Schmauder, R., Daugherty, K., Rudolph, M., Hutson, B., McCarthy, K., McCauley, S., & Eagle, D. (2025). The art and science of item writing: A review of established guidelines for multiple-choice questions. *Intersection: A Journal at the Intersection of Assessment and Learning, 6(3), 222-239.*

# The Art and Science of Item Writing: A Review of Established Guidelines for Multiple-Choice Questions

**René Schmauder, Ph.D., M.B.A., Kimberly K. Daugherty, Pharm.D., Ph.D., Michael Rudolph, Ph.D., Bryant Hutson, Ph.D., Kelly McCarthy, Ph.D., M.B.A., Sarah McCauley, M.A., Dina Eagle, M.A.**

**Author Note**

René Schmauder, https://orcid.org/0009-0005-5097-9186
Kimberly K. Daugherty, https://orcid.org/0000-0003-2488-2098
Dina Eagle, https://orcid.org/0009-0003-2862-8756
Bryant Hutson, https://orcid.org/0000-0003-3180-563X
Kelly McCarthy, https://orcid.org/0009-0003-2992-5170
Sarah McCauley, https://orcid.org/0009-0000-0985-1184

We have no conflicts of interest to disclose

Intersection: A Journal at the Intersection of Assessment and Learning
*Vol 6, Issue 3, 2025*

**Abstract:** This review of guidelines for writing high-quality multiple-choice questions (MCQs) focuses on using MCQs to assess higher-order cognitive skills when designed effectively. While MCQs are widely used due to efficiency, poorly constructed items compromise assessment validity and student performance. Drawing from interdisciplinary research, we highlight best practices for item construction, including content alignment, stem clarity, distractor plausibility, and formatting consistency. We indicate common pitfalls and offer practical strategies to enhance the reliability, fairness, and diagnostic value of MCQs. By adhering to evidence-based principles, educators can create assessments that not only streamline assessment but also meaningfully measure student learning outcomes and higher-order cognitive skills.

**Keywords:** *multiple-choice questions, assessment design, higher-order cognitive skills, item-writing guidelines, educational measurement*

## Introduction

As class sizes rise and curricular responsibilities increase, instructors seek efficient approaches to assessment (Flaherty, 2020). Among the efficient options, multiple-choice questions (MCQs) stand out for their practicality, broad coverage of content, and ease of grading (Pate & Caldwell, 2014). When well-designed MCQs are utilized, they not only streamline administration and provide prompt feedback but also support the evaluation of higher-order cognitive skills (Mate & Weidenhofer, 2021). Research has shown that carefully constructed MCQs can assess complex cognitive abilities such as problem-solving, analysis, and application (Brady, 2005; Palmer & Devitt, 2007; Simkin & Keuchler, 2005; Zaidi et al., 2017). Moreover, MCQs can enhance content retention and improve performance on future assessments (Adesope, et al., 2017; Marsh, et al., 2007; Roediger & Karpicke, 2006). Students often

appreciate the familiar format of MCQs as they help clarify concepts, boost confidence, and reduce anxiety (Douglas et al., 2012; Loepp, 2021).

Despite these advantages, MCQs are not without criticism. Some concerns stem from poorly constructed items that fail to follow established guidelines for quality design (Burton, 2005; Downing, 2002; Downing, 2006; Palmer & Devitt, 2007). Other concerns address whether MCQs truly measure deep understanding of course material (Mehrens, 1987; Ozuru et al., 2013). These critiques highlight the importance of thoughtful question design to ensure that MCQs fulfill their potential as effective assessment tools.

While numerous resources outline principles for writing MCQs, this article synthesizes the best practices across disciplines and addresses common pitfalls often overlooked in everyday use. It offers instructors practical, research-informed strategies to enhance the validity, reliability, and fairness of assessments, ensuring that MCQs serve not only as efficient tools but also as meaningful measures of student learning.

