https://doi.org/10.62560/dk.2025.2.02

LÁSZLÓ BOGNÁR *–ANTAL JOÓS **

The Impact of Allowing ChatGPT on Student Scores and Test Completion Time in a Mid-Term Math Exam

* *University of Dunaújváros, Institute of Computer Engineering, Department of Mathematics*
Email: bognarl@uniduna.hu

** *University of Dunaújváros, Institute of Computer Engineering, Department of Mathematics*
Email: joosa@uniduna.hu

Abstract This study explores the impact of ChatGPT usage on student performance in a mid-term mathematics exam, focusing on two key research questions. First, whether there is a significant difference in the average total scores between students using ChatGPT and those not using it. The findings indicate that students who used ChatGPT scored significantly lower on average than their non-GPT-using peers, with fewer high-achieving students and a higher proportion failing to meet the minimum performance threshold. Second, the study examines whether there is a significant difference in the average time taken to complete the test between the two groups. Surprisingly, no notable difference in completion time was observed, challenging the assumption that ChatGPT users would either complete the exam faster or spend time comparing their answers with those generated by GPT. These results highlight the need for further investigation into the role of AI tools like ChatGPT in education, particularly their effectiveness in enhancing learning outcomes in mathematics.

Keywords: ChatGPT; AI in education; student performance; mathematics exam; test completion time; AI tools; mid-term exam; learning outcomes; educational technology; AI-assisted learning.

Absztrakt: Ez a tanulmány a ChatGPT használatának hatását vizsgálja a diákok teljesítményére egy félévközi matematika vizsgán, két fő kutatási kérdésre összpontosítva. Először is, hogy van-e szignifikáns különbség a ChatGPT-t használó és nem használó diákok átlagos összpontszámai között. Az eredmények azt mutatják, hogy a ChatGPT-t használó diákok átlagosan jelentősen alacsonyabb pontszámot értek el, mint a ChatGPT-t nem használó társaik, kevesebb volt a kiemelkedő teljesítményt nyújtó diákok száma, és nagyobb arányban nem érték el a minimális teljesítmény küszöböt. Másodszor, a tanulmány azt vizsgálja, hogy a két csoport között van-e szignifikáns különb-

[1] Gulwani, S.–Polozov, O.–Singh, R. (2017): Program Synthesis. *Foundations and Trends® in Programming Languages,* 4., (1–2.), pp. 1–119.

[2] Miller, L. A. (1981): Natural Language Programming: Styles, Strategies, and Contrasts. *IBM Systems Journal,* 20., (2.), pp. 184–215.

[3] Dohmke, T. (2022): *GitHub Copilot Is Generally Available to All Developers.* Retrieved from https://github.blog/2022-06-21-github-copilot-is-generally-available-to-all-developers/

[4] OpenAI. (2022): *Introducing ChatGPT.* Retri-eved from https://openai.com/blog/chatgpt

[5] Hu, K. (2023): *ChatGPT Sets Record for Fastest-Growing User Base – Analyst Note. Reuters.* Retrieved from https://www.reuters.com/technology/chatgpt-sets-recordfastest-growing-user-base-analyst-note-2023-02-01/

[6] Mehdi, Y. (2023): *Reinventing Search with a New AI- powered Microsoft Bing and Edge, Your Copilot for the Web.* Retrieved from https://blogs.microsoft.com/blog/2023/02/07/rein-venting-search-with-a-new-ai-pow-ered-microsoft-bing-and-edgeyour-copilot-for-the-web/

[7] Pichai, S. (2023): *An Important next Step on Our AI Journey.* Retrieved from https://blog.google/technology/ai/bard-google-ai-search-updates/

[8] Buolamwini, J.–Gebru, T. (2018): Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In: *Conference on Fairness, Accountability and Transparency,* pp. 77–91. PMLR.

[9] Liang, P. P.–Wu, C.–Morency, L.-P.–Salakhutdinov, R. (2021): Towards understanding and mitigating social biases in language models. In: *International Conference on Machine Learning,* pp. 6565–6576.

