Journal of Learning Development in Higher Education ISSN: 1759-667X
Issue 27 April 2023

On ChatGPT: what promise remains for multiple choice assessment?

Chahna Gonsalves
Kings College London, UK

Abstract

Multiple-choice quizzes (MCQs) are a popular form of assessment. A rapid shift to online assessment during the Covid-19 pandemic in 2020, drove the uptake of MCQs, yet limited invigilation and wide access to material on the internet allow students to solve the questions via internet search. ChatGPT, an artificial intelligence (AI) agent trained on a large language model, exacerbates this challenge as it responds to information retrieval questions with speed and a good level of accuracy. In this opinion piece, I contend that while the place of MCQ in summative assessment may be uncertain, current shortcomings of ChatGPT offer opportunities for continued formative use. I outline how ChatGPT's limitations can inform effective question design. I provide tips for effective multiple-choice question design and outline implications for both academics and learning developers. This piece contributes to emerging debate on the impact of artificial intelligence on assessment in higher education. Its purpose is threefold: to (1) enhance academics' understanding of effective MCQ design, (2) promote shared understanding and inform dialogue between academics and learning developers about MCQ assessment, and (3) highlight the potential implications on learning support.

Keywords: multiple-choice quizzes; question design; formative assessment; artificial intelligence; large language models; ChatGPT; GPT-3.

Introduction

Multiple choice quizzes (MCQs) are widely used in online assessment in higher education, for both formative and summative purposes (Jin, Siu and Huang, 2022). While formative assessment can identify learning gains and gaps and provide feedback to learners to motivate them to improve their learning, summative assessments serve to measure, quantify, and accredit achievement against the learning outcomes (Mate and Weidenhofer, 2022). With the recent release of ChatGPT, an artificial intelligence (AI) agent trained on a large language model that responds quickly and largely accurately to multiple choices, is there still a place for multiple choice assessment? I argue that there is, at least for formative assessment.

Multiple-choice question design

A multiple-choice question consists of a stem, options, and auxiliary information (additional content in the stem or options, such as text, images, or audio) (Shin, Guo and Gierl, 2019). A stem is the context, or question being asked. The options typically comprise two-five possible answers, with four options found to be optimal (Gierl et al., 2017; Raymond, Stevens and Bucak, 2019). The options include the correct answer (or answers) amongst several plausible and partially correct but misleading options, which serve to distract students who do not have complete understanding of the subject (Shin, Guo and Gierl, 2019).

Distractors can be generated manually and by machine, by quantitatively and qualitatively analysing responses to open-ended questions (Haladyna and Rodriguez, 2013). More commonly though, it is up to subject matter experts to create a list of plausible but incorrect alternatives based on common misunderstandings and misconceptions (Haladyna and Rodriguez, 2013). For now, this expertise is key to multiple choice question writing because the educator can design questions based on a nuanced understanding of where and how these errors occur. While designing questions that test higher-order cognitive, algorithmic, and conceptual thinking skills is difficult, several guides exist to inform question design (Scalise and Gifford, 2006; Haladyna and Rodriguez, 2013; Domyancich, 2014; Haladyna, 2022) and question mapping against the levels of Bloom’s taxonomy (Allanson and Notar, 2019; Mate and Weidenhofer, 2022).

Benefits and challenges of MCQs

MCQs are well-favoured among academics because they enable objective testing, they scale easily for large cohorts, can be automated, and are quick to grade (Butler, 2018; Riggs, Kang and Rennie, 2020). Statistically analysing MCQ results can ensure the reliability and validity of the test (Stevens, Palocsay and Novoa, 2022), to guide curriculum and pedagogy adjustments and improve the quality of questions presented in the test. While research shows that MCQs are effective assessment tools that promote student learning (Butler, 2018), conducting these assessments online raises questions about their validity and academic integrity (Mate and Weidenhofer, 2022; Noorbehbahani, Mohammadi and Aminazadeh, 2022). For example, in a survey of undergraduate criminal justice students (n = 119), Burgason, Sefiha and Briggs (2019) found that the most common form of cheating was the use of multiple online sources during test taking. Given the limited empirical study accounting for intention to cheat, the level of student, and what is being assessed when the student has access to online sources, the impact of such access to online sources and the role of invigilation on the utility and validity of MCQ for summative assessment remain unclear.