## Methodology

This review employed a structured literature synthesis approach to identify and consolidate established guidelines for constructing high-quality multiple-choice items. The methodology was guided by principles of replicability and relevance to contemporary assessment contexts. A comprehensive search of peer-reviewed journals, educational measurement texts, and discipline-specific assessment literature. Particular attention was paid to sources within medical education, pharmacy education, business, and STEM fields, given their robust traditions of standardized testing and item development.

The selection process prioritized literature that offered evidence on effective item-writing practices, integration of cognitive theory into question design, and considerations of validity and reliability in assessment. Resources were evaluated for relevance to higher education and applicability to diverse learner populations. In addition to synthesizing guidelines, the review also examined theoretical frameworks and psychometric analyses that inform MCQ construction with a goal to bridge practical recommendations with foundational assessment principles.

## Cognitive Load Theory

The design of instructional activities and assessments plays a critical role in shaping a learner’s ability to understand and respond to information accurately and meaningfully. Consider John Sweller’s (1988) Cognitive Load Theory (CLT) foundational framework. Sweller suggests that human working memory has a limited capacity for processing new information, and that instructional materials and assessments should be designed to optimize this capacity. There are three types of loads a learner can experience. Intrinsic load measures a learner's interactivity with the information they are learning; it can range from low to high depending on the complexity of information or the task the learner is undertaking in relation to the amount of knowledge they possess on that subject (Mayer & Fiorella,

2021). Extraneous load results from activities that are not relevant to the learning at hand and are often the direct result of poorly designed instruction or assessments (Sweller et al., 2019). Germane load is the effort the learner makes to fit new information into existing cognitive frameworks. Both germane and intrinsic load are critical to the creation of new knowledge, commitment of information into long-term memory, and the learner's ability to retrieve it later (Krieglstein et al., 2023).

To enhance cognitive efficiency, both instructional materials and assessments need to be designed with CLT in mind. When MCQs are poorly constructed, like those with convoluted wording or irrelevant visuals, they can introduce extraneous cognitive load. This unnecessary mental effort diverts a student's attention away from the core task of retrieving the correct answer and applying their knowledge (Sweller, 1988). Therefore, the design of the question, including its stem and distractors, should aim to minimize this extraneous load, allowing students to focus on the essential task(s) (Krieglstein et al., 2023).

Another crucial principle in the application of learning is retrieval practice. MCQs, when well-designed, can serve as a powerful tool for reinforcing learning through active recall. The act of retrieving information from memory strengthens neural pathways, making that information more accessible in the future (Roediger & Butler, 2011). However, this benefit is only realized if the question requires genuine retrieval rather than simple recognition or guessing. The quality of the distractors is central to this process. If distractors are too easy to eliminate, the question becomes a low-stakes recognition task. Conversely, plausible distractors force students to engage in deeper cognitive processing and genuine retrieval.

### *Cognitive Taxonomies*

To ensure appropriate cognitive load, item development can be guided by Bloom's Taxonomy; a hierarchical model used for classifying levels of learning rigor. Traditional Bloom's Taxonomy (Bloom, 1956) outlines six categories of cognitive processes: knowledge, comprehension, application, analysis, synthesis, and evaluation. The cognitive process categories progress from lower- to higher-order cognition. Anderson and Krathwohl's (2001) revised taxonomy updated both structure and language to reflect more current understanding of cognition. By shifting to use of active verbs, including remember, understand, apply, analyze, evaluate and create, this revision emphasizes that learning is dynamic rather than static knowledge acquisition. Anderson and Krathwohl (2001) also differentiated between factual, conceptual, procedural, and metacognitive knowledge. Applying Anderson and Krathwohl's (2001) revision to item design encourages instructors to consider both the type of knowledge at issue as well as the cognitive processes being assessed. Regardless of the taxonomy or cognitive model employed (e.g. Fink, 2013; Marzano & Kendall, 2007), assessments must be aligned with the appropriate cognitive level of the skills and abilities being evaluated.