[10] Bender, E. M.–Gebru, T.–McMillan-Major, A.–Shmitchell. S. (2021): On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency,* pp. 610–623.

ség a teszt kitöltéséhez szükséges átlagos idő tekintetében. Meglepő módon a teszt kitöltési ideje nem különbözött számottevően, ami megkérdőjelezi azt a feltételezést, hogy a ChatGPT-felhasználók vagy gyorsabban teljesítik a vizsgát, vagy időt töltenek a GPT által generált válaszok összehasonlításával. Ezek az eredmények rávilágítanak arra, hogy tovább kell vizsgálni a ChatGPT-hez hasonló mesterséges intelligencia eszközök szerepét az oktatásban, különösen a matematika tanulási eredményeinek javításában való hatékonyságukat.

Kulcsszavak: ChatGPT; mesterséges intelligencia az oktatásban; tanulói teljesítmény; matematika vizsga; teszt kitöltési ideje; AI-eszközök; félévközi vizsga; tanulási eredmények; oktatási technológia; mesterséges intelligenciával támogatott tanulás.

Introduction

For decades, research areas such as neural networks, programming synthesis [1], and natural language programming [2] have been advancing steadily, but it is only recently that these technologies have entered the public spotlight through major commercial releases. In June 2022, GitHub Copilot, an AI-powered code generation tool, was launched after a year of private beta testing [3]: Shortly thereafter, in November 2022, OpenAI introduced ChatGPT [4], which gained 100 million users within just two months, setting a record for the fastest-growing app [5]: By early 2023, both Microsoft and Google integrated conversational AI like ChatGPT into their web search platforms [6; 7]: The rapid adoption of AI tools has sparked a range of concerns, including bias [8; 9], ethics [10], misinformation

[11], privacy [12], energy consumption [13], and the consolidation of corporate power [14]: In the education sector, particularly, educators are questioning the impact of AI tools on student learning and performance, wondering whether these tools enhance or undermine educational outcomes [15]:

In this study, we examine the effects of ChatGPT on student performance during a mid-term mathematics exam by focusing on two primary research questions:

Research Question 1: Is there a significant difference between the average total scores of students using ChatGPT and those not using it?

Research Question 2: Is there a significant difference in the time taken to complete the test between the two groups?

By addressing these questions, this research aims to offer insights into the role of AI tools like ChatGPT in academic assessments and their broader implications for educational environments.

Conditions of the experiment

The experiment was conducted with foreign students of *Mathematics 1* and *Engineering Mathematics 1* at the University of Dunaújváros, in Hungary. In the following we will refer to these subjects as Mathematics 1 only. In Mathematics 1, students write two tests in the Moodle Learning Management System during the semester. The final grade is determined by the sum of the scores achieved on these two mid-term tests. In this paper, we look only at student results in the first test. In this test, each student was asked 5 questions, one from each of the 5 sub-topics. The questions differed only in that they contained random parameters generated by the Moodle system. A total of 140 students submitted valid tests, 22 of which self-reported using ChatGPT. Hence, for these 140 tests, a total of 140x5=700 different questions were asked, and of these, 22x5=110 questions were solved using ChatGPT.

Learning material for the test covers the basic topics of linear algebra, which includes an introduction to matrices, matrix operations, calculating

[11] Kreps, S.–McCain, R. M.–Brundage, M. (2022): All the News That's Fit to Fabricate: AI-generated Text as a Tool of Media Misinformation. *Journal of Experimental Political Science,* 9., (1), pp. 104–117.

[12]Butterick, M. (2022): *GitHub Copilot Inv.estigation.* Joseph Saveri Law Firm & Matthew Butterick. Retrieved from https://githubcopilotinvestigation.com/

[13] Strubell, E.–Ganesh, A.–McCallum, A. (2019): *Energy and Policy Considerations for Deep Learning in NLP.* arXiv preprint arXiv:1906.02243.

