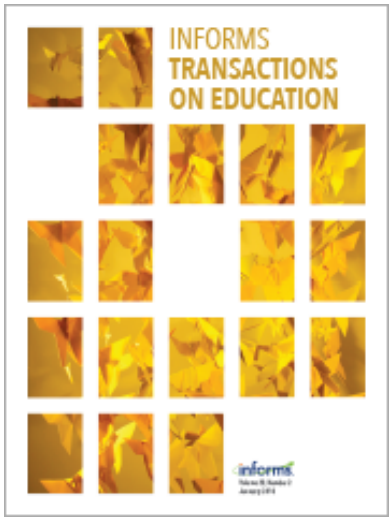


INFORMS Transactions on Education

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Students' Use of Generative AI Tools in Multiple Choice Exams

Nicoleta Serban, Tuba Ketenci, Joel Sokol

To cite this article:
Nicoleta Serban, Tuba Ketenci, Joel Sokol (2026) Students' Use of Generative AI Tools in Multiple Choice Exams. INFORMS Transactions on Education

Published online in Articles in Advance 11 May 2026

. https://doi.org/10.1287/ited.2025.0166

https://pubsonline.informs.org/journal/ited

INFORMS TRANSACTIONS ON EDUCATION
Articles in Advance, pp. 1–10
ISSN 1532-0545 (online)

Students' Use of Generative AI Tools in Multiple Choice Exams

Nicoleta Serban,[a,*] Tuba Ketenci,[a] Joel Sokol[a,b]

[a] H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, Georgia 30332; [b] College of Lifetime Learning, Georgia Institute of Technology, Atlanta, Georgia 30332

*Corresponding author

Contact: nserban@isye.gatech.edu, https://orcid.org/0000-0002-5813-7435 (NS); tuba.ketenci@isye.gatech.edu (TK); jsokol@isye.gatech.edu (JS)

Received: August 14, 2025
Revised: December 1, 2025
Accepted: March 7, 2026
Published Online in Articles in Advance: May 11, 2026

https://doi.org/10.1287/ited.2025.0166

Abstract. The integration of generative artificial intelligence (GenAI) tools into higher education presents both opportunities and challenges, particularly in assessment contexts. This study investigates the impact of GenAI usage on student performance in a graduate-level time series analysis course, focusing on multiple choice exams. Through a controlled experiment involving two student cohorts, each having access to GenAI tools during one of two midterms, we analyze performance outcomes and engagement behaviors using multimodal screen recordings. Statistical analyses reveal that GenAI access correlates with improved scores but only when students are adequately prepared to use the tools effectively. Frequent or prolonged GenAI usage alone did not predict better outcomes, highlighting the importance of AI literacy. Additionally, traditional course materials remained strong predictors of performance across both cohorts. These findings suggest that GenAI can enhance learning when integrated thoughtfully and accompanied by instructional support. The study contributes to the evolving discourse on AI in education by offering empirical insights into its role in assessments and proposing implications for instructional design and policy.

Funding: This research has been supported by the Master of Science in Analytics degree at Georgia Tech but we don't have a specific funding other than this internal support.

Supplemental Material: The online appendix is available at https://doi.org/10.1287/ited.2025.0166.

Keywords: generative AI tools • graduate data analysis course • multiple choice exam

1. Introduction

The rapid technological progress and widespread adoption of generative artificial intelligence (GenAI) tools have introduced both promising opportunities and notable challenges in higher education. GenAI tools offer personalized learning experiences, enhance productivity, and support academic skill development, from writing and translation to creative ideation and problem-solving (Chan and Hu 2023, Aldulaijan and Almalki 2025). However, their integration into educational settings also raises concerns about ethical use, overreliance, and the potential erosion of foundational skills (Gruenhagen et al. 2024, Kim et al. 2025). As education institutions address challenges in how to best incorporate these technologies, it becomes increasingly important to understand not only their benefits, but also their limitations.

Multiple streams of research have explored GenAI's role in teaching, learning, and assessment design (Goel and Joyner 2017, Bucchiarone et al. 2024, Mozelius 2024, Khlaif et al. 2025, Sohail et al. 2025), and in the support of student learning through personalized feedback and guidance (Wu and Yu 2024). Many studies to date have focused on surveys and qualitative methods to examine student and teacher perceptions and usage behaviors (Zhao et al. 2024). These studies have identified different usage styles, such as treating GenAI as an intellectual partner versus using it as an information browser (Aldulaijan and Almalki 2025), and have highlighted both enthusiasm and apprehension among students and faculty. Other approaches have focused at a higher level: on the collaborations required between teachers and GenAI specialists (and sometimes policymakers as well: Otundo 2025) and on ethics and policy in implementation (Kilinç 2024, Ogunleye et al. 2024, Nguyen et al. 2025). However, few studies have directly observed students during assessments, especially in rigorous, technical graduate-level courses. This gap limits our understanding of how GenAI tools influence actual performance in academic settings.

Despite growing adoption, the integration of GenAI into formal assessments remains contentious. Avidov

et al. (2025) observe that GenAI tools can push toward standardization rather than flexibility and openness when used in situational judgment tests. Some scholars have even proposed that AI could discourage the use of multiple choice assessments by enabling systems that require students to explain their reasoning (Klymkowsky and Cooper 2024). Written assignments also are not AI-proof; Combrinck and Loubser (2025) suggest using self-reflection in conjunction with an AI detector when giving written assessments. Studies have shown that students often do not perceive GenAI-assisted work as a breach of academic integrity (Gruenhagen et al. 2024), that faculty and students alike express mixed attitudes toward its impact on learning competencies (Kim et al. 2025), and that it can negatively impact student/teacher trust (Luo 2025). Moreover, the lack of consistent institutional policies and guidance has led to confusion and uneven implementation across courses (Johnston et al. 2024, Stritto et al 2024). On the other hand, Farrokhnia et al. (2025) suggest that GenAI can enhance the quality of assessments, and Perkins et al. (2023) suggest a framework for ethical inclusion of GenAI in assessments.

