Humanities & Social Sciences Communications

ARTICLE

https://doi.org/10.1057/s41599-023-02269-7 OPEN

Check for updates

ChatGPT-3.5 as writing assistance in students' essays

Željana Bašić[1], Ana Banovac[1]✉, Ivana Kružić[1] & Ivan Jerković[1]

ChatGPT-3.5, an AI language model capable of text generation, translation, summarization, and question-answering, has recently been released for public use. Studies have shown it can generate abstracts, research papers, and dissertations, and create quality essays on different topics. This led to ethical issues in using ChatGPT in academic writing, AI authorship, and evaluating students' essays. However, it is still unknown how ChatGPT performs in students' environments as a writing assistant tool and if it enhances students' essay-writing performance. In the present study, we examined students' essay-writing performances with or without ChatGPT as an essay-writing assistance tool. The average essay grade was C for both control (traditional essay-writing, $n = 9$) and experimental (ChatGPT-assisted essay-writing, $n = 9$) groups. None of the predictors affected essay scores: group, writing duration, study module, and GPA. The text unauthenticity was slightly higher in the experimental group, but the similarity among essays was generally low in the overall sample. In the experimental group, the AI classifier recognized more potential AI-generated texts. Our results demonstrate that the ChatGPT group did not perform better in either of the indicators; the students did not deliver higher quality content, did not write faster, nor had a higher degree of authentic text. We anticipate that these results can relieve some concerns about this tool's usage in academic writing. ChatGPT-assisted writing could depend on the previous knowledge and skills of the user, which might, in certain instances, lead to confusion in inexperienced users and result in poorer essay writing performance.

[1] University Department of Forensic Sciences, University of Split, Ruđera Boškovića 33, 21000 Split, Croatia. ✉email: ana.banovac@forenzika.unist.hr

Introduction

November 30, 2022, will go down in history as the date when a free version of the AI language model created by OpenAI called ChatGPT-3.5 (OpenAI, 2022) (in further text ChatGPT) was made available for public usage. This language model's functions encompass text generation, answering questions, and completing tasks such as translation and summarization (Agomuoh, 2023).

ChatGPT can be employed as assistance in the world of academia. It can improve writing skills since it is trained to deliver feedback on style, coherence, and grammar (Aljanabi et al., 2023), extract key points, and provide citations (Aydin and Karaarslan, 2022). This could increase the efficiency of researchers, allowing them to concentrate on more crucial activities (e.g., analysis and interpretation). This has been supported by studies showing that ChatGPT could generate abstracts (Gao et al., 2023; Ma et al., 2023), high-quality research papers (Kung et al., 2023), dissertations, and essays (Aljanabi et al., 2023). Previous studies showed that ChatGPT could create quality essays on different topics (Hoang, 2023; Hoang et al., 2023; Nguyen and La; 2023; Nguyen and Le, 2023a, Nguyen and Le, 2023b, Susnjak, 2023). For example, this program, in conjunction with DaVinci-003, generated high-quality short-form essays on Physics, which would be awarded First Class, the highest grade in the UK higher education system (Yeadon et al., 2023). It also led to questions on the ethics of using ChatGPT in different forms of academic writing, the AI authorship (Bishop, 2023; Grimaldi and Ehrler, 2023; Kung et al., 2023; Pourhoseingholi et al., 2023; Xiao, 2023), and raised issues of evaluating academic tasks like students' essays (Stokel-Walker, 2022; Whitford, 2022). Unavoidable content plagiarism issues were discussed, and solutions for adapting essay settings and guidelines were revised (Cotton et al., 2023; Hoang, 2023; Lo, 2023; Sallam, 2023; Stokel-Walker, 2022; Yeadon et al., 2023). A recent SWOT analysis of ChatGPT's impact on education comprehensively analyzed all the mentioned issues. Strengths included advanced natural language generation, self-improvement, and personalized feedback, with potential benefits in information accessibility, personalized learning, and reduced teaching workload. Weaknesses encompassed limited understanding of the topic, inability to critically evaluate information, response quality evaluation challenges, bias risks, and a lack of higher-order thinking. Threats included contextual limitations, academic integrity risks, discrimination perpetuation, increased plagiarism, etc. (Farrokhnia et al., 2023).

