Cognitive Offloading in the Classroom: Evidence from Written and Oral Statistics Exams Before and After ChatGPT

Marija Ivanković

Zagreb School of Economics and Management, Croatia, mivankov1@zsem.hr

Abstract

Since the introduction of OpenAI's ChatGPT to the masses in late 2022, its effect on education and learning has been both debated and researched. Quantitative subjects, like mathematics and statistics, have traditionally relied on written tests and task-solving to examine knowledge, which makes them particularly exposed to a tool capable of solving such tasks in seconds. This study analyses the effect of ChatGPT on learning by examining written and oral exam results from the Statistics 1 course at the Zagreb School of Economics and Management, across eight academic years (2017/18–2025/26, excluding 2021/22 due to missing data), split into pre-ChatGPT (2017/18–2020/21) and post-ChatGPT (2022/23–2025/26) periods. Using the oral exam as a comparison condition less exposed to direct AI assistance, a difference-in-differences design was applied to isolate a ChatGPT-specific effect from general trends and two identified controls. Results show that written exam performance declined relative to oral exam performance following ChatGPT's emergence (did = −0.211, $p < 0.001$). Event-study estimates show this decline as a sharp shift concentrated at the point of ChatGPT's release rather than a gradual pre-existing trend. The pattern is consistent with a cognitive offloading mechanism, whereby AI-assisted preparation produces a shallower understanding that becomes visible specifically once the tool is no longer available, though individual-level AI usage was not directly observed.

Keywords: *ChatGPT; difference-in-differences; oral examination; cognitive offloading; academic integrity; quantitative assessment*

ISSN 2991-9371

1. Introduction

The rapid development of AI in recent years has both been a blessing and a curse. On one hand, research has shown positive sides of using AI in education.[For example, ChatGPT can enhance learning approaches, like flipped learning where students prepare for lessons by examining materials before classes, where AI serves as a virtual tutor in assisting students with independent studies by answering their questions [1], [2]. On the other hand, faculty has faced many challenges. First one relates to the accuracy and reliability of these kinds of Large Language Models (LLMs) as they are trained on large data sets and has been proved to be biased [3] and provided incorrect information [4], [5]. This fact can lead to less and lower quality learning. But also, AI-powered tools like ChatGPT can answer complex questions correct, even providing explanations. With the increasing availability of these tools, students may be tempted to use them during exams or assignments to bypass traditional methods of studying and problem-solving [6]. This undermined the basic principles of academic integrity and goals, which are honesty, knowledge and learning.

As a results of this, educational institutions have been forced to reengineer existing methods of examinations, shifting more to in-person tests and oral exams. These formats are more resistant to AI use as they are conducted under direct supervision of proctors as students have no access to external devices or influences. Oral exams serve as an additional safe switch as students are forced to articulate and defend their reasoning in real-time, which makes shallow level or memorized answers easier to detect. This still however does not guarantee AI free results. Exam scores have traditionally been used as a tool of assessment, but the validity of those scores is tested in multiple ways. First way, while less likely, is still a direct threat as the possibility of AI assistance during the assignments itself is possible even despite the prohibitions. The second is more difficult to detect but has long-term lasting consequences, is that the tests fail to reflect genuine learning if AI tools are used heavily during preparation. This phenomenon is often called cognitive offloading [7] which can reduce the depth of engagement that consequently leads to shallower conceptual understanding that is visible only when the tool is no longer assessable, like in the proctored in-person exams or oral exams. Hence, the assumption is that this shallow understanding will be reflected in some way in the student's results. To test this, written and oral exam results from 340 Statistics 1 students at ZSEM (2017/18–2025/26, excluding 2021/22) were split into pre- and post-ChatGPT periods and compared using distributional and difference-in-differences analysis, with the oral exam serving as a comparison condition less exposed to direct AI assistance. The following section reviews the literature underpinning this design, covering both the effects of AI tools on learning and the role of oral versus written assessment in a post-ChatGPT context.

ISSN 2991-9371

2. Literature

2.1 AI, Learning and Cognitive Offloading

As mentioned previously, the introduction of AI into various spheres of our existence, including education has polarizing results. On one hand, ChatGPT can serve as a tutor, as research shows that ChatGPT assisted learning produced statistically significant learning gains compared to a no-help [8], which is specifically good for low-income students that can't afford real human tutors. Additionally, ChatGPT produced higher learning gains than human tutor authored hints, however the differences were not statistically significantly separable, and that there were no differences between human and AI tutorship [8]. Several other studies have shown similar results, as ChatGPT has caused moderately positive impact on students' academic achievement [9]. Furthermore, ChatGPT's responses were described as highly concordant, such that a human learner could easily follow the internal language, logic, and directionality of relationships contained within the explanation text [10].

