EXAMINATIONS OF ONLINE EXAM SCORES IN HIGHER EDUCATION BEFORE AND AFTER THE ARTIFICIAL INTELLIGENCE (AI) TOOLS MADE PUBLICLY AVAILABLE

Michael R. Santos, Sonoma State University
Jinglin Jiang, Sonoma State University
Vincent Richman, Sonoma State University

ABSTRACT

This paper compares online exam scores from an online asynchronous introductory finance course before and after January 1, 2023 to determine if the widely available AI tools, such as ChatGPT and Google Gemini, have any significant effect on the average exam scores of the students. The research findings show that there was not statistically significant change in the average exam scores after the AI tools were publicly available. Perhaps, this result is due to the limited success of the AI tools for the introductory corporate finance exams.

Keywords: *Artificial intelligence (AI), ChatGPT, Copilot, Google Gemini, Generative AI*

INTRODUCTION

The introduction of ChatGPT (also through Copilot of Microsoft) and Google Gemini to the public at no cost created substantial excitement and fear among the academic community and business world. The literature on Artificial Intelligence (AI) has been growing over the last 10 years, but after 2023, the interest and number of academic publications around the world increased exponentially. However, the current literature on AI is too fragmented due to the contributions from all sciences exploring its implications for higher education and the business world.

Many academic researchers focused on indirect measurements of AI by using student surveys or interviews to assess the perspectives of graduate and undergraduate students. Overall, the research in this area indicates that the AI framework at the universities is not well-defined and therefore creates confusion and uncertainty among students and educators for the potential use of AI tools in the classroom environment.

One of the well-known AI tools, ChatGPT, was first made available to the public on November 30, 2022. The other, Bard (currently known as Gemini), was introduced on March 21, 2023. This paper compares the average online and asynchronous exam scores from an introductory finance course before and after January 1, 2023 to determine if the widely available AI tools, such as ChatGPT and Google Gemini, made any difference in the average exam scores.

Due to the significant developments in AI technology that happened in the years 2023 and 2024, this paper focuses on the most recent literature.

Among the peer-reviewed published articles, (Schneider and Haried, 2024) surveyed 328 undergraduate students to investigate trust, anxiety, and the benefits of AI within the university environment. They find that students and instructors are already using AI and the survey participants believe Generative AI offers extensive benefits by improving work quality and efficiency by contributing to personalized learning for students. However, there are concerns about transparency and plagiarism. Further, (Song, 2024) claims that higher educational institutions (HEIs) are challenged with academic integrity because of the generative artificial intelligence (AI) causing an increase in plagiarism and cheating. Song suggests that HEIs need to improve AI literacy among students and faculty and establish policies for AI use to make it equitable, inclusive, and ethical.

Also, a study by (Hornberger et al., 2023) shows a multiple-choice AI literacy test to learn about the AI literacy among German university students. They find that AI literacy is higher among students who study or work in technical fields. Therefore, they propose AI courses for all students studying at different fields to understand the basics of AI. Further, a survey by (Vecchiarini and Somia, 2023) assess students' awareness as well as the benefits and limitations of ChatGPT for entrepreneurship courses. They find that AI can support student learning in the entrepreneurship field and increase engagement through curiosity and inquiry in the course assignments. In addition, (Ratten, 2023) investigates how ChatGPT affects family business management education. According to her, there was an initial panic among family business management educators proposing a ban for the use of AI. However, Ratten suggests open discussions should take place about the advantages and disadvantages of ChatGPT. Also, (Olga et al., 2024) use a case study to measure the impact of employing Generative AI tools and cyber-social teaching methods to assess AI literacy among the students. They analyze 37 participants for three 8-week graduate courses by teaching them how to use GenAI reviewer and GenAI image generator platforms. As a result, Olga et al. claim that the AI image generation tools make the participating students to be active learners and contribute to the development of their AI literacy.

In another note, (Benuyenah and Dewnarain, 2024) presents a qualitative study with the student interviews to understand their perceptions of ChatGPT. They show that the majority of students enrolled in the undergraduate and postgraduate business programs support ChatGPT to help improve student learning and the quality of assessments. They find that the majority of the students believe ChatGPT can eliminate certain careers in academia and the business world.

Also, (Sundberg and Holmström, 2024) advocates an approach with the no-code AI in a course to help the university students studying social sciences who could benefit from the no-code AI courses. In a different approach, (Memarian and Doleck, 2023) reviewed 33 publications from SCOPUS and Web of Science (WoS) for fairness, accountability, transparency, and ethics (FATE) in AI and higher education. They find that the FATE is defined more descriptive than technical and it is studied more in quantitative rather than qualitative means.