The use of GenAI in learning and assessment has been studied in general (Farrokhnia et al. 2025), in non-STEM areas (Law 2024), and at the K-12 level (Tang et al. 2024). However, except for a preliminary study involving undergraduates (Pesovski et al. 2024), there has been limited research targeted at whether AI can directly assist students during assessments. Building on this prior work, additional research should explore how students themselves interact with GenAI tools during assessments in graduate-level courses. Research by Nikolic et al. (2024) is an exception, testing multiple GenAIs on assessments from 10 engineering subjects, but that study focused on the GenAIs' capability and the ethical implications rather than on the effect of GenAI on student usage; Luo (2024) dealt with the issue of whether GenAI-assisted student work is their original work.

This paper addresses that gap by presenting findings from a controlled experiment in a graduate-level data analysis course, where students were allowed to use GenAI tools during one of two midterm multiple choice exams. By analyzing both performance outcomes and engagement behaviors, captured through multimodal screen recordings, we assess whether access to GenAI tools correlates with improved scores and identify the conditions under which it may be most beneficial.

Our study builds on recent work by Xiao et al. (2025), which demonstrated that access to advanced GenAI tools led to significant performance gains in graduate bioprocess engineering exams. Their findings showed that students with GenAI support improved scores by an average of 36%, with even greater benefits observed among students with disabilities and those lacking domain-specific backgrounds. Importantly, the study emphasized that GenAI usage requires practice and literacy: Students who learned to formulate effective prompts and critically evaluate AI-generated responses performed better over time.

Informed by these insights, our study not only evaluates the impact of GenAI on exam performance but also explores how students engage with GenAI tools during assessments. Similar to Xiao et al. (2025), we find that GenAI's effectiveness depends not only on its capabilities but also on students' preparedness and strategic use. Our findings contribute to the evolving discourse on GenAI in education by offering empirical evidence from a real-world classroom setting and by proposing implications for instructional design, assessment practices, and AI literacy development.

2. Experimental Design and Data Collection

In this section, we describe the setting, experimental design, and data collected from an asynchronous graduate course at the Georgia Institute of Technology. The objective of the experiment was to observe how students use GenAI tools for assistance during an exam and how such use relates to exam performance.

2.1. Course and Exam Description

The course selected for this study is a graduate course titled Time Series Analysis. This graduate-level course is part of the Masters of Analytics (online and on-campus) and attracts students from diverse quantitative backgrounds, including engineering, computer science, and applied statistics. The learning outcomes emphasize both conceptual understanding and practical implementation of time series analysis. By the end of the course, students are expected to (i) identify and formulate time series modeling problems, (ii) select appropriate time series models, (iii) implement these models using R and Python, and (iv) interpret results in the context of forecasting and diagnostics. Instruction combines theoretical foundations with hands-on data analysis, preparing students to rigorously evaluate model assumptions and performance. Although the course is statistically intensive, it is designed for students with prior exposure to regression and programming rather than domain-specific cohorts such as nursing or finance. This diversity in student backgrounds is reflected in the inclusion of degree program as a covariate in the regression models later in the paper.

The class included 112 students taught by a single instructor, which we acknowledge may introduce confounding effects and limit independence among observations. Although these factors constrain generalizability, they reflect a realistic classroom setting

and provide valuable insights into GenAI integration in technical graduate courses.

The first author of this paper was the instructor for the course in which the study was conducted. She has taught this course for many years, consistently refining its structure and content to align with evolving pedagogical practices and technological advancements. This experience provided a strong foundation for implementing and evaluating the controlled crossover experimental design described in the study.

In this experiment, students completed two proctored multiple choice midterm exams, with some questions allowing more than one correct answer and several items based on the interpretation of sample data analysis outputs. Exams were administered under a set of guidelines described in the syllabus and detailed in Online Appendix A.1. (The full exam questions are provided in Online Appendix A.2 and Online Appendix A.3.) The exam guidelines emphasized AI assistance rather than AI reliance. Students in GenAI-permitted conditions could consult GenAI tools but were expected to critically review any AI-generated output and remained fully responsible for errors arising from its use. To discourage submitting AI-generated responses without meaningful student input, copy-and-paste functionality was disabled for all students. For students in the cohort not allowed to use GenAI 2.2, any use of GenAI or other external tools beyond the official course materials was considered an honor code violation.

2.2. Experimental Design

Participants in this study were students registered for an asynchronous course (Time Series Analysis course) in Spring 2025. The students came from different degree programs, including the Master of Science in Analytics degree online (OMSA) and in-person (MSA Atlanta), the Online Master of Science in Computer Science (OMSCS), and other Masters programs. The largest cohort was from the OMSA program. This study has been approved by the Institutional Review Board (No. IRB2025-763).

Before the first midterm, all students completed a pre-experiment survey on their previous experience with GenAI tools and their perceptions of the use of GenAI. The survey questions are provided in Online Appendix B.1 along with some summary statistics.

