education sciences

MDPI

Article

An Effectiveness Study of Generative Artificial Intelligence Tools Used to Develop Multiple-Choice Test Items

Toni A. May [1,*], Yiyun Kate Fan [1], Gregory E. Stone [2], Kristin L. K. Koskey [1], Connor J. Sondergeld [3], Timothy D. Folger [1], James N. Archer [4], Kathleen Provinzano [1] and Carla C. Johnson [5]

[1] Community Schools, College of Community and Public Affairs, SUNY—Binghamton University, Binghamton, NY 13902, USA; yfan3@binghamton.edu (Y.K.F.); kkoskey@binghamton.edu (K.L.K.K.); tfolger@binghamton.edu (T.D.F.); kprovinzano@binghamton.edu (K.P.)
[2] Educational Studies, Judith Herb College of Education, University of Toledo, Toledo, OH 43606, USA; gregory.stone@utoledo.edu
[3] MetriKs Amerique, Oswego, NY 13126, USA; c.sondergeld@metriks.com
[4] International Business Machines Corp. (IBM), New York, NY 10022, USA; j.laplante@ibm.com
[5] Science, Technology, Engineering, and Mathematics Education, College of Education, North Carolina State University, Raleigh, NC 27695, USA; carlacjohnson@ncsu.edu
* Correspondence: tmay3@binghamton.edu

Abstract: Generative artificial intelligence (GenAI) tools developed to support teaching and learning are widely available. Trustworthiness concerns, however, have prompted calls for researchers to study their effectiveness and for educators and educational researchers to be involved in their creation and piloting processes. This study investigated one type of GenAI created to support educators: multiple-choice question generators (MCQ GenAI). Among the nine MCQ GenAI tools investigated, a variety of useful options were available, but only one indicated teacher involvement and none mentioned testing experts in development processes. MCQ GenAI-created items (n = 270) were coded based on MCQ quality item-writing guidelines. Results showed 80.00% of items (n = 216) violated at least one guideline, with 73.70% (n = 199) likely to produce major measurement error (should not use without revision), 6.30% (n = 17) likely to elicit minor measurement error (consider modifying), and 20.00% (n = 54) acceptable (usable as created). Implications suggest multidisciplinary teams are needed in educational GenAI tool development.

Keywords: generative artificial intelligence; educational assessment; multiple-choice tests; validity evidence

Academic Editors: Aleksandra Klašnja-Milićević, Boban Vesin and Dunja Vrbaški

Received: 20 December 2024
Revised: 16 January 2025
Accepted: 22 January 2025
Published: 24 January 2025

Citation: May, T. A., Fan, Y. K., Stone, G. E., Koskey, K. L. K., Sondergeld, C. J., Folger, T. D., Archer, J. N., Provinzano, K., & Johnson, C. C. (2025). An Effectiveness Study of Generative Artificial Intelligence Tools Used to Develop Multiple-Choice Test Items. *Education Sciences*, *15*(2), 144. https://doi.org/10.3390/educsci15020144

1. Introduction

United States (U.S.) investment in artificial intelligence (AI) is accelerating rapidly (Maslej et al., 2024). Substantial efforts have been dedicated to the development of AI tools for use within educational spaces (Cope et al., 2020). Still, a significant but largely overlooked area of AI use is "extending the ability of teachers to implement challenging but well-proven pedagogical strategies that require extensive work to implement" (Mollick & Mollick, 2023, p. 2). Cognitive assessment development is one such underdeveloped area. Educators across grade levels often lack the skills and training necessary to create high-quality assessments to most effectively measure student learning (e.g., Caldwell & Pate, 2013; Coombs et al., 2018; Kruse et al., 2020; Sondergeld, 2014, 2018) and generative AI (GenAI) tools offer the capacity to provide needed support (Mollick & Mollick, 2023). Yet, teachers' use of AI in the classroom has been coupled with skepticism and valid concerns about the trustworthiness of product outputs Office of Educational Technology (OET, 2023).

Recently, the U.S. Department of Education Office of Educational Technology (OET, 2023) published an extensive report on the future of teaching and learning with AI. In their report, they cautioned that all AI is based on a model similar to a mathematical or financial model, and each AI model is an incomplete estimate of reality which can thereby result in the production of false facts. Therefore, if AI tools, including GenAI assessment tools, are to be successfully employed to support teaching and learning at any grade level, it is critical to determine where these models fall short and in what ways (OET, 2023; Selwyn, 2022; Williamson et al., 2023).

The *Standards for Educational and Psychological Testing* (*Standards*; AERA et al., 2014) outlined five sources of validity evidence to be used to measure against newly developed assessments. It is crucial that GenAI-produced tests be held to the same standard as those developed by humans and undergo rigorous validity evaluation (Kaldaras et al., 2024). While assessment types are vast, MCQ tests are the most commonly administered across nearly all grade levels due to their versatility (Brookhart & Nitko, 2008; Caldwell & Pate, 2013; Jia et al., 2020; McMillan, 2011; Miller et al., 2021; Popham, 2014), making them an ideal launching point for evaluating GenAI effectiveness in assessment development. By examining nine generative AI tools available online for educators to create multiple-choice questions (MCQs) from educational content, the purpose of this research was to facilitate an understanding of the AI landscape within assessment development in terms of user features and content validity evidence (alignment with best practices). Two research questions were addressed in this study:

RQ1. What were the characteristics of the nine MCQ GenAI used in this study?

RQ2. To what extent do MCQ GenAI-produced items align with MCQ quality item-writing guidelines?

2. Literature Review

2.1. High-Quality Assessment Importance and Development

Assessment is a fundamental element of effective pedagogy as it plays a vital role in the teaching and learning process (e.g., Black & Wiliam, 1998; Brookhart, 2020; Shepard, 2000). In practice, however, this is not an easy task. Developing high-quality assessments of student learning is a science that requires expertise across multiple domains to produce valid and reliable outcomes (AERA et al., 2014; Kisker & Boller, 2014; NRC, 2001). Rather than having content experts create assessments in isolation, "the assessment design process must be a truly multidisciplinary and collaborative activity with educators, cognitive scientists, subject matter specialists, and psychometricians informing one another during the design process" (NRC, 2001, p. 314). Previous assessment studies highlight the profound benefits and capacity to achieve high-quality outcomes through the use of multidisciplinary team collaborations in the development and validation of cognitive assessments (e.g., Bostic et al., 2017; Brijmohan et al., 2018; Severino et al., 2018; Sondergeld & Johnson, 2019). While the benefits of collaboration are manifold, the NRC (2001), suggested that disciplinary boundaries may be one of the greatest assessment design challenges to overcome.

Educators across grade levels, who are experts in their respective content areas, are routinely tasked with independently creating classroom assessments to measure their students' learning without the proper training and skills to do this effectively (Caldwell & Pate, 2013; Coombs et al., 2018; Kruse et al., 2020; Sondergeld, 2014, 2018). Creating teacher-made assessments that produce valid and reliable outcomes is paramount to effectively contribute to teaching and learning (Brookhart, 2020; Brookhart & Nitko, 2008; McMillan, 2011; Popham, 2014). In the context of educational assessment, strong validity refers to effectively measuring the intended learning objectives such that the inferences drawn from assessment results promote sound educational decision-making, while high reliability

suggests consistency of results (AERA et al., 2014; Kane, 2013). Without these qualities, assessment data may be inaccurate and could potentially lead to misguided actions related to instructional approaches or student learning. Therefore, educators must prioritize the development of high-quality assessments to ensure effective teaching and learning practices (Brookhart & Nitko, 2008; McMillan, 2011; Popham, 2014). It is possible that AI tools may offer a means of providing educators with support for assessment development and help fill the gap between educator pedagogical needs and ability (Mollick & Mollick, 2023). This can only be true, however, if AI tools demonstrate strong content validity evidence by producing effective items that are well-aligned with both educational content and best practices in assessment item-writing.

2.2. GenAI in Education and Assessment

GenAI is a promising form of artificial intelligence that specializes in creating content. It has been defined as "a technology that leverages deep learning models to generate human-like content (e.g., images, words) in response to complex and varied prompts (e.g., languages, instructions, questions)" (Lim et al., 2023, p. 3). Applications of GenAI are varied in educational settings spanning learning, teaching, assessment, administration, and research (Alasadi & Baiz, 2023; Chiu et al., 2023; Michel-Villarreal et al., 2023). For assessment in particular, GenAI has demonstrated clear application and capability to produce assessment items (Kurdi et al., 2020) and answer assessment questions (Gilson et al., 2023). Moreover, it has shown promises in facilitating formative assessments without significantly increasing teacher workload by (a) innovating assessment techniques and formats, (b) providing timely grading and continuous feedback, (c) adapting to learner ability and knowledge, and (d) improving accessibility and inclusivity of assessments (Boscardin et al., 2024; Mao et al., 2024; OET, 2023).

