Will ChatGPT pass the online quizzes? Adapting an assessment strategy in the age of generative AI

Damien Raftery[1]
South East Technological University

Abstract

As generative AI (artificial intelligence) technologies, such as ChatGPT, become increasingly available, traditional online assessments must be re-evaluated to maintain their educational value. Open-book online quizzes have long been an effective tool for engaging students, effectively supporting learning, and reinforcing fundamental knowledge and skills. However, the ease of using AI to complete these quizzes may undermine their intended purpose.

This article explores the initial findings of using ChatGPT versions 3.5 and 4 to answer twelve online quizzes used for continuous assessment in two first-year quantitative techniques modules on business programmes in an Irish technological university. ChatGPT, along with suitable plugins, is increasingly accurate in answering the online quizzes, with results as follows: ChatGPT-3.5 achieving an average percentage score of 35%, ChatGPT-4 scoring 47% and ChatGPT-4 with Wolfram plugin 78% (the *percentage score* is the total marks out of 100 marks for each quiz). Most of the incorrect responses are due to calculation errors; if these are corrected by simply checking the arithmetic with a calculator, the averages increase to ChatGPT-3.5 scoring 72%, ChatGPT-4 76% and ChatGPT-4 with Wolfram plugin 80%. Thus, the online quizzes on these modules can be quickly completed with the assistance of ChatGPT with a high level of success. The implications of this for using online quizzes as an assessment strategy are discussed; potential assessment redesigns are outlined, including how to integrate generative AI into the learning and assessment process in an ethical and constructive manner. Although generative AI provides a challenge to traditional online quizzes, it also has the potential to aid student comprehension and learning, and the skills of prompt engineering are likely to become increasingly relevant and useful.

1. Introduction

Open-book online quizzes are an effective tool for engaging students, particularly with fundamental knowledge and skills (Angus and Watson, 2009; Lyng and Kelleher, 2019). They support learning through retrieval practice, interleaving and spaced practice (Lang, 2021; Karpicke and Roediger, 2008; Rohrer, Dedrick, and Stershic, 2015), and encourage metacognition and self-regulation (Brame, 2019). For each *Quantitative Techniques* module, there are six online quizzes that students can flexibly take each up to five times over about five days, with the best attempt counting (after completion, the workings for the best attempt should be uploaded). After each attempt, the virtual learning environment (VLE) provides the correct answer and, for any mistakes, students are encouraged to seek help from me, our university maths support centre, and their peers.
Early in 2023, I began to check how good generative AI, in particular ChatGPT (https://openai.com/), was at completing the online quizzes my students were doing. There

[1] Corresponding author damien.raftery@setu.ie

was substantial media coverage of ChatGPT successfully completing online exams, such as passing an MBA exam at Wharton (Rosenblatt, 2023). In this case, the reality is a little more complicated, with Terwiesch (2023) checking that ChatGPT could answer a few final exam questions and extrapolating from this that it would get a B- to B on his course. Mitchell (2023) reviewed Terwiesch's exam questions, showing the hit-and-miss nature of ChatGPT, noting when it fails to correctly answer a similar question with superficial rewordings.

With technology changing rapidly, Dawson, Nicola-Richmond, and Partridge (2023, p.2) commented on the need "to revisit types of openness in online examinations and their consequences for assessment", noting that when the conditions for an assessment change, "judgements about student performance become less valid" (p.3). In February 2023, Newton (2023a) reviewed the literature on the performance of ChatGPT on multiple-choice questions (MCQs), concluding that it performed modestly. At the time, ChatGPT struggled with higher-level problem-solving, whilst also having difficulty with calculations and images. By the end of March, Newton (2023b) was refreshingly blunt and honest that this was no longer the case; with the release of ChatGPT-4, Newton (2023b) commented that it was now "really good" at MCQs.

Over the past few months, when informally discussing ChatGPT with my students, there is a small but growing awareness with little admitted use. Malmström, Stöhr, and Ou (2023), in a survey of Swedish university students undertaken mostly in April 2023, found that almost all are familiar with ChatGPT with more than a third using it regularly. Thus, by next September it is likely that most of my students will be able to use ChatGPT. Whilst noting the importance of the ethical issues in using generative AI – such as equality of access, data privacy, energy use, intellectual property, lack of regulation, outsourcing of moderation for toxic content, and embedded bias (Sabzalieva and Valentini, 2023; Hillier, 2023), there are the pragmatic considerations that my students are likely to use generative AI and indeed need to learn the skills of prompt engineering. Prompt engineering is a term describing how to write good prompts to program Large Language Models such as ChatGPT. This skill is becoming important for graduates, with Felten, Raj, and Seamans (2023, p.1) commenting that "highly-educated, highly-paid, white-collar occupations may be most exposed to generative AI".

To summarise, online quizzes can effectively support learning. In general, generative AI is improving its performance on online quizzes, and students are increasingly becoming aware of, and are using, ChatGPT. The use of generative AI through prompt engineering is an important skill. Thus, the dilemma is how to retain the learning benefits of using online quizzes if ChatGPT performs well on them. If students are permitted, supported, and/or encouraged to use generative AI tools, will important learning be bypassed? Alternatively, if ChatGPT use is prohibited, will any such use be considered *unauthorised content generation* (Foltynek *et al.*, 2023), a breach of academic integrity due to undeclared technological assistance? And with students missing out on important learning opportunities that generative AI may enable?

