and pitfalls of LLMs as feedback providers: A study of prompt engineering and the quality of AI-driven feedback*. https://doi.org/10.31219/osf.io/cr257

Jaeger, A., Buratto, L. G., Pompeia, S., & Ekuni, R. (2024). How can retrieval practice improve educational achievement in Brazil? *Journal of Applied Research in Memory and Cognition, 13*(1), 57-62. https://doi.org/10.1037/mac0000129

Jaeger, A., Eisenkraemer, R. E., & Stein, L. M. (2014). Test-enhanced learning in third-grade children. *Educational Psychology*, *35*(4), 513-521. https://doi.org/10.1080/01443410.2014.963030

Jamieson, J. P., Peters, B. J., Greenwood, E. J., & Altose, A. J. (2016). Reappraising stress arousal improves performance and reduces evaluation anxiety in classroom exam situations. *Social Psychological and Personality Science*, *7*(6), 579-587. https://doi.org/10.1177/1948550616644656

Javaid, M., Haleem, A., Singh, R. P., Khan, S., & Khan, I. H. (2023). Unlocking the opportunities through ChatGPT Tool towards ameliorating the education system. *BenchCouncil Transactions on Benchmarks, Standards and Evaluations*, *3*(2), 100115. https://doi.org/10.1016/j.tbench.2023.100115

Jensen, J. L., McDaniel, M. A., Woodard, S. M., & Kummer, T. A. (2014). Teaching to the test … or testing to teach: Exams requiring higher order thinking skills encourage greater conceptual understanding. *Educational Psychology Review*, *26*, 307-329. https://doi.org/10.1007/s10648-013-9248-9

Kang, H. (2021). Sample size determination and power analysis using the G*Power software. *Journal of Educational Evaluation for Health Professions*, *18*(17). https://doi.org/10.3352/jeehp.2021.18.17

Karpicke, J. D. (2017). Retrieval-based learning: A decade of progress. In J. H. Byrne (Ed.), *Learning and memory: A comprehensive reference* (2nd ed.). Academic Press. https://doi.org/10.1016/B978-0-12-809324-5.21055-9

Karpicke, J. D., & Roediger, H. L. (2007). Expanding retrieval practice promotes short-term retention, but equally spaced retrieval enhances long-term retention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(4), 704-719. https://doi.org/10.1037/0278-7393.33.4.704

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., … Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, *103*, 102274. https://doi.org/10.1016/j.lindif.2023.102274

Kikalishvili, S. (2023). Unlocking the potential of GPT-3 in education: Opportunities, limitations, and recommendations for effective integration. *Interactive Learning Environments*, *32*(9), 5587-5599. https://doi.org/10.1080/10494820.2023.2220401

Kim, T. K., & Park, J. H. (2019). More about the basic assumptions of t-test: Normality and sample size. *Korean Journal of Anesthesiology*, *72*(4), 331-335. https://doi.org/10.4097/kja.d.18.00292

Kiryakova, G., & Angelova, N. (2023). ChatGPT – A challenging tool for the university professors in their teaching practice. *Education Sciences, 13*(10), 1056. https://doi.org/10.3390/educsci13101056

Kıyak, Y. S., Coşkun, Ö., Budakoğlu, I. İ., & Uluoğlu, C. (2024). ChatGPT for generating multiple-choice questions: Evidence on the use of artificial intelligence in automatic item generation for a rational pharmacotherapy exam. *European Journal of Clinical Pharmacology*, *80*(5), 729-735. https://doi.org/10.1007/s00228-024-03649-x

Kıyak, Y. S., & Emekli, E. (2024). ChatGPT prompts for generating multiple-choice questions in medical education and evidence on their validity: A literature review. *Postgraduate Medical Journal*, *100*(1189), 858-865. https://doi.org/10.1093/postmj/qgae065

Krathwohl, D. R. (2002). A revision of Bloom's taxonomy: An overview. *Theory into Practice*, *41*(4), 212-218. https://doi.org/10.1207/s15430421tip4104_2

Kubik, V., Gaschler, R., & Hausman, H. (2021). PLAT 20(1) 2021: Enhancing student learning in research and educational practice: The power of retrieval practice and feedback. *Psychology Learning & Teaching*, *20*(1), 1-20. https://doi.org/10.1177/1475725720976462

Linacre, J. M. (2025). *A user's guide to Winsteps Ministep Rasch-model measurement computer program.* https://www.winsteps.com/winman/copyright.htm

