Multiple-Choice Questions Reimagined: Exploring the Ethical and Pedagogical Implications of GenAI for Higher Education

Manish Malik[1], Rehan Shah[2], Georgina Zimbittas[1] and Soumya Manna[1].

[1] *Canterbury Christ Church University*, [2] *Queen Mary University of London*

This work focuses on the re-imagination of the process of creating multiple choice questions by augmenting the skills academic bring, with Generative Artificial Intelligence (GenAI) tools such as ChatGPT. Multiple choice questions can help students practise and internalise concepts and help staff in making classrooms interactive and knowing where their students are in their learning journey and provide remedial support. However, the design process can be challenging, time consuming and rely on the knowledge and experience of staff, in particular their knowledge of any relevant cognitive conflicts or common misconceptions students harbour in a topic. This paper presents some initial results from a study on the integration of GenAI in the design of multiple-choice questions in mathematics within STEM disciplines. The paper shares some pedagogically inspired prompts used with GenAI to co-produce outputs. The paper concludes with recommendations for future research and ethical considerations, particularly concerning copyright issues.

Multiple-choice question design, ChatGPT, Generative AI, human in the loop.

Introduction

Generative Artificial Intelligence (GenAI) tools have proliferated practices within many industries and within a short space of time (Gupta, Ding, Guan, and Ding, 2024). These include, studies exploring the capabilities of ChatGPT within the area of human genetics (Duong and Solomon, 2023), and report writing within radiology (Jeblick et. al, 2023) and comparing creative outputs of ChatGPT, apla.ai, Copy.ai and YouChat with human outputs (Haase and Hanel, 2023) and likewise within many more disciplines (Gupta et al., 2024). A common thread in all such developments is the use of GenAI to augment humans, which encourages us to ascertain what it means for higher education and the role of academics. Many academics engage in the development of multiple-choice questions (MCQs). MCQs are useful in many different pedagogical settings such as formative self-assessment, aiding reflective learning through feedback, peer instruction, summative assessments, diagnostic tests and concept inventories for research as well as tailoring teaching methods for students (Nicol, 2007; Little and Bjork, 2015; Crouch and Mazur, 2001; Chien, Chang and Chang, 2016). This paper looks at how co-designing these can be done in a reliable, ethical and efficient way and reports on the initial findings. The rest of the paper highlights the development of high-quality multiple-choice questions as described in research literature (*ibid)*, and outlines how GenAI can be prompted in a way inspired by pedagogic theories. We share some examples of Mathematics questions created to be used within Science, Technology, Engineering and Mathematics (STEM) disciplines. Lastly, we discuss some ethical issues around this and future research considerations.

Multiple Choice Questions (MCQs) - their use and misuse within higher education

Multiple choice questions can help students practice and internalise concepts (Little and Bjork, 2015), or are used by staff in-class using signature pedagogies like peer instruction (Crouch and Mazur, 2001) or when using clicker systems (Chien, Chang and Chang, 2016). Well-designed questions with suitable distractors may help give insights into students' abilities to navigate common misconceptions and apply their knowledge in different settings (Gierl, Bulut, Guo, and Zhang, 2017; Little and Bjork, 2015; and Morrison and Free, 2001). In machine-orchestrated learning systems and in teacher-orchestrated classroom settings, answering the questions can trigger formative feedback and or remedial teaching that can help improve student understanding (Crouch and Mazur, 2001; Malik and Sime, 2022). MCQs can also be used as concept inventories, as commonly used within some disciplines (Steif and Dantzler, 2005; Hestenes, Wells and Swackhamer, 1992) to test student knowledge for assessment of learning gain. In a large cohort, they can be an efficient way to assess students. Some pitfalls include not being able to prevent a student from guessing an answer. Clarity in the wordings used is yet another area where MCQs can go wrong, due to learner interpretations. Their suitability for assessing higher-order thinking skills assessment (Bloom et al., 1956), may be more sensitive to the quality of questions (Javaeed, 2018; Liu et al., 2023).

Reimagining MCQ creation by augmenting the human in the loop

When augmented by GenAI chat tools, which have access to vast amounts of information, including pedagogical theories, a human user could design robust questions. Including, creating questions that assess student knowledge and understanding of solving linear equations, by juxtaposing options that represent common mistakes or cognitive conflicts, which can trigger reflections in students and encourage learning the topic. It also makes it simple to avoid options that are too obvious by simply prompting it to replace those with plausible answers. It is these abilities of GenAI tools that attracted us to explore creating MCQs. But GenAI can produce outputs that are incorrect, sometimes referred to as *hallucinations* (Ji et. al, 2023). It can learn from feedback given to it on the produced output (outcome supervision) and extra guidance given within the prompt (process supervision) and produce better outputs this way (Brown et al., 2020; Siontis et al., 2024). However, ultimately human agency is still needed in refining the final output. This central idea is summarised in the quote from Gemini (GenAI from Google) as well as graphically in Figure 1.