As argumentative essays are one of the most advanced students' tasks in higher education, and as such pose a challenge for students (Latifi et al., 2021), one of the ways where ChatGPT could be tested is essay writing. Such essays empower students' ability to give an argument and build confidence in their knowledge preparing them not only for the academic environment but also for real-life situations (Valero Haro et al., 2022; Heitmann et al., 2014). A previous study showed that students need further development of argumentation competencies, as they demonstrated externalization issues with argumentation that did not differ if they worked in groups or individually. The results suggest that students experience problems in externalizing their argumentation knowledge both at the individual (argumentative essay) and collaborative levels (argumentative discourse), and that they need to further develop their argumentation competence (Banihashem et al., 2023a; Banihashem et al., 2023b; Kerman et al., 2023; Ranjbaran et al., 2023). However, it is still unknown how ChatGPT performs in students' environment as a writing assistant tool and does it enhance students' performance. Thus, this research investigated whether ChatGPT would improve students' essay grades, reduce writing time, and affect text authenticity.

Materials and methods

We invited the second-year master's students from the University Department of Forensic Sciences, to voluntarily participate in research on essay writing as a part of the course Forensic Sciences seminar. Out of 50 students enrolled in the course, 18 applied by web form and participated in the study. Before the experiment, we divided them into two groups according to the study module and the weighted grade point average (GPA) to ensure a similar composition of the groups. The control group ($n = 9$, GPA = 3.92 ± 0.46) wrote the essay traditionally, while the experimental group ($n = 9$, GPA = 3.92 ± 0.57) used ChatGPT assistance, version 2.1.0. (OpenAI, 2022).

We explained the essay scoring methodology (Schreyer Institute for Teaching Excellence (2023)) to both groups, with written instructions about the essay title (The advantages and disadvantages of biometric identification in forensic sciences), length of the essay (800–1000 words in a Croatian language), formatting, and citation style (Vancouver). We introduced the experimental group to the ChatGPT tool which included a brief explanation of the tool, and an example of entering the prompt about their essay-unrelated issue. They were instructed to use the tool freely, without any limitations (e.g., for creating a complete essay, for concept drafting, for specific topic-related questions, for corrections and suggestions, etc.). We did not demand students to submit the prompts they used and the responses they received. All students had four hours to finish the task and could leave whenever they wanted. The control group was additionally supervised to ensure they did not use the ChatGPT. The students' names were coded to assure the individual and group anonymity and prevent grading bias.

Two teachers graded the essays (ŽB, associate professor, and IJ, assistant professor). The teachers compared the grades, and if their scoring differed the final grade was decided by the consensus. We used the essay rubrics from the Schreyer Institute for Teaching Excellence, Pennsylvania State University (http://www.schreyerinstitute.psu.edu/pdf/suanne_general_resource_WritingRubric.pdf), that included the following criteria (mechanics, style, content, and format) and grades from A to D (Schreyer Institute for Teaching Excellence (2023)). We converted categorical grades to numbers (A = 4, B = 3, C = 2, D = 1) for further analysis. For each student, we recorded writing time.

We checked the authenticity of each document using PlagScan (2022), and conducted the pairwise comparison for document similarity using R studio (ver. 1.2.5033) and package Textreuse (Mullen, 2015) using the Jaccard similarity index. We checked the content using an AI text classifier to test if a human or an AI created the text. According to this classifier, text was scored as very unlikely, unlikely, unclear, possibly, and likely that it was AI-generated (OpenAI, 2023). We opted for this package after similar programs (OpenAI, 2022; Goal and ChatGPT, 2023; Debut et al., 2023) did not recognize a ChatGPT-generated text in a non-English language as AI-assisted text.

Statistical analysis and visualization were conducted using Excel (Microsoft Office ver. 2301) and R Studio (ver. 1.2.5033). The final essay score was calculated as an average of four grading elements (mechanics, style, content, and format). The linear regression was used to test the effects of group, writing duration, module, and GPA on overall essay scores. The level of statistical significance was set at $P \leq 0.05$.