We employed a crossover design with two cohorts. Each student took both multiple choice midterm exams under standard testing conditions (Online Appendix A.1), but access to GenAI tools varied by cohort and exam:

Cohort 1: allowed to use GenAI tools during Midterm 1, but not during Midterm 2.

Cohort 2: allowed to use GenAI tools during Midterm 2, but not during Midterm 1.

Thus, every student had the opportunity to use GenAI tools on exactly one of the two midterms, and each exam included both a GenAI-permitted group and a GenAI-restricted control group. In all conditions, students could access non-GenAI resources such as course materials and personal notes, ensuring that GenAI was not the only source of support.

The assignment of students to cohorts was conducted using weighted random assignment, with weights based on students' degree programs, to preserve the distribution of the four degree programs across both cohorts.

2.3. Engagement Factors

In this study, we captured a variety of student test-taking indicators and behaviors to help us understand whether and how they used GenAI and whether it was helpful. The data we collected for each student (Table 1) include the tools used (if any), what the student was primarily looking at on the screen, the timing of the student's progress through the exam, the number of times the student used each type of resource (GenAI or non-GenAI resources), and the student's score on each question. This approach allows us to consider the decision-making process as shown on screen, offering insight into how students engage with the available resources rather than looking only at the outcomes. In the education literature, this would be classified as a *social semiotic multimodal analysis* using the visual, spatial and temporal modes (Ho 2021).

2.4. Data Collection

The exams were proctored using the Honorlock proctoring tool (O'Brien 2020) embedded within the Canvas learning management system (Instructure 2025). Students' activity during the entire exam was thus video recorded, allowing us to learn about how students used resources during the exam. Both multiple choice examinations were fully recorded for each student using the Honorlock proctoring system (Honorlock 2025). This system captured detailed screen activity, including cursor movements, application usage, browser interactions, as well as webcam video and audio. These recordings provided rich behavioral data on how students interacted with exam questions and accessed both course materials and external resources, including GenAI tools.

The specific data in Table 1 were collected by human encoding of screen recordings captured by Honorlock. Three trained coders independently reviewed all screen recordings and applied the coding scheme using a standardized rubric. Any disagreements were discussed and resolved through collaborative review sessions, which helped ensure consistency and coding reliability. The videos were visualized at a 2× speed for faster completion when possible.

We also discretized the time-based factors. This discretization was motivated by both modeling and

Table 1. Data Collected Per Student Per Question and Per Midterm

Variable	Explanation	Codes
Was GenAI Used?	Whether the student engaged with a GenAI tool during the test	Yes, No
GenAI Tools Used	Specific GenAI tool(s) used by the student	BERT, ChatGPT, Claude, Co-Pilot, DeepSeek, Gemini, Grammarly, None
Screen Composition	Primary window on a student's screen during each question	Canvas, GenAI Tool, Local Files/Notes, Stack Overflow, Test Interface, Other
Time Spent on Question	How long the student spent answering	<10 sec, 10 sec–1 min, 1–3 min, >3 min
Time-Based Engagement	Timing of answer submission	Fast completion, long pauses before answering, frequent rewrites, multiple attempts, even response-time distribution
External Resource Usage	How many times a student consulted GenAI or other external tools	0 times, 1–3 times, >3 times
Course Resource Usage	How many times a student referred to official course materials or personal notes	0 times, 1–3 times, >3 times
Manner of GenAI usage	Manner in which GenAI was used to respond to a question	Asked problem verbatim, Reworded problem, General concept explanation, Asked clarifying questions, Not applicable
Score	Points awarded for each question	0–3
Correctness	Whether question was answered correctly	Correct, Not Correct

interpretability considerations. First, this approach captures nonlinear effects without strong parametric assumptions. Discretization provided a parsimonious way to approximate these nonlinearities without using a nonlinear form or without overfitting with high-order polynomials. Second, discretization also addresses robustness to heavy tails. For example, time-based measures are often right-skewed and thus more robust to extreme values. Last, it can improve interpretability and reporting. Prespecified, semantically meaningful ranges (e.g., "< 1 minute," "1–3 minutes," "> 3 minutes") yield immediately interpretable contrasts. These categories might align with decision thresholds instructors actually use (e.g., short versus long engagement).

3. Research Hypothesis and Methods

In this section, we present the statistical learning approaches for assessing a series of hypotheses on the students' performance in the two exams, contrasting the students with and without access to GenAI tools.

3.1. Research Hypothesis

We distinguish between two complementary sets of analyses: one focusing on overall exam performance and the other on the factors associated with the correctness of individual exam question responses.

The first analysis examines students' overall performance on midterm exams, with particular attention to whether they had access to and utilized GenAI tools. The primary hypothesis is that *access to GenAI tools during the exam does not lead to a statistically significant improvement in students' overall performance.* More specifically, we hypothesize the following.

1. *GenAI use may help students avoid very low scores;* and

2. *GenAI use may hinder students from achieving the highest scores.*

The second analysis leverages engagement data derived from Honorlock video recordings of the exam sessions. This analysis investigates whether patterns of resource use, particularly the frequency and intensity of GenAI or other tool usage, are associated with the correctness of students' responses to individual exam questions. The primary hypothesis here is that *the correctness of exam responses is not significantly associated with the extensive or frequent use of GenAI or other external resources.*

3.2. Statistical Learning Methods

The students in the course were divided into two cohorts: Cohort 1 had access to GenAI on the first midterm, but not the second, and Cohort 2 had access to GenAI on the second midterm, but not the first. The cohorts were initially of the same size, but because not every student completed the course, the Cohort 1 data set has 53 students and the Cohort 2 data set has 59 students. The two midterms had 26 and 25 questions, respectively.