This emerging technology, however, still has notable limitations. The prevalence of hallucinations (i.e., production of false facts) demands human oversight, and the potential lack of content understanding means output generated may not reflect the true intricacies of the underlying domain (Ramos et al., 2024). Also, most GenAI tools developed for use in educational spaces have been created by individuals lacking pedagogical expertise (Luckin & Cukurova, 2019). Ethical considerations represent another significant challenge associated with GenAI applications in educational assessments, encompassing issues such as academic integrity, authenticity, potential biases, and privacy concerns (Baidoo-Anu & Ansah, 2023; Michel-Villarreal et al., 2023; Preiksaitis & Rose, 2023; Rahman & Watanobe, 2023). Over-reliance on GenAI and lack of guidance for its use in assessment practices may exacerbate these limitations. As a countermeasure, policies and guidelines have been established starting in higher education spaces and international organizations, promoting core principles such as transparency, justice and fairness, non-maleficence, responsibility, privacy, beneficence, and autonomy (Jobin et al., 2019; Adams et al., 2023). Many research-intensive institutions in the U.S., for instance, have guidelines in place to regulate GenAI use for educational purposes, with a focus on guiding instructors in their GenAI assessment design and practices (Moorhouse et al., 2023). Presently, practical ways to mitigate ethical concerns at the AI development and use levels are limited in the literature. Considering these limitations, there is a need for greater insight into how to construct effective assessments using GenAI (Chiu, 2024) and for the evaluation of AI-generated assessment validity evidence (Chiu, 2024; Kaldaras et al., 2024).

OET's (2023) comprehensive report on the future of AI in teaching and learning outlined perceived benefits, limitations, and seven constituent-informed policy recommendations for using AI in educational contexts. Recommendations summarized from listening sessions with more than 700 constituents include:

1. Emphasize Humans in the Loop—Central to all other recommendations is that educators should be co-designers and co-evaluators of AI systems, teach students about AI use, and serve as integral and responsible users of AI in teaching and learning.
2. Align AI Models to a Shared Vision for Education—Educational policy and decision-makers must keep learners' educational needs and priorities central in evaluating the appropriateness of an AI model for use in teaching and learning.
3. Design Using Modern Learning Principles—AI products should be designed using current evidence-based practices in education and be inclusive for diverse learners.
4. Prioritize Strengthening Trust—Build trust among constituents for how AI can support (not replace) educators in innovative teaching and learning.
5. Inform and Involve Educators—Prioritize engaging educators and researchers in evaluating AI through outreach efforts, involvement in the design and development of AI, and systematic, thoughtful integration of AI into existing programming.
6. Focus R&D on Addressing Context and Enhancing Trust and Safety—Researchers need to advance AI models to be more responsive to diverse learners and settings to broaden use and digital equality, strengthening the trust and safety of AI systems.
7. Develop Education-Specific Guidelines and Guardrails—Policymakers should work in parallel with constituents across levels of the educational system to inform local and federal laws addressing emergent privacy and security issues with AI systems.

One key component of Recommendation 5 that is relevant to educational research and evaluation is for researchers to study AI tools intended for classrooms to expand trust and form a better understanding of the conditions under which AI educational tools function well or break down. As such, the present study seeks to answer the OET call by investigating the landscape and effectiveness of one type of existing AI tool (MCQ GenAI) that has been developed by numerous entities for use in educational settings.

3. Theoretical Frameworks

3.1. Evaluating Assessment Validity Evidence

Within the social sciences, the *Standards* (AERA et al., 2014) are widely recognized as the definitive set of guidelines aimed at promoting fairness, validity, and reliability in the development, administration, and use of assessments. Integrating multiple sources of validity evidence to form a sound evidence-based validity argument in support of the specific uses and interpretations of test scores for a designated purpose is essential (Kane, 2006, 2013). The *Standards* highlight five key sources of validity evidence as important for gathering and evaluating data: content (Do items align with content they are intended to measure?), response processes (Do participants interpret items as developers intended?), consequences of testing (Do participants experience any adverse effects from completing the test?), internal structure (Do items function reliably as a unidimensional construct?), and relation to other variables (Are test outcomes associated with other variables as hypothesized?) (AERA et al., 2014). While the *Standards* have been widely used for supporting the development and validation processes of human-made assessments, recent calls have been made for GenAI-produced assessments to uphold the same practices (Kaldaras et al., 2024).

Examining multiple forms of validity evidence is necessary for building a robust validity argument (AERA et al., 2014; Kane, 2006, 2013). Collection and evaluation of content validity evidence serves as the foundational starting point in this process (AERA et al., 2014). Content validity evidence establishes the basis for subsequent validity studies by ensuring a test accurately represents the construct. As such, our study of GenAI-produced MCQ tests focused on the investigation of content validity evidence by looking for item alignment with both the academic content and best practices in MCQ item-writing.

3.2. MCQ Item-Writing Best Practices

A primary goal of classroom-based assessment is to evaluate students' knowledge, skills, and abilities (Chatterji, 2003). Classroom assessments should serve to differentiate between students who have mastered the content and those who have not. Discrimination for any other reason is unjust and inappropriate. No test can perfectly measure an individual's ability, however, because test scores consist of two components: an individual's true ability and measurement error (Thorndike, 2005). Measurement error, often referred to as "noise," represents information unrelated to a person's true ability and can result in scores that are not fully reflective of their ability. The degree and sources of measurement error can vary. For instance, measurement error may arise from factors unrelated to an individual's ability, such as fatigue, anxiety, motivation, emotional state, or health issues (Crocker & Algina, 2008). Unfortunately, test developers and administrators cannot anticipate or control these personal circumstances during testing.

Measurement error can also result from poorly written test items. Unlike test-taker-related factors, error brought into the test score equation from poorly written items falls squarely within the purview of test developers, who have an ethical responsibility to minimize error (AERA et al., 2014). Measurement error introduced from poorly written items works in one of two ways. It either provides an unfair advantage or disadvantage for correctly answering an item. An unfair advantage may occur if an item provides test-takers who do not know the content with information that allows them to guess the answer correctly. And an unfair disadvantage could arise if an item is worded in a confusing or misleading manner that causes test-takers who do have the requisite knowledge to answer the item correctly to select an incorrect response. If the goal of an assessment is to gain an accurate understanding of test-taker knowledge and learning, then measurement error (both favorable and unfavorable for examinees) must be reduced.

Operationally defined in Table 1 is a set of well-established research-based MCQ item-writing guidelines informed by multiple sources (see Brookhart & Nitko, 2008; McMillan, 2011; Miller et al., 2021; Popham, 2014). These guidelines focus on two principal aspects of an item, its academic content in relation to curricular materials and a test-taker's potential interpretation of the item. To apply the 16 guidelines in the current study, the research team classified the extent to which violating each guideline was deemed a major or minor issue (see Table 1). Major issues introduce fundamental measurement error that will likely lessen the meaning of the test-taker's performance outcome. Minor issues are structural item-writing errors that make inconsistent test-taker interpretation a possibility, but significant measurement error is less likely. By following these guidelines when developing MCQ test items, tests are more likely to generate psychometrically sound outcomes with minimal measurement error, such that examinee scores are more closely reflective of their true ability.

Table 1. MCQ Item-Writing Guidelines.

Guideline	Operational Definition	Relevance: Academic Content or Test-Taker Interpretation	Measurement Error Concern: Minor or Major
Content Aligned Stem	Item reflects the content presented within the text in a way that test-takers can answer the item.	Academic Content	Major
Plausible Distractors	Answer options are related to the content presented in the item and represent logically possible answers.	Academic Content	Major
One Correct Answer	There is a single, clearly correct answer. All other options are incorrect.	Academic Content	Major
Mutually Exclusive Options	Answer options do not include overlapping content (e.g., substantive terms which appear in more than one option).	Test-Taker Interpretation	Major