Results

The duration of the essay writing for the ChatGPT-assisted group was 172.22 ± 31.59, and for the control, 179.11 ± 31.93 min. ChatGPT and control group, on average, obtained grade C, with a

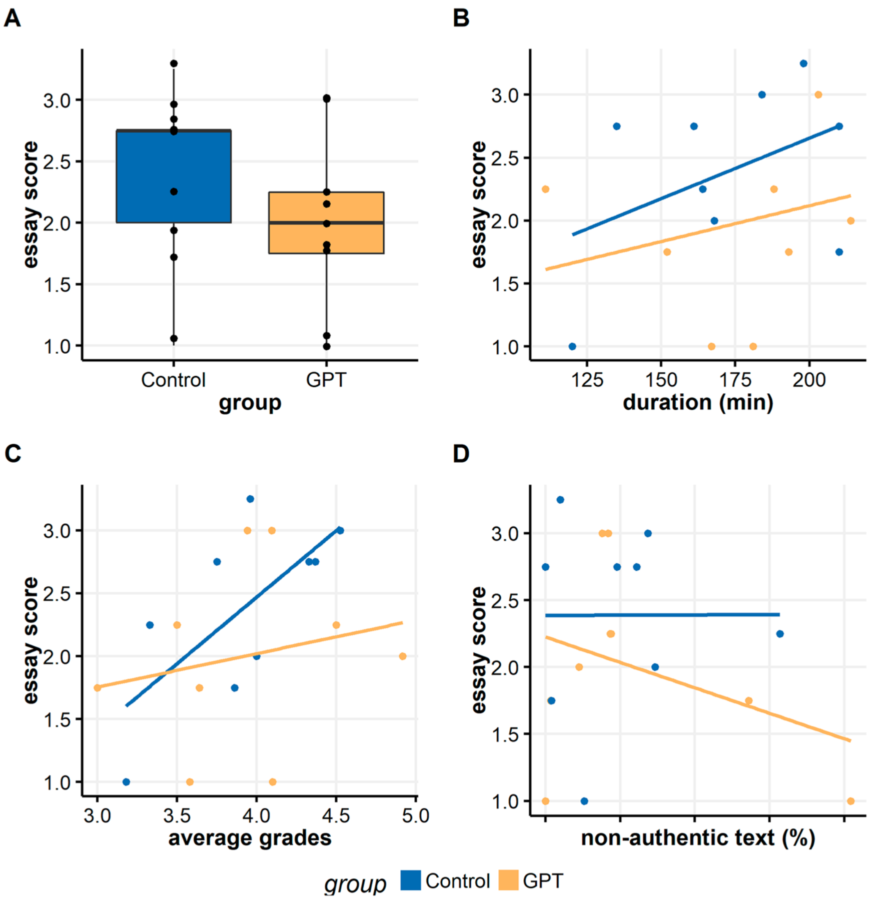

Fig. 1 Essay scores by variable for control and the ChatGPT-assisted group. **A** Average essay scores, **B** Duration and essay scores, **C** GPA and essay scores, **D** Text authenticity and essay scores.

slightly higher average score in the control (2.39 ± 0.71) than in the ChatGPT group (2.00 ± 0.73) (Fig. 1A). The mean of text unauthenticity was 11.87% ± 13.45 in the ChatGPT-assisted group and 9.96% ± 9.81% in the control group. The text similarity in the overall sample was low (Supplementary Table 1), with a median value of the Jaccard similarity index of 0.002 (0–0.054). The AI text classifier showed that, in the control group, two texts were possibly, one likely generated by AI, two were unlikely created by AI, and four cases were unclear. The ChatGPT group had three possible and five cases likely produced by AI, while one case was labeled as unclear.

Figure 1B, C implies a positive association between duration and GPA with essay scores. Students with higher GPAs in the control group achieved higher scores than those in the ChatGPT group. The association of essay scores and non-authentic text proportion (Fig. 1D) was detected only in the ChatGPT group, where the students with more non-authentic text achieved lower essay scores.

The linear regression model showed a moderate positive relationship between the four predictors and the overall essay score ($R = 0.573$; $P = 0.237$). However, none of the predictors had a significant effect on the outcome: group ($P = 0.184$), writing duration ($P = 0.669$), module ($P = 0.388$), and GPA ($P = 0.532$).

Discussion

As we are aware, this is the first study that tested ChatGPT-3.5 as an essay-writing assistance tool in a student population sample. Our study showed that the ChatGPT group did not perform better than the control group in either of the indicators; the students did not deliver higher quality content, did not write faster, nor had a higher degree of authentic text.