On the other hand, one study examined the midterm exams of mathematics and engineering students on the University Dunaújváros, in Hungary. The results show that students that used ChatGPT had significantly lower average total scores compared to non-ChatGPT counterparts [11]. Additionally, students that used ChatGPT also showed lower engagement levels in online learning activities. Another study reveals also showed mixed results as ChatGPT had a positive impact on exam scores, but it did not improve assignment performance significantly [12]. Additionally, when assignments are customized to students, a performance gap arises as the non-AI group outperformed experimental group, which suggests that tailored, context-rich tasks are effective in minimizing AI-assisted cheating [12]. While ChatGPT supported performance in well-structured tasks, its utility was limited when the student needed to apply flexible reasoning or interpret problems independently, emphasizing the decline in critical thinking abilities.

These finding by itself is not as interesting as the reason behind why this occurs. Many researchers have examined the impact of AI tools on students' cognitive engagement during academic writing tasks and concluded that AI tools encourage students to become "lazy thinkers" [13]. Furthermore, students that did not use AI were more likely to report engaging in deep processing, reflecting a stronger commitment to understanding the task beyond surface-level comprehension. They also indicated exerting greater mental effort, suggesting higher levels of self-regulated thinking [13]. Absence of AI tools appeared support engagement, with participants demonstrating greater strategic flexibility by exploring multiple approaches to the problem. These findings suggest that, rather than enhancing mental effort, AI tools like ChatGPT may reduce learners' cognitive engagement when thoughtfully integrated into academic tasks [13]. The phenomenon described as cognitive offloading, i.e. the use of physical action to alter the information processing requirements of a task to reduce cognitive demand [7]. For this study, it can be interpreted as "outsourcing thinking" to AI to reduce the mental strain. Cognitive offloading can be beneficial in some cases, for example, individuals with impaired unaided cognitive ability [7], which includes memory and focus problems. However, cognitive offloading in education is beneficial only if the cognitive thinking is allocated productively and for a good purpose [7]. This is a clear warning on the allocation of thinking. While relief of learning can be

alleviated, the risk of outsourcing thinking completely to AI carries dangerous consequences. This is supported by [14] research that showed that people who do not use any form of AI exhibited the strongest, widest ranging neural networks, i.e. highest value of cognitive engagement and load, while LLM group showed the lowest results. No AI (i.e. Brain group) demonstrated higher memory recall, and re engagement, while the LLM group also fell behind in their ability to quote from the essays they wrote just minutes prior [14]. This and much other similar research point to an urgent problem in education of decreasing learning skills and abilities of critical thinking which is highly important, especially in quantitative subjects like mathematics and statistics.

2.2 Academic evaluation

Findings described above prove even further that educational institutions need to shift the way in which the students are both educated and examined. Many institutions have already made those changes by increasing the importance of oral exams. Oral examination offers a plethora of benefits. For instance, there is evidence that this dialogic form of assessment may help students express their relational understanding, especially in STEM subjects [15], [16]. In addition, in the right circumstances (for example introducing peer questioning, increased intrinsic motivation and familiarity with the structure) helps to adopt deep learning approaches [17]. Oral assessment can also mitigate the risk of academic integrity as it has this capacity, often missing in other forms of standardised or paper-based assessments that motivate students to engage in academic misconduct. [18], [19]. Students have also expressed the support for this kind of examination as this assessment method was better at bringing the material together and was also perceived as a learning experience, additionally students perceived oral examination to help more with conceptual understanding, written exams were predominantly perceived to assess computational methods. [20].

The importance of written exams is not to be forgotten as enunciation of complex abstract ideas might be easier in writing, especially for mathematics and statistics courses [21]. To conclude, oral assessments specifically for developing and measuring reflective skills, whilst abstract conceptual thinking should be assessed using traditional written formats [21]. Researchers also agree that oral exams tend to cause more anxiety, while written test are not as suitable for students with learning impairments, like dyslexia or dysgraphia [20], [21]. Hence why this research uses both written and oral examinations for the data analysis.

2.3 Hypothesis and research questions

Based on the conducted literature review and available data, the main research questions arise:

> *RQ1: Did the emergence of ChatGPT in November 2022 influence students' performance of proctored written exams (midterms) and oral exams in the same course Statistics 1?*
> *RQ2: Is the pattern of change consistent with a cognitive offloading mechanism, i.e. a decline concentrated in assessment formats without real-time AI access?*

Since the analysis for the research is quantitative in nature, a hypothesis needs to be stated and tested. Similarly to the research question, from the literature the hypothesis is drawn as:

> *H0: Introduction of ChatGPT did not change the performance on proctored written exams and oral exams.*
> *H1: Introduction of ChatGPT decreased the performance on proctored written exams and oral exams.*

3. Methodology

3.1 Data collection and research design

The data set consists of students' results from the Statistics 1 course at Zagreb School of Economics and Management (ZSEM) covering eight academic years (2017/18-2020/21, 2022/23-2025/26), with academic year 2021/22 excluded due to missing records. Three grade assessment components were collected for each student, two written midterm exams (MT1, MT2) and one oral examination (OE). The final grade is calculated based on the performance on these three components. All scores, both on written and oral examinations, were provided pre-scaled [0,1] interval. A small number of values exceeded 1.0 due to bonus points awarded during grading; these were capped at 1.0 before analysis. Since ChatGPT became widely available to the public in November 2022, and since Statistics 1 is taught in the summer semester (February–June), the sample was split at the academic-year level into a pre-ChatGPT period (2017/18–2020/21) and a post-ChatGPT period (2022/23–2025/26). This means the entire 2022/23 academic year already falls after ChatGPT's public release. Each period covers four academic years, which allows for a balanced comparison.