Besides being quick to administer and grade, MCQs are ideal for providing students with timely or even immediate formative feedback on their learning, academics with data on achievement, and learning developers with data on student engagement. In this regard, the value of MCQs and the motivations to use them remain unchanged. However, creating multiple choice questions can be a difficult and time-consuming process. Therefore, many academics source questions from textbook test banks (Stevens, Palocsay and Novoa, 2022). Test bank questions have been criticised for testing only the most basic levels of understanding and recall, promoting rote memorisation (Simkin and Kuechler, 2005). Moreover, since test bank questions and their answers are widely available to academics, the probability that they can be found quickly through Google Search or on websites such as Chegg, CourseHero, or Spark Notes which provide students with solutions to textbook problems is high (Nguyen, Keuseman and Humston, 2020). While it is possible to create MCQs that assess higher-order levels of thinking, such as analysis and evaluation (Scully, 2017), which can mitigate against use of the internet to resolve the questions (Nguyen, Keuseman and Humston, 2020), academics receive limited training in MCQ design and find these types of questions challenging to write (Haladyna, 2022).

ChatGPT and the threat of AI

On 30 November 2022, the OpenAI foundation (https://openai.com/) made its AI chatbot, known as ChatGPT, available to the public for free. ChatGPT is an application of the latest version of GPT-3 (Generative Pretrained Transformer 3), a state-of-the-art large language processing model. Unlike traditional chatbots, the GPT-3 uses deep learning algorithms to generate human-like responses to user prompts. The GPT-3 model was trained using 570GB of data from web texts and books, equating to a text database of approximately 300 billion words. As a general-purpose dialogic agent with access to deep domain knowledge, ChatGPT provides detailed responses to a range of user prompts. For example, it can answer questions on a wide range of topics, provide detailed explanations, suggest ways to solve problems, respond to optimisation queries and write code (Haque et al., 2022). Specifically, ChatGPT's capacity to respond to multiple choice questions might just have annihilated the learning and assessment value of information retrieval MCQs.

A recent study found that ChatGPT could correctly answer up-to over 60% of MCQ test bank questions for topics covered in a United States Medical Licensing Examination – the standard exam used to licence doctors in the United States – indicating that ChatGPT could perform at the expected level of a third-year medical student (Gilson et al., 2022). The study also revealed that ChatGPT's response included information featured in the question stem in more than 90% of correct and incorrect responses, and correct answers were likely to contain information external to the question stem significantly more frequently than incorrect responses. This means avoiding repetition of terms in the question and answer options effectively hinders the AI in using a matching strategy to select the correct answer. Moreover, question stems that probe knowledge of material studied in class, as opposed to that found in the textbook, are likely to be more effective as ChatGPT can only answer questions related to data within its corpus.

Current limitations of ChatGPT's language model offer insight into how academics might write multiple choice questions that promote engagement and thinking – at least until Open.AI and other providers release the next generation of large language models. For example, Chat GPT cannot read visual media, only text. The current language model is only trained on data up to 2021 and ChatGPT cannot browse the web (Deng and Lin, 2022). Therefore, it cannot answer questions which include, or require the use of up-to-date information. Moreover, acknowledging its lack of current information which it considers to be important in responding to contextual questions, ChatGPT overlooks the concept or principle being assessed. In addition, when all options are plausible answers to a question and can be rationalised, ChatGPT struggles to make an evaluation, commenting on its need for further information to determine an answer. User prompts with many parts can be too complicated for the model, in which case, it may ignore some parts of the prompt entirely. The model can generate incorrect information due to misunderstanding the prompts and therefore returning incorrect responses. Whether ChatGPT selects the correct or incorrect answer, the model provides a very convincing logical explanation for the option selected. Suggestions for question design are presented below.