In the next section, the methods are described for this investigation into using ChatGPT to answer online quiz questions and for a survey of students' experiences of online quizzes. This is followed by the presentation of the findings of both the investigation and the survey. These results are then discussed in the context of the above dilemma and themes, followed by some concluding remarks.

2. Methods

The main investigation, undertaken in early May 2023, explored how accurately ChatGPT can answer questions in the online quizzes in two quantitative techniques modules undertaken by first-year business students. The results of this investigation are complemented by results of a survey of students completed in June 2021 into their experiences of, and attitudes to, online quizzes.

Following some informal ongoing use of ChatGPT-3 from December 2022 to April 2023 to complete the online quizzes that my students were doing at the time, in May 2023 a systematic testing plan was devised and undertaken. For the *Quantitative Techniques* module in each semester, there are six online quizzes. Five to seven questions are randomly picked from pools of questions, with most questions of the calculation type (each time the question is presented, different numbers are used for selected variables). There are a small number of MCQs and questions incorporating images (such as hot spot questions with an image of a chart or a table). Individual questions are worth differing number of marks and the *percentage scores* presented below are the overall score out of 100 marks for each quiz.

Each quiz was taken once within the virtual learning environment (VLE). For each question, the text of the question was copied and pasted into ChatGPT, with the answer as presented by ChatGPT copied and pasted back into the VLE. Other than the question, the only other text added to the prompt was to show the method, if required. For calculation-type questions, the answer was a number. For MCQs, an option is selected. For hot spot questions involving images, an element of judgement was required; the text of the question was pasted into ChatGPT and the response read. If the response very clearly explained how to select an area in the image, then this was done so (and, if correct, the marks for this question were awarded). The VLE marked the responses and provided the correct answer. It took on average less than five minutes to complete each quiz.

Each of the twelve quizzes was initially taken using ChatGPT-3.5, each quiz using one chat session. To provide an audit trail, the responses from ChatGPT-3.5 were recorded, as were the questions and the scoring and correct answers from the VLE. For each quiz, the same questions were then used to test ChatGPT-4 and ChatGPT-4 with Wolfram plugin[2]. In addition, as the most common error that ChatGPT makes is to be inaccurate in calculations, each quiz was re-scored with updated answers after checking the arithmetic with a handheld calculator; this was done by simply following the answer as presented by ChatGPT, entering and manipulating the numbers in the steps as clearly presented by ChatGPT. This was done for ChatGPT-3.5, ChatGPT4 and for ChatGPT-4 with Wolfram plugin. Note the limitation that each quiz was attempted once. As questions are randomly drawn from pools and ChatGPT can be inconsistent in its response (Mitchell, 2023), scores on the quizzes should be viewed as indicative rather than definitive.

Following appropriate ethical approval, students taught primarily online during the academic year 2020-2021 due to COVID-19 lockdowns and restrictions, were invited to complete questionnaires about the learning, teaching and assessment on the *Quantitative Techniques* module (Raftery, 2021). This included specific questions about their experiences of, and attitudes to, the online quizzes used during the module. In a final survey in June 2021, 25

[2] Note that ChatGPT Plus, the paid version of ChatGPT, gives access to ChatGPT-4 and enables optional plugins such as Wolfram (Wolfram, 2023), which provides accurate computations within ChatGPT output.

students responded, this is an approximate response rate of 30% of those 84 students who completed the module. Note the limitation of potential sample bias, that those students who responded may differ in their views from those who did not.

3. Findings

The results of the investigation into the accuracy of using ChatGPT to answer online quiz questions are presented first, followed by the survey findings of students' experiences of online quizzes.

3.1 Percentage scores for ChatGPT-3.5 and ChatGPT-4, presented by semester

There are six online quizzes for each module, taught over one semester. Initially, the results for each semester are presented, along with the overall average (mean) score. Figure 1 below shows the percentage scores (i.e. score out of 100 marks for each quiz) for the quizzes in semester 1 when attempted with ChatGPT-3.5 and ChatGPT-4, as well as the improved scores when the calculations are checked with a calculator and the updated number used as the answer. Note that ChatGPT-4 outperforms ChatGPT3.5, with an overall average of 50% compared to 43%. When calculations are checked, the average score has increased, with both having similar averages, 64% and 66% respectively.