## The Importance of High-Quality MCQs

While taxonomies provide the theoretical foundation for aligning assessments with cognitive rigor, multiple-choice questions represent one of the most widely used tools for putting these principles into practice. Multiple-choice questions can effectively measure a wide range of constructs, including psychological traits, academic achievement, intelligence, knowledge, and skills. Their structured format

supports efficient test construction and grading, while also offering familiarity and clarity for students (Douglas et al., 2012; Haladyna, 2004; Pate & Caldwell, 2014). Beyond assessing factual recall, well-designed MCQs can evaluate the application of knowledge; determining whether students can transfer learned information to new situations or contexts (Anderson & Krathwohl, 2001). Moreover, MCQs can assess higher-order cognitive skills such as analysis, evaluation, and inference (Zheng et al., 2008). MCQs also serve as diagnostic tools, where distractor analysis can reveal common misconceptions, highlighting areas where instructional improvement is needed (Bridgeman, 1992; Haladyna et al., 2002). Additionally, performance data across MCQ subdomains can provide valuable feedback to both educators and learners, identifying specific strengths and weaknesses (Thissen & Mislevy, 2000).

## Significance of Quality Item Construction

While there is substantial evidence supporting the effectiveness of MCQ assessments, their utility depends heavily on thoughtful design and careful implementation. Poorly constructed MCQs can undermine assessment validity, leading to the misclassification of students as non-proficient. This may not be due to a lack of knowledge or skills, but rather because of flaws in the test items themselves (Breakall et al., 2019; Downing, 2005; Pate & Caldwell, 2014; Stagnaro-Green & Downing, 2006; Tarrant et al., 2006; Tarrant & Ware, 2008).

Research across disciplines has highlighted the consequences of such flaws. For instance, Downing (2005) reported that in a medical education context, 10-15% of students were incorrectly classified as failing due to flawed MCQs. These flawed items were also found to be 0-15% more difficult than well-constructed items. Similarly, Moncada and Moncada (2010) demonstrated that in business education, the quality of MCQs significantly influenced the accuracy of learning assessments, with poorly written items leading to distorted evaluations of student performance. Pate and Caldwell's (2014) study in pharmacy education found that 51.8% of 187 test items over four exams violated at least one area of item writing best practice rules, leading to a statistical difference in the percentage of correct responses (adherent 83.7% and non-adherent 76.3%, $p = 0.01$). They posited that such a larger number of flawed items may have diminished students' overall scores by half a letter grade. Tarrant and Ware (2008) conducted an analysis of ten summative assessments within a nursing program and discovered that between 28% and 75% of the items on each exam contained errors. Overall, 47.3% of all test items reviewed were identified as flawed. The results of these studies indicate that question flaws are common and require deeper attention from item writers.

## Challenges with Item Writing

Despite the widespread use of multiple-choice questions in educational assessment, faculty often encounter significant challenges when tasked with constructing high-quality items. One of the most persistent issues is the occurrence of common item-writing flaws such as ambiguous stems, implausible distractors, and cues that inadvertently reveal the correct answer. Developing distractors that reflect common misconceptions or errors in reasoning requires deep content expertise and insight into learner thinking, which can be difficult to achieve under time constraints. Item-writing flaws can undermine the validity and reliability of assessments, yet they frequently appear in faculty-generated items due to limited formal training in assessment design (Przymuszala et al., 2020). Once an item is drafted, it should undergo a thorough vetting process. This should ideally involve both the original

writer and one or more peers to ensure clarity, alignment with learning objectives, and appropriate cognitive rigor (Haladyna et al., 2002). This collaborative review helps identify potential flaws and strengthens the overall quality of the assessment.

Time constraints and competing professional responsibilities may further complicate the item-writing process. Faculty often juggle teaching, research, and administrative duties. This can leave limited time and/or energy for the iterative refinement process that item writing requires. As a result, item construction may be rushed, leading to inconsistencies in cognitive level, alignment with learning objectives, and distractor quality (Abdulghani et al., 2017).

