EXCELLENCIA
INTERNATIONAL MULTI-DISCIPLINARY JOURNAL OF EDUCATION
Volume 4, Issue 1, 2026 ISSN (E): 117-121

Beyond the Bubble Sheet: How Generative AI May Finally Liberate Assessment from Multiple-Choice Testing

Sarvar Umirov
Senior lecturer at Tashkent State University of Law
E-mail: sarvar_work@inbox.ru

Abstract:
Multiple-choice questions (MCQs) have dominated educational assessment for decades, particularly in large-scale and high-stakes examinations. Their popularity rests on perceived efficiency, reliability, and objectivity. However, extensive research has highlighted their serious limitations, including construct underrepresentation, susceptibility to guessing, shallow measurement of understanding, and negative washback on learning. This article argues that recent advances in generative artificial intelligence (GenAI), such as large language models, create a realistic opportunity to move beyond MCQs toward open-ended, constructed-response assessment at scale. GenAI systems have the ability to preserve administrative efficiency by automating the evaluation of student-generated answers while significantly increasing validity and insight into student thinking. In this article, we discuss why MCQs remain deeply entrenched in higher education, the pedagogical and epistemic costs of this dependence, and how GenAI-enabled assessment is likely to transform the nature of student preparation and the assessment of learning. The article concludes by describing the circumstances under which GenAI-based assessment can be responsibly implemented and the issues that remain to be worked through.

Keywords: : multiple-choice questions, assessment, generative AI, automated scoring, higher education, learning-oriented assessment

Introduction

For much of modern educational history, multiple-choice questions (MCQs) have been the default instrument for large-scale assessment. From university entrance examinations to professional certification tests, the MCQ has been treated as a pragmatic solution to the challenges of testing large numbers of students efficiently and consistently. This is especially true for high-stakes use cases that require reliability, standardization, and the ability to pledge [1].

But during the same period, frustration with MCQs has been both unshakable and universal. For

decades, critics from all fields have targeted their failure to reflect deep understanding, reasoning processes, or meaningful learning outcomes. However, in spite of this criticism, MCQs are still a staple of educational systems across the world.

This article considers whether recent advances in generative artificial intelligence (GenAI) provide a true tipping point for assessment practice. Such selected-response formats have borne the brunt of a long-standing assumption that they are the only question types that can be reliably automated, but technologies such as large language models have now made it technically feasible to assess open-ended student responses at scale [2]. Which brings us to the essential question: if computer delivery methods can score constructed-response answers as well as humans and in the majority of cases more reliably and rapidly, why are we still so heavily dependent on MCQs for test and assessment purposes?

The article proceeds as follows. It first summarizes the fundamental shortcomings of MCQs with respect to measurement and learning. Third, it says that MCQs continue to be a staple of higher education and high-stakes testing in spite of these shortcomings. Third, it investigates how GenAI can facilitate an evolution towards open-ended assessment and why this evolution might change the way students learn. Last, it addresses challenges and safeguards required for

The Limitations of Multiple-Choice Questions

1. Guessing and the Illusion of Knowledge

One of the most fundamental problems with MCQs is that they permit correct answers without knowledge or understanding. A student with no mastery of the content can still select the correct option through random guessing or test-taking strategies. This issue is well documented in the assessment literature [3].

From an epistemic perspective, MCQs obscure student thinking. When a student selects an option, the assessor cannot know whether the choice reflects genuine understanding, partial knowledge, elimination strategies, or chance. As a result, the test score provides little diagnostic insight into what the student actually knows or misunderstands.

2. Construct Underrepresentation

MCQs are inherently limited in the types of constructs they can measure. While they may efficiently assess factual recall or recognition, they struggle to capture complex cognitive processes such as synthesis, argumentation, explanation, or problem-solving in authentic contexts [4].

In many disciplines, including law, medicine, and engineering, professional competence depends not on selecting a correct option but on generating, justifying, and applying knowledge. MCQs often reduce these complex abilities to simplified proxies, weakening construct validity.