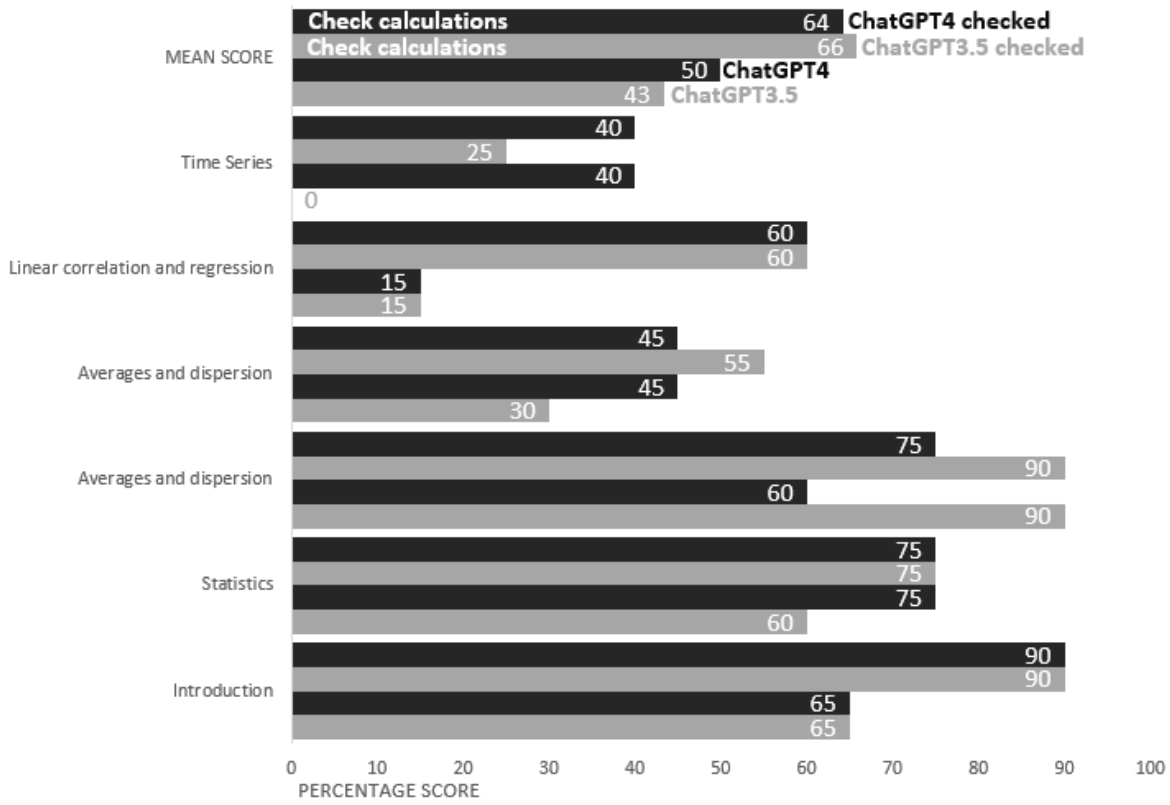

Figure 1: ChatGPT percentage scores on Semester 1 quizzes, along with scores when the answer is updated after the calculations checked with a calculator.

Figure 2 below shows the percentage scores for Semester 2 quizzes. Note that ChatGPT-4 again outperforms ChatGPT3.5, with an overall average of 43% compared to 26%. When calculations are checked, there is a substantial increase in the average percentage score, 88% and 78% respectively. This increase is primarily due to increases in the scores for the investment mathematics quizzes (*Compound Interest*, *Regular Payments*, and *NPV, IRR and Depreciation*), where most of the solutions involve calculating powers which ChatGPT performs poorly at.