The overall essay score was slightly better in the control group, which could probably result from the students in the experimental group over-reliance on the tool or being unfamiliar with it. This was in line with Fyfe's study on writing students' essays using ChatGPT-2, where students reported that it was harder to write using the tool than by themselves (Fyfe, 2022). This issue is presented in the study of Farrokhnia et al., where the authors pointed out the ChatGPT weakness of not having a deep understanding of the topic, which, in conjunction with students' lack of knowledge, could lead to dubious results (Farrokhnia et al., 2023). Students also raised the question of not knowing the sources of generated text which additionally distracted them in writing task (Fyfe, 2022). It is noteworthy that both groups obtained an average grade of C, which can be explained by other studies that argued that students' writing lacks solid argumentation both when writing in general or when writing argumentative essays (Banihashem et al., 2023a; Banihashem et al., 2023b; Kerman et al., 2023; Farrokhnia et al., 2023; Ranjbaran et al., 2023). This demanding task could have been even more difficult when using ChatGPT, which could stem from several already mentioned issues like unfamiliarity when using ChatGPT and additional time requirements to link ChatGPT-created content and/or information with real literature sources.

Some studies did show more promising results (Hoang, 2023; Hoang et al., 2023; Nguyen and La; 2023; Nguyen and Le, 2023a, Nguyen and Le, 2023b, Susnjak, 2023; Yeadon et al., 2023), but unlike our study, they were mainly based on ChatGPT and experienced researcher interaction. This could be a reason for the lower performance of our ChatGPT group, as the experienced researchers are more skilled in formulating questions, guiding the program to obtain better-quality information, and critically evaluating the content.

The other interesting finding is that the use of ChatGPT did not accelerate essay writing and that the students of both groups required a similar amount of time to complete the task. As expected, the longer writing time in both groups related to the better essay score. This finding could also be explained by students' feedback from Fyfe's (2022) study, where they specifically reported difficulties combining the generated text and their style. So, although ChatGPT could accelerate writing in the first phase, it requires more time to finalize the task and assemble content.

Our experimental group had slightly more problems with plagiarism than the control group. Fyfe (2022) also showed that his students felt uncomfortable writing and submitting the task since they felt they were cheating and plagiarizing. However, a pairwise comparison of essays in our study did not reveal remarkable similarities, indicating that students had different reasoning and styles, regardless of whether they were using ChatGPT. This could also imply that applying the tool for writing assistance produces different outcomes for the same task, depending on the user's input (Yeadon et al., 2023).

The available ChatGPT text detector (Farrokhnia et al., 2023) did not perform well, giving false positive results in the control group. Most classifiers are intended for English and usually have disclaimers for performance in other languages. This raises the necessity of improving existing algorithms for different languages or developing language-specific ones.

The main concern of using ChatGPT in academic writing has been the unauthenticity (Cotton et al., 2023; Susnjak, 2023; Yeadon et al., 2023), but we believe that such tools will not increase the non-originality of the published content or students' assignments. The detectors of AI-generated text are developing daily, and it is only a matter of time before highly reliable tools are available. While our findings suggest no immediate need for significant concern regarding the application of ChatGPT in students' writing, it is crucial to acknowledge that this study's design reflects real-life situations of using ChatGPT as a convenient and rapid solution to submit assignments, potentially at the expense of the overall quality of their work. This issue remains an important consideration when assessing the broader implications of our study.

The main drawback of this study is the limited sample size (9 per group) which does not permit the generalization of the findings or a more comprehensive statistical approach. One of the limitations could also be language-specificity (students wrote in native, non-English language for their convenience), which disabled us from the full application of AI detection tools. We should also consider that ChatGPT is predominantly fed with English content, so we cannot exclude the possibility that writing in English could have generated higher-quality information. Lastly, this was our students' first interaction with ChatGPT, so it is possible that lack of experience as well as inadequate training in using AI language models also affected their performance. Therefore, it is crucial to exercise caution when generalizing these findings, as they may not necessarily reflect the experiences of a broader range of ChatGPT users, who often report rapid draft generation. Future studies should therefore expand the sample size, number, and conditions of experiments, include students of different profiles, and extend the number of variables that could generally relate to writing skills. Also, it would be useful to conduct a study that would analyze the quality and depth of the students' prompts to ChatGPT, as it seems that the question type and the feedback provided by the user could remarkably affect the final result (Farrokhnia et al., 2023).

However, the academia and media concern about this tool might be unjustified, as, in our example, the ChatGPT was found to perform similarly to any web-based search: the more you know —the more you will find. In some ways, instead of providing structure and facilitating writing, it could distract students and make them underperform.