Table 1. *Cont.*

Guideline	Operational Definition	Relevance: Academic Content or Test-Taker Interpretation	Measurement Error Concern: Minor or Major
Avoid Use of Negatives	Answer options are phrased positively (e.g., do not include "except" or "not").	Test-Taker Interpretation	Major
Avoid All or Nothing	Choices do not include "all of the above" or "none of the above".	Test-Taker Interpretation	Major
Parallel Response Options	Answer options are similar content (if the correct answer is a type of transportation, then all options are modes of transport).	Test-Taker Interpretation	Major
Similar Option Length	All answer options should be similar in length, both literally and conceptually (refer to the same number of concepts).	Test-Taker Interpretation	Major
Avoid Ambiguous Language	Items should not include language easily open to interpretation (e.g., often, rarely, frequently).	Test-Taker Interpretation	Major
Present a Problem in Stem	Items should ask the test-taker a question not simply present a list of statements in options where the test-taker must read each option before knowing what is being asked (e.g., avoid "which of the following statements is true" items).	Test-Taker Interpretation	Major
Do Not Clue the Answer	Answer options should not include terms or symbols which suggest the correct answer by virtue of their similarity.	Test-Taker Interpretation	Major
Order Options Logically	Answer options should be arranged in an order that naturally follows (e.g., size, numerically).	Test-Taker Interpretation	Minor
Avoid Repeating Words in Options	Answer options should not include identical phrases that might be easy to include in the stem.	Test-Taker Interpretation	Minor
Do Not Teach in Stem	Items should present only information necessary for the test-taker to answer correctly; teaching is carried out during a lesson.	Test-Taker Interpretation	Minor
Simple, Direct, and Concise Language	Use language appropriate for the grade level and content (e.g., avoid the use of superfluously complicated language).	Test-Taker Interpretation	Minor
Appropriate Structure	Blanks do not belong within the stem; answer options should complete or answer the stem. Use consistent and appropriate grammar and punctuation.	Test-Taker Interpretation	Minor

Notes. Relevant Term Definitions: Academic Content = pertaining to a measurement error that may arise from a disconnection between information in an instructional text (criterion) and the information assessed in an item; Test-taker Interpretation = pertaining to an error that may arise due to the structure or wording of an item which may contribute to test-taker interpretation of what is being asked as more or less difficult. Measurement Error Concerns Defined: Major = a fundamental measurement error that will likely lessen the meaning of the student performance outcome; Minor = structural errors that make inconsistent test-taker interpretation of the item a possibility, but less likely to produce significant measurement error.

4. Methods

4.1. Research Design

Drawing from the field of medicine, this research utilized an effectiveness study design to investigate the use of MCQ GenAI-produced tests in real-world contexts rather than under optimal conditions used in efficacy studies (Gartlehner et al., 2006). Effectiveness studies are conducted under routine conditions with practical outcomes examined and can be used to inform both practice and policy. Further, we employed a parallel mixed methods research data transformation variant approach (Creswell & Plano Clark) to address the second research question. Specifically, the qualitative data (item content) were quantitized (Sandelowski et al., 2009) as detailed in the corresponding Data Collection and Analysis section.

4.2. Data Collection and Analysis

4.2.1. MCQ GenAI Landscape (RQ1)

Three researchers conducted an online search for MCQ GenAI tools that would allow the upload of content in PDF format because many educators have access to online textbooks or other curricular resources in PDF form. A total of 20 MCQ GenAI tools were identified and 9 were included in our sample as they allowed for PDF content uploads and produced accessible MCQ items with keys. A review of the nine MCQ GenAI tool websites was conducted using a document analysis approach as a systematic procedure for reviewing and evaluating documents with minimal researcher intervention (Bowen, 2009; Corbin & Strauss, 2008). Researchers first skimmed through all content (e.g., text, images, video clips) on the nine AI tool websites (i.e., electronic documents) to gain a basic understanding of the information provided. A structured coding scheme was established collaboratively by the team based on initial impressions and the research purpose. Relevant information from websites was selected for detailed re-reading, compared across documents, and classified using the predetermined coding scheme within three broad categories: (1) Descriptives—MCQ GenAI tool basic information; (2) Input Options—possible information users could input into the tool; and (3) Outputs—products derived from the tool for users. Qualitative data were graphically organized and descriptively synthesized to describe the current landscape of MCQ GenAI tools used in this study.

4.2.2. Evaluating MCQ GenAI Items (RQ2)

Researchers uploaded three different lessons (PDF format) from a college-level classroom assessment course into each of the nine MCQ GenAI tools and requested the production of test items. Lessons used in this effectiveness study were created by two of the researchers on this project and have been delivered in classroom assessment courses at multiple universities. Three distinctly different academic content lessons were selected for uploads to provide academic content variety in terms of complexity (i.e., less and more cognitively challenging for students) and format (i.e., less or more text-based, mathematical concepts, images included, example problems provided):

(1) *Standardized Testing*: high cognitive complexity requiring students to use synthesis and evaluation skills to master—multiple mathematical equations, many figures/graphs, context-based scenarios, and intricate decision-making based on synthesis of numerous datapoints;
(2) *Learning Objectives and Taxonomies*: moderate cognitive complexity requiring students to apply learned skills to demonstrate mastery—mostly text-based, involves nuanced terminology and application of these terms; and
(3) *Validity and Reliability:* low cognitive complexity requiring students to employ knowledge and comprehension skills to master—primarily text-based, focused on mostly straightforward definitions, limited and simplistic mathematical concepts.

For consistency purposes, 10 items were produced from each of the MCQ GenAI tools for the 3 academic content areas (30 total items per test generator), resulting in a total of 270 items for content validity evidence evaluation (90 items total per academic content area).

Items were collaboratively rated by three research pairs with at least one of the partners being a PhD trained psychometrician who has taught college-level classroom assessment courses using the same lessons (content expertise). Raters were trained in a group coding session using a set of items for calibration purposes. Study items were then pair-coded to enhance credibility (Lincoln & Guba, 1985). MCQ GenAI items were first evaluated for their alignment with best practices in MCQ item-writing (see Table 1) using *a priori* dichotomous codes of Met Criteria (0) or Did Not Meet Criteria (1) for each of the 16 guidelines. Next, individual items were assigned one of three holistic ratings based on the sum and type of

criteria not met: Acceptable = all guidelines followed; Minor Issues = at least 1 guideline with minor measurement error concern broken but 0 major; and Major Issues = at least 1 guideline with major measurement error concern broken. Descriptive statistics were computed to examine the extent to which the GenAI tools adhered to the guidelines. Graphical representations of the quantitized (Sandelowski et al., 2009) coded data were constructed to illustrate patterns in the GenAI outputs. Exemplar items were selected, presented, and annotated to bolster trustworthiness of the findings (Lincoln & Guba, 1985).

5. Results

5.1. RQ1—MCQ GenAI Characteristics

5.1.1. Descriptives

Fundamental tool information collected from the nine MCQ GenAI websites that met the inclusion criteria are depicted in Table 2. Each tool generally offered two to four plans with annual pricing ranging from USD 0 to USD 1200. Slightly more than half (n = 5, 55.56%) provided a free version. All indicated offering user support through email or an online contact form. Three (33.33%) listed brief technical information about the specific AI methods adopted. For instance, QuizWhiz described its workflow as "using natural language processing and machine learning algorithms . . . [which] extracts key information, identifies relationships between concepts, and generates multiple choice questions". Two-thirds (n = 6, 66.67%) of the tools did not mention that users should review AI outputs, while three (33.33%) tools cautioned users about the accuracy of outputs with a warning message such as "AI-generated answers and questions may be incorrect" (Exam Generator) and "Please make sure you have read through them [items] before using them in your materials" (QuestionWell). Only one (11.11%) MCQ GenAI website mentioned teacher involvement in the tool development process, and none indicated psychometrician support.

Table 2. MCQ GenAI Tool Descriptives.

MCQ GenAI Tool Name	# Plans: Annual Pricing	User Support	AI Tech Info Provided	AI Caution Noted	Teacher Involved in Development	Psychometrician Involved in Development
Exam Generator	2 Plans: USD 0–USD 144	Yes	No	Yes	Not Indicated	Not Indicated
Questgen	2 Plans: USD 0–USD 100	Yes	No	No	Not Indicated	Not Indicated
QuestionWell	4 Plans: USD 0–TBD	Yes	No	Yes	Yes	Not Indicated
Quizbot	3 Plans: USD 120–USD 1200	Yes	Yes	Yes	Not Indicated	Not Indicated
Quizgecko	3 Plans: USD 0–USD 79	Yes	No	No	Not Indicated	Not Indicated
QuizWhiz	3 Plans: USD 0–USD 384	Yes	Yes	No	Not Indicated	Not Indicated
Quizizz	2 Plans: USD 0–TBD	Yes	No	No	Not Indicated	Not Indicated
Revisely Quiz Maker	4 Plans: USD 60–USD 588	Yes	No	No	Not Indicated	Not Indicated
Testportal AI	4 Plans: USD 348–TBD	Yes	Yes	No	Not Indicated	Not Indicated

Notes. All results were based on an online search conducted in December 2024. # = Total number.

5.1.2. Input Options

Table 3 lists a variety of input options for the nine MCQ GenAI tools in this study. Due to the sampling approach, all tools accepted PDF uploads. All tools also allowed text entry.