[14] Xiang, C. (2023): *OpenAI Is Now Everything It Promised Not to Be: Corporate, Closed-Source, and For-Profit.* Retrieved from https://www.vice.com/en/article/5d3naz/openaiis-now-everything-it-promised-not-to-be-corporate-closed-source-and-forprofit

[15] Johnson, A. (2023): *ChatGPT In Schools: Here's Where It's Banned–And How It Could Potentially Help Students.* Retrieved from https://www.forbes.com/sites/ariannajohnson/2023/01/18/chatgpt-in-schools-heres-where-its-banned-and-how-it-could-potentially-help-students/

determinant and inverse, operations with vectors such as scalar multiplication, vector multiplication and mixed multiplication, calculating the angles of vectors, and addition, subtraction, multiplication and division of complex numbers given in algebraic form. More specifically:

- The first question (Q1) was on addition, subtraction, multiplication, transposition, determinant, adjoint and inverse calculus with matrices.
- The second question (Q2) was on addition and subtraction of vectors, multiplication by scalar, linear independence of vectors, base of vectors, rank of matrix, scalar multiplication of vectors and solution of linear system of equations.
- The third question (Q3) was about mixed product of vectors, scalar product of vectors, equation of a plane, equation of a line, length of a vector, product of vectors and closed angle of vectors.
- The fourth question (Q4) was taken from the same set as the first question.
- The fifth question (Q5) was taken from the topics: real and imaginary parts of complex numbers, sum of complex numbers, difference of complex numbers, multiplication of complex numbers and absolute value of complex numbers.

Students have 45 minutes to complete the test. At the end of the 45 minutes, any test that has not been submitted by the student will be automatically submitted. The student will immediately see the result of their test and the correct answers. Teaching was face-to-face, but the test was online. Students were only told on the morning of the test that they could use ChatGPT for the test. The test included a self-report question about whether they had used ChatGPT during the test. There were no penalties or rewards for using ChatGPT. There was no use of ChatGPT at all in the mathematics lessons.

The mathematical problem-solving capabilities of ChatGPT and its pitfalls were not demonstrated. Students were allowed to use not only ChatGPT, but also other similar tools.

Comparison of Total scores of ChatGPT users and non-users

Figure 1. shows the distribution of the students’ Total scores. The Total score is the sum of the scores for the 5 questions of the test. The mean Total score was 16.57, the standard deviation was 5.018. It seems that most students passed the mid-term test, with many scoring the maximum 20 points. Only 14 students (10%) failed to reach the minimum performance of 60%, i.e., 12 points. However, if we split the students by GPT use, we get a more nuanced picture.

Figure 1. The distribution of students' Total scores

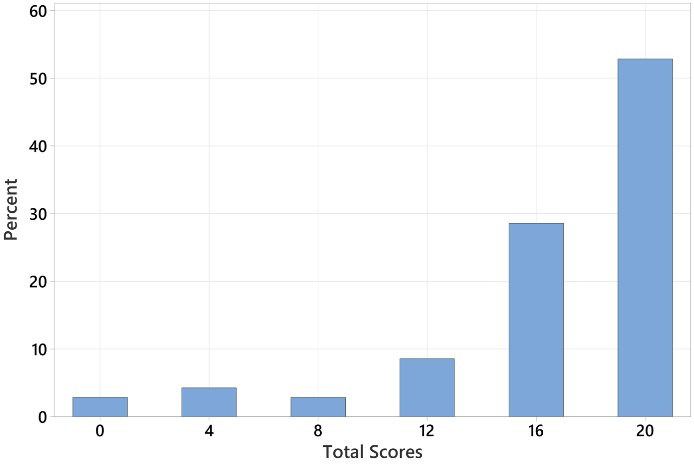
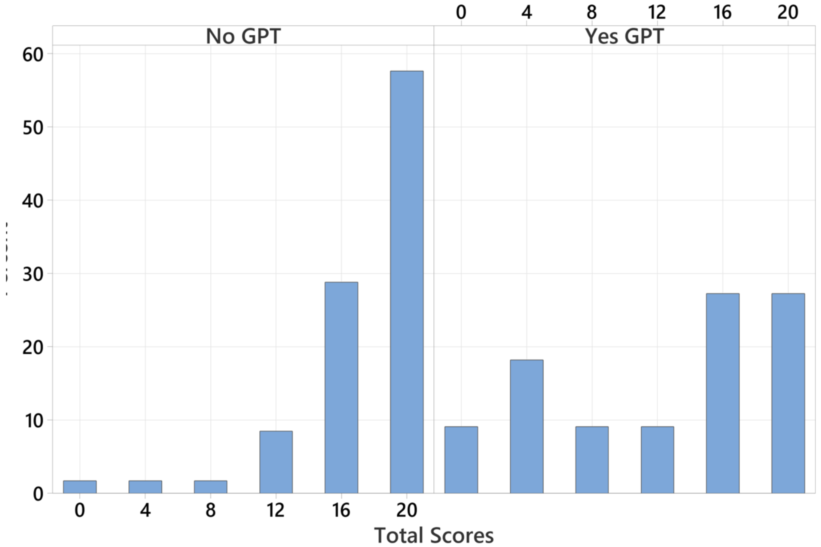