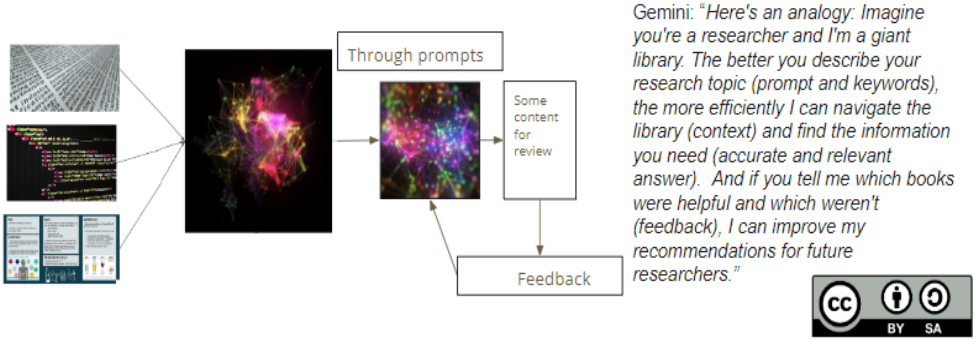

Figure 1. Iterative Process - Augmenting the human in the loop

To generate high quality MCQs, a lecturer may start with a learning outcome they want their students to demonstrate. A solved example in the case of mathematics or related questions in other subject domains could be used to trigger learning in the model and improve its outcome (Brown et al., 2020; Siontis et al., 2024). Defining the context and giving feedback on the outcome informs any future interactions within a chat thread (a single set of interaction). This gives it the opportunity to apply the learnt knowledge within the correct context. Figure 2 shows this process in further detail.

Prompt: Do you know how to solve linear equations with one variable?

Read through everything it says and find any errors. If there are errors, then prompt it to correct the same and recheck until you are happy, else ask if it can share some common mistakes students make in this topic.

Prompt: What are the common mistakes first year students make when solving these equations.

Check for any errors and give your own workings related to the mistakes identified if you agree and give some workings of mistakes students make that have not yet been identified but you know from your experience.

Prompt: Ok, now tabulate 5 questions on basic algebra (solving for x) and make sure that one column has the correct answer.

Again check for errors and use the prompt below adjusted for errors as needed.

Prompt: Show workings for Q2 first and tabulate again all 5 questions with corrected answers.

Next, generate distractors as below:

Prompt: For the same 5 questions use the misconception 1 highlighted above and workout the wrong answer, clearly showing the steps. I want you to produce wrong answers or distractors now. This will help later in the MCQ design.

Do the same for 2 other distractors and then ask it to produce the MCQs using the prompt:

Prompt: Now tabulate the question, the correct answer, three distractors (as you calculated above) based on the common mistakes students make.

One final check and manually correct any mistakes to finish the process. You could ask it to randomise the location of the correct answer.

Figure 2. Prompts and iteration outline for creating MCQs based on common misconceptions.

Examples of work produced

Figure 3 shows the output from the iterative process in Figure 2.

Question	Correct Answer	Distractor 1	Distractor 2	Distractor 3
1. Solve for x: $2x + 5 = 11$	$x = 3$	$x = 6$	$x = 8$	$x = -3$
2. Solve for x: $3(x - 4) = 21$	$x = 11$	$x = 33$	$x = 10$	$x = -2$
3. Solve for x: $\frac{1}{2}x - 3 = 5$	$x = 16$	$x = 8$	$x = -2$	$x = 20$
4. Solve for x: $4 - 3x = 13$	$x = -3$	$x = 9$	$x = -6$	$x = 4$
5. Solve for x: $\frac{x}{4} + 7 = 10$	$x = 12$	$x = -18$	$x = 20$	$x = 5$

Figure 3. ChatGPT 3.5, for creating MCQs for solving variable linear equations (some errors left deliberately).

WolframAlpha (WRA) Chat-enabled Cloud Notebooks may be used to address calculation errors where these persist. We tried this and Figure 4 shows the questions

that were created by WRA. This version requires paid access to ChatGPT. As is evident from Figure 4, human review is still needed.

Question	Correct Solution	Distractor 1	Distractor 2	Distractor 3
$3x+5=17$	$x=4$	$x=\frac{17}{3}$	$x=12$	$x=5$
$2(x+4)=16$	$x=4$	$x=8$	$x=3$	$x=5$
$4x-7=5x+3$	$x=-10$	$x=-3$	$x=-2$	$x=-4$
$2(x-3)=10-x$	$x=\frac{16}{3}$	$x=10-x+6$	$x=16-x$	$x=\frac{11}{2}$
$6-2x=4x+8$	$x=-\frac{1}{3}$	$6=4x+8+2x$	$6=6x+8$	$x=-\frac{5}{2}$

Figure 4. ChatGPT 3.5 and WRA combined for creating MCQs for solving variable linear equations (some errors left deliberately).