Another challenge lies in the application of cognitive theory to item design. Although Bloom's taxonomy and related frameworks offer valuable guidance for targeting higher-order thinking, faculty may struggle to translate these principles into practice. Studies have shown that without targeted support, educators tend to default to lower-level recall questions, missing opportunities to assess application, analysis, or synthesis skills (Abdulghani et al., 2017). Faculty development programs have demonstrated promise in addressing these challenges, but their effectiveness depends on sustained engagement and institutional support. Even with training, faculty may require ongoing feedback and examples to internalize best practices and avoid common pitfalls (Gupta et al., 2020; Rauf & Sultana, 2021). For institutions without dedicated teaching and learning centers, and for faculty with limited time to participate in sustained professional development, item-writing guides offer a practical, accessible resource for enhancing the quality of multiple-choice question construction.

## Purpose of Item Writing Guidelines

To lessen common pitfalls in item writing, researchers and practitioners have developed evidence-based guidelines to support instructors and test developers in crafting and reviewing assessment items prior to administration (Haladyna, 1997; Haladyna, 2004; Haladyna et al., 2002; Moreno et al., 2006; Suskie, 2009). These guidelines are instrumental in reducing the prevalence of flawed items and ensuring that assessments accurately measure student learning. Research supports the use of such guidelines as an effective strategy for minimizing the influence of "construct-irrelevant" factors such as test-taking strategies that can distort assessment outcomes (Downing, 2005; Pate & Caldwell, 2014). By adhering to established item-writing principles, faculty can avoid frequent errors in MCQ design, including unfocused or ambiguous stems, negatively worded items, the use of problematic options like "all of the above" or "none of the above," and overly complex formats that may confuse rather than challenge students (Downing, 2005; Hansen & Dexter, 1997).

Some of the most widely cited item-writing guidelines were developed by Haladyna et al. (2002) and later expanded by Haladyna (2004). These works emphasize five key areas: content relevance, formatting consistency, stylistic clarity, and best practices for crafting both stems and answer choices. Building on this foundation, Xu et al. (2016) introduced additional strategies aimed at enhancing assessment quality. Their recommendations address fairness for test-takers, effective feedback mechanisms, content and formatting improvements, and safeguards against academic dishonesty.

## Item Writing Guideline Recommendations

Item-writing guidelines generally fall into two categories: test item formats and quality assurance checklists. The first category focuses on identifying which MCQ formats are most effective for assessing student learning. Formats such as single-best-answer and extended matching items are known to enhance validity, while problematic formats like "all of the above" or "none of the above" tend to promote guessing and diminish item quality (Downing, 2005; Haladyna et al., 2002; Xu et al., 2016). The second category includes structured checklists and guiding principles designed to help educators craft clear, fair, and effective items. When appropriate item formats are combined with well-established guidelines, assessments are more likely to yield valid measures of student knowledge and cognitive ability.

### *Content Choice Recommendations*

When writing or evaluating an MCQ, careful attention to content is essential. Many scholars agree that each item should target a specific content area and assess a single cognitive process, as outlined in a test blueprint (Haladyna, 2004; Moreno et al., 2006; Suskie, 2009; Towns, 2014; Zimmaro, 2016). Items should focus on critical concepts aligned with course objectives or learning outcomes, rather than trivial details or minutiae (Brady, 2005; McCoubrie, 2004; Zimmaro, 2016). To promote deeper learning and reduce the likelihood of students relying on quick internet searches (during unsecure exams) items should emphasize the application of knowledge and skills rather than simple factual recall (Moreno et al., 2006; National Board of Medical Examiners, 2020; Xu et al., 2016). See Table 1 for a revised MCQ with improved assessment of higher-order cognitive skills.