Figure 2. shows the distribution of Total scores for students who do not use ChatGPT and students who use ChatGPT separately. The mean Total score for students not using ChatGPT was 17.356 with a standard deviation of 4.110, and the mean Total score for students using ChatGPT was 12.36 with a standard deviation of 7.26. On the one hand, it is striking how much lower the average score of those who used ChatGPT was, and on the other hand, the distribution of scores in the two cases varies considerably. Among the ChatGPT users the highest scorers are few, and while 36.3% of them failed to reach the minimum 12 points, only 5.1% of non-users failed.

To check that this trend is not only apparent in these samples, but that a similar result would be obtained from the whole population, we performed a two-sample t-test for considering the difference of the means of the Total scores. Let μ_1 be the population mean of the questions Total score when GPT was not used and let μ_2 be the population mean of the questions Total score when GPT was used. Let the null hypothesis be $\mu_1-\mu_2=0$ and the alternative hypothesis be $\mu_1-\mu_2>0$. Now the observed t-value is 4.04, and the P-value is less than 0.0001. This means that the difference between the mean Total scores is highly significant.

Figure 2. The distribution of students' Total scores for students not using ChatGPT and students using ChatGPT



As the distributions of the two samples are quite different, and thus the application of the t-test is questionable, we performed a Mood's Median Test as well. Here the observed Chi-square-value is 3.43, and the P-value is 0.064.

This means that the difference between the medians is also significant at this level. Both tests support our finding that, for the given experimental set-up and conditions, there is a significant difference between the mean of the Total scores of students who use GPT and those who do not use GPT. Those using ChatGPT score significantly lower on average.

It is quite surprising. One would think that the GPT would be an extra help in solving tasks and therefore GPT users would have a higher score. Further research questions will address this phenomenon, among others. What are the causes and factors behind:

Question 4. How do students who used ChatGPT differ from other students? Maybe they are the ones who studied less for the test, and this is reflected in the results?

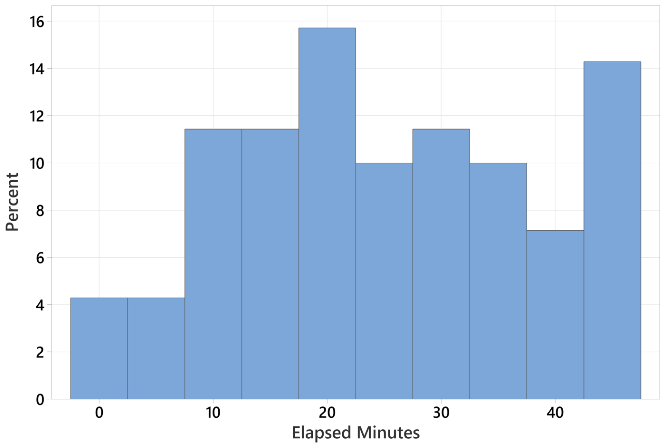
Question 5. Even if they studied less, why didn't the use of ChatGPT make up for it? How correct are the answers given by ChatGPTs?

Question 6. How consistent are the answers given by ChatGPTs? Do all GPTs give the same answer to the test questions in all cases?

Comparison of time used by GPT users and non-users

Figure 3. shows the distribution in minutes of the time students spent solving the test. The average number of minutes was 24.64 with a standard deviation of 12.9 minutes.