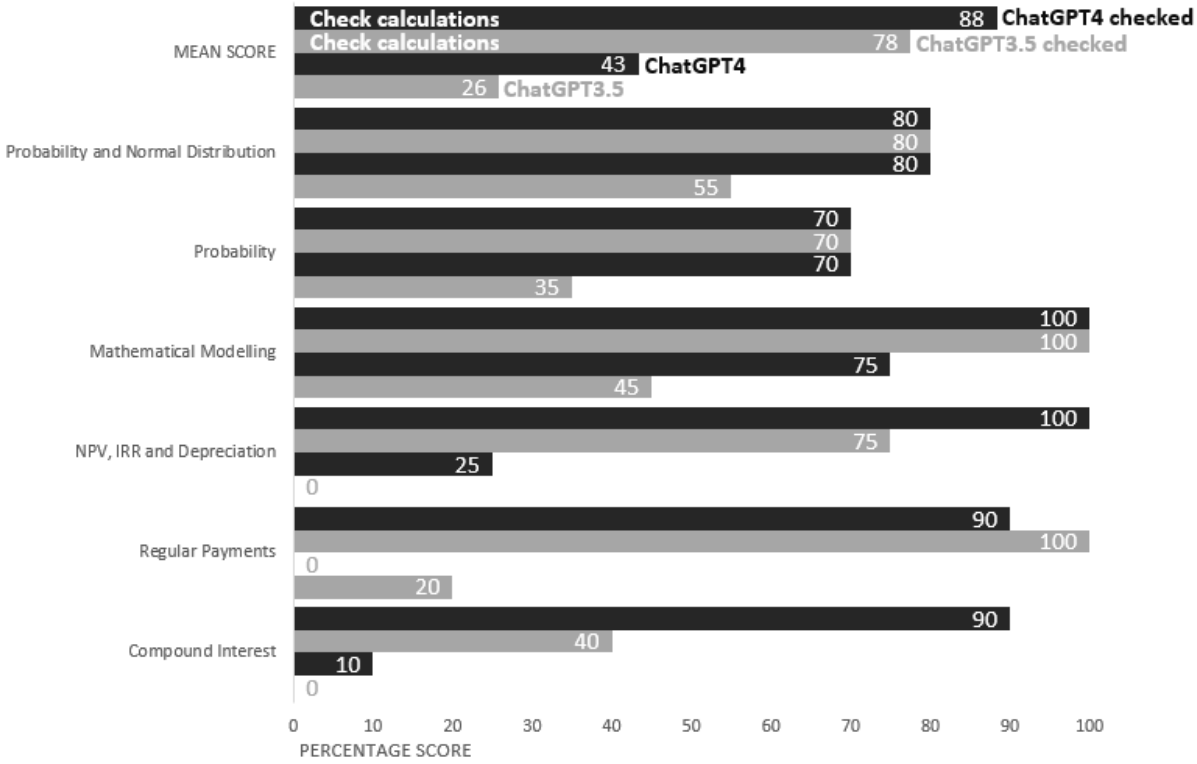

Figure 2: ChatGPT percentage scores on Semester 2 quizzes, along with scores when the answer is updated after the calculations checked with a calculator.

3.2 Percentage scores for ChatGPT-4 and ChatGPT-4 with Wolfram plugin, presented by semester

Figure 3 below shows the percentage scores for the quizzes in semester 1 when attempted with ChatGPT-4 and the improved scores when the calculations are checked (as previously presented in Figure 1), with the scores achieved by ChatGPT-4 with Wolfram plugin. Note that ChatGPT-4 with Wolfram plugin slightly outperforms ChatGPT4 with the calculations check, with an overall average of 71% compared to 64%.

https://doi.org/10.22554/ijtel.v7i1.114

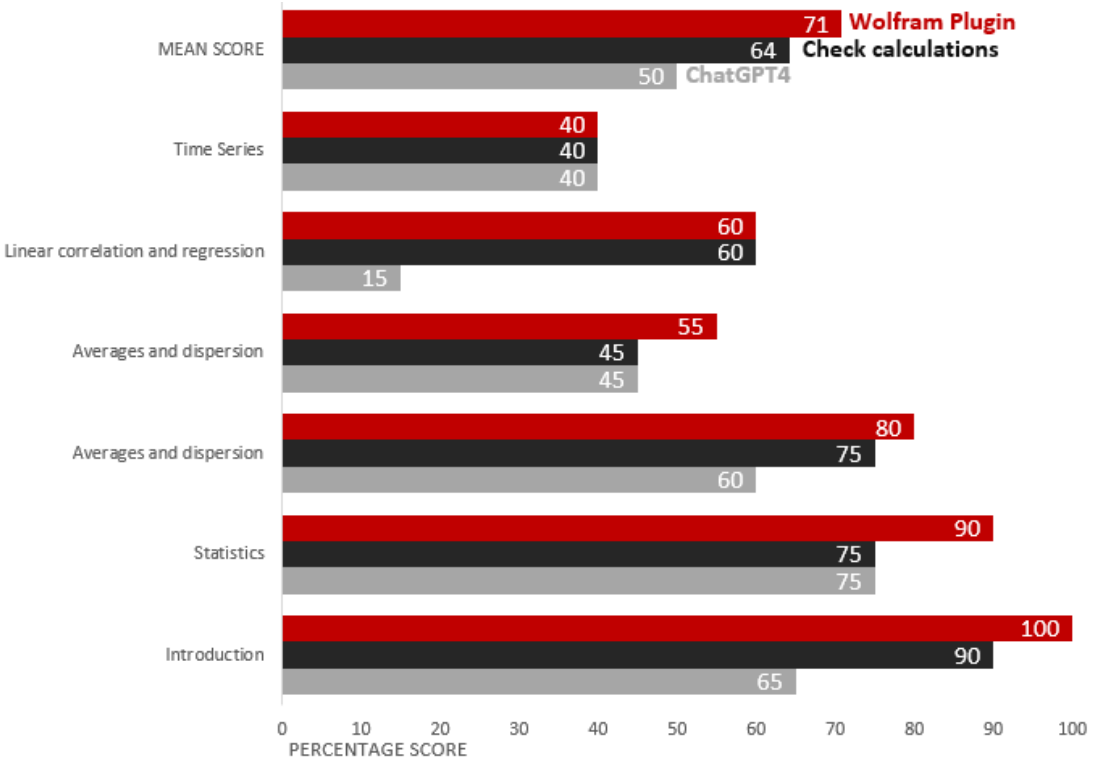

Figure 3: ChatGPT-4 percentage scores on Semester 1 quizzes, along with scores with calculations checked and scores with ChatGPT-4 with the Wolfram plugin enabled.

Figure 4 below shows the percentage scores for the quizzes in semester 2 when attempted with ChatGPT-4 and the improved scores when the calculations are checked (as previously presented in Figure 2), with the scores achieved by ChatGPT-4 with Wolfram plugin. Note that ChatGPT-4 with Wolfram plugin performs similarly to ChatGPT4 with the calculations check, with an overall average of 85% compared to 88%. This difference disappears when the calculations for ChatGPT-4 with Wolfram plugin are checked, with both scoring an average of 88%.