3. Negative Washback on Learning

Assessment shapes learning behavior. When students know that exams consist primarily of MCQs, they adapt their study strategies accordingly. Previous studies of washback have found that MCQ-heavy assessment promotes surface learning strategies such as memorisation, pattern recognition and testwise strategies rather than conceptual understanding [5].

In universities around the world, professors is often found posting large banks of MCQs, sometimes with hundreds or thousands of questions, and stating that a subset will be included in the exam [6]. In such contexts, even students who achieve perfect scores may demonstrate minimal long-term learning.

4. Lack of Insight into Student Reasoning

Perhaps the most serious limitation of MCQs is their silence about student reasoning. An incorrect option provides no explanation of *why* a student failed, and a correct option provides no evidence of *how* the student arrived there. This lack of transparency limits the formative value of assessment and weakens feedback loops between teaching and learning [7].

Why MCQs Remain Dominant in Higher Education

Given these limitations, the persistence of MCQs may seem paradoxical. However, their dominance is not accidental; it is driven by structural, administrative, and legal factors.

1. Efficiency and Scalability

MCQs are inexpensive to administer and score. Optical scanners and digital platforms can process thousands of responses within minutes, making MCQs attractive for large cohorts. In contrast, constructed-response assessment has traditionally required extensive human labor, making it costly and slow.

2. Perceived Objectivity and Legal Defensibility

In high-stakes contexts, exam administrators prioritize assessments that appear objective and defensible. MCQs offer clear scoring rules: an answer is either correct or incorrect [8]. This binary logic allows institutions to claim impartiality and consistency, reducing the risk of legal disputes or appeals.

In contrast, human-scored open-ended responses are often criticized for subjectivity, inter-rater variability, and potential bias. Even when such concerns are exaggerated, they shape institutional risk management strategies.

3. Administrative Convenience Over Pedagogical Value

In many universities, assessment design is influenced more by logistical constraints than by learning theory. MCQs align well with centralized exam administration, large lecture formats, and limited grading resources. As a result, pedagogical concerns are often subordinated to operational convenience [9].

Generative AI and the Possibility of Constructed-Response Assessment at Scale

1. What Has Changed Technologically?

Recent advances in natural language processing have produced systems capable of analyzing, interpreting, and generating human-like text. Research on automated scoring of constructed responses predates GenAI by several, but earlier systems were limited in flexibility and transparency [10].

Large language models represent a qualitative shift. They can evaluate coherence, relevance, argument structure, and conceptual accuracy across a wide range of prompts. While they are not infallible, their performance has reached a level that makes large-scale constructed-response assessment technically plausible [11].

2. **From Selecting Answers to Producing Knowledge**

GenAI enables a fundamental redesign of assessment tasks. Instead of asking students to choose from predefined options, educators can ask them to:

1. Explain concepts in their own words.
2. Apply knowledge to novel scenarios.
3. Justify decisions or solutions.
4. Summarize, critique, or synthesize information.

Automated evaluation of such responses makes it possible to assess what students actually *know*, rather than what they can recognize or guess.

3. Transforming Washback and Study Behavior

Simply put, if exams test students on generating answers, as opposed to recognising them, then the study habits must change. Memorizing answer patterns becomes useless. Instead, students should be building conceptual knowledge, clarity of expression, and ability to reason (do science).

And this change has far-surpassing implications for learning. It would also mean assessment would reward not pattern recognition but real understanding. This could change curricula, pedagogy, and student expectations over time as the emphasis shifted from test scores to learning [12].

Addressing Concerns and Limitations of GenAI-Based Assessment

1. Validity and Reliability

Automated scoring systems must be rigorously validated. Scores must reflect the intended constructs, and model outputs must be monitored for systematic errors. GenAI should support, not replace, principled assessment design [13].