Lo, C. K. (2023). What is the impact of ChatGPT on education? A rapid review of the literature. *Education Sciences, 13*(4), 410. https://doi.org/10.3390/educsci13040410

Luckin, R., & Holmes, W. (2016). *Intelligence unleashed: An argument for AI in Education.* UCL Knowledge Lab. https://discovery.ucl.ac.uk/id/eprint/1475756/

Lydersen, S. (2018). Balanced or imbalanced samples? *Tidsskrift for Den Norske Legeforening, 138*. https://doi.org/10.4045/tidsskr.18.0539

Ma, X., Li, T., Duzi, K., Li, Z. Y., Ma, X., Li, Y., & Zhou, A. B. (2020). Retrieval practice promotes pictorial learning in children aged six to seven years. *Psychological Reports*, *123*(6), 2085-2100. https://doi.org/10.1177/0033294119856553

Maeda, M. (2021). Exam cheating among Cambodian students: When, how, and why it happens. *Compare: A Journal of Comparative and International Education*, *51*(3), 337-355. https://doi.org/10.1080/03057925.2019.1613344

Malaysian Qualifications Agency. (2018). *Code of practice for programme accreditation* (2nd ed.). https://www2.mqa.gov.my/qad/garispanduan/COPPA/COPPA%202nd%20Edition%20%282017%29.pdf

Mariano, G. J., Allwardt, D. E., Raptis, P. R., & Stilwell, K. (2024). Reintroducing the oral exam: Finding out what your students really know in the age of ChatGPT. *Currents in Teaching & Learning*, *16*(1), 59-69. https://webcdn.worcester.edu/currents-in-teaching-and-learning/wp-content/uploads/sites/65/2024/09/Reintroducing-the-Oral-Exam.pdf

McDougall, S., & Gruneberg, M. (2002). What memory strategy is best for examinations in psychology? *Applied Cognitive Psychology*, *16*(4), 451-458. https://doi.org/10.1002/acp.800

Meier, J., & Löfqvist, A. (2024). *Retrieval practice through ChatGPT: Implications for learning, attitudes and transfer* [Bachelor thesis, Umeå University, Sweden]. https://urn.kb.se/resolve?urn=urn:nbn:se:umu:diva-226602

Metu, I. C., Agu, N. N., & Eleje, L. I. (2024). Students' perceptions of and preferences for assessment feedback in higher education: Implication for evaluators. *Futurity of Social Sciences*, *2*(2), 21-37. https://doi.org/10.57125/FS.2024.06.20.02

Ministry of Digital. (2010). *The Personal Data Protection Act 2010 (Act 709).* https://www.pdp.gov.my/ppdpv1/en/akta/pdp-act-2010/

Modran, H. A., Chamunorwa, T., Ursuțiu, D., & Samoilă, C. (2024). Integrating artificial intelligence and ChatGPT into higher engineering education. In M. E. Auer, U. R. Cukierman, E. Vendrell Vidal, & E. Tovar Caro (Eds.), *Towards a hybrid, flexible and socially engaged higher education* (pp. 499-510). Springer. https://doi.org/10.1007/978-3-031-51979-6_52

Morano, S. (2019). Retrieval practice for retention and transfer. *TEACHING Exceptional Children*, *51*(6), 436-444. https://doi.org/10.1177/0040059919847210

Moreira, B. F. T., Pinto, T. S. S., Starling, D. S. V., & Jaeger, A. (2019). Retrieval practice in classroom settings: A review of applied research. *Frontiers in Education*, *4*(5). https://doi.org/10.3389/feduc.2019.00005

Morris, C. D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning and Verbal Behavior*, *16*(5), 519-533. https://doi.org/10.1016/S0022-5371(77)80016-9

Naik, D., Naik, I., & Naik, N. (2024). Sorry, I am an AI language model: Understanding the limitations of ChatGPT. In N. Naik, P. Jenkins, S. Prajapat, & P. Grace (Eds.), *Contributions Presented at the International Conference on Computing, Communication, Cybersecurity & AI* (pp. 26-42). Springer. https://doi.org/10.36227/techrxiv.173121393.30482183/v1

National Science Council. (2020). *The Malaysian code of responsible conduct in research* (2nd ed.). Academy of Sciences Malaysia. https://www.akademisains.gov.my/asm-publication/the-malaysian-code-of-responsible-conduct-in-research-2nd-edition/