Other lesser-reported upload formats included image, video, audio, web link, PowerPoint, and Microsoft Word. Upload limits ranged from 3000 words (QuizWhiz) to unlimited character uploads (Testportal AI). For all tools, there was some level of user input related to the number of items generated, ranging from ability to specify the exact number of items on a test or a range (e.g., "Fewer, Moderate, More"—QuizWhiz). Two (22.22%) of the tools allowed users to indicate the number of selectable response options ranging from three to five for MCQs. A third (n = 3, 33.33%) of the MCQ GenAI tools gave users the ability to determine the "blueprint" for the test generated through additional instructions such as adding "learning outcomes" (QuestionWell) and/or selecting "the best study area to focus on" (Exam Generator) from AI provided options after reading the uploaded text. More commonly, most (n = 7, 77.78%) of the tools enabled users to establish the cognitive level of items tested in some way. While Exam Generator allowed for the indication of the specific Bloom's (1956) Taxonomic level (a hierarchical educational model used to classify learning objectives) for individual items, all others with this feature provided an option to select a broad range of cognitive levels such as "Easy, Medium, Hard" (Questgen) or the intended grade level for the test (QuestionWell and Quizizz).

Table 3. MCQ GenAI Tool Input Options.

MCQ GenAI Tool Name	Upload Formats	Upload Limit	Upload Blueprint Option	Select Number of Items	Select Number of Options	Select Item Cognitive Levels
Exam Generator	Text, PDF, DOC, PowerPoint, URL	10,000 characters	Yes	Yes	No	Yes—Bloom's Taxonomy
Questgen	Text, PDF, URL, Video, Audio	100,000 words	No	Yes	Yes	Yes—Broad Range
QuestionWell	Text, PDF, DOC, PowerPoint, Image, URL, YouTube link	10,000 words	Yes	Yes	No	Yes—Grade Level
Quizbot	Text, PDF, DOC, Image, Video, Audio, URL	70 pages	No	Yes	Yes	Yes—Broad Range
Quizgecko	Text, PDF, DOC, URL	100,000 characters	Yes	Yes	No	Yes—Broad Range
QuizWhiz	Text, PDF, Image, URL	3000 words	No	Yes	No	Yes—Broad Range
Quizizz	Text, PDF, DOC, PowerPoint, image, URL, YouTube link	25 mb or 30 pages	No	Yes	No	Yes—Grade Level
Revisely Quiz Maker	Text, PDF, DOC, Image, Video	60,000 characters	No	Yes	No	No
Testportal AI	Text, PDF, DOC	No character limit	No	Yes	No	No

Notes. All results were based on an online search conducted in December 2024.

5.1.3. Outputs

Given the inclusion criteria, all nine MCQ GenAI tools provided a set of questions and selectable options, identified a correct answer, and were downloadable in various standard formats such as PDF, Microsoft Word, and JSON. Four (44.44%) of the tools in this sample provided analytic and reporting services for tests taken directly through their website, such as automatic grading (Exam Generator, Questgen, Revisely, and Testportal AI).

5.2. RQ2—MCQ GenAI Item Alignment with Assessment Best Practices

Evaluation of 270 items created across the nine MCQ GenAI tools and three academic content areas revealed 80.00% (n = 216) of items had at least one (and as many as eight) academic content- or structural-related issue that would likely lead to some varying degree

of measurement error if given to students in their original format. On average, individual MCQ GenAI tools produced 24 items (out of 30 items) that violated item-writing guidelines, with a range of 17–28 items (*SD* = 3.67 items). Most item-writing issues identified were holistically classified as "major" (*n* = 199, 73.70%), potentially leading to invalidation of student performance measures due to one or more significant error concerns. In practical terms, items with a major guideline violation should not be used without revision. Far fewer items were identified as possessing at least one "minor" (*n* = 17, 6.30%) problem and no major issues, making consistent student interpretation of the item challenging, yet unlikely to invalidate results. These items could be used, but revision(s) should be considered. A total of 54 (20.00%) items were holistically classified as "acceptable" by meeting all guideline criteria and could therefore be used in their MCQ GenAI-created format.

Figure 1 details which MCQ Item-Writing Guidelines were broken and ordered from most to least frequently observed in the data set. Great variation in the occurrence of guideline violations was found, ranging from 0.37% (1 item) to 30.00% (81 items). Exemplars from different MCQ GenAI tools are provided for the five most commonly identified guideline problems as each represented more than 10% of the total items that would likely elicit measurement error from poor wording. Some of the examples illustrate more than one item-writing guideline violation and are also discussed in these instances.

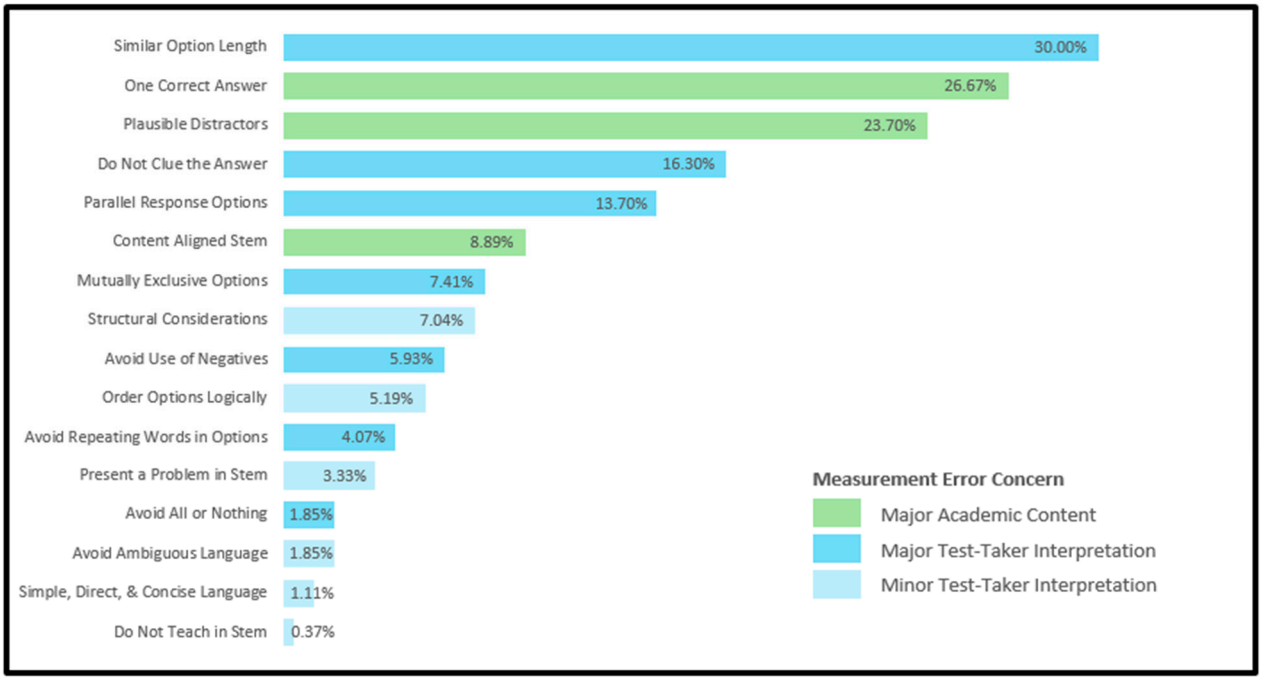

Figure 1. Percentage of Items Misaligned with MCQ Item-Writing Guidelines.

Similar Option Length was the least adhered to item-writing guideline. Nearly a third (*n* = 81, 30.00%) of the items evaluated across lessons and MCQ GenAI tools were coded as failing to meet this guideline. This item-writing guideline violation is quite commonly found in human-made assessments, both high-stakes and low-stakes, because test developers often find it easier to write more about the correct response than the distractors. Figure 2 is an example illustrating how a correct answer is both literally and conceptually longer than the incorrect response options (distractors). Correct answers that are longer than the distractors are oftentimes viewed as more attractive to test-takers who do not know the content. These test-takers may see the longer answer as being more specific or complex, and thus more credible. This can thereby lead to a potentially unfair guessing advantage for students who do not possess the knowledge needed to answer the

item correctly. Additionally, the example item in Figure 2 failed to meet criteria for *Avoid Use of Negatives* as the word "not" was in one of the response options. Negatives in the stem or response options can lead to measurement error for a number of reasons, including test-taker confusion, misinterpretation, cognitive overload, or misreading.