An important issue with this design is that the oral exam, used here as a comparison condition less exposed to direct AI assistance since it happens live and under supervision, was not a fully stable benchmark throughout the sample period. Two things changed alongside, independently from the emergence of ChatGPT. First, the weight of the oral exam in the final grade increased from 10% (2017/18–2023/24) to 20% (2024/25–2025/26). Second, the examiner responsible for the oral exam changed during the 2022/23 and 2023/24 academic years. Both changes could plausibly affect oral exam outcomes on their own, regardless of any AI-related effect, so both were built directly into the statistical models (Section 3.2) as control variables. This follows standard practice in quasi-experimental education research, where concurrent confounding events are modelled explicitly rather than assuming a clean natural experiment [22].

3.2 Statistical Analysis

A large share of students scored exactly the minimum (0) or maximum (1) value on their exams, which violates the assumptions of standard continuous distributions like the Normal distribution. This pattern was pronounced in the data: for instance, up to 41% of oral exam scores in a single year were exactly 1.0, and a non-trivial share of students across all three exam types scored exactly 0, likely reflecting non-attendance rather than an actual performance outcome. Instead of excluding these boundary cases, which would throw away meaningful information about non-participation and ceiling performance, scores were modelled using a zero-and-one-inflated Beta (ZOIB) distribution, following the general framework of [23], [24]. This approach has previously been applied to model non-selective university entrance exam scores [25], and to model bounded introductory-statistics exam scores containing a large share of perfect scores [26], which is more related to this research. For each combination of academic year and exam type, the ZOIB model was fitted via maximum likelihood and compared against a set of standard bounded and unbounded distributions using the Akaike Information Criterion (AIC) [27], to confirm the inflated model fit better than conventional alternatives before drawing any conclusions from it.

To test whether the shape of the score distribution differed between the pre- and post-ChatGPT periods, three tests were applied to each exam type separately, each answering a slightly different question. The two-sample Kolmogorov-Smirnov is used to compare the full empirical distribution of two samples as it is sensitive to any kind of distributional difference [28], as the ChatGPT-related effect could show up as a shift in the mean, in the variance, or in the share of extreme (0 or 1) scores, but not necessarily all three at once. Since this test does not specify the differences, the Mann-Whitney U test was additionally used to test directly whether the central tendency (median) differed between the pre- and post-ChatGPT periods. This test was chosen over the more common independent t-test because it does not assume normally distributed data, which fits the boundary-inflated, non-normal shape of the scores already confirmed by the ZOIB fitting step [29]. Finally, since a change in variance is itself a meaningful outcome and not just a nuisance to control for (for example, a narrowing spread of results could indicate that AI use is compressing the gap between weaker and stronger students), the Levene's test was used to test for equality of variances between the two periods, again without requiring the normality assumption an F-test would need [30].

To isolate a change attributable to the emergence of ChatGPT from a general secular trend affecting all exam types equally, a difference-in-differences (DiD) design was applied [22], [31] treating the two written midterms as the exposed condition and the oral exam as the comparison condition. The model used can be expressed like:

$$score = \beta_0 + \beta_1(treated) + \beta_2(post) + \beta_3(treated \times post) + \beta_4(oral_weight) + \beta_5(oral_prof_change) + \varepsilon \quad (1)$$

where *treated* indicates a written exam (MT1 or MT2), *post* indicates the post-ChatGPT period, *oral_weight* is the oral exam's weight in the final grade for that academic year, and *oral_prof_change* is an indicator for the 2022/23–2023/24 academic years during which the oral examiner changed. The coefficient of interest, β_3 *(the treated × post interaction)*, captures the differential change in written-exam performance relative to oral-exam performance following ChatGPT's

emergence, net of the general post-period trend and both identified control variables (*oral_weight* and *oral_prof_change*). Standard errors were clustered at the student level to account for the non-independence of a given student's three exam scores. The same specification was re-estimated with a binary zero/one outcome (logit link) to test whether the probability of a failing (0) or maximal (1) score changed differentially between exam types.

Full academic-year fixed effects were deliberately excluded from this specification, as *post* and *oral_weight* are themselves determined by academic year and would therefore be collinear with a saturated year-fixed-effects term. As a complementary, fully saturated robustness check, an event-study specification was performed. The score is regressed on variable *treated*, a full set of academic-year dummies, and their interaction was estimated to trace the *treated–control* gap separately for each academic year, without imposing a single pre/post breakpoint. This allows visual and statistical confirmation that any estimated effect coincides with the timing of ChatGPT's emergence rather than reflecting a pre-existing trend. As an additional robustness check, the core DiD model was re-estimated excluding the 2022/23–2023/24 academic years entirely, to confirm that the main result does not depend on the years affected by the oral examiner change. All statistical analyses were conducted in Python.