**Table 1**

*Example of MCQ Revision for Conceptual Understanding (Biology)*

| Issue | Flawed MCQ | Revised MCQ |
|---|---|---|
| Stem | Photosynthesis occurs in which part of the plant cell? | A researcher applies an herbicide that disrupts chloroplast function. What impact will this most likely have on the plant? |
| Options | A. Mitochondria<br>B. Chloroplast<br>C. Nucleus<br>D. Cell wall | A. Decreased energy production via photosynthesis<br>B. Enhanced protein synthesis<br>C. Increased cellular respiration<br>D. Increased root absorption of nutrients |
| Issues & Improvements | ● Factual recall<br>● Weak distractors<br>● No context | ● Scenario-based application<br>● All options plausible<br>● Encourages conceptual reasoning |

### *General Item Formatting Recommendations*

Researchers offer specific formatting guidelines for item stems and answer choices to ensure that items are no more complex than necessary and that students are not penalized due to formatting issues rather than gaps in content knowledge. Items should be presented vertically and use clear, content-specific vocabulary (Haladyna et al., 2002; National Board of Medical Examiners, 2020). Unnecessary wording should be eliminated to reduce cognitive load and minimize reading time. A key benefit of using concise, straightforward stems and answer options is that it helps level the playing field between faster and slower readers, an important consideration, as most exam objectives do not assess speed of response (Breakall et al., 2019; Haladyna et al., 2002; McCoubrie, 2004). Additionally, all items should be carefully proofread for grammar, spelling, and punctuation. Even minor errors can distract students or, in more serious cases, lead to confusion or misinterpretation of the question (Haladyna et al., 2002; McCoubrie, 2004; National Board of Medical Examiners, 2020). See Table 2 for an example of a revised MCQ with improved formatting.

**Table 2**

*Example of MCQ Revision with Improved Formatting (Mathematics)*

| Issue | Flawed MCQ | Revised MCQ |
|---|---|---|
| Stem | A triangle has side lengths of 3 cm, 4 cm, and 5 cm, choose the best answer: | A triangle has side lengths of 3 cm, 4 cm, and 5 cm. Which of the following statements is correct? |
| Options | A. 6cm² and no it's not a right triangle<br>B. Area is 7.5 and it is a right triangle<br>C. It's not a right triangle and area equals 7.5cm²<br>D. Yes, it is right triangle and the area is 6 | A. The triangle is a right triangle, and its area is 6 cm².<br>B. The triangle is a right triangle, and its area is 7.5 cm².<br>C. The triangle is not a right triangle, and its area is 6 cm².<br>D. The triangle is not a right triangle, and its area is 7.5 cm². |
| Issues & Improvements | ● Grammatically awkward and doesn't present the stem in a question format<br>● Inconsistent structure across distractor options<br>● Disjointed presentation introduces extraneous cognitive load | ● Stem is presented clearly and in a question format<br>● Answer choices are formatted similarly, and like choices are grouped together |

### *Question Stem Recommendations*

Instructors or item developers should dedicate time to reviewing both the exam instructions and the question stem carefully. Instructions must be directly relevant to the question, clearly worded, and easy to follow. The stem itself should be concise, with the central idea embedded within it rather than dispersed among the answer choices (DiBattista & Kurzawa, 2011; Haladyna et al., 2002; McCoubrie, 2004). Each item should be distinct to avoid overlapping with others, which can inadvertently provide clues to the correct answer (Moreno et al., 2006; National Board of Medical Examiners, 2020). Items should also be fact-based, not opinion-driven, and free from unnecessary complexity or misleading phrasing (Haladyna, 2004; McCoubrie, 2004; Moreno et al., 2006).

Precision in language is critical; vague terms such as “may,” “could be,” or “usually” should be avoided (McCoubrie, 2004; National Board of Medical Examiners, 2020). Each question should present only one clearly correct answer. Negatively worded stems such as “Which of the following is NOT correct?” or “All of the following are true EXCEPT” should generally be avoided, as they can confuse test-takers and increase item difficulty unnecessarily (Breakall et al., 2019; Haladyna, 2004; Moreno et al., 2006; National Board of Medical Examiners, 2020; Suskie, 2009). However, if negative stems are necessary, such as in disciplines where identifying incorrect actions is essential, keywords like NOT or EXCEPT should be bolded and underlined to ensure visibility (Downing, 2006; Suskie, 2009).