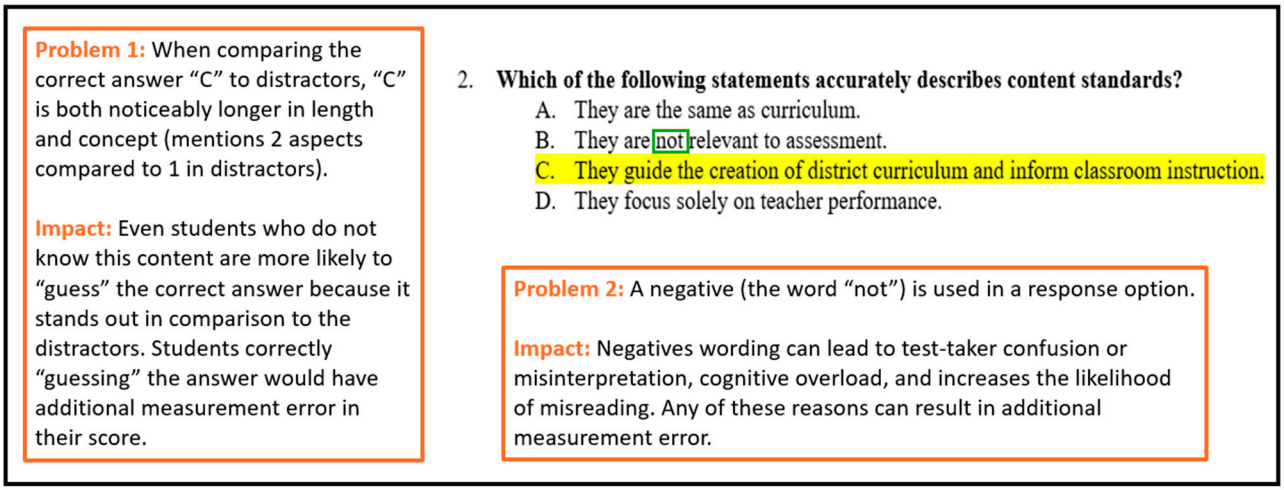

Figure 2. Similar Option Length and Avoid Use of Negatives Guidelines—Annotation of a Sample Misaligned Item. Note. MCQ item created from May & Stone's *Learning Objectives and Taxonomies* lesson uploaded into MCQ GenAI QuestionWell.

Having *Only One Correct Answer* was the second most common MCQ guideline that items did not follow (n = 72, 26.67%). This issue was observed in three ways. First, items with more than one correct answer, as presented in the Figure 3 exemplar, were coded for this violation as they did not have only one correct answer. Second, items with no correct answer from the possible response options fell into this category as they also did not have any correct answers. A third way items were identified for this error was if an incorrect response option was keyed as correct, but a correct answer existed as a distractor and was keyed as incorrect. Regardless of the classification reason, if revisions were not made to these items before administering, an unfair item would be distributed resulting in significant measurement error.

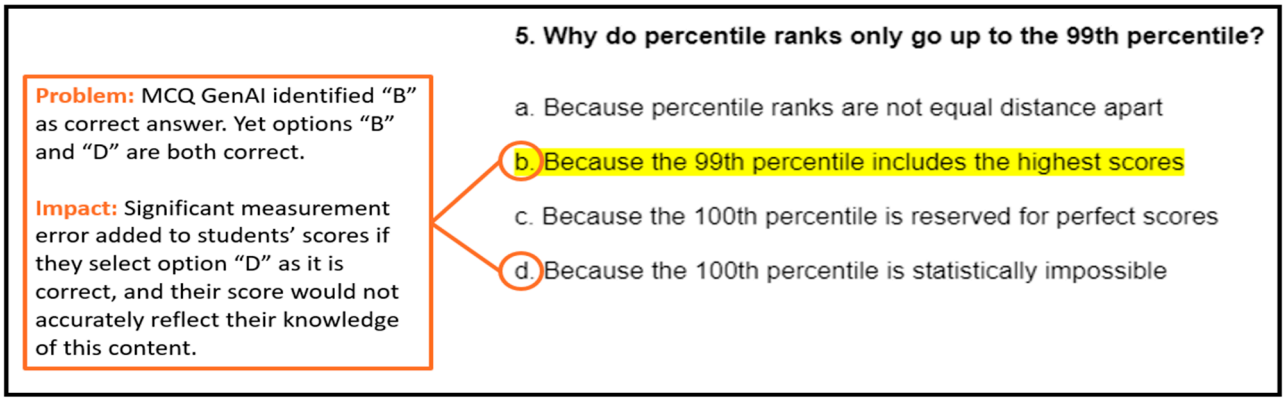

Figure 3. Only One Correct Answer Guideline—Annotation of a Sample Misaligned Item. Note. MCQ item created from May & Stone's *Standardized Testing* lesson uploaded into MCQ GenAI Exam Generator.

Plausible Distractors was another MCQ item-writing guideline that was not adhered to in almost a quarter of items (n = 64, 23.70%). While some of the items had only one distractor (or response option) that was not plausible, Figure 4 depicts a sample item with

no plausible options. This item asks a question that should render a "student" as the correct response. However, given response options given were all content areas "Math, Reading, Science, Social Studies". It is important to understand that as plausible distractors decrease, the likelihood for measurement error increases. Because this item had no plausible response options, it was also coded as violating the *Only One Correct Answer* guideline. Further, the item referred to a "given figure" that was not presented with the item, resulting in not meeting the criteria for *Content Aligned Stem* as participants would not be able to successfully approach the item based on information (or lack thereof) in the stem. To provide additional context, the MCQ GenAI tool that produced this item was focusing on content from a complex figure presented as an image embedded within the lesson. While all MCQ GenAI tools indicated they could read images in uploaded text, this particular program seemed to struggle with interpreting and applying the image's complex content. In its current form, this item would offer nothing other than measurement error to a test-taker's score.

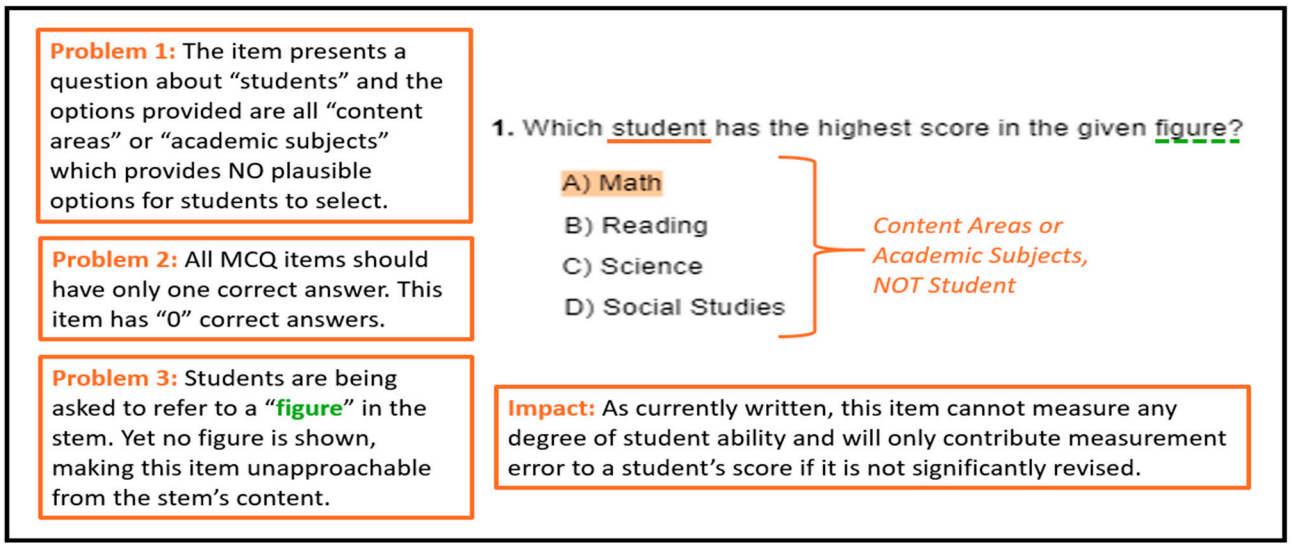

Figure 4. Plausible Distractors, Only One Correct Answer, and Content Aligned Stem Guidelines—Annotation of a Sample Misaligned Item. Note. MCQ item created from May & Stone's *Standardized Testing* lesson uploaded into MCQ GenAI QuizWhiz.

Do Not Clue the Answer was an item-writing violation that represented 16.30% (n = 44) of the problematic items flagged. Figure 5 shows that wording from the stem of the item is also in the correct answer ("inter" and "consistency"). Similar language is not used in the distractors, making the correct answer seem most viable due to its phrasing "clang" with the stem. As such, a test-taker who does not know the correct answer would have an unfair guessing advantage, thereby introducing significant measurement error.

Rounding out the top five major errors, 37 items (13.70%) generated across GenAI MCQ tools failed to meet the *Parallel Response Options* item-writing guideline. Parallel response options means if an item is asking about vegetables and the correct answer is a type of vegetable, then all distractors must be types of vegetables. In the Figure 6 example item, the correct answer is a component of a learning objective according to the lesson uploaded. All of the distractors are literacy skills and not general learning objective aspects. In this case, the MCQ GenAI tool that produced this item pulled direct language (pronunciation, spelling, phonemic awareness) from literacy learning objective examples presented in the lesson uploaded. However, the program failed to understand the nuanced difference between language used when writing learning objectives and more general learning objective components or structural terminology. Thus, students who do not know

the correct answer will likely have a higher probability of guessing correctly and increasing measurement error.