3.2.1. Cohort-Level Performance Comparison. The first, simplest analysis is a comparison of the grades of the two cohorts separately on each midterm. For robustness, we compare the median of the grades between the two cohorts using a Mann-Whitney *U* test, with the null hypothesis being that the median grade of Cohort 1 is equal to the median grade of Cohort 2 on each midterm. We also used bootstrap sampling to obtain empirical distributions of the first and third quartiles and again apply a Mann-Whitney U test to compare these quantiles of the two cohorts.

Because each student also took one midterm exam with GenAI tools and one without, we also use a paired Wilcoxon test to compare each student's performance with and without GenAI. (The midterm grades are scaled before doing this test, because the two midterms' score distributions might be different.) For each cohort, the null hypothesis is that the median score on Midterm 1 is equal to the median score on Midterm 2. We keep the two cohorts separate for this analysis in case there is a learning effect (e.g., after learning about Cohort 1 students' experience using GenAI on Midterm 1 and practicing with such tools, Cohort 2 students might have prepared differently for their GenAI usage on Midterm 2).

3.2.2. Student-Level Performance Analysis. To analyze student performance, we use a regression approach. For each student and each midterm, the response is the student's midterm grade, and the predictors are the aggregated engagement factors (summed over all questions). The modeling approach includes traditional multiple regression models with regularized model selection, as well as gradient boosting machine (GBM) models. The regression models are applied to evaluate variable importance in the associates of the engagement factors to the grades, while the GBM models are applied to demonstrate prediction accuracy. In this analysis, we control for two factors: each student's registration section (i.e., each student's degree program) and each student's prior experience with GenAI as captured by intake questions "Have you ever used AI tools at work, in your studies, or in your personal life to automate tasks or improve productivity?" and "What is the primary purpose of using generative AI tools in creative tasks?"

3.2.3. Question-Level Analysis. We perform a more granular analysis using the detailed question-level data. For each student and each midterm and each question, we use the binary indicator (whether the student answered the question correctly or not) as the response, and we use the same engagement factors as in the student-level analysis but (where applicable) on a question-by-question basis. Because the engagement factors capture similar aspects of GenAI usage, there will be multicollinearity; to address that multicollinearity, we employ regularized approaches for variable selection. Similar to the student-level analysis, we also use GBM models to evaluate the consistency in results across different modeling approaches. We control for two factors in this analysis: each student's overall performance in the course (represented by the student's overall course grade) and the difficulty of each question (measured by the aggregate grade on each question among the cohort without access to GenAI tools).

4. Summary of the Results

In this section, we present the results from the detailed statistical modeling described in Section 3.2.

4.1. Student Cohorts

The distribution of students in the two cohorts is not equal; we have complete data for 53 students in Cohort 1 and 59 students in Cohort 2. The distribution of students over the four different programs and the prior exposure to AI use are similar between the two cohorts, as shown in Figures B.1 and B.2 in Online Appendix 6.

Our first analysis compares the distribution of the scores between the GenAI and non-GenAI cohorts in both midterm exams. Figure 1 presents a comparison of the participants' midterm grades for the GenAI and non-GenAI cohorts on the two exams. The two plots on the left show the distributions for Midterm 1, where both groups appear to follow a roughly normal distribution centered around the 30–35 range, with the non-GenAI group showing slightly higher overall counts at the upper end. In contrast, the two right plots for Midterm 2 show overall improvement. Notably, the GenAI group shows a more uniform increase, whereas the non-GenAI group has a sharper concentration near the maximum score, suggesting a stronger performance skew.

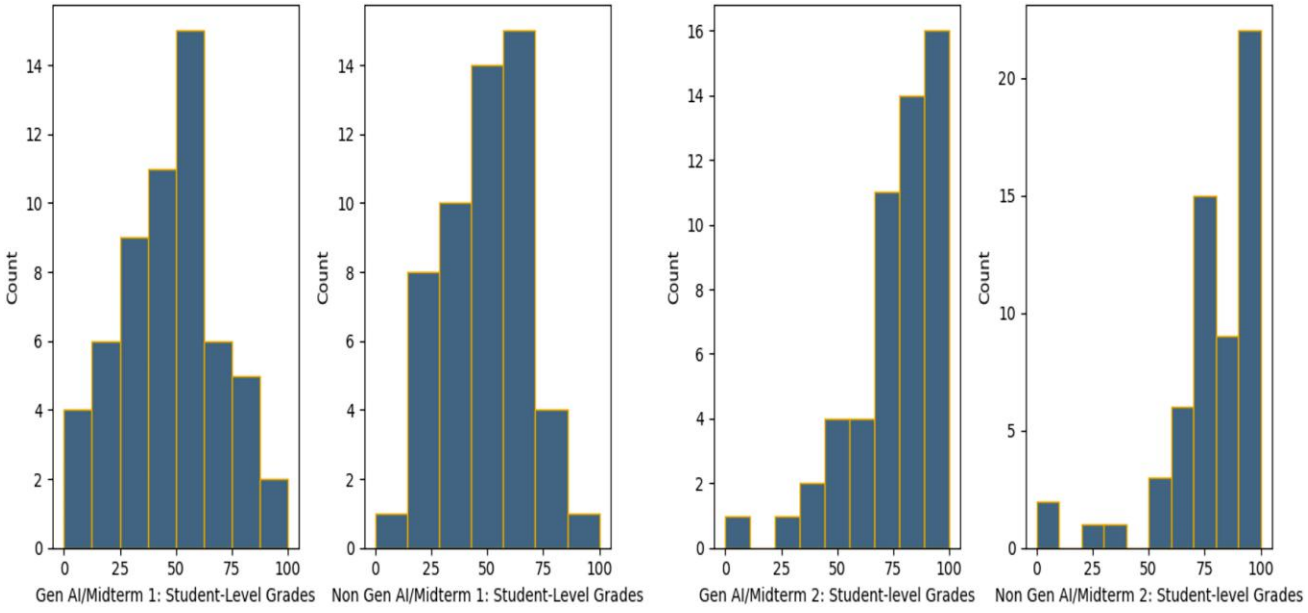
We applied the paired Wilcoxon test to evaluate whether students have performed significantly better on the exams in which they were allowed to use GenAI tools. The test is applied separately to each of the two cohorts comparing the grades between Midterm 1 and Midterm 2 and we applied the test with and without scaling the scores within each midterm to account for differences within each exam. The p-values are all very small (smaller than 0.005), indicating significant differences. However, the direction of the difference is the same for both cohorts, indicating better grades for students in Midterm 2. Thus, the use of GenAI tools does not significantly impact the score of each individual student compared between the two midterms.

