112 Exploring the Impact of a disruptive technology on Higher Education assessment design: The case of ChatGPT

Alexander Kofinas[1], Crystal Tsay[2], David Pike[1]

[1]University of Bedfordshire, Luton, United Kingdom. [2]University of Greenwich, London, United Kingdom

Research Domains

Digital University and new learning technologies (DU)

Abstract

This exploratory study looks at the extent to which a generative AI writing tool, ChatGPT, impacts on student performance in Higher Education assessments and how effectively educators are able to differentiate AI-authored and human work. A two-phased, within-subjects experiment, involving paired academics from two UK business schools, was conducted. Preliminary findings revealed a struggle for participants in differentiating AI and human work. AI enhancements proved to be neutral to both originally high-quality and subpar student work. However, work that was crafted by ChatGPT only was of very high quality, often getting the highest grades, and it was particularly difficult to identify. It is suggested that ChatGPT's effectiveness varied according to assessment type, showing greater impacts on traditional than on authentic assessments. These findings pose more questions than answers in redefining academic integrity and re-exemplifying academic misconduct. Future research should explore what would constitute effective assessment strategies in Higher Education.

Full paper

Assessments play a vital role in Higher Education (HE), providing a measurement of a student's learning (Cilliers et al., 2012). Summative assessments are seen to drive student's learning as students often aim to achieve a high grade. The task for educators is to design effective assessments that instil deep learning in students.

The advent of technology enables innovative assessment types but also challenges academic integrity. ChatGPT, or Generative Pretraining Transformer, is a large language model developed by OpenAI and released in November 2022. It uses machine learning to generate human-like text by predicting the probability of a word given the words that came before it. ChatGPT offers promising applications for HE, including language practice and virtual tutoring for students and teaching, learning, and assessment design for educators (Millick & Mollick, 2023). Nevertheless, its role in assessment design is complex, balancing benefits of diverse, creative assignments against the risk of student misuse and academic misconduct (Cotton et al., 2023).

We ask two research questions (RQ):

RQ1: How effectively would markers be able to differentiate between human-authored assessments and ChatGPT modified/generated assessments?

RQ2: Does ChatGPT generate better versions of the assessment so that it has a substantial impact on students' grades?

We conducted a two-phased, within-subjects experimental design to compare AI writing to human writing (Charness, Gneezy, & Kuhn, 2012). Institutional ethics approval was obtained. Two undergraduate degree programmes at two post-92 business schools in the United Kingdom were first identified. Upon receiving academics' agreements, one module at each undergraduate academic level (levels 4 to 6) within a programme was selected for the generation of writing samples. Two academics from each programme, who are familiar with the modules but not involved in the marking of the selected writing samples, were invited to mark the samples based on the established assessment rubrics.

Four pairs of two academics participated in the experiment. For each programme, a total of 21 writing samples were prepared for each pair of two academics, marked by the academics and followed by live interviews with a research

team member. All identifying information and feedback were removed from the writing samples. The writing samples at each academic level contained three pieces of work written at three grade bands (i.e., <40, low 50s, and 70s) by the original human authors, three pieces of ChatGPT-modified version of the same assessments, and one ChatGPT-generated work.

In Phase 1, each pair of participants were presented with a mix of randomly ordered original, GPT-modified, and GPT-generated scripts to a total of four scripts. They were asked to assign grades against the marking criteria and identify the ChatGPT submission(s). In Phase 2, participants were asked to grade the remaining scripts, and then identify the ChatGPT submission(s).

After each phase, a debrief and interview is conducted where the assessment identities were revealed, and the participants were asked to compare the original and GPT-modified and generated assessments. The participants' overall impressions, grading differences, and reflections were gathered. The process was carefully designed to minimise the order effect and sources of bias (Charness et al., 2012).

Data analyses from interviews revealed interesting preliminary findings. To answer RQ1, most participants were unable to distinguish between student-written and AI-written content, demonstrating the sophistication of the AI's output. To answer RQ2, ChatGPT enhancements did not significantly improve the original student-written work across all grade bands. Furthermore, utilising ChatGPT to generate completely original work, was mostly graded at the upper second level, and was effective for traditional assessments such as literature review essays. In contrast, the results were mixed when the assessments were deemed more authentic.

Interestingly, the fact that academics could not distinguish between human- and AI-authored work poses a few interesting questions that are worth future examinations. For example, to what extent should we encourage the use of disruptive technologies like ChatGPT in student assessment preparation, as opposed to using current technologies such as studiosity and grammerly? How would academic misconduct/integrity be redefined or examplied if disruptive technologies are embedded as part of HE teaching and learning?

To conclude, this research suggests that there are significant challenges for academic integrity and the need for assessment adaptations. Future studies must explore ethical AI usage, possibly redefining academic misconduct in the context of emerging technologies.

References

Charness, G., Gneezy, U., & Kuhn, M. A. (2012). Experimental methods: Between-subject and within-subject design. Journal of Economic Behavior & Organization, 81(1), 1-8.

Cilliers, F. J., Schuwirth, L. W., Herman, N., Adendorff, H. J. & Van Der Vleuten, C. P. (2012). A model of the pre-assessment learning effects of summative assessment in medical education. Advances in Health Sciences Education, 17, 39-53.

Cotton, D. R., Cotton, P. A. & Shipway, J. R. J. I. I. E. (2023). Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. Innovations in Education Teaching International, 1-12.

Newton, P. E. (2007). Clarifying the purposes of educational assessment. Assessment in Education, 14, 149-170.

Mollick, E. and Mollick, L. (2023) 'Let ChatGPT Be Your Teaching Assistant: Strategies for Thoughtfully Using AI to Lighten Your Workload', Harvard Business Publishing Education. Available at: https://hbsp.harvard.edu/inspiring-minds/let-chatgpt-be-your-teaching-assistant?cid=email%7Cmarketo%7C2023-05-02-the-faculty-lounge%7C10083656%7Cthought-lead-faculty-lounge%7Ceducator%7Cvarious%7Cmay2023&acctID=20207275&mkt_tok=ODU1LUFUWi0yOTQAAAGLex1Mq40JLmiV2r3nbId903gQ4_wtILwWf7GEqMYBYDpDvVb_MqeQSpO6_bwvZwhKi0JqjKwee3qC8QrGsq2B7P2-HtuoqXSejQm9eEhpvII. Access: 08 May 2023.