Nazaretsky, T., Mejia-Domenzain, P., Swamy, V., Frej, J., & Käser, T. (2024). AI or human? Evaluating student feedback perceptions in higher education. In R. Ferreira Mello, N. Rummel, I. Jivet, G. Pishtari, & J. A. Ruipérez Valiente (Eds.), *Technology enhanced learning for inclusive and equitable quality education* (pp. 284-298). Springer. https://doi.org/10.1007/978-3-031-72315-5_20

OpenAI. (2023). *GPT-4*. https://openai.com/index/gpt-4-research/

Otaki, B., & Lindwall, O. (2024). Generative AI and the human touch: Investigating the changing landscape of feedback in higher education. In R. Lindgren, T. I. Asino, E. A. Kyza, C. K. Looi, D. T. Keifert, & E. Suárez (Eds.), *Proceedings of the 18th International Conference of the Learning Sciences* (pp. 1099-1102). International Society of the Learning Sciences. https://doi.org/10.22318/icls2024.230928

Özbay, Y. (2024). Evaluation of ChatGPT as a multiple-choice question generator in dental traumatology. *Medical Records*, *6*(2), 235-238. https://doi.org/10.37990/medr.1446396

Pavelea, A. M., & Moldovan, O. (2020). Why some fail and others succeed: Explaining the academic performance of PA undergraduate students. *The NISPAcee Journal of Public Administration and Policy*, *13*(1), 109-132. https://doi.org/10.2478/nispa-2020-0005

Racsmány, M., Szőllősi, Á., & Marián, M. (2020). Reversing the testing effect by feedback is a matter of performance criterion at practice. *Memory & Cognition*, *48*, 1161-1170. https://doi.org/10.3758/s13421-020-01041-5

Rasul, T., Nair, S., Kalendra, D., Robin, M., de Oliveira Santini, F., Ladeira, W. J., Sun, M., Day, I., Rather, R. A., & Heathcote, L. (2023). The role of ChatGPT in higher education: Benefits, challenges, and future research directions. *Journal of Applied Learning and Teaching*, *6*(1), 41-56. https://doi.org/10.37074/jalt.2023.6.1.29

Rasyid, A. R., Al Yakin, A., Muthmainnah, M., Zulfiqar Bin Tahir, S., & Obaid, A. J. (2024). Revolutionize the potential of ChatGPT as teaching material to engage students in learning. *Lentera Pendidikan: Jurnal Ilmu Tarbiyah Dan Keguruan*, 27(1), 1-14. https://doi.org/10.24252/lp.2024v27n1i1

Richardson, J. T. (2015). Coursework versus examinations in end-of-module assessment: A literature review. *Assessment & Evaluation in Higher Education*, *40*(3), 439-455. https://doi.org/10.1080/02602938.2014.919628

Ritchie, S. J., Della Sala, S., & McIntosh, R. D. (2013). Retrieval practice, with or without mind mapping, boosts fact learning in primary school children. *PLoS One*, *8*(11), https://doi.org/10.1371/journal.pone.0078976

Rivera-Rosas, C. N., Calleja-López, J. T., Ruibal-Tavares, E., Villanueva-Neri, A., Flores-Felix, C. M., & Trujillo-López, S. (2024). Exploring the potential of ChatGPT to create multiple-choice question exams. *Educación Médica*, *25*(4), 100930. https://doi.org/10.1016/j.edumed.2024.100930

Roediger, H. L., & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences*, *15*(1), 20-27. https://doi.org/10.1016/j.tics.2010.09.003

Roediger, H. L., & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, *17*(3), 249-255. https://doi.org/10.1111/j.1467-9280.2006.01693.x

Rudolph, M. J., Daugherty, K. K., Ray, M. E., Shuford, V. P., Lebovitz, L., & DiVall, M. V. (2019). Best practices related to examination item construction and post-hoc review. *American Journal of Pharmaceutical Education*, *83*(7), 7204. https://doi.org/10.5688/ajpe7204

Ruiz-Primo, M. A. (2011). Informal formative assessment: The role of instructional dialogues in assessing students' learning. *Studies in Educational Evaluation*, *37*(1), 15-24. https://doi.org/10.1016/j.stueduc.2011.04.003

Rusticus, S. A., & Lovato, C. Y. (2014). Impact of sample size and variability on the power and Type I error rates of equivalence tests: A simulation study. *Practical Assessment, Research, and Evaluation*, *19*(1), 11. https://doi.org/10.7275/4s9m-4e81