When using media, it should serve a clear instructional purpose and directly support the question: unnecessary visuals or unrelated information should be excluded (National Board of Medical Examiners, 2020). After editing, a helpful strategy for evaluating clarity is to cover the answer choices and assess whether the question can be reasonably answered based solely on the stem. Ideally, test-takers should be able to derive the correct response without relying on the options. See Table 3 for a revised MCQ with an improved question stem.

**Table 3**

*Example of MCQ Revision with an Improved Question Stem (English Literature)*

| Issue | Flawed MCQ | Revised MCQ |
|---|---|---|
| Stem | What is the theme of the poem, which talks about nature and feelings and is written by a famous poet who lived in the 19th century and used lots of metaphors and similes to describe things? | Which of the following best expresses the central theme of the poem "To Autumn" by John Keats? |

| | | |
|---|---|---|
| Options | A. The beauty and abundance of the natural world<br>B. The importance of industrial progress<br>C. The inevitability of death and decay<br>D. The isolation of the individual in modern society | A. The beauty and abundance of the natural world<br>B. The importance of industrial progress<br>C. The inevitability of death and decay<br>D. The isolation of the individual in modern society |
| Issues & Improvements | ● Overly complex stem<br>● Background information doesn't provide enough support to answer the question<br>● Uses vague and colloquial language | ● Concise stem<br>● Uses discipline-relevant terminology<br>● Targets higher-level skills such as interpretation and synthesis of the poem |

*Answer Choice Recommendations*

In addition to stem development, answer choices should be carefully reviewed prior to administering an assessment, as several common issues can compromise item quality. MCQs typically include three to five response options, and the position of the correct answer should vary across items to avoid patterns such as consistently placing the correct answer as option "C" (Haladyna et al., 2002; National Board of Medical Examiners, 2020; Xu et al., 2016). When applicable, answer choices should be arranged in a logical or numerical order to prevent confusion, especially when dealing with ranges or sequences, so that students do not overlook an option due to disorganized formatting (Breakall, 2019; DiBattista & Kurzawa, 2011; Xu et al., 2016). Instructors and test administrators should also be mindful of how electronic testing platforms or learning management systems handle answer randomization, ensuring that formatting and logic remain intact. Each answer choice must be mutually exclusive, particularly when using numerical ranges (e.g., "less than 1.0," "1.0 to 1.9," "2.0 or more") to avoid overlap that could confuse test-takers (National Board of Medical Examiners, 2020).

To maintain fairness and avoid unintentional clues, all answer options should be similar in content, grammatical structure, and length. This prevents students from identifying the correct answer based on superficial differences (Breakall et al., 2019; National Board of Medical Examiners, 2020). Problematic formats such as "all of the above," "none of the above," or complex K-type options (e.g., "A and B," "B and C") should generally be avoided as they can introduce ambiguity and encourage guessing (Downing, 2006; Suskie, 2009). Distractors should fall into the same category, or category groups, as the correct answer (e.g., data types, parts of speech, diagnoses, instruments, historical events, etc.). Authors should avoid using "double options" (e.g., do W and X; do Y because of Z) unless the correct answer and all the distractors are double options (National Board of Medical Examiners, 2020).

All distractors should be plausible and relevant. Implausible or obviously incorrect options can be easily dismissed, reducing the item's ability to discriminate between students who understand the material

and those who do not. A distractor that is never selected provides little diagnostic value to the instructor (DiBattista & Kurzawa, 2011; Haladyna, 2004; Haladyna et al., 2002; National Board of Medical Examiners, 2020). All distractors should be grammatically consistent and of the same (relative) length as the correct answer. The author should review the item to ensure the text from the vignette is not included in any of the distractors, to avoid cueing. To ensure readability, the author should read each option immediately following the vignette to ensure the language is a good fit. See Table 4 for a revised MCQ with improved distractor options. See Appendix A for a checklist on constructing high-quality MCQs.