4. Findings

4.1 Distributional Fit

For each combination of academic year and exam type, the ZOIB model was compared against a set of standard distributions using AIC. The standard Beta distribution outperformed the ZOIB model in every single year-by-exam combination. This does not mean the boundary concentration described in Section 3.2 was absent from the data, but rather that the Beta distribution with shape parameters below one is itself able to produce a U-shaped density that piles up probability mass near both 0 and 1 without needing a separate inflation term. Since the Beta distribution does this with two parameters instead of the four required by ZOIB it was penalized less under AIC and won the comparison despite the extra flexibility of the inflated model. This is a useful finding as it suggests the boundary concentration in these results is closer to a continuous U-shape (many students clustering near, but not necessarily exactly at, the extremes) than to a distinct subgroup of true non-participants mixed with a separate performing group.

Academic Year	Exam	n	Mean	Variance	π_0	π_1
17/18	MT1	22	0.552	0.115	0.182	0.000
18/19	MT1	32	0.712	0.095	0.125	0.062
19/20	MT1	41	0.686	0.120	0.171	0.049
20/21	MT1	21	0.774	0.035	0.000	0.048
22/23	MT1	48	0.527	0.121	0.208	0.021
23/24	MT1	53	0.325	0.091	0.245	0.019
24/25	MT1	57	0.506	0.125	0.246	0.000
25/26	MT1	70	0.486	0.087	0.171	0.000
17/18	MT2	22	0.485	0.105	0.182	0.000
18/19	MT2	32	0.623	0.110	0.156	0.125
19/20	MT2	41	0.533	0.119	0.244	0.024
20/21	MT2	21	0.643	0.079	0.048	0.048
22/23	MT2	48	0.455	0.100	0.208	0.042
23/24	MT2	53	0.335	0.079	0.264	0.000
24/25	MT2	57	0.377	0.088	0.263	0.018
25/26	MT2	70	0.363	0.069	0.243	0.000
17/18	OE	22	0.332	0.117	0.409	0.045
18/19	OE	32	0.703	0.140	0.188	0.406
19/20	OE	41	0.572	0.189	0.341	0.268
20/21	OE	19	0.795	0.051	0.000	0.368
22/23	OE	48	0.842	0.125	0.146	0.812
23/24	OE	53	0.811	0.153	0.189	0.811
24/25	OE	57	0.509	0.167	0.351	0.211
25/26	OE	70	0.686	0.121	0.186	0.214

Note: π_0 and π_1 denote the empirical proportion of students scoring exactly 0 and exactly 1, respectively. Across all year–exam combinations, a standard Beta distribution was preferred over the zero-and-one-inflated Beta (ZOIB) specification by AIC.

That said, the raw proportions of exact 0 and exact 1 scores (π_0 and π_1) are still informative descriptively. For MT1 and MT2, π_0 (proportion scoring exactly zero) tends to increase in the post-ChatGPT years — for instance, MT1 goes from 0% (2020/21) up to around 21–25% in most post-ChatGPT years. The oral exam shows a very different pattern: π_1 (proportion scoring exactly the maximum) jumps sharply to 81% in both 2022/23 and 2023/24, before dropping back down to around 21% in 2024/25–2025/26. As discussed in Section 3.1, this spike coincides with, and is most plausibly explained by, the change in oral examiner during those two years rather than by ChatGPT. Figure 1 visualizes these trajectories across the full sample period.

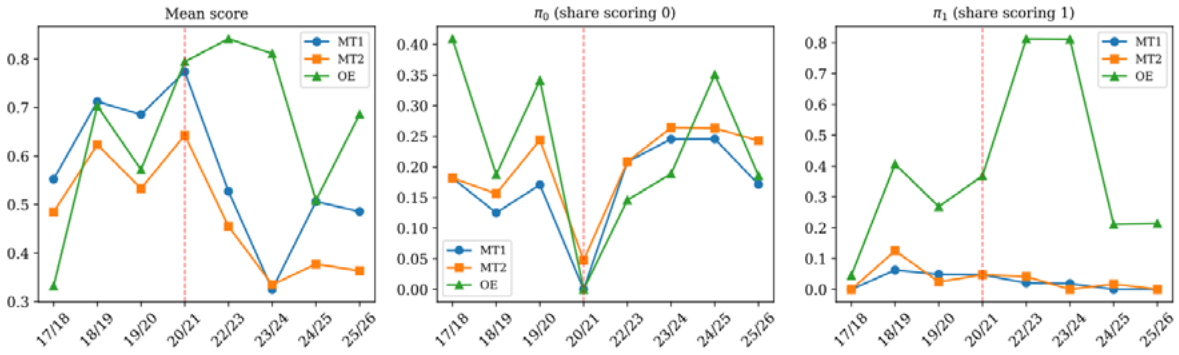


Figure 1. Trajectories of parameters

4.2 Pre and Post Chat GPT Comparison

Table 2 summarizes the distributional comparison between the pre- and post-ChatGPT periods, separately for each exam type. For both written midterms, the mean score dropped considerably: MT1 fell from 0.684 (pre) to 0.462 (post), and MT2 fell from 0.569 to 0.379. Both drops are highly statistically significant across every test applied — the Kolmogorov-Smirnov test (MT1: $p < 0.001$; MT2: $p < 0.001$), the Mann-Whitney U test (MT1: $p < 0.001$; MT2: $p < 0.001$), and the Welch t-test (MT1: $p < 0.001$; MT2: $p < 0.001$). Levene's test shows a significant increase in variance for MT1 ($p = 0.014$, variance rising from 0.103 to 0.111), while the variance change for MT2 was not significant ($p = 0.277$).