Sadler, D. R. (1998). Formative assessment: Revisiting the territory. *Assessment in Education: Principles, Policy & Practice*, *5*(1), 77-84. https://doi.org/10.1080/0969595980050104

Sain, Z. H., Vasudevan, A., Thelma, C. C., & Asfahani, A. (2024). Harnessing ChatGPT for effective assessment and feedback in education. *Journal of Computer Science and Informatics Engineering*, *3*(2), 74-82. https://doi.org/10.55537/cosie.v3i2.856

Sana, F., & Yan, V. X. (2022). Interleaving retrieval practice promotes science learning. *Psychological Science*, *33*(5), 782-788. https://doi.org/10.1177/09567976211057507

Shadish, W. R., Galindo, R., Wong, V. C., Steiner, P. M., & Cook, T. D. (2011). A randomized experiment comparing random and cutoff-based assignment. *Psychological Methods*, *16*(2), 179-191. https://doi.org/10.1037/a0023345

Silseth, K., & Furberg, A. (2024). Bridging group work and whole-class activities through responsive teaching in science education. *European Journal of Psychology of Education*, *39*(3), 2155-2176. https://doi.org/10.1007/s10212-023-00770-w

Singh, K. (2019). Lecturer's feedback and its impact on student learning: A study of a public university in Sarawak, Malaysia. *Asian Journal of University Education*, *15*(3), 83-91. https://doi.org/10.24191/ajue.v15i3.7562

Skaik, Y. (2015). The bread and butter of statistical analysis "t-test": Uses and misuses. *Pakistan Journal of Medical Sciences*, *31*(6), 1558–1559. https://doi.org/10.12669/pjms.316.8984

Smolinsky, L., Marx, B. D., Olafsson, G., & Ma, Y. A. (2020). Computer-based and paper-and-pencil tests: A study in calculus for STEM majors. *Journal of Educational Computing Research, 58*(7)*,* 1256-1278. https://doi.org/10.1177/0735633120930235

Steiss, J., Tate, T., Graham, S., Cruz, J., Hebert, M., Wang, J., Moon, Y., Tseng, W., Warschauer, M., & Olson, C. B. (2024). Comparing the quality of human and ChatGPT feedback of students' writing. *Learning and Instruction*, *91*, 101894. https://doi.org/10.1016/j.learninstruc.2024.101894

Stobart, G. (2012). Validity in formative assessment. In J. Gardner (Ed.), *Assessment and learning* (pp. 233–243). Sage.

Strasser, A. (2024). Pitfalls (and advantages) of sophisticated large language models. In S. Caballé, J. Casas-Roma, & J. Conesa (Eds.) *Ethics in online AI-based systems* (pp. 195-210). Academic Press. https://doi.org/10.1016/B978-0-443-18851-0.00007-X

Sullivan, G. M., & Artino, A. R., Jr. (2013). Analyzing and interpreting data from Likert-type scales. *Journal of Graduate Medical Education*, *5*(4), 541-542. https://doi.org/10.4300/JGME-5-4-18

Tan, A. Y. T. (2020). *Technology enhanced assessment and feedback among educators across disciplines* [Doctoral dissertation, Monash University]. https://doi.org/10.26180/5e9037fda21e8

Theobald, E. J., Eddy, S. L., Grunspan, D. Z., Wiggins, B. L., & Crowe, A. J. (2017). Student perception of group dynamics predicts individual performance: Comfort and equity matter. *PLoS ONE*, *12*(7), e0181336. https://doi.org/10.1371/journal.pone.0181336

Trumbull, E., & Lash, A. (2013). *Understanding formative assessment: Insights from learning theory and measurement theory*. WestEd. https://www2.wested.org/www-static/online_pubs/resource1307.pdf

Tyson, J. (2023). Shortcomings of ChatGPT. *Journal of Chemical Education*, *100*(8), 3098-3101. https://doi.org/10.1021/acs.jchemed.3c00361

Wood, D., Bruner, J. S., & Ross, G. (1976). The role of tutoring in problem solving. *Journal of Child Psychology and Psychiatry*, *17*(2), 89-100. https://doi.org/10.1111/j.1469-7610.1976.tb00381.x