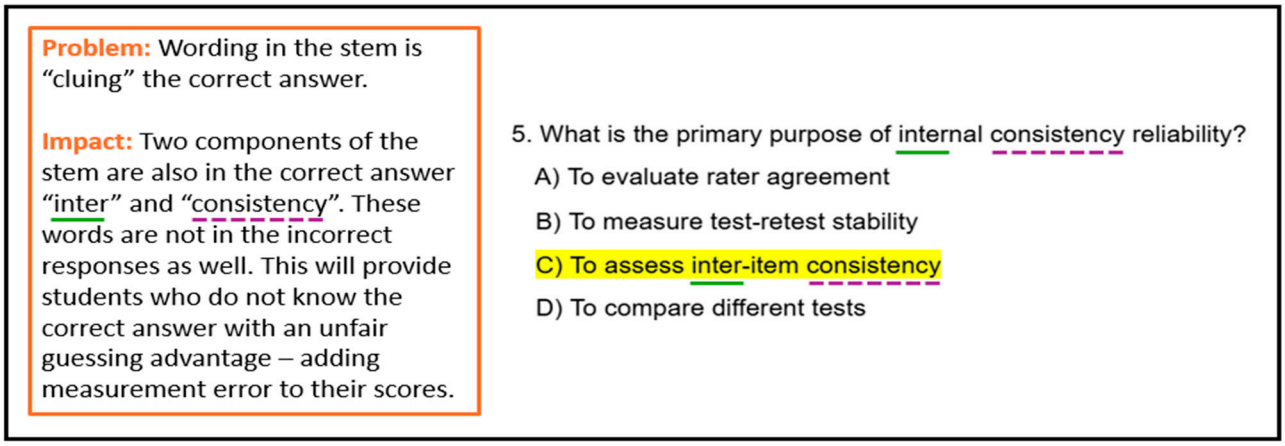

Figure 5. Do Not Clue the Answer Guideline—Annotation of a Sample Misaligned Item. Note. MCQ item created from May & Stone's *Validity and Reliability* lesson uploaded into MCQ GenAI Questgen.

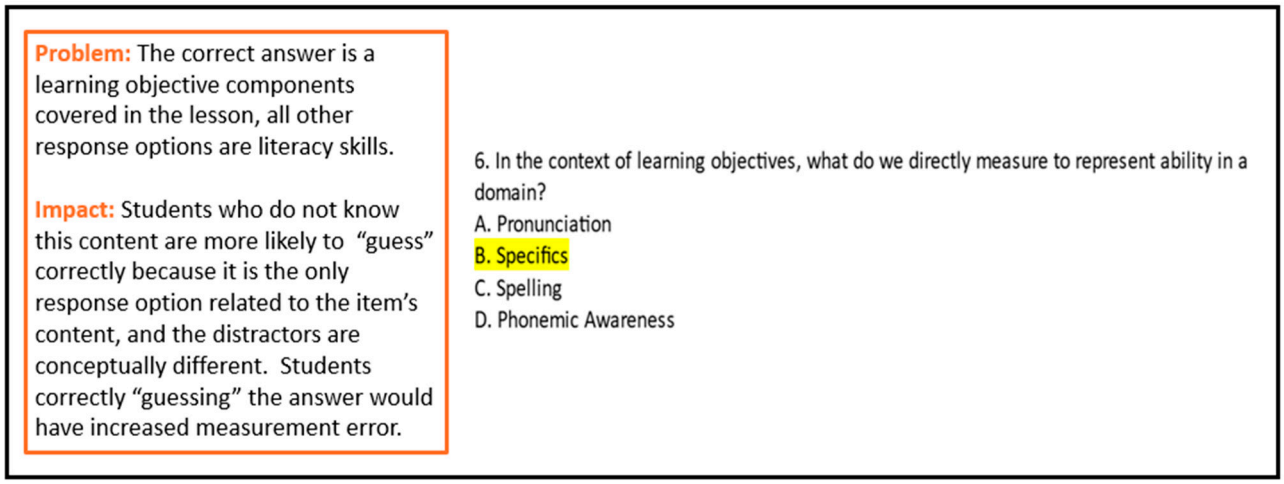

Figure 6. Parallel Response Options Guideline—Annotation of a Sample Misaligned Item. Note. MCQ item created from May & Stone's *Learning Objectives and Taxonomies* lesson uploaded into MCQ GenAI Questgen.

6. Discussion

This MCQ GenAI effectiveness study directly addressed calls to explore how such tools can both support educators in implementing challenging pedagogical strategies (Mollick & Mollick, 2023) and better understand where these tools may excel and fall short (OET, 2023). Findings from the landscape analysis component of this study revealed that MCQ GenAI tools are widely available and offer numerous useful features. However, when the generated output (MCQ tests and keys) was investigated for content validity evidence, as called for by the field (Kaldaras et al., 2024), results were less favorable due to significant measurement error that poorly written items would likely introduce into student ability calculations (Thorndike, 2005).

It is important to note that this study was purposefully delimited in three ways: (a) it examined only MCQ GenAI tools because MCQ tests are the most widely implemented across contexts; (b) the study employed only GenAI tools that accepted PDFs and could generate tests with corresponding keys to ensure practical applicability for educators; and (c) only lessons drawn from our team's intellectual property that were developed for a single college-level course were included to align with content expertise and avoid any

potential copyright concerns. Future research on GenAI-created assessment features and effectiveness in education should broaden the scope by including diverse content areas and investigating additional assessment types. Yet, even with our intentionally narrow focus, valuable insights can be gleaned related to evidence-based areas of MCQ GenAI tool promise and needed growth in educational settings.

6.1. User Accessibility and Utility

Availability of multiple, free MCQ GenAI tools which were straightforward to use and offered various options relevant to practicing educators is promising. In particular, common beneficial features, including the ability to select the number of items created and indicating the cognitive level at which test items should be crafted (e.g., grade level, Bloom's Taxonomy), may assist educators in generating appropriate assessments. Most MCQ GenAI tools, however, did not allow users to provide a content blueprint identifying which areas of the uploaded materials to emphasize nor to specify the number of response options for items. Both of these aspects are important for educators to consider when developing an assessment of their students' learning. Use of an assessment blueprint allows for a more tailored product based on the specific academic content covered (Brookhart & Nitko, 2008), rather than the content deemed important or easy to write items about by the MCQ GenAI tool, and stands to increase content validity evidence. While some research suggests three response options as optimal (cf., Rodriguez, 2005), there is no required number for an MCQ item. The number of response options is instead determined by multiple factors including test-taker characteristics (e.g., grade level, reading ability) and content appropriateness. Some content may require more or less than the common four response options most of the MCQ GenAI tools in the current study produced to align with the *Plausible Distractors* guideline. For example, if posing a question about the main types of rock, only three responses would most likely be plausible and appropriate (igneous, sedimentary, and metamorphic), regardless of the grade level. Yet, if presenting an item where days of the week were being assessed, a test-creator might want to include seven options.

Some of the MCQ GenAI websites claimed to perform automatic grading, run analytics, and present reports if the tests were administered in their platform. This full-service model of test development, scoring, analyzing, and reporting may indeed be of interest to some users. However, it is worth noting that none of the tools reportedly implementing these services indicated teachers were involved in their development. While studying the functionality of features such as these was outside the scope of this study, they should be investigated further to determine their effectiveness and if the outputs are in fact useful to end-users such as practicing educators (OET, 2023).

6.2. Transparency and Human-Centeredness

Themes of trust and transparency were threaded throughout the OET's (2023) report on the future of AI in teaching in learning. The authors explicitly called for AI developers of classroom tools "to be forthcoming about the [AI] models they use so that the marketplace can function on the basis of information about AI models and not only by the claims of their benefits" (OET, 2023, p. 56). With this appeal for action in mind, we intentionally investigated the websites associated with each of the MCQ GenAI tools in our study for language related to AI technical information. Unfortunately, most fell short in this area at the time of this review, with only three of the nine tools investigated sharing even limited information about the AI language or technologies used in production of their model. We could speculate on any number of reasons why this information was lacking, ranging from proprietary concerns to a desire to not overwhelm users without technical backgrounds. Regardless of the justification, these tools are not meeting basic calls for demonstrating

transparency and building trust with end-users, which can be easily fixed with the addition of accessible technical language on websites.