4.2. Student-Level Performance Analysis

Applying the Mann-Whitney U test to compare the medians of the score between the GenAI and non-GenAI cohorts, we find that the p-value for Midterm 1 is 0.25 and for Midterm 2 is 0.02, indicating that the medians are similar in Midterm 1 but not in Midterm 2; in the second midterm, the GenAI cohort had a significantly higher median score than the non-GenAI cohort.

We also apply the bootstrap approach to generate empirical distributions for the first and third quartiles of the midterm scores and then estimate the confidence

Figure 1. Exam Scores



Notes. (Left) GenAI vs. non-GenAI for Midterm 1. (Right) GenAI vs. non-GenAI for Midterm 2.

intervals (CIs) based on the bootstrap samples as provided in Table 2. When comparing the CIs of the first quartile, they overlap for Midterm 1, but do not overlap for Midterm 2, with the values in the CI of the first quartiles of the scores for the GenAI being higher than for NonGen AI cohort. Similar results are noted for the third quartile. Thus, overall, the lower and upper tails of the scores are different for Midterm 2 but not for Midterm 1, with the scores being consistently higher for the GenAI cohort in Midterm 2.

Last, we consider an association analysis to factors identified as important in explaining the variability in students' performance, where the factors describe GenAI and other resources used during the exams along with two controlling factors as described in Section 3.2.

We first evaluate the marginal correlation (see Figure B.6 in Online Appendix B.6). Based on this analysis, the students' grades in Midterm 1 have small marginal association with most factors (excluding controlling factors). The exceptions for the GenAI cohort are frequent use of the course material (negative correlation) and frequent rewrites, and the exceptions for the Non-GenAI cohort are occasional use of course material, frequent rewrites, and spending between one and three minutes of time (all with negative correlations). For Midterm 2, there are additional high correlations with the students' grades, including the use of CoPilot and long pauses for the GenAI cohort, and spending less than one minute (only positive correlation) and long pauses for the non-GenAI cohort.

To evaluate how all factors conditionally explain the variability in the students' grades, we also develop multiple linear regression models applied separately to the GenAI and non-GenAI cohorts and for the two midterms. The full model (including all predicting factors) of the GenAI cohort shows no statistical significance for Midterm 1 and only statistical significance for frequent use of the course materials, long pauses, and the purpose of AI prior use for Midterm 2. However, the GenAI models, although good fit, also show multicollinearity. The full model of the non-GenAI cohort shows statistical significance for AI prior use under Midterm 1 and both spending under one minute and frequent rewrites for Midterm 2. The multicolinearity is less of a concern for the non-GenAI models.

Because of the limited statistical significance in the full models and to address multicollinearity, we further perform variable selection using two classic but

Table 2. Confidence Intervals for First and Third Quartiles of the Midterm Scores

	Midterm 1		Midterm 2	
Quartile	GenAI	Non-GenAI	GenAI	Non-GenAI
First quartile	[62.3, 71.7]	[58.5, 70.0]	[61.9, 71.4]	[64.0, 75.0]
Third quartile	[77.7, 88.1]	[75.6, 82.9]	[87.1, 95.9]	[89.4, 93.3]

effective approaches: stepwise regression and lasso regularized regression. The results from variable selection are provided in Table 3. Consistently across all models, both the use of course materials and the student's degree program are statistically significantly associated with the midterm grades. Frequent rewrites, time to respond, long pauses, and multiple attempts also were shown to be significant for some models. These factors all point to one aspect of students' behavior, their challenges to respond to the exam questions, possibly due to stress or limited knowledge about the course material.

4.3. Question-Level Analysis

To evaluate how all factors conditionally explain the variability in the students' ability to respond correctly or not to each individual question, we develop logistic regression models applied separately to the GenAI and Non-GenAI cohorts and to the two midterms. The statistical significance of the full model (including all predicting factors) of the GenAI cohort cannot be evaluated because of multicollinearity for both midterms. The full model of the non-GenAI cohort shows statistical significance for the controlling factors, student's overall grade, and question difficulty for both midterms, as well as for the degree program in Midterm 2. The multicollinearity is less of a concern for the non-GenAI models.