YeckehZaare, I., Resnick, P., & Ericson, B. (2019). A spaced, interleaved retrieval practice tool that is motivating and effective. *Proceedings of the ACM Conference on International Computing Education Research* (pp. 71-79). Association for Computing Machinery. https://doi.org/10.1145/3291279.3339411

Yu, H. (2023). Reflection on whether Chat GPT should be banned by academia from the perspective of education and teaching. *Frontiers in. Psychollogy, 14*, 1181712. https://doi.org/10.3389/fpsyg.2023.1181712

Yusof, I. J. (2024). *Research literacy: A guiding path for postgraduate students*. Penerbit UTM Press. https://penerbit.utm.my/booksonline/research-literacy-a-guiding-path-postgraduate-students/

Yusof, I. J., & Ismail, L. H. (2023). Innovating assessment: An exploration of ChatGPT's role in assisting students with multiple choice question items development. In L. Indiran, M. K. Ishak, & Y. Subramaniam (Eds.), *e-Proceeding of New Academia Learning Innovation Exhibition & Competition* (pp. 225-229). Universiti Teknologi Malaysia. http://tiny.cc/u0zyzz

Zimmerman, B. J., & Schunk, D. H. (2011). Self-regulated learning and performance: An introduction and an overview. In D. H. Schunk & B. J. Zimmerman (Eds.), *Handbook of self-regulation of learning and performance* (1st ed., pp. 15-26). Routledge.

Zuckerman, M., Flood, R., Tan, R. J., Kelp, N., Ecker, D. J., Menke, J., & Lockspeiser, T. (2023). ChatGPT for assessment writing. *Medical Teacher*, *45*(11), 1224-1227. https://doi.org/10.1080/0142159X.2023.2249239

APPENDIX

STUDENT FEEDBACK QUESTIONNAIRE ON CHATGPT-ASSISTED RETRIEVAL PRACTICE

Statement	Agreement scale 1. Strongly disagree 2. Disagree 3. Agree 4. Strongly agree			
Section A: Perceptions of Retrieval Practice	**1**	**2**	**3**	**4**
Retrieval practice helped me retain information better for the final exam.				
Retrieval practice improved my understanding of key course concepts.				
Retrieval practice sessions were a valuable addition to my overall learning process.				
I feel more confident about answering exam questions after engaging in retrieval practice.				
Section B: Preference for Student-Developed vs. Lecturer-Developed Questions	**1**	**2**	**3**	**4**
I preferred lecturer-generated questions because they provided a clearer structure for retrieval practice.				
Student-developed questions were more relatable and aligned with my learning needs.				
Lecturer-generated questions were well-structured and closely aligned with the format of the final exam.				
I found it easier to learn from retrieval practice sessions that used lecturer-generated questions.				
Section C: Students' Experience with ChatGPT in Developing Questions	**1**	**2**	**3**	**4**
ChatGPT helped me create better questions for retrieval practice.				
ChatGPT was useful in helping me structure questions that aligned with the learning objectives.				
I felt more confident in the quality of the questions I developed with ChatGPT's assistance.				
ChatGPT made it easier to generate diverse types of questions for retrieval practice.				
Section D: Experience with ChatGPT Feedback	**1**	**2**	**3**	**4**

Statement	Agreement scale 1. Strongly disagree 2. Disagree 3. Agree 4. Strongly agree			
ChatGPT provided sufficient feedback for me to understand my mistakes.				
I feel less confident relying solely on ChatGPT feedback.				
ChatGPT's feedback was accessible whenever I needed support for reviewing my answers.				
ChatGPT's feedback sometimes lacked the depth needed for complex questions.				
Section E: Feedback's Preference	**1**	**2**	**3**	**4**
I found lecturer feedback to be more comprehensive than ChatGPT's feedback.				
ChatGPT's feedback was quicker and more convenient than receiving feedback from the lecturer.				
I trust lecturer feedback more than ChatGPT feedback for ensuring accurate understanding.				
I would like a combination of feedback from both ChatGPT and the lecturer in future learning sessions.				

AUTHOR

Ibnatul Jalilah Yusof is a senior lecturer and the Program Coordinator for Measurement and Evaluation in Education in the Department of Educational Studies and Behavioral Sciences, Faculty of Educational Sciences and Technology, Universiti Teknologi Malaysia (UTM). She holds both a Ph.D. and a Master of Education in Measurement and Evaluation in Education. Her research interests and areas of expertise focus on educational assessment, research literacy, risk assessment, and quantitative research methodology.