Exam	Mean (pre)	Mean (post)	Var (pre)	Var (post)	KS p	MWU p	Levene p
MT1	0.684	0.462	0.103	0.111	6.01e-10***	3.81e-10***	0.0135*
MT2	0.569	0.379	0.111	0.085	4.34e-09***	4.39e-08***	0.277
OE	0.600	0.704	0.163	0.158	0.00384**	0.00352**	0.253

Note: KS = two-sample Kolmogorov–Smirnov test; MWU = Mann–Whitney U test. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table 2. Comparison of pre- and post-ChatGPT periods, by exam type

The oral exam moves in the opposite direction: the mean actually increased, from 0.600 (pre) to 0.704 (post), and this shift is also statistically significant (KS: $p = 0.004$; Mann-Whitney: $p < 0.001$; Welch t-test: $p = 0.025$), though the variance did not change significantly (Levene: $p = 0.253$). Given what is already known about the examiner change in 2022/23–2023/24, this increase in oral exam performance should not, on its own, be read as evidence against a ChatGPT effect as it can be driven by the same confound the DiD design was built to address. Looking at the boundary proportions specifically, the change in π_1 (proportion scoring the maximum) on the oral exam is highly significant (χ^2 test, $p < 0.001$), which lines up with the spike already noted in Section 4.1. The change in π_0 (proportion scoring zero) is not significant for any of the three exam types at the 5% level, though it is close to significance for MT1 ($p = 0.075$).

4.3 Difference-in-Differences (DiD) Results

The core DiD regression (Table 3) gives a treated × post coefficient of −0.211 ($p < 0.001$). This is the key result of the analysis **and it points that the** written exam performance dropped by an additional 0.211 points (on the [0,1] scale), relative to the oral exam, after the emergence of ChatGPT (over and above whatever general shift happened across all exam types in the same period, and after controlling for both the oral exam's changing weight in the final grade and the change in oral examiner). Neither the treated term (0.027, $p = 0.313$) nor the post term (0.003, $p = 0.958$) is significant on its own, **meaning** there was no meaningful baseline difference between written and oral exams before ChatGPT, and no general post-period shift affecting all exam types equally. The effect is specific to the interaction, exactly as the difference-in-differences design is meant to isolate. The *oral_prof_change* control is itself highly significant (0.221, $p <$

0.001), confirming that the examiner change had a real, independent effect on oral exam scores, separate from anything related to ChatGPT.

	Score (OLS)	$P(score = 0)$ (Logit)	$P(score = 1)$ (Logit)
Intercept	0.592***	−0.773*	−0.729
	(0.059)	(0.304)	(0.428)
treated	0.027	−0.653*	−2.059***
	(0.026)	(0.283)	(0.373)
post	0.003	0.190	−0.255
	(0.057)	(0.316)	(0.359)
did (treated × post)	−0.211***	0.455	−1.161†
	(0.038)	(0.366)	(0.625)
oral_weight	0.072	−3.025	−2.117
	(0.457)	(2.154)	(3.742)
oral_prof_change	0.221***	−0.635†	2.715***
	(0.052)	(0.339)	(0.348)
N	1030	1030	1030

Note: Standard errors in parentheses; clustered at the student level for the OLS model.
†$p < 0.10$, *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

Table 3. Difference-in-differences regression results

The same specification estimated on the probability of a zero score shows a positive but non-significant did coefficient (0.455, p = 0.215). This means that once the professor **changes** and other controls are accounted for, there is no strong evidence that complete failure became relatively more likely on written exams after ChatGPT. The equivalent model for the probability of a perfect score shows a negative, marginally significant did coefficient (−1.161, p = 0.063). Perfect scores became somewhat less likely on written exams relative to oral, after ChatGPT, though this result should be treated with caution given the p-value sits just above the conventional 5% threshold.

4.4 Robustness Checks

Excluding the two years affected by the oral examiner change (2022/23 and 2023/24) entirely and re-estimating the core model gives a *treated* × *post coefficient* of −0.201 (p < 0.001) — nearly identical to the −0.211 obtained in the full model. This means the main finding does not depend on those two disputed years and is not a result of the examiner change. The event-study specification, which estimates the treated effect (written minus oral) separately for every academic year instead of collapsing everything into a single pre/post-split, adds further support. In the pre-ChatGPT years, the *treated* effect is moderate in size: −0.222 (2018/19), −0.149 (2019/20), and −0.273 (2020/21), all statistically significant.