This paper relates to (Song, 2024) and the research papers suggesting the establishment of AI policies in higher education. However, it differs from the previous research because the

data is provided by an instructor who did not have an AI policy given to his students whether to use the AI tools or not during the exams. Therefore, the utilization of the AI tools during the exams is not alarming and it cannot be classified as cheating or unethical. This paper is motivated to understand if the students were using publicly available AI tools during online exams for an introductory asynchronous corporate finance course without ethical judgement of the students.

DATA AND METHODOLOGY

Table 1 below summarizes the details of this study. There were three exams during each semester with 40 questions made up with multiple-choice and true-false questions. All exams were randomized and therefore each student received an exam with a different set of questions.

Over the years from 2020 to 2024, the structure of the exam stayed identical and the students took exams online and asynchronously. Further, the exams were given through Canvas LMS during the fall and spring semesters starting from Fall 2020 to Spring 2024. Assuming that the AI tools [ChatGPT and Google Bard (Gemini)] widely become available to the public at the beginning of 2023, there might be some effects on the average exams scores for year 2023 and 2024.

Table 1
Summary of BUS 370 Introductory Corporate Finance Course Exam Specifications

Course	# Of Exams	# Of Questions	Question Types*	Questions in the Exams (Randomized/Fixed)*	# Of Students Enrolled	Semester
BUS 370-1	3	40	MC/TF	Randomized	40	Spring 2024
BUS 370-2	3	40	MC/TF	Randomized	37	Spring 2024
BUS 370	3	40	MC/TF	Randomized	43	Fall 2023
BUS 370-1	3	40	MC/TF	Randomized	42	Spring 2023
BUS 370-2	3	40	MC/TF	Randomized	50	Spring 2023
BUS 370	3	40	MC/TF	Randomized	43	Fall 2022
BUS 370	3	40	MC/TF	Randomized	100	Spring 2022
BUS 370	3	40	MC/TF	Randomized	68	Fall 2021
BUS 370	3	40	MC/TF	Randomized	50	Spring 2021
BUS 370	3	40	MC/TF	Randomized	56	Fall 2020

*MC = Multiple Choice, TF = True or False

** Randomized questions in the exam provide a unique set of questions to each student. In contrast, non-randomized (fixed) questions provide the same questions being asked to all students taking the exam.

Table 2 exhibits the average exam scores (Exam 1, Exam 2, and Exam 3) from the introductory corporate finance classes from Fall 2020 to Spring 2024 semesters. A cursory look at the average exam scores on Table 3 from the period starting from Fall 2020 to Fall 2022 and Spring 2023 to Spring 2024 shows no discernable change of averages.

Table 2
Three Average Online Asynchronous Average Exam Scores for the Semesters Starting from Fall 2020 to Spring 2024

COURSE	# OF STUDENTS	SEMESTER	AVERAGE EXAM 1 SCORE*	AVERAGE EXAM 2 SCORE*	AVERAGE EXAM 3 SCORE*	GRAND AVERAGE**
BUS 370-1	40	Spring 2024	81.06	79.87	80.88	**80.60**
BUS 370-2	37	Spring 2024	80.95	77.09	80.63	**77.09**
BUS 370	43	Fall 2023	80.47	81.57	82.86	**81.63**
BUS 370-1	42	Spring 2023	79.51	78.72	78.10	**78.78**
BUS 370-2	50	Spring 2023	76.50	75.77	77.14	**76.47**
BUS 370	43	Fall 2022	75.24	75.12	79.45	76.60
BUS 370	100	Spring 2022	76.18	71.94	77.16	75.09
BUS 370	68	Fall 2021	81.47	77.61	82.74	80.61
BUS 370	50	Spring 2021	77.55	75.26	78.65	77.15
BUS 370	56	Fall 2020	76.64	76.50	81.32	78.15

*Highest achievable score is 100 points.
**Grand Average refers to the average of the three exam (Exam 1, Exam 2, and Exam 3) averages.

In order to confirm this cursory observation that there are no significant differences in two periods, an ANOVA Single Factor test is applied to determine if there were any statistical difference between the average exam scores of the students BEFORE (fall 2020 to fall 2022) and AFTER (spring 23 to spring 24).