Staff perspectives

So far, we have trained over 10 staff, some of whom have reported that an initial investment in time is needed to perfect the prompts. The process can then be repeated to create several more questions on the same topic. Reuse of these prompts can also make creating MCQs for other topics efficient, as long as the platform version is the same.

"I am glad that I have attended this training session" (Participant 1)

Initial thoughts [regarding topic 1]: was lengthy to get it correct, but once the commands have been played around with to ensure ChatGPT is creating MCQs of sufficient quality then it is easy to replicate to produce new MCQs on differing (sic) topics. (Participant 2)

Conclusion

GenAI was able to augment human users in producing MCQs from an ethical starting point by using a learning outcome, which is already available in the public domain. Human review and update are still needed to make the questions suitable for use by students. We notice that these tools seem to cut corners and take the *path of least effort* to deliver what was prompted. From our own experience, asking for 3 to 5 questions gives the tool less to do within its allocated quota of parameters, which may result in more accurate outcomes.

References

Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *Taxonomy of educational objectives: The classification of educational goals. Handbook* 1: Cognitive domain (pp. 1103-1133). New York: Longman.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., …Amodei, D. (2020). Language models are few-shot learners. Advances in neural information processing systems, 33, 1877-1901.

Chien, Y. T., Chang, Y. H., & Chang, C. Y. (2016). Do we click in the right way? A meta-analytic review of clicker-integrated instruction. *Educational Research Review*, 17, 1-18. https://doi.org/10.1016/j.edurev.2015.10.003

Crouch, C. H., & Mazur, E. (2001). Peer instruction: Ten years of experience and results. *American journal of physics*, 69(9), 970-977. https://doi.org/10.1119/1.1374249

Duong, D., & Solomon, B. D. (2023). Analysis of large-language model versus human performance for genetics questions. *European Journal of Human Genetics*, 1-3.

Gierl, M. J., Bulut, O., Guo, Q., & Zhang, X. (2017). Developing, analyzing, and using distractors for multiple-choice tests in education: A comprehensive review. *Review of Educational Research*, 87(6), 1082-1116. https://doi.org/10.3102/0034654317726529

Gupta, P., Ding, B., Guan, C., & Ding, D. (2024). Generative AI: A systematic review using topic modelling techniques. Data and Information Management, 100066. https://doi.org/10.1016/j.dim.2024.100066

Haase, J., & Hanel, P. H. (2023). Artificial muses: Generative artificial intelligence chatbots have risen to human-level creativity. *Journal of Creativity*, 33(3), 100066. https://doi.org/10.1016/j.yjoc.2023.100066

Hestenes, D., Wells, M., & Swackhamer, G. (1992). Force concept inventory. The physics teacher, 30(3), 141-158.

Javaeed, A. (2018). Assessment of higher ordered thinking in medical education: multiple choice questions and modified essay questions. *MedEdPublish*, 7. https://doi.org/10.15694%2Fmep.2018.0000128.1

Jeblick, K., Schachtner, B., Dexl, J., Mittermeier, A., Stüber, A. T., Topalis, J., ... & Ingrisch, M. (2023). ChatGPT makes medicine easy to swallow: an exploratory case study on simplified radiology reports. *European radiology*, 1-9.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., and Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12), 1-38.

Little, J. L., & Bjork, E. L. (2015). Optimizing multiple-choice tests as tools for learning. *Memory & Cognition*, 43, 14-26.

Liu, Q., Wald, N., Daskon, C., & Harland, T. (2023). Multiple-choice questions (MCQs) for higher-order cognition: Perspectives of university teachers. *Innovations in Education and Teaching International*, 1-13.

Nicol, D. (2007). E-assessment by design: using multiple-choice tests to good effect. Journal of Further and higher Education, 31(1), 53-64.

Malik, M., & Sime, J. A. (2022). Orchestrating Learning Together and Development of Team-Trust in Neurologically Typical and Neurologically Atypical Students: A Multicase Study. *IEEE Transactions on Education*, 65(3), 320-330.

Morrison, S., & Free, K. W. (2001). Writing multiple-choice test items that promote and measure critical thinking. *Journal of Nursing Education*, 40(1), 17-24.

Steif, P. S., & Dantzler, J. A. (2005). A statics concept inventory: Development and psychometric analysis. *Journal of Engineering Education*, 94(4), 363-371.

Reed-Rhoads, T., & Imbrie, P. K. (2008, October). Concept inventories in engineering education. In *Proceedings from Evidence on Promising Practices in Undergraduate Science, Technology, Engineering, and Mathematics (STEM) Education Workshop* (Vol. 2, pp. 13-14).

Siontis, K. C., Attia, Z. I., Asirvatham, S. J., & Friedman, P. A. (2024). ChatGPT hallucinating: can it get any more human like?. *European Heat Journal*, 45(5), 321-323.