Data availability

The authors confirm that the data supporting the findings of this study are available within the article [and/or] its supplementary materials.

Received: 23 February 2023; Accepted: 11 October 2023;
Published online: 28 October 2023

References

Agomuoh F (2023) ChatGPT: how to use the viral AI chatbot that took the world by storm. Digital Trends. https://www.digitaltrends.com/computing/how-to-use-openai-chatgpt-text-generation-chatbot/. Accessed 10 Oct 2023

Aljanabi M, Ghazi M, Ali AH et al. (2023) ChatGpt: Open Possibilities. Iraqi J Comput Sci Math 4(1):62–64. https://doi.org/10.52866/20ijcsm.2023.01.01.0018

Aydin Ö, Karaarslan E (2022) OpenAI ChatGPT generated literature review: digital twin in healthcare. Emerg Comput Technol 2:22–31. https://doi.org/10.2139/ssrn.4308687

Banihashem SK, Noroozi O, den Brok P et al. (2023a) Identifying student profiles based on their attitudes and beliefs towards online education and exploring relations with their experiences and background. Innov Educ Teach Int 1–15. https://doi.org/10.1080/14703297.2023.2227616

Banihashem SK, Noroozi O, den Brok P et al. (2023b) Modeling teachers' and students' attitudes, emotions, and perceptions in blended education: Towards post-pandemic education. Int J Manag Educ 21(2):100803. https://doi.org/10.1016/j.ijme.2023.100803

Bishop LA (2023) Computer wrote this paper: what ChatGPT means for education, research, and writing. Res Writ. https://doi.org/10.2139/ssrn.4338981

Cotton DRE, Cotton PA, Shipway JR (2023) Chatting and cheating: ensuring academic integrity in the era of ChatGPT. Innov Educ Teach Int 00:1–12. https://doi.org/10.1080/14703297.2023.2190148

Debut L, Kim JW, Wu J (2023) RoBERTa-based GPT-2 Output Detector from OpenAI. https://openai-openai-detector.hf.space/. Accessed 10 Oct 2023

Farrokhnia M, Banihashem SK, Noroozi O et al. (2023) A SWOT analysis of ChatGPT: implications for educational practice and research. Innov Educ Teach Int 1–15. https://doi.org/10.1080/14703297.2023.2195846

Fyfe P (2022) How to cheat on your final paper: assigning AI for student writing. AI Soc 38:1395–1405. https://doi.org/10.17613/0h18-5p41

Gao CA, Howard FM, Markov NS et al. (2023) Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers. NPJ Digit Med. https://doi.org/10.1038/s41746-023-00819-6

Goal D, ChatGPT (2023) GPT3 content detector. https://detector.dng.ai/. Accessed 10 Oct 2023

Grimaldi G, Ehrler B (2023) AI et al.: machines are about to change scientific publishing forever. ACS Energy Lett 8(1):878–880. https://doi.org/10.1021/acsenergylett.2c02828

Heitmann P, Hecht M, Schwanewedel J et al. (2014) Students'argumentative writing skills in science and first-language education: Commonalities and differences. Int J Sci Educ 36(18):3148–3170. https://doi.org/10.1080/09500693.2014.962644

Hoang G (2023) Academic writing and AI: Day-5 experiment with cultural additivity. https://osf.io/u3cjx/download

Hoang G, Nguyen M, Le T (2023) Academic writing and AI: Day-3 experiment with environmental semi-conducting principle. https://osf.io/2qbea/download

Kerman NT, Banihashem SK, Noroozi O (2023) The relationship among students' attitude towards peer feedback, peer feedback performance, and uptake. in the power of peer learning: fostering students' learning processes and outcomes. Springer, p. 347–371. https://doi.org/10.1007/978-3-031-29411-2_16

Kung TH, Cheatham M, Medenilla A et al. (2023) Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health. https://doi.org/10.1371/journal.pdig.0000198

Latifi S, Noroozi O, Talaee E (2021) Peer feedback or peer feedforward? Enhancing students' argumentative peer learning processes and outcomes. Br J Educ Technol 52:768–784. https://doi.org/10.1111/bjet.13054

Lo CK (2023) What is the impact of ChatGPT on education? A rapid review of the literature. Educ Sci 13(4):410. https://doi.org/10.3390/educsci13040410