Period 1: Average exam scores BEFORE (fall 2020 to fall 2022).
Period 2: Average exam scores AFTER (spring 23 to spring 24).

$$H_0 = \bar{X}_{BEFORE} - \bar{X}_{AFTER} = 0$$
$$H_1 = \bar{X}_{BEFORE} - \bar{X}_{AFER} \neq 0$$

H_0 (Null hypothesis) claims that the differences of the average exam scores BEFORE (fall 2020 to fall 2022) and AFTER (spring 23 to spring 24) periods are equal to zero. That means if the students were using the AI tools during AFTER (spring 23 to spring 24) period, the effect of the AI tools on the average student scores are negligible and it is not statistically significant. And, H_1, the alternative hypothesis claims otherwise: the differences of average exam scores BEFORE (fall 2020 to fall 2022) and AFTER (spring 23 to spring 24) periods are not equal to zero and therefore the effect of the AI tools on the average exam scores is statistically significant.

Table 3 provides the details of the F-test statistics. According to the ANOVA estimations, F-statistics value (1.068578) is below the F-critical value (5.317655), and therefore we accept the null hypothesis at 5% level that the differences of average exam scores for the BEFORE (fall 2020 to fall 2022) and AFTER (spring 23 to spring 24) periods are zero. Thus, H_0 (null

hypothesis) is accepted indicating that the average exam scores of two periods are statistically similar.

Table 3
ANOVA Single Factor for Two Groups: BEFORE and AFTER*

Groups	Count	Sum	Average	Variance		
BEFORE	5	387.6	77.52	4.2083		
AFTER	5	394.57	78.914	4.88433		
Source of Variation	**SS**	**df**	**MS**	**F**	**P-value**	**F crit**
Between Groups	4.85809	1	4.85809	1.068578	0.331502	5.317655
Within Groups	36.37052	8	4.546315			
Total	41.22861	9				

*BEFORE refers to the period from fall 2020 to fall 2022 and AFTER refers to the period from spring 23 to spring 24.

Even though there is no way to know whether the students are using the AI tools during the online asynchronous exams during years 2023 and 2024, the results indicate that the AI tools made no difference for the average exams scores in the introductory corporate finance classes.

In the last several years, it has been frequently reported in the popular media that ChatGPT achieved high marks on the medical and law school exams [Murphy Kelly, S., 2023, January 26) and (Cole, S., 2023, March 14)]. After reading the sensational news about the achievements of the AI tools in the exams, it is a question whether the students in this study used these tools to increase their grades. Is it possible that the students were using the AI tools after 2023 but the AI tools were not effective or useful to answer questions during the exams? To understand the effectiveness of the AI tools in the introductory corporate finance course, the same three exams were given to the publicly available AI tools (Copilot as an approximation to ChatGPT and Google Gemini).

Table 4 below shows the AI exam results. The exam scores of the AI tools are lower than the averages of the students at all semesters. The AI tools scored ranging from a low score of 62.50/100 to a high score of 75/100. These exam scores are not too impressive and raise questions about their effectiveness for the introductory corporate finance course exams. Therefore, it is possible that the students tried the AI tools during their online asynchronous exams but did not benefit sufficiently to improve their grades.

About the limitations of the AI tools, (Martínez, 2024) and (Hicks, 2024) claim that the success of the ChatGPT is inflated especially by the popular media resources.

Table 4
Giving the Same Randomized Exams to MS Copilot and Google Gemini in Year 2024

MS COPILOT	**MC**	**MC CALC**	**TF**	**TOTAL**	**SCORE**
Exam 1 (Correct/# of Questions)	16/16	1/12	10/12	27/40	**67.50**
Exam 2 (Correct/# of Questions)	14/15	5/13	9/12	28/40	**70.00**
Exam 3 (Correct/# of Questions)	6/11	8/17	11/12	25/40	**62.50**
GOOGLE GEMINI	**MC**	**MC CALC**	**TF**	**TOTAL**	**SCORE**
Exam 1 (Correct/# of Questions)	14/16	5/12	11/12	30/40	**75.00**
Exam 2 (Correct/# of Questions)	10/15	10/13	9/12	29/40	**72.50**
Exam 3 (Correct/# of Questions)	2/11	15/17	11/12	28/40	**70.00**

*The exams are online and have 40 questions with the highest achievable score being 100 points.

**MC = Multiple-Choice questions, MC CALC = Multiple-Choice calculation questions, TF = True-False questions, TOTAL = # of Correct Answers/(MC + MC CALC + TF).