Tips for writing multiple-choice questions that stump ChatGPT:

1. Present questions using images, figures, or charts as auxiliary information, and a non-specific question as stem (for examples, see Mate and Weidenhofer, 2022). For example, 'which section of the figure below demonstrates. . . ?'
2. Present questions with auxiliary visuals as hotspot questions where the student must click on an area of the image to indicate the correct answer (for examples, see Joshi et al., 2020). For example, 'select the area on the image which shows . . .'
3. Present questions using a series of images, or a video accompanied with conditional logic branching questions. Conditional logic branching questions typically include several questions related to the same topic. MCQs with conditional logic branching questions are completed quickly as the student serially selects the correct response options but require reattempts at incorrect questions until the student demonstrates mastery of the content (Castro, 2018). For example, 'at this point in the interaction, which question should you ask the customer? Your response is incorrect, try again'.
4. Present questions that require the student to apply a concept or principle to an up-to-date scenario or case study. For example, 'the Higher Education Freedom of Speech Bill passed in the House of Commons in June 2020, but changes have been made and a clause removed while the Bill is under consideration at the House of Lords. Without this clause, what is the implication for a guest speaker who had been invited to give a lecture but found that a section had been censored in the recording?'
5. Use distractors that are all plausible, consistent in content and structure, and share important information with the correct option (Haladyna and Rodriguez, 2013; Gierl et

al., 2017). The plausibility of all distractors increases the need for the evaluation of all options to identify the correct answer.

Conclusions

This opinion piece first sought to evaluate the remaining value of MCQ assessment given the public accessibility of ChatGPT. I argued that ChatGPT has disrupted summative MCQ assessment practice in un-invigilated contexts. However, MCQs continue to be valuable for formative assessment providing they assess higher- rather than lower-order cognitive thinking skills, as these are the questions ChatGPT is currently challenged to answer. Shifting the focus of MCQs to test higher levels of cognitive skills has implications for the provision of support for students. Students may require greater support developing learning strategies and approaches that meet the needs of increasingly complex assessment.

Second, this opinion piece sought to outline how ChatGPT’s limitations can inform effective question design with the provision of examples. A shared understanding of ChatGPT’s capabilities can enhance learning developers’ perspective on assessment and their understanding of how MCQ might be used across assessment contexts. Such sharing of knowledge also contributes to the fostering of dialogue between learning developers and academics to understand effective approaches to supporting learning.

Third, motivations to use MCQ to measure and inform learning are unlikely to change, despite the threats access to the internet and ChatGPT pose in un-invigilated contexts. It may not be long before natural language processing (NLP) models become so intelligent that we can no longer exploit their weaknesses. Moreover, Google and Microsoft have yet to make their proprietary NLP models available which boast more flexibility and advanced features than ChatGPT (Rahaman et al., 2023). Nevertheless, better understanding of how these large language models work can bide us a little more time to use MCQ for formative assessment and improve our abilities accordingly.

References

Allanson, P. and Notar, C. (2019) 'Writing multiple choice items that are reliable and valid', *American International Journal of Humanities and Social Science*, 5(3), pp.1-9.

Burgason, K. A., Sefiha, O. and Briggs, L. (2019) 'Cheating is in the eye of the beholder: an evolving understanding of academic misconduct', *Innovative Higher Education*, 44(3), pp.203-218. https://doi.org/10.1007/s10755-019-9457-3.

Butler, A. C. (2018) 'Multiple-choice testing in education: are the best practices for assessment also good for learning?', *Journal of Applied Research in Memory and Cognition*, 7(3), pp. 323-331. https://doi.org/10.1016/j.jarmac.2018.07.002.

Castro, S. (2018) 'Google forms quizzes and substitution, augmentation, modification, and redefinition (SAMR) model integration', *Issues and Trends in Educational Technology*, 6(2), pp.4-14. https://www.learntechlib.org/p/188257/ (Accessed: 04 January 2023).

Deng, J. and Lin, Y. (2022) 'The benefits and challenges of ChatGPT: an overview', *Frontiers in Computing and Intelligent Systems*, 2(2), pp.81-83. https://doi.org/10.54097/fcis.v2i2.4465.

Domyancich, J. M. (2014) 'The development of multiple-choice items consistent with the AP Chemistry curriculum framework to more accurately assess deeper understanding', *Journal of Chemical Education*, 91(9), pp.1347-1351. https://doi.org/10.1021/ed5000185.

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

Gilson, A., Safranek, C., Huang, T., Socrates, V., Chi, L., Taylor, R. A. and Chartash, D. (2022) 'How well does ChatGPT do when taking the medical licensing exams?

The implications of large language models for medical education and knowledge assessment', *medRxiv*. https://doi.org/10.1101/2022.12.23.22283901.

Haladyna, T. (2022) 'Creating multiple-choice items for testing student learning', *International Journal of Assessment Tools in Education*, 9, pp.6-18. https://doi.org/10.21449/ijate.1196701.