As at the student level, because of the limited statistical significance and multicollinearity in the full models, we further perform variable selection using stepwise regression and lasso regularized regression. The results from variable selection are provided in Table 4. Consistently across all models, student's overall midterm grade is selected. Question difficulty is selected for both midterms in the GenAI models but only selected for Midterm 1 in the non-GenAI models. Responding fast to questions is also selected across most models, but not for the Midterm 2 non-GenAI model. For the non-GenAI cohort, dummy variables for time spent on questions are also selected for both midterms along with other factors related to time-based engagement (e.g., long pauses and multiple attempts). Last, the degree program is also selected for some GenAI and non-GenAI models, although not consistently across the variable selection methods or across midterms.

5. Discussion

Contrary to our initial hypotheses, the use of GenAI tools had a measurable overall positive impact on student performance, particularly evident in GenAI cohort's results on the second midterm. Although we anticipated that GenAI might help lower-performing students avoid basic errors and potentially hinder top-performing students on more nuanced questions, the data revealed a broader positive effect. Specifically, although there was no significant difference between the cohorts on Midterm 1, on Midterm 2, students in the cohort with access to GenAI demonstrated significantly higher scores in both the lower

Table 3. Selected Factors for the Regression Models with Individual Student Grades as the Target Variable Applied to GenAI and Non-GenAI Cohorts and for Both Midterms

Midterm	Method	Selected variables
	Panel A: GenAI: Variable selection results	
Midterm 1	Stepwise regression using BIC	*Frequent use of course material, Degree program*
	Regularized regression	*Frequent use of course material, Frequent rewrites, Degree program*
Midterm 2	Stepwise regression using BIC	*Frequent rewrites, Perceived AI purpose, Long pauses Frequent use of course of course material Time between 1 and 3 min, Degree program (AIC) Use of Co-Pilot (AIC)*
	Regularized regression	*Use of ChatGPT, Use of Co-Pilot, Frequent use of course material, Frequent rewrites, Time between 1 and 3 min, Time more than 3 min Multiple attempts, Long pauses, prior AI use Degree program*
	Panel B: Non-GenAI: Variable selection results	
Midterm 1	Stepwise regression (BIC)	*AI purpose, Occasional use of course material, Time between 1 and 3 min (AIC), Degree program (AIC), Multiple attempts (AIC)*
	Regularized regression	*Occasional use of course material, Time between 1 and 3 min, Frequent rewrites, AI purpose, Degree program*
Midterm 2	Stepwise regression (BIC)	*Frequent rewrites, Time less than 1 min, Time between 1 and 3 min (AIC), Occasional and Frequent use of course material (AIC)*
	Regularized regression	*Occasional use of course material, Frequent rewrites, Time less than 1 min*

Table 4. Selected Factors for the Regression Models with Question-Level (Binary) Response as the Target Variable Applied to GenAI and Non-GenAI Cohorts and for Both Midterms

Midterm	Method	Selected variables
Panel A: GenAI: Variable selection results		
Midterm 1	Stepwise regression (BIC)	*Student grade, Question difficulty, Time-based engagement (fast), Degree program*
	Regularized regression	*Student grade, Question difficulty*
Midterm 2	Stepwise regression (BIC)	*Student grade, Question difficulty, Time-based engagement (fast), Prior perceived AI use (math)*
	Regularized regression	*Student grade, Question difficulty*
Panel B: Non-GenAI: Variable selection results		
Midterm 1	Stepwise regression (BIC)	*Student grade, Question difficulty, Time-based engagement (fast)*
	Regularized regression	*Student grade, Question difficulty, Time-based Engagement (Fast), No use of course material, Time less than 1 min, Multiple attempts*
Midterm 2	Stepwise regression (BIC)	*Student grade*
	Regularized regression	*Student grade, No course material usage, Long pauses, Time less than 1 min, Time greater than 3 min), Degree program*

and upper quartiles compared with their counterparts who did not have GenAI access.

From a learning-theoretic perspective, these findings should be interpreted in light of the cognitive demands of the assessment format. Both midterms in this study consisted exclusively of multiple choice questions in a graduate-level time series analysis course. Although multiple choice questions can be designed at various levels of complexity, they are most commonly aligned with lower- to midlevel cognitive processes such as remembering, understanding, and routine application of procedures, corresponding to the lower levels of Bloom's taxonomy (Forehand 2010). Consequently, the performance gains we observe for the GenAI cohort on the second midterm are best understood as improvements in accuracy on these lower-tier cognitive tasks rather than as direct evidence of enhanced higher-order skills such as analyzing model assumptions, evaluating alternative modeling strategies, or creating new approaches.

At the same time, GenAI tools are often positioned as technologies that could support higher-order cognition by helping students compare alternative solution paths, critique model outputs, justify modeling choices, or generate and refine explanations. In our study, however, the multiple choice format inherently limited students' opportunities to externalize such higher-order reasoning and our outcome measure; correctness on individual items was not designed to capture deeper conceptual understanding or critical evaluation. This mismatch between the theoretical potential of GenAI to scaffold higher-order thinking and the primarily lower-order nature of our assessment represents an important theoretical limitation of the present work.

With this context in mind, it is also important to consider how students approached the exam itself. Several engagement factors were consistently associated with exam performance. Indicators such as time spent on questions, long pauses, multiple attempts, and frequent rewrites likely reflect cognitive effort, stress, or uncertainty factors, all known to influence test outcomes. Additionally, students' degree programs emerged as a significant predictor of performance, suggesting that prior domain knowledge and academic background play a role in how effectively students engage with course content and assessment formats. Different degree programs might also have different admissions cutoffs, which could also lead to differentiated performance.