Figure 3. The distribution of elapsed minutes

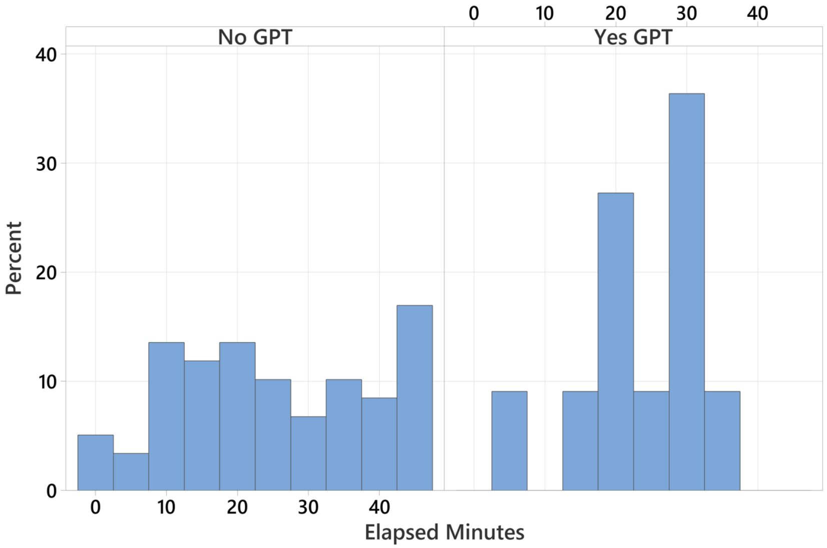


This average value, this distribution, suggests that the test was not difficult in this sense, with only a few students taking the maximum 45 minutes allowed. About half of the students used between 15 and 35 minutes.

Figure 4. shows the relative frequency histogram of elapsed minutes for students not using ChatGPT and students using ChatGPT. The mean of elapsed minutes of students not using ChatGPT was 24.88 with standard deviation 13.6 and mean 23.36 standard deviation 8.18 of students using ChatGPT. The average times used by the two groups are very close. Although the distribution of the data shows some differences in the two samples, the normality test for both samples confirmed that they could be from a normally distributed population, so there is reason to believe that these average times would not be significantly different in the populations.

We performed a two-sample t-test to prove this. For the null hypothesis $\mu_1-\mu_2=0$, and the alternative hypothesis $\mu_1-\mu_2\neq0$ we obtained the observed t-value of 0.50 and the P-value of 0.622. These values support our finding that the average durations used by the two groups are not significantly different.

Figure 4. The distribution of elapsed minutes for students not using ChatGPT and for those using ChatGPT



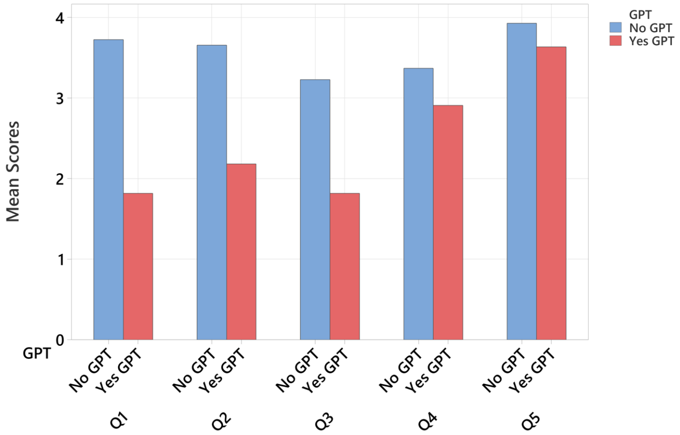
This may also be considered surprising. On the one hand, we might have thought that those who use GPT would finish sooner. On the other hand, we might have thought that someone who uses GPT would get away with comparing the GPT answer with their own answer. This is not the case.

Comparison of scores for non-GPT users and GPT users for questions on different topics

Figure 5. shows the mean scores of the questions on different topics for those who do not use ChatGPT and those who use GPT. It seems that those without GPT have a more balanced performance on different questions, on different topics. For each question, the average score obtained was between 3.2 and 3.9 points.

The same cannot be said for those using GPT, where the results were quite extreme. Question 1, which included calculating the inverse of matrices, had the lowest score of 1.8, while the highest score of 3.6 was for question 5 on complex numbers. It is particularly surprising that their average scores for the first and fourth questions are so different (1.8 and 2.9): Both questions were on the same topic, different questions from the same question bank. If both questions were based on ChatGPT answers, the question arises again whether ChatGPT gives the correct answer in all cases.