	Score (OLS)
Intercept	0.552***
	(0.066)
treated	0.027
	(0.026)
post	−0.019
	(0.062)
did (treated × post)	−0.201***
	(0.033)
oral_weight	0.476
	(0.543)
N	727

Note: Standard errors (clustered at student level) in parentheses. ***$p < 0.001$. The 2022/23–2023/24 academic years, affected by the oral examiner change, are excluded entirely. The *did* coefficient (−0.201) is nearly identical to the full-sample estimate (−0.211 in Table 3), indicating the main finding does not depend on these two years.

Table 4. Robutness check: core model excluding 2022/23 2023/24

Immediately after ChatGPT's emergence, the gap widens sharply: −0.537 (2022/23) and −0.668 (2023/24). It narrows somewhat in the following years but never returns to pre-ChatGPT levels: −0.254 (2024/25) and −0.448 (2025/26). This pattern of a stable gap before the emergence of ChatGPT, followed by a sudden widening exactly at the point of transition is consistent with a level shift tied to ChatGPT's emergence, rather than a gradual pre-existing trend that would have shown up gradually across the earlier years as well.

As a further robustness check, an event-study specification was estimated, tracing the written-versus-oral exam gap separately for each academic year rather than collapsing all post-ChatGPT years into a single average effect. Figure 2 plots the resulting coefficients, each representing the difference between written and oral exam performance each year, relative to the 2017/18 baseline. In the years preceding ChatGPT's emergence, this gap is relatively stable, ranging from -0.15to -0.27, with overlapping confidence intervals across all three years. Immediately following ChatGPT's public release, the gap widens sharply to -0.54(2022/23) and -0.67(2023/24), clearly outside the range observed in any pre-ChatGPT year. In the two most recent academic years, the gap narrows somewhat (-0.25 and -0.45, respectively) but does not return to its pre-ChatGPT level. This pattern is very informative. A gradual, pre-existing trend unrelated to ChatGPT would be expected to produce a steadily changing gap across the entire period, whereas the data instead show a flat trend followed by a sudden shift concentrated exactly at the point of transition. This is consistent with a discrete change coinciding with ChatGPT's emergence, rather than a continuation of an existing trajectory. It should be noted that part of the widening in 2022/23–2023/24 coincides with the previously discussed change in oral examiner during those same years. This is precisely why the model excluding these two years (Table 4.) is an important complement to Figure 2 and its near-identical estimate (-0.201 vs. -0.211) indicates that the overall pattern is not solely an artifact of that control variable.

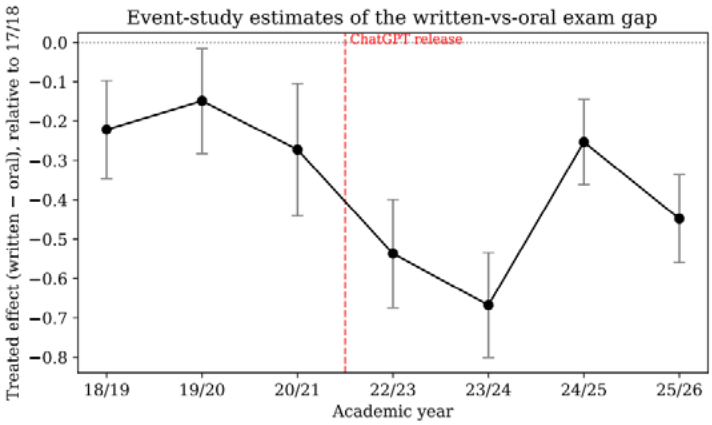


Figure 2. Event-study estimates of the written-vs-oral exam gap

5. Conclusion

This study examined whether the emergence of ChatGPT in November 2022 influenced student performance in Statistics 1 at the Zagreb School of Economics and Management, using written midterm exams and an oral exam as two assessment formats with different levels of exposure to AI-assisted preparation. The results show a change that is not uniform across formats, which directly addresses RQ_1. Written exam performance declined significantly relative to oral exam performance after ChatGPT's emergence (did = −0.211, $p < 0.001$), a result that holds after controlling for two identified control variables (a change in the oral exam's grade weight and a change in oral examiner during 2022/23–2023/24) and remains essentially unchanged when the control years are excluded entirely (did = −0.201). Event-study estimates reinforce this conclusion as the written-versus-oral gap was stable for three years before ChatGPT's release and widened sharply exactly at the point of transition, rather than following a trend already building beforehand. On this basis, H_0 is rejected for the written exams. The oral exam, in contrast, showed an increase in mean score after ChatGPT, but this is best explained by the examiner change rather than any AI-related effect, since the *oral_prof_change* control absorbs a large, significant share of the shift. H_1, which assumed a decrease in both formats, therefore holds only partially. The performance declined on written exams but not on the oral exam once its own control is addressed. This asymmetry is itself the more informative result, since it is exactly this differential pattern, rather than a uniform decline, that the design was built to detect. The pattern also addresses, RQ_2 as the decline is concentrated precisely in the format where real-time AI access during preparation was assumed to matter most, while the oral exam, requiring live, unaided articulation of reasoning, shows no equivalent decline once its effect is accounted for. This is what a cognitive offloading account predicts. A reliance on ChatGPT during preparation can produce a shallower level of understanding that becomes visible specifically once the tool is no longer available, as in a proctored exam [7], [14]. It is consistent with prior findings that ChatGPT-using students scored lower on midterm exams than their peers [11], and that AI assistance can support performance on well-structured tasks while undermining