Interestingly, even for the exam where the cohort with access to GenAI performed significantly better than the non-GenAI cohort, frequent or prolonged use of GenAI tools was not selected as a significant predictor in any of the models. This suggests that simply accessing or using GenAI tools does not guarantee improved performance; rather, the effectiveness of GenAI depends on how strategically and thoughtfully students engage with it.

A key observation is the difference in performance between the two cohorts across the two midterms. Although both cohorts reported similar levels of GenAI familiarity in the pre-experiment survey, only Cohort 2 showed a significant performance boost when using GenAI. This discrepancy can be attributed to timing and preparation: Cohort 2 had the advantage of learning from Cohort 1's experience, potentially leading to a more effective use of GenAI tools. This finding underscores the importance of GenAI literacy; students must not only have access to these tools but also understand how to use them effectively in academic contexts.

The consistent association between course material usage and exam performance across both cohorts further reinforces the value of traditional learning resources. For GenAI to yield similar benefits, students

must be both proficient in its use and able to critically evaluate its outputs.

One of the limitations of our study is that, although we asked students to self-report their AI usage, we did not directly assess their GenAI skills (e.g., prompting, evaluation of GenAI output, etc.) before the course began. It could be that such a question might explain some or all the difference in Cohort 1's and Cohort 2's performance.

Although the findings of this study provide valuable insights into the use of GenAI in multiple choice examination in a graduate-level time series analysis course, we acknowledge that they may not directly generalize to all operations research (OR)/management science (MS) courses, particularly introductory courses, or to undergraduate settings. The specialized nature of the course and the technical skills required limit the scope of applicability because using GenAI tools may struggle to produce accurate responses for highly specialized or advanced questions, which discourages students from relying on them in such cases. However, the experimental design and methodological framework offer a foundation that can be adapted to other OR/MS courses, particularly those involving quantitative modeling and computational tools. Importantly, the implications for instructional design—such as the need for AI literacy, structured integration of generative AI tools, and strategies for fostering critical evaluation of AI outputs—are broadly relevant across OR/MS curricula. Future research should explore these principles in diverse OR/MS contexts to assess their impact on learning outcomes at different levels.

6. Conclusions

This study provides empirical evidence on the impact of GenAI tools in graduate-level technical courses, specifically within the context of multiple choice assessments. Our findings suggest that GenAI can enhance student performance, but its effectiveness is contingent on students' preparedness and ability to use the tools strategically.

The observed performance gains in the second midterm, where students had more time to prepare and potentially learn from peers' GenAI experience, highlight the importance of GenAI literacy. Simply granting access to AI tools is not sufficient; students must understand how to engage with these tools critically and effectively. This parallels the use of other academic resources, such as calculators or reference materials, which require instruction and practice to be beneficial. Moreover, the consistent association between course material usage and performance underscores the continued relevance of traditional resources, even in AI-enhanced learning contexts.

A natural next step is to design assessments that explicitly span different levels of cognitive complexity and to analyze the impact of GenAI as a function of that complexity. Assessment items could be classified using a cognitive taxonomy such as Bloom's (Forehand 2010) and grouped into lower-level (remember, understand) and higher-level (analyze, evaluate, create) categories. Future studies could then test whether GenAI usage is more strongly associated with correctness on higher-complexity items than on lower-complexity items, thereby directly estimating interaction effects between GenAI access and cognitive demand. In parallel, we are planning a follow-up study using an open-ended data analysis exam, which will allow us to examine higher-order aspects of learning and stress-related behaviors that can trigger the use of GenAI during cognitively demanding tasks. Understanding these dynamics will be essential for designing effective and forward-looking educational practices in an AI-integrated academic landscape.

Acknowledgments

We thank the supporting team of online teaching assistants Matias Sacoto Molina, Hannah Quintal and Elly Konjkav, who have helped in collecting much of the data from this study.

References

Aldulaijan AT, Almalki SM (2025) The impact of generative AI tools on postgraduate students' learning experiences: New insights into usage patterns. *J. Inform. Tech. Ed. Res.* 24(3):1–29.

Avidov D, Ezra O, Cohen G, Cohen A, Bronshtein A (2025) Beyond scenario creation: Human and GenAI synergy in situational judgment test response development. *Ubiquity Proc.* 6(1):9.

Bucchiarone A, Cicchetti A, Azquez-Ingelmo V, Adami A, Schiavo F, Garc G, Ia-Holgado A, et al. (2024) Designing and generating lesson plans combining open educational content and generative AI. Wimmer M, Egyed A, Combemale B, Chechik M, eds. *Proc. ACM/IEEE 27th Internat. Conf. Model Driven Engrg. Languages Systems* (Association for Computing Machinery (ACM), New York), 78–86.

Chan CKY, Hu W (2023) Students' voices on generative AI: Perceptions, benefits, and challenges in higher education. *Internat. J. Ed. Tech. High. Ed.* 20:43.

Combrinck C, Loubser N (2025) Student self-reflection as a tool for managing GenAI use in large class assessment. *Discovery Ed.* 4:72.

Farrokhnia M, Soleimani S, Noroozi O (2025) 3: Generative AI in higher education: Transformative tools for research, teaching, and assessment. Sabbaghan S, ed. *Navigating Generative AI in Higher Education* (Edward Elgar Publishing, Cheltenham, UK), 33–53.

Forehand M (2010) Bloom's taxonomy. Orey M, ed. *Emerging Perspectives on Learning, Teaching, and Technology* (University of Georgia, Athens, Greece), 41–47.