Haladyna, T. M. and Rodriguez, M. C. (2013) *Developing and validating test items.* New York: Routledge.

Haque, M. U., Dharmadasa, I., Sworna, Z. T., Rajapakse, R. N. and Ahmad, H. (2022) '"I think this is the most disruptive technology": exploring sentiments of ChatGPT early adopters using twitter data'. *arXiv preprint*. https://doi.org/10.48550/arXiv.2212.05856.

Jin, K. Y., Siu, W. L. and Huang, X. (2022) 'Exploring the impact of random guessing in distractor analysis', *Journal of Educational Measurement*, 59(1)**,** pp.43-61. https://doi.org/10.1111/jedm.12310.

Joshi, A., Virk, A., Saiyad, S., Mahajan, R. and Singh, T. (2020) 'Online assessment: concept and applications', *Journal of Research in Medical Education & Ethics*, 10(2)**,** pp. 49-59. https://doi.org/10.5958/2231-6728.2020.00015.3.

Mate, K. and Weidenhofer, J. (2022) 'Considerations and strategies for effective online assessment with a focus on the biomedical sciences', *Faseb Bioadvances*, 4(1)**,** pp.9-21. https://doi.org/10.1096/fba.2021-00075.

Nguyen, J. G., Keuseman, K. J. and Humston, J. J. (2020) 'Minimize online cheating for online assessments during COVID-19 pandemic', *Journal of Chemical Education*, 97(9)**,** pp.3429-3435. https://doi.org/10.1021/acs.jchemed.0c00790.

Noorbehbahani, F., Mohammadi, A. and Aminazadeh, M. (2022) 'A systematic review of research on cheating in online exams from 2010 to 2021', *Education and Information Technologies***,** pp.1-48. https://doi.org/10.1007/s10639-022-10927-7.

Rahaman, M. S., Ahsan, M. T., Anjum, N., Rahman, M. M. and Rahman, M. N. (2023) 'The AI race is on! Google's Bard and OpenAI's ChatGPT head to head: an opinion article', *SSRN*. https://doi.org/10.2139/ssrn.4351785.

Raymond, M. R., Stevens, C. and Bucak, S. D. (2019) 'The optimal number of options for multiple-choice questions on high-stakes tests: application of a revised index for detecting nonfunctional distractors', *Advances in Health Sciences Education*, 24(1)**,** pp.141-150. https://doi.org/10.1007/s10459-018-9855-9.

Riggs, C. D., Kang, S. and Rennie, O. (2020) 'Positive impact of multiple-choice question authoring and regular quiz participation on student learning', *CBE—Life Sciences Education*, 19(2)**,** pp.ar16-9. https://doi.org/10.1187/cbe.19-09-0189.

Scalise, K. and Gifford, B. (2006) 'Computer-based assessment in e-learning: a framework for constructing "intermediate constraint" questions and tasks for technology platforms', *The Journal of Technology, Learning and Assessment*, 4(6). Available at: https://ejournals.bc.edu/index.php/jtla/article/view/1653 (Accesed: 14 March 2023).

Scully, D. (2017) 'Constructing multiple-choice items to measure higher-order thinking', *Practical Assessment, Research, and Evaluation*, 22(1)**,** pp.4. https://doi.org/10.7275/swgt-rj52.

Shin, J., Guo, Q. and Gierl, M. J. (2019) 'Multiple-choice item distractor development using topic modeling approaches', *Frontiers in Psychology*, 10**,** pp.825. https://doi.org/10.3389/fpsyg.2019.00825.

Simkin, M. G. and Kuechler, W. L. (2005) 'Multiple-choice tests and student understanding: what is the connection? *Decision Sciences Journal of Innovative Education*, 3(1)**,** pp.73-98. https://doi.org/10.1111/j.1540-4609.2005.00053.x.

Stevens, S. P., Palocsay, S. W. and Novoa, L. J. (2022) 'Practical guidance for writing multiple-choice test questions in introductory analytics courses', *INFORMS Transactions on Education*. https://doi.org/10.1287/ited.2022.0274.

Author details

Chahna Gonsalves is a Lecturer in Marketing (Education), where her research and scholarship focuses on teacher and student assessment and feedback literacy, communication, and message impact. She is a Senior Fellow of Advance HE.

Licence