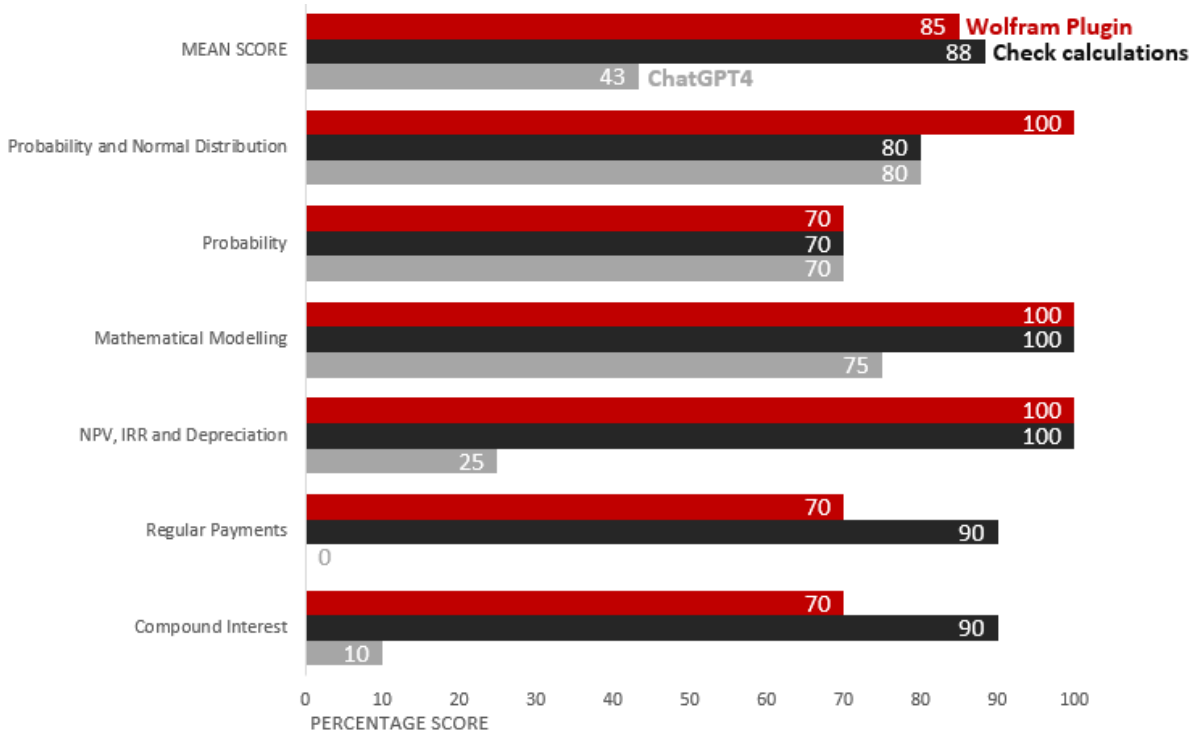

Figure 4: ChatGPT-4 percentage scores on Semester 2 quizzes, along with scores with calculations checked and scores with ChatGPT-4 with the Wolfram plugin enabled.

3.3 Survey of student experiences of online quizzes

A previous cohort of students was surveyed about their experiences studying the *Quantitative Techniques* module, with 25 responses out of 84 students. As can be seen in Figure 5 below, 88% of the 25 respondents rated online quizzes as being *Extremely useful* to their learning, with the remaining 12% choosing *Very useful*. Of many learning resources and activities, online quizzes were viewed most positively along with access to video solutions to problem questions.

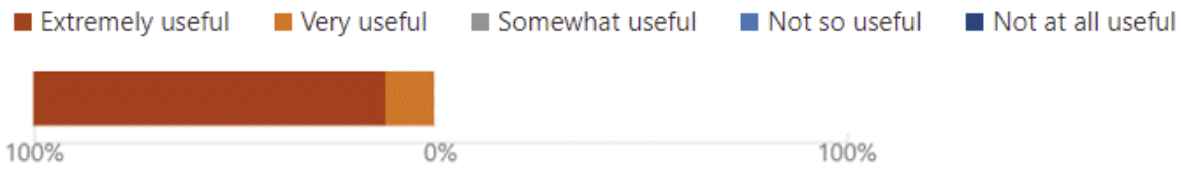

Figure 5: Students' rating of the usefulness of online quizzes to their learning.

As can be seen in Figure 6 below, the same cohort of respondents were very positive about online quizzes, in that the quizzes encouraged them to study and try questions, to get help if they made a mistake and were useful to their learning, disagreeing only that the online quizzes were too time-consuming.