deeper conceptual engagement [12]. The answer to RQ_2 is therefore a qualified yes. The pattern is consistent with cognitive offloading, but since individual students' actual ChatGPT usage was not measured, this remains a plausible, literature-consistent interpretation rather than a directly tested mechanism.

Several results call for caution rather than overstatement. This is an observational, course-level analysis, not an experiment, so an identical pattern could in principle arise from other unobserved changes coinciding with the same years, such as changes in how the written exams were written or graded, or in the profile of enrolling students. The fact that the effect is specific to the exam-type × period interaction, rather than a general post-period shift, makes such alternatives less likely but does not rule them out. The oral exam itself, intended as a stable benchmark, also proved less stable than assumed, a reminder that even assessment formats chosen for their resistance to AI-related effects require case-by-case verification rather than being assumed stable by design. The study's main limitations follow from its scope as it relies on aggregate exam outcomes rather than individual-level AI usage data, which limits conclusions about the mechanism at the student level. Furthermore, it covers a single course at a single institution, limiting generalizability. Additionally, the exclusion of 2021/22 leaves a gap in the pre-ChatGPT period, though the remaining three years still allow a reasonably balanced comparison. Future work could address the first limitation directly by collecting individual-level usage data, allowing a more direct test of the cognitive offloading mechanism proposed here. Replicating this design across other quantitative courses, institutions, or countries would help establish how general the observed pattern is. Finally, a longer post-ChatGPT observation window would help clarify whether the partial narrowing of the written-oral gap in 2024/25–2025/26 reflects genuine adaptation — for instance, through updated assessment design or explicit AI-literacy instruction — or is simply noise around a persistently elevated gap.

6. Acknowledgements

During the preparation of the paper, the author used Claude for help with generating the exploratory code for data clearing, formatting and analysis. The authors independently verified all outputs and results.

References

[1] M. S. Aslam and S. Nisar, Artificial Intelligence Applications Using ChatGPT in Education: Case Studies and Practices. in Advances in Educational Technologies and Instructional Design. IGI Global, 2023. doi: 10.4018/978-1-6684-9300-7.

[2] C. K. Lo, "What Is the Impact of ChatGPT on Education? A Rapid Review of the Literature," Education Sciences, vol. 13, no. 4, p. 410, Apr. 2023, doi: 10.3390/educsci13040410.

[3] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?, " in Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, Virtual Event Canada: ACM, Mar. 2021, pp. 610–623. doi: 10.1145/3442188.3445922.

[4] F. M. Megahed, Y.-J. Chen, J. A. Ferris, S. Knoth, and L. A. Jones-Farmer, "How generative AI models such as ChatGPT can be (mis)used in SPC practice, education, and research? An exploratory study," Quality Engineering, vol. 36, no. 2, pp. 287–315, Apr. 2024, doi: 10.1080/08982112.2023.2206479.

[5] J. Qadir, "Engineering Education in the Era of ChatGPT: Promise and Pitfalls of Generative AI for Education," in 2023 IEEE Global Engineering Education Conference (EDUCON), Kuwait, Kuwait: IEEE, May 2023, pp. 1–9. doi: 10.1109/EDUCON54358.2023.10125121.

[6] Z. Djebbari, "A New Era of Exam Cheating: The Use of ChatGPT and the Rise of Academic Dishonesty," vol. 7, no. 2, 2025.

[7] E. F. Risko and S. J. Gilbert, "Cognitive Offloading," Trends in Cognitive Sciences, vol. 20, no. 9, pp. 676–688, Sep. 2016, doi: 10.1016/j.tics.2016.07.002.

[8] Z. A. Pardos and S. Bhandari, "ChatGPT-generated help produces learning gains equivalent to human tutor-authored help on mathematics skills," PLoS ONE, vol. 19, no. 5, p. e0304013, May 2024, doi: 10.1371/journal.pone.0304013.

[9] Z. Liu, H. Zuo, and Y. Lu, "The Impact of CHATGPT on Students' Academic Achievement: A Meta-Analysis," Computer Assisted Learning, vol. 41, no. 4, p. e70096, Aug. 2025, doi: 10.1111/jcal.70096.

[10] T. H. Kung et al., "Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models," PLOS Digit Health, vol. 2, no. 2, p. e0000198, Feb. 2023, doi: 10.1371/journal.pdig.0000198.