Ma Y, Liu J, Yi F (2023) Is this abstract generated by AI? A research for the gap between AI-generated scientific text and human-written scientific text. Preprint at *arXiv*. https://doi.org/10.48550/arXiv.2301.10416

Mullen L (2015) Package 'textreuse'. https://mran.revolutionanalytics.com/snapshot/2016-03-22/web/packages/textreuse/textreuse.pdf. Accessed 10 Oct 2023

Nguyen M, Le T (2023a) Academic writing and AI: Day-2 experiment with Bayesian Mindsponge Framework. https://osf.io/kr29c/download. Accessed 10 Oct 2023

Nguyen M, Le T (2023b) Academic writing and AI: Day-1 experiment. https://osf.io/kr29c/download. Accessed 10 Oct 2023

Nguyen Q, La V (2023) Academic writing and AI: Day-4 experiment with mindsponge theory. OSF Prepr awysc, Cent Open Sci. https://osf.io/download/63c551a4774ea80319ad67ba/. Accessed 10 Oct 2023

OpenAI (2022) Optimizing language models for dialogue. https://openai.com/blog/chatgpt/. Accessed 10 Oct 2023

OpenAI (2023) AI text classifier. https://platform.openai.com/ai-text-classifier. Accessed 10 Oct 2023

PlagScan (2022) http://www.plagscan.com/plagiarism-check/. Accessed 10 Oct 2023

Pourhoseingholi MA, Hatamnejad MR, Solhpour A (2023) Does chatGPT (or any other artificial intelligence language tools) deserve to be included in authorship list? chatGPT and authorship. Gastroenterol Hepatol Bed Bench 16(1):435–437

Ranjbaran F, Babaee M, Akhteh Khaneh MP et al. (2023) Students' argumentation performance in online learning environments: Bridging culture and gender. Int J Technol Educ 6:434–454. https://doi.org/10.46328/ijte.460

Sallam M (2023) ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. Healthcare. https://doi.org/10.3390/healthcare11060887

Schreyer Institute for Teaching Excellence. Writing rubric example. http://www.schreyerinstitute.psu.edu/pdf/suanne_general_resource_WritingRubric.pdf. Accessed 10 Oct 2023

Stokel-Walker C (2022) AI bot ChatGPT writes smart essays—should professors worry? Nature. https://doi.org/10.1038/d41586-022-04397-7

Susnjak T (2023) ChatGPT: the end of online exam integrity? Preprint at *arXiv*. https://doi.org/10.48550/arXiv.2212.09292

Valero Haro A, Noroozi A, Biemans O et al. (2022) Argumentation Competence: students' argumentation knowledge, behavior and attitude and their relationships with domain-specific knowledge acquisition. J Constr Psychol 135(1):123–145. https://doi.org/10.1080/10720537.2020.1734995

Whitford E (2022) Here's how Forbes got the ChatGPT AI to write 2 college essays in 20 min Forbes. https://www.forbes.com/sites/emmawhitford/2022/12/09/heres-how-forbes-got-the-chatgpt-ai-to-write-2-college-essays-in-20-minutes/?sh=7be402d956ad. Accessed 10 Oct 2023

Xiao Y (2023) Decoding authorship: is there really no place for an algorithmic author under copyright law? International Rev Intellect Prop Compet Law 54:5–25. https://doi.org/10.1007/s40319-022-01269-5

Yeadon W, Inyang O, Mizouri A et al. (2023) The death of the short-form physics essay in the coming AI revolution. Phys Educ 58(3):035027. https://doi.org/10.1088/1361-6552/acc5cf

Author contributions

All authors have contributed equally.

Competing interests

The authors declare no competing interests.

Ethical approval

The study was retrospectively approved by the Ethical Committee of the University Department of Forensic Sciences, University of Split, Croatia (053-01/23-01/12; 1, 3/8/2023). The research was performed in accordance with the principles of the Declaration of Helsinki. Research participants were not subjected to harm in any way whatsoever. Respect for the dignity of research participants was prioritized.

Informed consent

Full consent was obtained from the participants. Before the study participants signed the informed consent and were given a separate sheet to write their names and password, which enabled anonymity while grading essays and further analysis of student-specific variables. The protection of the privacy of research participants has been ensured.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1057/s41599-023-02269-7.

Correspondence and requests for materials should be addressed to Ana Banovac.

Reprints and permission information is available at http://www.nature.com/reprints