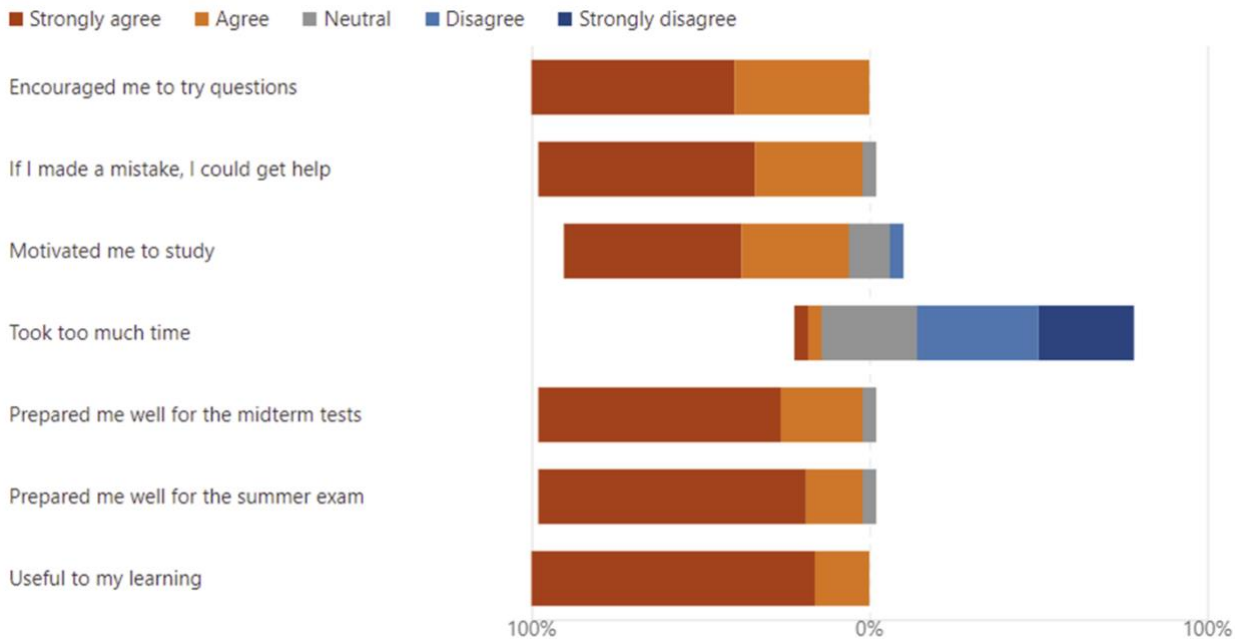

Figure 6: Students' level of agreement with statements relating to online quizzes.

In response to an open-ended question, two students responded:

- They allowed you to try each question and learn from your mistakes.
- They kept me on top of my work every week and I had time to do it as it was given to us for a week so I didn't feel under pressure so it made it a lot less stressful.

The overall implications of the above survey findings, both the quantitative ratings of statements and the open-ended comments, show the students' positive disposition to online quizzes and the importance that they attach to online quizzes for helping them to learn, be motivated, and prepare for summative examinations.

4. Discussion

The findings from the survey support that students are positively disposed to online quizzes, finding them useful to help them learn. The results of the investigation show the ability of ChatGPT to successfully answer many of the online quiz questions. In Figure 7 below, for all twelve online quizzes, the average (mean) score with ChatGPT-3.5 is 35%, ChatGPT-4 47% and ChatGPT-4 with Wolfram plugin 78%. The improved scores when the calculations are checked for ChatGPT-3.5 and ChatGPT-4 are close to those achieved by ChatGPT-4 with Wolfram plugin. Indeed, with the calculations checked, the average score for ChatGPT-3.5 is 72%, ChatGPT-4 76% and ChatGPT-4 with Wolfram plugin 80%.

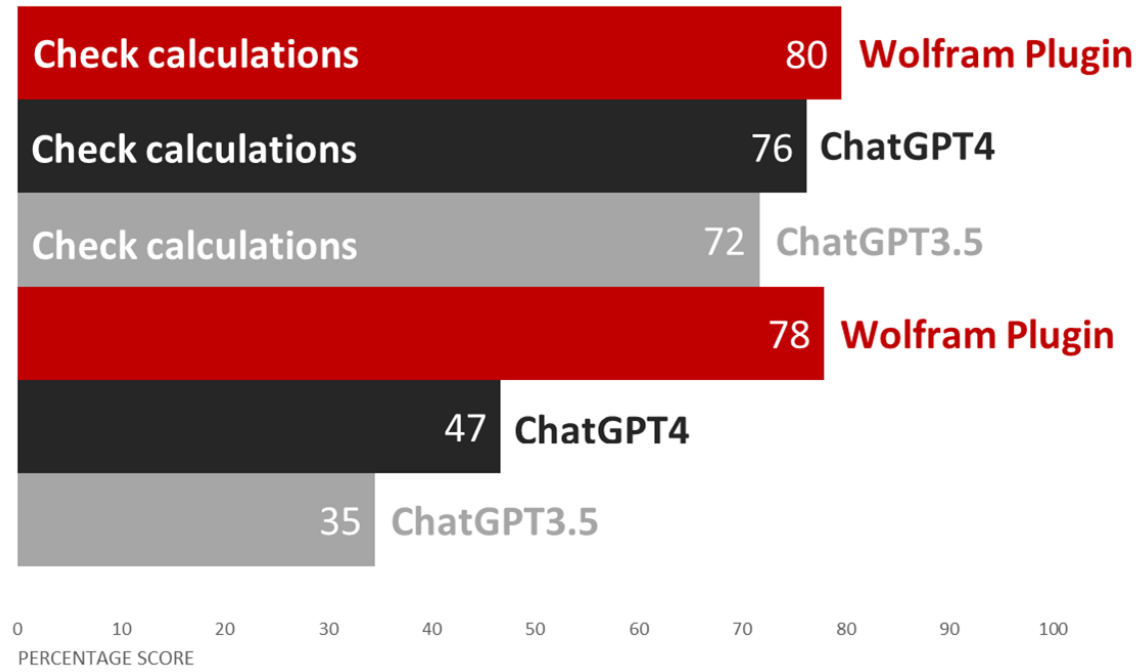

Figure 7: ChatGPT-4 percentage scores on all twelve quizzes, along with scores with calculations checked and scores with ChatGPT-4 with the Wolfram plugin enabled.