Another method of gaining user trust is by acknowledging AI tool limitations related to hallucinations (i.e., inaccurate or misleading output) (OET, 2023; Ramos et al., 2024). To evaluate the MCQ GenAI tools in this study for a more human-centered approach to using AI, our research team explicitly looked for AI cautionary language on tool websites. Again, only a third of the websites or generated tests had this type of language posted at the time of our study. We find this particularly concerning because 80% of the items produced across MCQ GenAI generators were flawed to some degree when evaluated for their alignment with best practices in MCQ item-writing. Further, nearly three-quarters of the items created had major item-writing violations. Academic content issues (*Only One Correct Answer*, *Plausible Distractors*, and *Content Aligned Stem*) cumulatively accounted for more than 59% of major violations, indicating inaccurate content (or misinformation) was produced by these tools as they failed to fully understand the instructional domains (Ramos et al., 2024), resulting in significant measurement error introduced into a test-taker's score. The importance of informing end-users of the need to verify output for accuracy and not over-trust the system must not be understated, especially in situations where the intent is to use AI-generated output to measure test-taker learning and inform educational decisions. Again, greater levels of transparency may be achieved by simply incorporating language on websites, or the actual assessments produced, indicating the need for human review. Small actions, such as this, will lay the foundation for building trust between GenAI tool developers and end-users.

7. Recommendations for Test Developers and Educational Policy Makers

Findings from our study demonstrate that simply selecting or purchasing an MCQ GenAI platform to use, uploading instructional content, and trusting these tools to produce high-quality items for effectively measuring test-taker learning would be a mistake at this point in time. Test items derived from MCQ GenAI tools may certainly serve as a starting point for test developers, saving both time and resources, but it is clear that further refinements to programming algorithms are necessary. As such, users must be cautioned of potential MCQ GenAI tool pitfalls and engage in thorough review of the output (assessments) produced.

Our experienced assessment development and validation team believes at its core that individuals designing tests—whether solely through human power or with assistance from GenAI—should complete extensive assessment development training or consult a test development expert (psychometrician). We also recognize that our lofty aspirations for all test developers is not feasible. In the absence of intensive assessment development and validation training, or access to expert support, we offer practical suggestions for reviewing MCQ GenAI-produced tests that are rooted in evidence and will likely reduce measurement error introduced from poorly written items.

(1) Question the Content: Nearly 60% of the MCQ GenAI test item-writing errors identified in our study were related to a discrepancy between GenAI's understanding of the academic content uploaded (as demonstrated through the item it produced) and the actual meaning of the content. At a minimum, users of MCQ GenAI tools should ask four basic questions relevant to each item's stem and response options produced to ensure an MCQ GenAI tool fully grasped the content. (a) Is the item's overall content aligned with what was taught in the classroom, module, or lesson? (b) Is the item keyed correctly? (c) Is there one, and only one, correct answer? (d) Are all of the response options logical?

(2) Look at Length: Approximately a third of the items in this study failed to adhere to the item-writing guideline of providing options similar in length. As previously stated, extensive prior work has highlighted this as an error found frequently in tests created by humans, regardless of whether they are for high- or low-stakes purposes. MCQ GenAI test development tools appear to also fall short in addressing this issue. Option length dissimilarity is particularly easy to spot. Look to see if the correct answer is noticeably longer than the distractors and adjust so all options are more similar in length. Additionally, look for the word "and" in the correct response but not in the distractors to ensure responses are conceptually similar in length.

(3) Be Mindful of your Materials: MCQ GenAI test generators examined in this study indicated they were capable of reading content from figures and images, and our data confirm this. However, our results suggested that these tools often struggled to correctly make sense of the material within these figures and images when the content was complex. Further, the test generators often produced items that referred to these figures or images without including them for test-takers to view, rendering the item difficult or even impossible to answer. Nuanced or higher-level content in lessons also posed a challenge for GenAI to understand. At this time, we suggest test developers carefully consider the content they plan to upload into MCQ GenAI tools. Simpler, more straightforward text may yield more favorable results when compared to abstract or nuanced material with complex figures or images, as GenAI may find it difficult to interpret and learn from these types of lessons.

8. Concluding Thoughts: Collaboratively Moving GenAI Assessment Development Forward

GenAI technology advancement offers the promise of assisting educators in the successful development and implementation of effective pedagogical strategies that are often viewed as difficult or time-consuming to create and employ (Mollick & Mollick, 2023; OET, 2023). Generation of high-quality assessments to evaluate student learning falls into this category (see Caldwell & Pate, 2013; Coombs et al., 2018; Kruse et al., 2020; Severino et al., 2018; Sondergeld, 2014). Yet, our present research demonstrated numerous and significant shortcomings in the output from the MCQ GenAI tools studied, highlighting a strong need for improvement before tests created from these tools could be implemented in classrooms and have a hope of producing quality measures of student learning. We do not attribute this failure to a lack of technological advancement in GenAI because tools in this study did produce a small percentage (20%) of acceptably written items. This alone illustrates the groundwork for programming MCQ GenAI tools exists. Current AI models, however, must be better programmed and refined based on the well-established and long-researched best practices in MCQ item-writing (see Brookhart & Nitko, 2008; McMillan, 2011; Miller et al., 2021; Popham, 2014).

Findings from our study suggest that MCQ GenAI companies simply may not be aware of these best item-writing practices, as none indicated having a psychometrician involved in their item generator development team. While it would seemingly be easy for MCQ GenAI companies to have programmers program their tools with the guidelines presented in our manuscript or any other classroom assessment textbook, we propose forming multidisciplinary teams of key constituents including educators, psychometricians, and programmers (OET, 2023) to produce the highest-quality assessment outcomes (Luckin & Cukurova, 2019; NRC, 2001; OET, 2023). Such collaborations would allow for the testing of produced assessments to move beyond a foundational content validity evidence review and into more robust testing of other forms of validity evidence in the age of AI (AERA et al., 2014; Kaldaras et al., 2024). Without this careful attention to quality, the benefits of

speed and automation will continue to be undermined by outputs that fail to effectively measure student learning, rendering these assessments meaningless at best and harmful at worst, depending on educator use.

Author Contributions: Conceptualization, T.A.M., G.E.S., C.J.S., K.L.K.K., J.N.A., K.P., and C.C.J.; methodology, T.A.M., G.E.S., and K.L.K.K.; validation, J.N.A., C.C.J., and K.P.; formal analysis, T.A.M., G.E.S., Y.K.F., C.J.S., K.L.K.K., and T.D.F.; data curation, C.J.S. and Y.K.F.; writing—original draft preparation, T.A.M., Y.K.F., G.E.S., K.L.K.K., C.J.S., and J.N.A.; writing—review and editing, T.A.M., Y.K.F., G.E.S., K.L.K.K., T.D.F., J.N.A., K.P., and C.C.J.; visualization, T.A.M., Y.K.F., and C.J.S.; supervision, T.A.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: A limited data set is presented within this manuscript as examples. Additional GenAI MCQ items produced from online tools used in this study are available upon request from the corresponding author.

Conflicts of Interest: Author James N. Connor J. Sondergeld was employed by the company MetriKs Amerique and Author James N. Archer was employed by the company International Business Machines Corp. (IBM). The remaining author declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
AERA	American Educational Research Association
DOC	Document
GenAI	Generative Artificial Intelligence
JSON	JavaScript Object Notation
MCQ	Multiple-Choice Question
N	Number
NRC	National Research Council
OET	Office of Educational Technology
PDF	Portable Document Format
RQ	Research Question
SD	Standard Deviation
URL	Uniform Resource Locator
US	United States

References

Adams, C., Pente, P., Lemermeyer, G., & Rockwell, G. (2023). Ethical principles for artificial intelligence in K-12 education. *Computers and Education: Artificial Intelligence*, *4*, 100131. [CrossRef]

Alasadi, E. A., & Baiz, C. R. (2023). Generative AI in education and research: Opportunities, concerns, and solutions. *Journal of Chemical Education*, *100*(8), 2965–2971. [CrossRef]

American Educational Research Association, American Psychological Association & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Baidoo-Anu, D., & Ansah, L. O. (2023). Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning. *Journal of AI*, *7*(1), 52–62. [CrossRef]

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice*, *5*(1), 7–74. [CrossRef]

Bloom, B. S. (1956). *Taxonomy of educational objectives, handbook: The cognitive domain*. David McKay.

Boscardin, C. K., Gin, B., Golde, P. B., & Hauer, K. E. (2024). ChatGPT and generative artificial intelligence for medical education: Potential impact and opportunity. *Academic Medicine*, *99*(1), 22. [CrossRef] [PubMed]

Bostic, J. D., Sondergeld, T. A., Folger, T., & Kruse, L. (2017). PSM7 and PSM8: Validating two problem-solving measures. *Journal of Applied Measurement*, *18*(2), 151–162. [PubMed]

Bowen, G. A. (2009). Document analysis as a qualitative research method. *Qualitative Research Journal*, *9*(2), 27–40. [CrossRef]

Brijmohan, A., Khan, G. A., Orpwood, G., Standford Brown, E., & Childs, R. A. (2018). Collaboration between content experts and assessment specialists: Using a validity argument framework to develop a college mathematics assessment. *Canadian Journal of Education*, *41*(2), 584–600.