Figure 5. The average scores for questions on different topics for those who do not use ChatGPT and those who use GPT



An analysis of variance (ANOVA) was carried out to examine the significance of differences in the responses to the different topics. A two-factor interaction model was used. The „Question" factor has 5 levels (Q1; Q2; Q3; Q4; Q5) and the „GPT Use" factor has two levels (Yes GPT; No GPT), so that in total 5x2=10 factor level combinations were used to examine the differences in average response scores. Since not only the two factors but also the interaction term was found to be significant in the model, it is worth comparing the average scores for the 10 factor level combinations.

Table 1. shows the grouping using the Tukey Method and 95% confidence.

Table 1. Comparison of the average scores for the different questions

	Question*GPT Use	Mean	Grouping		
1.	*Q5 No GPT*	*3.92*	*A*		
2.	*Q1 No GPT*	*3.72*	*A*		
3.	*Q2 No GPT*	*3.65*	*A*		
4.	*Q5 Yes GPT*	*3.63*	*A*	*B*	
5.	*Q4 No GPT*	*3.36*	*A*	*B*	
6.	*Q3 No GPT*	*3.22*	*A*	*B*	
7.	*Q4 Yes GPT*	*2.90*	*A*	*B*	*C*
8.	*Q2 Yes GPT*	*2.18*		*B*	*C*
9.	*Q3 Yes GPT*	*1.81*			*C*
10.	*Q1 Yes GPT*	*1.81*			*C*

According to this notation, if we want to compare two factor level combinations, those that have at least one grouping symbol (A, B, ...) that appears in both are not significantly different. Thus, there is no significant difference between the cases in the first 7 rows of the table, because the group A symbol appears in each case. The case in row 8 (Q2-Yes GPT) is different from the cases in the first three rows. The cases in rows 9 and 10 are significantly different from the cases in the first six rows. This clustering also shows that, for those using GPT, the mean scores for questions 1 and 4 are significantly different, even though these questions are from the same topic and the same question bank. The rows in the table are arranged in descending order of average scores. Question 5 on complex numbers was among the easy questions for both GPT users and non-users. Among the GPT users, the lowest scores were for questions 1 and 3, i.e. the inverse of matrices and different types of multiplication of vectors were the most challenging.

Student learning activities

The students took the test in Moodle, where their learning activity was tracked. By activity, we mean the number of clicks made by the student. So, we could measure the number of times a student clicked on the file containing the course material or completed the practice test.

Figure 6. shows student activity by received scores. The first column shows that the average activity of students with a score of 0 was close to 50 clicks. In addition, the last column shows that the average activity of students with a score of 20 was close to 270 clicks. Obviously, we would expect that students with higher scores would have been more engaged in learning the course material.

Figure 6. The mean number of activities per students vs. Total scores

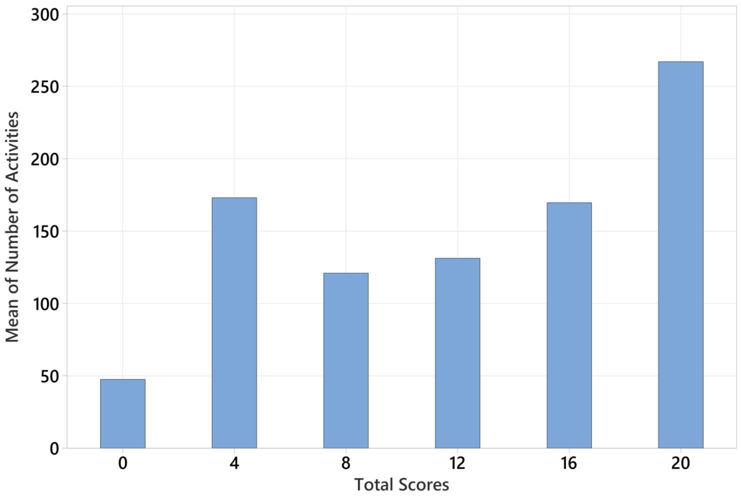


Figure 7. The mean number of activities per student vs. GPT usage

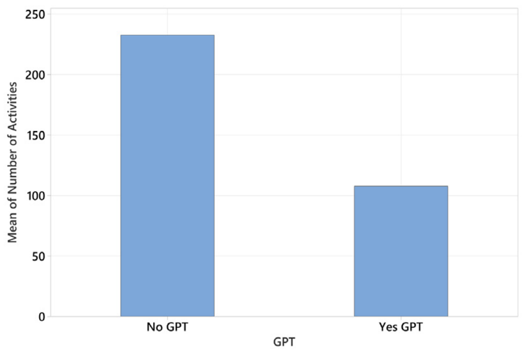


Figure 7. presents a comparison of the activity levels of students using GPT and those not using it. To determine whether there is a statistically significant difference between their activities, we conducted a two-sample two-sided t-test.