STUDY LIMITATIONS

This study has several limitations that should be acknowledged. First, the analysis is based on data from a single introductory finance course at one institution, which limits the generalizability of the findings to other disciplines, institutions, or educational contexts. Second, while the paper investigates the potential influence of generative AI tools on student exam performance, it does not control for other factors that may have affected student outcomes over time, such as changes in student demographics, study habits, instructor approaches, or institutional policies. These unobserved variables may confound the results and obscure any subtle effects of AI usage. Third, the study assumes that students may have had access to AI tools but lacks direct evidence of actual usage during the exams, making it difficult to draw firm conclusions about AI's role in student performance.

CONCLUSION

The bewildering speed in AI development created much excitement and concern in higher education and the business world. This paper investigates whether there were any AI effects on the average student scores for an introductory finance course after these tools became publicly available to all students.

The findings show that there was no statistically significant change in the average exam scores after the students had access to the AI tools (years 2023 and 2024) while they were taking the online asynchronous exams. Naturally, it cannot be claimed that the students did not try the AI tools during the exams. Further, this research tests the effectiveness of the AI tools such as Copilot or Google Gemini by making them take the exams. The scores obtained from the same randomized, online and asynchronous exams given to Copilot and Google Gemini were not impressive. Perhaps indicating that the AI tools are not as effective as they are claimed to be by social media.

Since the AI tools are easily available to students, higher education institutions may consider adapting these tools in education and incorporate in ways that the integrity of student

assessment can be maintained. Future research should consider the development of policies to guide educators and students to reduce ethical concerns.

REFERENCES

Benuyenah, V., & Dewnarain, S. (2024). Students' intention to engage with ChatGPT and artificial intelligence in higher education business studies programmes: An initial qualitative exploration. International Journal of Distance Education Technologies, 22(1), 1-21. https://doi.org/10.4018/IJDET.348061.

Cole, S. (2023, March 14). The new GPT-4 AI gets top marks in law, medical exams, OpenAI claims. Vice. Retrieved September 2, 2024, from https://www.vice.com/en/article/the-new-gpt-4-ai-gets-top-marks-in-law-medical-exams-openai-claims.

Hicks, M. T., Humphries, J., & Slater, J. (2024). ChatGPT is bullshit. Ethics and Information Technology, 26(38). https://doi.org/10.1007/s10676-024-09775-5.

Hornberger, M., Bewersdorff, A., & Nerdel, C. (2023). What do university students know about artificial intelligence? Development and validation of an AI literacy test. Computers and Education: Artificial Intelligence, 5, 100165. https://doi.org/10.1016/j.caeai.2023.100165.

Martínez, E. (2024). Re-evaluating GPT-4's bar exam performance. Artificial Intelligence and Law. https://doi.org/10.1007/s10506-024-09396-9.

Memarian, B., & Doleck, T. (2023). Fairness, accountability, transparency, and ethics (FATE) in artificial intelligence (AI) and higher education: A systematic review. Computers and Education: Artificial Intelligence, 5, 100152. https://doi.org/10.1016/j.caeai.2023.100152.

Murphy Kelly, S. (2023, January 26). ChatGPT passes exams from law and business schools. CNN Business. Retrieved September 2, 2024, from https://www.cnn.com/2023/01/26/tech/chatgpt-passes-exams/index.html.

Olga, A., et al. (2024). Combining human and artificial intelligence for enhanced AI literacy in higher education. Computers and Education: Open, 6, 100184. https://doi.org/10.1016/j.caeo.2024.100184.

Ratten, V. (2023). Editorial: ChatGPT, the metaverse and artificial intelligence: Implications for family business management education. Journal of Family Business Management, 13(4), 821-827. https://doi.org/10.1108/JFBM-12-2023-199.

Schneider, S., & Haried, P. (2024). Use of artificial intelligence in higher education with particular reference to automated content generation, trust, and anxiety. Journal of Business and Educational Leadership, 14(1), 4-21.

Song, N. (2024). Higher education crisis: Academic misconduct with generative AI. Journal of Contingencies and Crisis Management, 32(1), e12532. https://doi.org/10.1111/1468-5973.12532.

Sundberg, L., & Holmström, J. (2024). Teaching tip: Using no-code AI to teach machine learning in higher education. Journal of Information Systems Education, 35(1), 56-66. https://doi.org/10.62273/CYPL2902.

Vecchiarini, M., & Somia, T. (2023). Redefining entrepreneurship education in the age of artificial intelligence: An explorative analysis. The International Journal of Management Education, 21, 100879. https://doi.org/10.1016/j.ijme.2023.100879.