ChatGPT is good at answering the questions in the online quizzes – very good. Yet online quizzes are effective at supporting learning and the survey results support that students view them favourably. What are the implications?

Willingham states that "student learning is a complex system, and predicting the consequences of change to one part of that system is at best uncertain" (2019, p.23). I agree with Willingham that access to the Internet has not meant that students no longer need to learn facts. Christodoulou (2023) argues that students need to "work through problems that computers *can* do" to be able to "successfully grapple with problems computers *cannot* do" (*emphasis* in original). Access to generative AI does not mean students will no longer need to be able to solve problems, develop mathematical and statistical literacy, and do calculations. What is required is the thoughtful integration of generative AI tools into the process whereby students think about, and solve, problems.

There is potential for the output of generative AI to be used to help students learn, with the output useful in explaining how to solve problems. In ChatGPT, hitting *Regenerate response* gives alternative solutions and explanations. Using the VLE, the student can identify

incorrect answers and confirm the level of inaccuracy of any incorrect number, and thus potentially learn from mistakes made by ChatGPT. Learning how to write good prompts to guide generative AI is important. For example, for a base-rate probability question, adding the prompt "Solve the following problem by creating a contingency table with an overall total of 10,000" before the question *usually* results in a clear solution that mirrors the approach taken in class (see sample ChatGPT3.5 output with prompt and without prompt). This helps the student to understand all the steps to solve the problem, using the frequency format approach that is simpler to use and understand by both students and professionals than a more formal Bayesian approach (Gigerenzer, 1996). Another useful prompt is to guide the solution to the question by providing the formula to be used, simply copied from the notes. This results in the output from ChatGPT using notation similar to that which the student is familiar with (see sample ChatGPT3.5 output with prompt and without prompt). In these two modules and across their programmes, students will need to learn prompt engineering to improve the output from generative AI tools, including *prompt patterns* such as *Persona*, *Question Refinement*, and *Output Automater* (White *et al.*, 2023).

The planned assessment strategy for the next academic year is to retain the online quizzes as low-stakes continuous assessment worth 15%, with the permitted use of generative AI and students required to document any such use. To complement this, a two-stage exam based on the questions from the online quizzes may be introduced. Students learn through discussion and reflection whilst doing the group exam immediately after the individual exam (Nicol and Selvaretnam, 2022). However, Kinnear's (2021) study found little impact on long-term learning of this approach in mathematics, observing that it may be that "more procedural questions offer less opportunity for fruitful discussion than conceptual questions" (p.51), whilst still noting the potential value to foster student collaboration. More generally, Dawson, Nicola-Richmond, and Partridge (2023, p.9) highlight that restrictions introduced by changes to assessments have consequences for validity and it is important to consider "their impact on students' lives beyond the immediate act of assessment". From a student's perspective, a potentially negatively perceived change of reducing the weighting of the flexible option of online quizzes to increase the weighting of a much less flexible in-class exam may be partially offset by introducing the group exam as part of a two-stage process. This may also bring pedagogical benefits of learning through discussion.

5. Conclusion

This article has presented the initial findings of an investigation into the accuracy of using ChatGPT to answer online quizzes in two first-year quantitative techniques modules. It is very good, with ChatGPT-3.5 achieving an average percentage score of 35%, ChatGPT-4 47% and ChatGPT-4 with Wolfram plugin 78%. If calculation errors are corrected by simply checking the arithmetic with a calculator, the averages increase to ChatGPT-3.5 scoring 72%, ChatGPT-4 76% and ChatGPT-4 with Wolfram plugin 80%. The main finding is that the online quizzes on these modules can be quickly completed with the assistance of ChatGPT with a high level of success.

This has implications for the use of online quizzes for summative assessment and the process whereby students learn whilst completing them. Due to the benefits to student learning of online quizzes, and students' positive experiences and attitudes to them, I favour retaining online quizzes as a low-stakes element of continuous assessment with permitted and

documented use of generative AI. By thoughtfully embracing generative AI in an ethical and constructive manner, there is potential to enhance the student learning process, help them solve problems, and equip them with important skills for life and work. In this age of generative AI, I am looking forward to learning together with my students.

Acknowledgements

I would like to thank my colleague Dr. P.J. Wall for the fruitful discussions around generative AI, ethics and academic integrity. Thanks also to my colleague Sharon McDonald for collaborating on teaching and assessing the *Quantitative Techniques* modules. And to my students and colleagues for the ongoing conversations about generative AI.

References

Angus, S. D. & Watson, J. (2009). Does regular online testing enhance student learning in the numerical sciences? Robust evidence from a large data set. *British Journal of Educational Technology*, *40*(2), 255-272. https://doi.org/10.1111/j.1467-8535.2008.00916.x

Brame, C. J. (2019). *Science Teaching Essentials: Short Guides to Good Practice.* Elsevier Science & Technology.