Let μ_1 be the population mean of clicks of students not using GPT and μ_2 be the population mean of clicks of students using GPT. Let the null hypothesis be-. $\mu_1-\mu_2=0$ and the alternative hypothesis be $\mu_1-\mu_2\neq0$. Now the observed t-value is 2.33, and the P-value is 0.023. This means that the difference between the mean click counts is statistically significant.

Figure 8. shows the activity of students using GPT and students not using GPT by scores received. The average number of clicks for students not using GPT was extremely high at 4 scores. The activity of students not using GPT is generally higher. We can speculate that the less prepared students used a chatbot.

Figure 8. The mean number of activities per student vs. Total scores and GPT usage

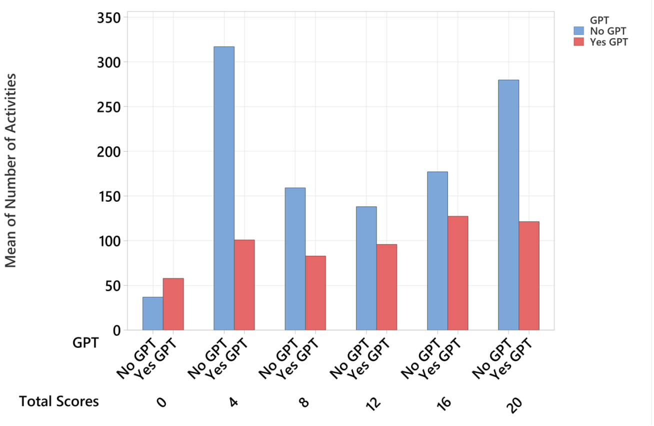
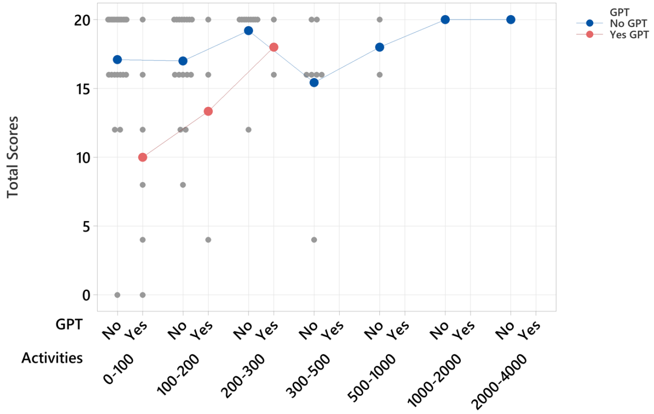


Figure 9. shows the individual value plot of the total scores and the trends. Again, it is obvious that the student who has practiced more will get a higher score.

This can be seen in the trend of the students who use GPT and in the trend of the students who do not use GPT.

Figure 9. The Total scores vs. the mean of the number of activities and GPT usage



Conclusions

In this study examining the impact of using ChatGPT in a mid-year math exam, several noteworthy findings emerged, shedding light on the complexities and nuances associated with integrating AI tools into educational settings.

Performance Discrepancy:

Contrary to expectations, students using ChatGPT demonstrated significantly lower average total scores compared to their non-GPT counterparts. The distribution of scores among ChatGPT users exhibited fewer high achievers, and a notable proportion failed to reach the minimum performance threshold.

Time Usage:

Surprisingly, the average time taken by students using ChatGPT and those who did not, was not significantly different. The assumption that ChatGPT users might finish sooner or engage in comparison with GPT-generated answers was not supported by the data.

Student Learning Activities:

Student activity, measured by the number of clicks, revealed a positive correlation between engagement and higher scores. Notably, students who scored lower were generally less active, indicating a potential link between engagement in learning activities and academic performance. Despite the lower learning activity among ChatGPT users, the question still remains as to why both ChatGPT and Copilot failed to compensate for this lack of engagement.