Brookhart, S. M. (2020). Feedback and measurement. In S. Brookhart, & J. McMillan (Eds.), *Classroom assessment and educational measurement* (pp. 63–78). Routledge.

Brookhart, S. M., & Nitko, A. J. (2008). *Assessment and grading in classrooms* (5th ed.). Pearson.

Caldwell, D. J., & Pate, A. N. (2013). Effects of question formats on student and item performance. *American Journal of Pharmaceutical Education*, *77*(4), 1–5. [CrossRef] [PubMed]

Chatterji, M. (2003). *Designing and using tools for educational assessment*. Pearson.

Chiu, T. K. F. (2024). Future research recommendations for transforming higher education with generative AI. *Computers and Education: Artificial Intelligence*, *6*, 100197. [CrossRef]

Chiu, T. K. F., Xia, Q., Zhou, X., Chai, C. S., & Cheng, M. (2023). Systematic literature review on opportunities, challenges, and future research recommendations of artificial intelligence in education. *Computers and Education: Artificial Intelligence*, *4*, 100118. [CrossRef]

Coombs, A., DeLuca, C., LaPointe-McEwan, D., & Chalas, A. (2018). Changing approaches to classroom assessment: An empirical study across teacher career stages. *Teaching and Teacher Education*, *71*, 134–144. [CrossRef]

Cope, B., Kalantzis, M., & Searsmith, D. (2020). Artificial intelligence for education: Knowledge and its assessment in AI-enabled learning ecologies. *Educational Philosophy and Theory*, *53*(12), 1229–1245. [CrossRef]

Corbin, J., & Strauss, A. (2008). *Basics of qualitative research: Techniques and procedures for developing grounded theory* (3rd ed.). SAGE.

Creswell, J. W., & Plano Clark, V. L. *Designing and conducting mixed methods research* (3rd ed.). SAGE.

Crocker, L., & Algina, J. (2008). *Introduction to classical and modern test theory* (2nd ed.). Cengage Learning.

Gartlehner, G., Hansen, R. A., Nissman, D., Lohr, K. N., & Carey, T. S. (2006). *Criteria for distinguishing effectiveness from efficacy trials in systematic reviews.* U.S. Department of Health and Human Services. Available online: https://www.ncbi.nlm.nih.gov/books/NBK44024/ (accessed on 1 January 2025).

Gilson, A., Safranek, C. W., Huang, T., Socrates, V., Chi, L., Taylor, R. A., & Chartash, D. (2023). How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? The Implications of Large Language Models for Medical Education and Knowledge Assessment. *JMIR Medical Education*, *9*, e45312. [CrossRef] [PubMed]

Jia, B., He, D., & Zhu, Z. (2020). Quality and feature of multiple-choice questions in education. *Problems of Education in the 21st Century*, *78*(4), 576–594. [CrossRef]

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, *1*(9), 389–399. [CrossRef]

Kaldaras, L., Akaeze, H. O., & Reckase, M. D. (2024). Developing valid assessments in the era of generative artificial intelligence. *Frontiers in Education*, *9*, 1399377. [CrossRef]

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). American Council on Education.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1–73. [CrossRef]

Kisker, E. E., & Boller, K. (2014). *Forming a team to ensure high-quality measurement in education studies* (REL 2014-052). U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Analytic Technical Assistance and Development. Available online: http://ies.ed.gov/ncee/edlabs (accessed on 1 January 2025).

Kruse, L., Impellizeri, W., Witherel, C., & Sondergeld, T. A. (2020). Evaluating the impact of an assessment course on preservice teachers' classroom assessment literacy and self-efficacy. *Mid-Western Educational Research Association*, *32*(2), 107–132.

Kurdi, G., Leo, J., Parsia, B., Sattler, U., & Al-Emari, S. (2020). A Systematic Review of Automatic Question Generation for Educational Purposes. *International Journal of Artificial Intelligence in Education*, *30*(1), 121–204. [CrossRef]

Lim, W. M., Gunasekara, A., Pallant, J. L., Pallant, J. I., & Pechenkina, E. (2023). Generative AI and the future of education: Ragnarök or reformation? A paradoxical perspective from management educators. *The International Journal of Management Education*, *21*(2), 100790. [CrossRef]

Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic inquiry*. Sage.

Luckin, R., & Cukurova, M. (2019). Designing educational technologies in the age of AI: A learning sciences-driven approach. *British Journal of Educational Technology*, *50*(6), 2824–2838. [CrossRef]

Mao, J., Chen, B., & Liu, J. C. (2024). Generative artificial intelligence in education and its implications for assessment. *TechTrends*, *68*(1), 58–66. [CrossRef]

Maslej, N., Fattorini, L., Perrault, R., Parli, V., Reuel, A., Brynjolfsson, E., Etchemendy, J., Ligett, K., Lyons, T., Manyika, J., Niebles, J. C., Shoham, Y., Wald, R., & Clark, J. (2024). *The AI index 2024 annual report*. Institute for Human-Centered AI, Stanford University. Available online: https://aiindex.stanford.edu/report/ (accessed on 1 January 2025).

McMillan, J. H. (2011). *Classroom assessment: Principles and practices for effective standards-based instruction* (5th ed.). Pearson.

Michel-Villarreal, R., Vilalta-Perdomo, E., Salinas-Navarro, D. E., Thierry-Aguilera, R., & Gerardou, F. S. (2023). Challenges and opportunities of generative AI for higher education as explained by ChatGPT. *Education Sciences*, *13*(9), 856. [CrossRef]

Miller, D. M., Linn, R. L., & Gronlund, N. E. (2021). *Measurement and assessment in teaching* (11th ed.). Pearson.

Mollick, E., & Mollick, L. (2023). *Using AI to implement effective teaching strategies in classrooms: Five strategies, including prompts*. Wharton School of the University of Pennsylvania & Wharton Interactive. [CrossRef]

Moorhouse, B. L., Yeo, M. A., & Wan, Y. (2023). Generative AI tools and assessment: Guidelines of the world's top-ranking universities. *Computers and Education Open*, *5*, 100151. [CrossRef]

National Research Council (NRC). (2001). *Knowing what students know: The science and design of educational assessment*. The National Academies Press. [CrossRef]

Office of Educational Technology (OET). (2023). *Artificial intelligence and the future of teaching and learning: Insights and recommendations*. U. S. Department of Education. Available online: https://www2.ed.gov/documents/ai-report/ai-report.pdf (accessed on 1 January 2025).

Popham, W. J. (2014). *Classroom assessment: What teachers need to know* (7th ed.). Pearson.

Preiksaitis, C., & Rose, C. (2023). Opportunities, challenges, and future directions of generative artificial intelligence in medical education: Scoping review. *JMIR Medical Education*, *9*, e48785. [CrossRef] [PubMed]

Rahman, M. M., & Watanobe, Y. (2023). ChatGPT for education and research: Opportunities, threats, and strategies. *Applied Sciences*, *13*(9), 5783. [CrossRef]

Ramos, L., Yan, B., Khandabattu, H., & Rigon, G. (2024). *When not to use generative AI*. Gartner.

Rodriguez, M. C. (2005). Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. *Educational Measurement: Issues and Practice*, *24*(2), 3–13. [CrossRef]

Sandelowski, M., Voils, C., & Knafl, G. (2009). On quantitizing. *Journal of Mixed Methods Research*, *3*(3), 208–222. [CrossRef] [PubMed]

Selwyn, N. (2022). The future of AI and education: Some cautionary notes. *European Journal of Education*, *57*(4), 620–631. [CrossRef]

Severino, L., DeCarlo, M. J., Sondergeld, T. A., Ammar, A., & Izzetoglu, M. (2018). A validation study of an eighth grade reading comprehension assessment. *Research in Middle Level Education*, *41*(10), 1–16.

Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational Researcher*, *29*(7), 4–14. [CrossRef]

Sondergeld, T. A. (2014). Closing the gap between STEM teacher classroom assessment expectations and skills. *School Science and Mathematics Journal*, *114*(4), 151–153. [CrossRef]

Sondergeld, T. A. (2018, August 19–24). *Engineering effective assessments of student learning: How psychometrics improve measurement capacity*. XXVII International Materials Research Congress, Cancun, Mexico.

Sondergeld, T. A., & Johnson, C. C. (2019). Development and validation of a 21st Century Skills assessment: Using an iterative multi-method approach. *School Science and Mathematics Journal*, *119*(6), 312–326. [CrossRef]

Thorndike, R. M. (2005). *Measurement and evaluation in psychology and education*. Pearson.

Williamson, B., Macgilchrist, F., & Potter, J. (2023). Re-examining AI, automation and datafication in education. *Learning, Media and Technology*, *48*(1), 1–5. [CrossRef]