Goel AK, Joyner DA (2017) Using AI to teach AI Lessons from an online AI class. *AI Magazine* 38(2):48–59.

Gruenhagen JH, Sinclair PM, Carroll J-A, Baker PR, Wilson A, Demant D (2024) The rapid rise of generative AI and its implications for academic integrity: Students' perceptions and use of chatbots for assistance with assessments. *Comput. Ed. Artificial Intelligence* 7:100273.

Ho WYJ (2021) 'I knew that you were there, so I was talking to you': The use of screen-recording videos in online language learning research. *Qualitative Res.* 21(1):120–139.

Honorlock (2025) Replace Yuja with Honorlock proctoring. Accessed August 6, 2025, https://honorlock.com/replace-yuja-proctoring/.

Instructure (2025) Canvas LMS by instructure: Simplify teaching and learning. Accessed August 6, 2025, https://www.instructure.com/landing/canvas.

Johnston H, Wells RF, Shanks EM, Boey T, Parsons BN (2024) Student perspectives on the use of generative artificial intelligence technologies in higher education. *Internat. J. Ed. Integrity* 20:2.

Khlaif ZN, Salama N, Alkouk WA, Eideh BA (2025) Redesigning assessments for AI-enhanced learning: A framework for educators in the generative AI era. *Ed. Sci. (Basel)* 15(2):174.

Kilinç S (2024) Comprehensive AI assessment framework: Enhancing educational evaluation with ethical AI integration. *J. Ed. Tech. Online Learn.* 7(4):521–540.

Kim J, Klopfer M, Grohs JR, Eldardiry H, Weichert J, Cox LA, Pike D (2025) Examining faculty and student perceptions of generative AI in university courses. *Innovative Higher Ed.* 50(4):1281–1313.

Klymkowsky M, Cooper MM (2024) The end of multiple choice tests: Using AI to enhance assessment. Preprint, submitted June 11, https://arxiv.org/abs/2406.07481.

Law L (2024) Application of generative artificial intelligence (GenAI) in language teaching and learning: A scoping literature review. *Comput. Edu Open* 6:100174.

Luo J (2024) A critical review of GenAI policies in higher education assessment: A call to reconsider the "originality" of students' work. *Assessment Evaluation Higher Ed.* 49(5):651–664.

Luo J (2025) How does GenAI affect trust in teacher-student relationships? Insights from students' assessment experiences. *Teaching Higher Ed.* 30(4):991–1006.

Mozelius P (2024) Generative AI and its impact on activities and assessment in higher education: Some recommendations from master's students. *Proc. 4th Internat. Conf. AI Res. (ICAIR 2024)*, vol. 4 (ACI Academic Conferences International, Bicester, UK), 289–295.

Nguyen A, Duong AT, Nguyen DTB, Lai VTT, Dang B (2025) Guidelines for learning design and assessment for generative artificial intelligence-integrated education: A unified view. *Inform. Learn. Sci.* 126(7–8):491–512.

Nikolic S, Sandison C, Haque R, Daniel S, Grundy S, Belkina M, Neal P (2024) ChatGPT, Copilot, Gemini, SciSpace and Wolfram versus higher education assessments: An updated multi-institutional study of the academic integrity impacts of generative artificial intelligence (GenAI) on assessment, teaching and learning in engineering. *Australasian J. Engrg. Ed.* 29:126–153.

O'Brien C (2020) Honorlock: What the online proctor does and doesn't do. Accessed April 23, 2026, https://news.wisc.edu/honorlock-what-it-does-and-doesnt-do/.

Ogunleye B, Zakariyyah KI, Ajao O, Olayinka O, Sharma H (2024) Higher education assessment practice in the era of generative AI tools. Preprint, submitted April 1, https://arxiv.org/abs/2404.01036.

Otundo J (2025) Navigating the implications of generative AI in education, learning, and assessment. *Internat. J. Res. Ed. Humanities Commerce* 6(5).

Perkins M, Furze L, Roe J, Vietnam JMBU, University D, Singapore JCU (2023) The artificial intelligence assessment scale (AIAS): A framework for ethical integration of generative AI in educational assessment. *J. University Teaching Learn. Practice* 21(6).

Pesovski I, Santos R, Henriques R, Trajkovik V (2024) ChatGPT in exam rooms: Preliminary insights into student performance with and without AI assistance. *Proc. Edulearn24* (IATED, Valencia, Spain), 5908–5914.

Sohail S, Parveen S, Dar T (2025) Investigating the role of generative AI in transforming teaching, learning, and assessment practices. *J. Appl. Linguistics TESOL* 8(4):159–174.

Stritto MED, Underhill GR, Aguiar NR (2024) Online students' perceptions of generative AI. Accessed April 23, 2026, https://ecampus.oregonstate.edu/research/study/ai-survey/.

Tang K-S, Cooper G, Rappa N, Cooper M, Sims C, Nonis K (2024) A dialogic approach to transform teaching, learning & assessment with generative AI in secondary education: A proof of concept. *Pedagogies* 19(3):493–503.

Wu R, Yu Z (2024) Do AI chatbots improve students learning outcomes? Evidence from a meta-analysis. *British J. Ed. Tech.* 55(1):10–33.

Xiao Z, Lee E, Yuan S, Ding R, Tang YJ (2025) Generative AI in graduate bioproces engineering exams: Is attention all students need? *Ed. Chemical Engrg.* 52:133–140.

Zhao J, Chapman E, Sabet PGP (2024) *Ed. Res. Perspectives* 51:124–155.