2. Transparency and Explainability

One challenge of GenAI is opacity. Institutions must ensure that scoring criteria are transparent and that students can understand how their responses are evaluated. Hybrid models, combining AI

scoring with human moderation, may be necessary, especially in high-stakes contexts.

3. Academic Integrity

Concerns about students using GenAI to generate answers are legitimate. However, this challenge is not unique to assessment; it reflects broader changes in knowledge production [14]. Task design, time constraints, in-class assessment, and oral follow-ups can mitigate misuse.

Implications for the Future of Assessment

The rise of generative AI challenges a long-standing assumption: that large-scale assessment must rely on selected-response formats to remain feasible and defensible. If constructed-response assessment can be automated responsibly, the pedagogical justification for MCQs weakens considerably.

This does not mean MCQs will disappear entirely. They may remain useful for limited purposes, such as rapid diagnostic testing or low-stakes checks of factual knowledge [15]. However, their dominance in high-stakes assessment is no longer technologically inevitable.

Conclusion

Multiple-choice questions have shaped educational assessment not because they are pedagogically ideal, but because they were administratively convenient. Their limitations, including guessing, shallow measurement, and negative washback, have long been recognized yet tolerated.

Generative AI offers a realistic opportunity to break this cycle. By enabling scalable evaluation of open-ended responses, it allows assessment to focus on what students truly know, understand, and can do. If implemented responsibly, this shift could transform not only assessment practices but also the very nature of student learning.

The question is no longer whether we *can* move beyond MCQs, but whether educational institutions are willing to prioritize meaningful learning over administrative convenience.

References

[1] D. Boud and N. Falchikov, "Rethinking Assessment in Higher Education," *Assess. Eval. High. Educ.*, vol. 32, no. 2, pp. 131–143, 2007.

[2] K. Scouller, "The Influence of Assessment Method on Students' Learning Approaches," *High. Educ.*, vol. 35, no. 4, pp. 453–472, 1998.

[3] T. M. Haladyna, S. M. Downing, and M. C. Rodriguez, "A Review of Multiple-Choice Item-Writing Guidelines," *Appl. Meas. Educ.*, vol. 15, no. 3, pp. 309–334, 2002.

[4] M. Birenbaum and K. K. Tatsuoka, "Open-Ended Versus Multiple-Choice Response Formats," *Appl. Psychol. Meas.*, vol. 11, no. 4, pp. 385–395, 1987.

[5] J. Biggs and C. Tang, *Teaching for Quality Learning at University*. McGraw-Hill, 2011.

[6] R. E. Bennett, "Formative Assessment: A Critical Review," *Assess. Educ.*, vol. 18, no. 1, pp. 5–25, 2011.

[7] H. D. Brown, *Language Assessment: Principles and Classroom Practices*. Pearson Education, 2004.

[8] A. J. Nitko and S. M. Brookhart, *Educational Assessment of Students*, 6th ed. Pearson, 2011.

[9] S. Messick, "Validity," in *Educational Measurement*, 3rd ed., R. L. Linn, Ed., Macmillan, 1989, pp. 13–103.

[10] M. D. Shermis and J. Burstein, *Handbook of Automated Essay Evaluation*. Routledge, 2013.

[11] S. M. Brookhart, *How to Create and Use Rubrics for Formative Assessment and Grading*. ASCD, 2013.

[12] L. W. Anderson and D. R. Krathwohl, *A Taxonomy for Learning, Teaching, and Assessing*. Longman, 2001.

[13] L. F. Bachman and A. S. Palmer, *Language Assessment in Practice*. Oxford University Press, 2010.

[14] D. M. Williamson, X. Xi, and F. I. Breyer, "A Framework for Evaluation and Use of Automated Scoring," *Educ. Meas. Issues Pract.*, vol. 31, no. 1, pp. 2–13, 2012.

[15] M. C. Rodriguez, “Three Options Are Optimal for Multiple-Choice Items,” *Educ. Meas. Issues Pract.*, vol. 24, no. 2, pp. 3–13, 2005.