Christodoulou, D. (2023, February 5). If we are setting assessments that a robot can complete, what does that say about our assessments? *No More Marking* blog. https://substack.nomoremarking.com/p/if-we-are-setting-assessments-that-a-robot-can-complete-what-does-that-say-about-our-assessments-cbc1871f502

Dawson, P., Nicola-Richmond, K., & Partridge, H. (2023). Beyond open book versus closed book: a taxonomy of restrictions in online examinations. *Assessment & Evaluation in Higher Education*. https://doi.org/10.1080/02602938.2023.2209298

Felten, E. W., Raj, M. & Seamans, R. (2023, April 10). Occupational Heterogeneity in Exposure to Generative AI. *SSRN*. http://dx.doi.org/10.2139/ssrn.4414065

Foltynek, T., Bjelobaba, S., Glendinning, I. *et al.* (2023). ENAI Recommendations on the ethical use of Artificial Intelligence in Education. *International Journal for Educational Integrity, 19***(**12). https://doi.org/10.1007/s40979-023-00133-4

Gigerenzer, G. (1996). The Psychology of Good Judgment: Frequency Formats and Simple Algorithms. *Medical Decision Making*, *16*(3), 273-280. https://doi.org/10.1177/0272989X9601600312

Hillier, M. (2023, March 30). A proposed AI literacy framework. *TECHE* blog, Macquarie University. https://teche.mq.edu.au/2023/03/a-proposed-ai-literacy-framework/

Karpicke, J. D. & Roediger, H. (2008). The Critical Importance of Retrieval for Learning. *Science*, 319(5865), 966-968. https://www.science.org/doi/10.1126/science.1152408

Kinnear, G. (2021). Two-Stage Collaborative Exams have Little Impact on Subsequent Exam Performance in Undergraduate Mathematics. *International Journal of Research in Undergraduate Mathematics Education*, *7*, 33–60. https://doi.org/10.1007/s40753-020-00121-w

Lang, J. M. (2021). *Small Teaching: Everyday Lessons from the Science of Learning.* John Wiley & Sons.

Lyng, C. & Kelleher, E. (2019). Engaging large cohorts of students in online formative assessment to reinforce essential learning for summative assessment. *AISHE-J*, *11*(1), 1-21. https://ojs.aishe.org/index.php/aishe-j/article/view/366

Newton, P. M. (2023a, February 21). ChatGPT performance on MCQ-based exams. https://doi.org/10.35542/osf.io/sytu3

Newton, P. M. (2023b, March 28). Online exams in the age of ChatGPT; now what? *QQI Webinar*. *https://www.youtube.com/watch?v=YloLWCO3qWY*

Nicol, D. & Selvaretnam, G. (2022). Making internal feedback explicit: harnessing the comparisons students make during two-stage exams. *Assessment & Evaluation in Higher Education*, *47*(4), 507-522. https://doi.org/10.1080/02602938.2021.1934653

Malmström, H., Stöhr, C. & Ou, A. W. (2023). Chatbots and other AI for learning: A survey of use and views among university students in Sweden. *Chalmers Studies in Communication and Learning in Higher Education*, *1*. https://doi.org/10.17196/cls.csclhe/2023/01

Mitchell, M. (2023, February 10). Did ChatGPT Really Pass Graduate-Level Exams? *AI: A Guide for Thinking Humans* blog. https://aiguide.substack.com/p/did-chatgpt-really-pass-graduate

Raftery, D. (2021). Attending: reflections on adapting a first-year module during COVID-19. *CETL-MSOR Conference 2021*, Coventry University, Sept 2-3.

Rohrer, D., Dedrick, R. F., & Stershic, S. (2015). Interleaved practice improves mathematics learning. *Journal of Educational Psychology, 107*(3), 900–908. https://doi.org/10.1037/edu0000001

Rosenblatt, K. (2023, January 23). ChatGPT passes MBA exam given by a Wharton professor. *NBC News*. https://www.nbcnews.com/tech/tech-news/chatgpt-passes-mba-exam-wharton-professor-rcna67036

Sabzalieva, E. & Valentini, A. (2023). ChatGPT and Artificial Intelligence in higher education: Quick start guide. *UNESCO*. https://www.iesalc.unesco.org/wp-content/uploads/2023/04/ChatGPT-and-Artificial-Intelligence-in-higher-education-Quick-Start-guide_EN_FINAL.pdf

Terwiesch, C. (2023). Would Chat GPT3 Get a Wharton MBA? A Prediction Based on Its Performance in the Operations Management Course. *Mack Institute for Innovation Management at the Wharton School, University of Pennsylvania*. https://mackinstitute.wharton.upenn.edu/wp-content/uploads/2023/01/Christian-Terwiesch-Chat-GTP.pdf

White, J., Fu, Q., Hays, S. *et al.* (2023). A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT. https://doi.org/10.48550/arXiv.2302.11382

Willingham, D. T. (2019). The digital expansion of the mind gone wrong in education. *Journal of Applied Research in Memory and Cognition*, *8*(1), 20–24. https://doi.org/10.1016/j.jarmac.2018.12.001

Wolfram, S. (2023, March 23). ChatGPT Gets Its "Wolfram Superpowers"! *STEPHEN WOLFRAM Writings* blog. https://writings.stephenwolfram.com/2023/03/chatgpt-gets-its-wolfram-superpowers/