[11] L. Bognár and A. Joós, "The Impact of Allowing ChatGPT on Student Scores and Test Completion Time in a Mid-Term Math Exam," DK, vol. 13, no. 2, pp. 11–24, Feb. 2025, doi: 10.63684/dk.2025.2.02.

[12] P. Rahman and S. Mehnaz, "Evaluating the Role of ChatGPT in Mathematics Education: Impact on Student Performance, Academic Integrity, and Perceptions in Higher Education," SSRN Journal, 2024, doi: 10.2139/ssrn.5054029.

[13] G. P. Georgiou, "ChatGPT produces more 'lazy' thinkers: Evidence of cognitive engagement decline".

[14] N. Kosmyna et al., "Your Brain on ChatGPT: Accumulation of Cognitive Debt when Using an AI Assistant for Essay Writing Task," 2025, arXiv. doi: 10.48550/ARXIV.2506.08872.

[15] J. Schoultz, R. Säljö, and J. Wyndhamn, "Conceptual knowledge in talk and text: What does it take to understand a science question?," Instructional Science, vol. 29, no. 3, pp. 213–236, May 2001, doi: 10.1023/A:1017586614763.

[16] A. J. Stylianides, "Secondary students' proof constructions in mathematics: The role of written versus oral mode of argument representation," Review of Education, vol. 7, no. 1, pp. 156–182, Feb. 2019, doi: 10.1002/rev3.3157.

[17] P. Iannone, C. Czichowsky, and J. Ruf, "The impact of high stakes oral performance assessment on students' approaches to learning: a case study," Educ Stud Math, vol. 103, no. 3, pp. 313–337, Mar. 2020, doi: 10.1007/s10649-020-09937-4.

[18] D. Kang et al., "Providing an Oral Examination as an Authentic Assessment in a Large Section, Undergraduate Diversity Class," ij-sotl, vol. 13, no. 2, May 2019, doi: 10.20429/ijsotl.2019.130210.

[19] S. Nallaya, S. Gentili, S. Weeks, and K. Baldock, "The validity, reliability, academic integrity and integration of oral assessments in higher education: A systematic review".

[20] P. Iannone and N. Vondrová, "The Novelty Effect on Assessment Interventions: a Qualitative Replication Study of Oral Performance Assessment in Undergraduate Mathematics," Int J of Sci and Math Educ, vol. 22, no. 2, pp. 375–397, Feb. 2024, doi: 10.1007/s10763-023-10368-9.

[21] M. Huxham, F. Campbell, and J. Westwood, "Oral versus written assessments: a test of student performance and attitudes," Assessment & Evaluation in Higher Education, vol. 37, no. 1, pp. 125–136, Feb. 2012, doi: 10.1080/02602938.2010.515012.

[22] D. Corral and M. Yang, "An introduction to the difference-in-differences design in education policy research," Asia Pacific Educ. Rev., vol. 25, no. 3, pp. 663–672, Sep. 2024, doi: 10.1007/s12564-024-09959-0.

[23] R. Ospina and S. L. P. Ferrari, "Inflated beta distributions," Stat Papers, vol. 51, no. 1, pp. 111–126, Jan. 2010, doi: 10.1007/s00362-008-0125-4.

[24] R. Ospina and S. L. P. Ferrari, "A general class of zero-or-one inflated beta regression models," Computational Statistics & Data Analysis, vol. 56, no. 6, pp. 1609–1623, Jun. 2012, doi: 10.1016/j.csda.2011.10.005.

[25] L. Masserini, M. Bini, and M. Pratesi, "Effectiveness of non-selective evaluation test scores for predicting first-year performance in university career: a zero-inflated beta regression approach," Qual Quant, vol. 51, no. 2, pp. 693–708, Mar. 2017, doi: 10.1007/s11135-016-0433-z.

[26] C.-H. Yu and A. Ebeling, "Active Learning with Bayesian Reasoning: A POGIL-Based Pedagogy in Introductory Statistics," 2026, arXiv. doi: 10.48550/ARXIV.2606.08660.

[27] H. Bozdogan, "Model Selection and Akaike's Information Criterion (AIC): The General Theory and its Analytical Extensions," Psychometrika, vol. 52, no. 3, pp. 345–370, Sep. 1987, doi: 10.1007/BF02294361.

[28] N. Smirnov, "Table for Estimating the Goodness of Fit of Empirical Distributions," Ann. Math. Statist., vol. 19, no. 2, pp. 279–281, Jun. 1948, doi: 10.1214/aoms/1177730256.

[29] H. B. Mann and D. R. Whitney, "On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other," Ann. Math. Statist., vol. 18, no. 1, pp. 50–60, Mar. 1947, doi: 10.1214/aoms/1177730491.

[30] H. Levene, "Robust Tests for Equality of Variances," In: Olkin, I., Ed., Contributions to Probability and Statistics, pp. 278–292, 1960.

[31] M. Lechner, "The Estimation of Causal Effects by Difference-in-Difference Methods," Foundations and Trends® in Econometrics, vol. 4, no. 3, pp. 165–224, Nov. 2011, doi: 10.1561/0800000014.