**Table 4**

*Example of MCQ Revision with an Improved Answer Choices (European History)*

| Issue | Flawed MCQ | Revised MCQ |
|---|---|---|
| Stem | Which of the following best explains how the system of alliances contributed to the outbreak of World War I? | Which of the following best explains how the system of alliances contributed to the outbreak of World War I? |
| Options | A. Archduke Franz Ferdinand was assassinated.<br>B. Regional tensions created global conflict.<br>C. Imperialism.<br>D. Military mobilization delayed the war. | A. It created a balance of power that prevented conflict.<br>B. It discouraged military mobilization among European powers.<br>C. It escalated regional tensions into a global conflict.<br>D. It led to diplomatic negotiations that delayed war. |
| Issues & Improvements | ● Inconsistent distractor formatting<br>● Tone of choices is inconsistent<br>● Choices are not alphabetized | ● Plausible distractors<br>● Consistent tone and structure<br>● Grammatically correct and alphabetized |

Consistent application of item-writing guidelines offers a practical and scalable pathway for improving assessment quality across disciplines. Application of item-writing checklists, like that provided in Figure 1, helps reduce ambiguity, minimize item-writing flaws, and promote fairness in test design. Such guidelines also support both instructional integrity and student success. Whether used by individual instructors, collaborative faculty teams, or curriculum committees, these evidence-based principles help ensure that multiple-choice questions are clear, cognitively appropriate, and aligned with learning objectives.

## Directions for Future Research

Practical challenges confront instructors developing MCQ assessments, especially when applying item-writing guidelines in practice. This is particularly true when questions are designed to assess higher-order cognitive processes such as application, analysis, and evaluation. Developing plausible distractors requires content expertise and an understanding of common student misconceptions. Constructing scenario-based stems demands time for creation and validation. Ideally, peer review and/or expert validation of items, pilot testing of new items, and revisions based on response data or other forms of feedback are incorporated into the development process. Although these steps are critical to ensure item quality, time and personnel constraints often make it difficult to implement such processes consistently.

Wide accessibility of generative artificial intelligence (AI) tools may offer a tempting solution to at least some of the challenges facing instructors creating MCQs, but caution is advised. Wu et al. (2025) conducted a blinded analysis of MCQ creation by GPT-4, novice human item writers, and expert human item writers. Expert writers were subject-matter experts, while novice items were created by non-subject-matter experts and were not reviewed or edited by subject-matter experts. Items were scored for content validity, structure of the MCQ item, item-writing flaws, scope, and cognitive skill level, and the quality of clinical reasoning, using a rubric developed by Wu et al. (2025). The rubric was based on Bloom's revised taxonomy and Haladyna et al.'s (2002) item writing guidelines.

Items were blinded as to authorship, and the rubric's scoring system was used by a human consensus panel to rate each item. Wu et al. (2025) found a small but statistically significant difference between expert items and AI items on content validity, clinical reasoning, and higher-order cognitive skill testing. AI items had a higher rating than expert items on biased answer positioning and incorrect "correct" answers. The results suggested that although GPT-4 produced MCQs for evaluating complex clinical medical concepts with structural quality comparable to the expert group, human review is required prior to using those AI-generated MCQs to catch errors. With the recent release of GPT-5 and other updated genAI models, there is an opportunity for updated research on the quality of AI-generated MCQs. The leading AI Reasoning models, including Gemini 2.5 Pro, Open AI O3 Pro, Claude 4 Opus, Grok 3, and DeepSeek R1, may also serve interesting to explore as tools for generating MCQ assessments or evaluating quality of MCQ assessments within disciplines.

Although MCQ-producing programs are proliferating, the degree to which software creates high-quality MCQs is uncertain, especially when considering assessment of higher-level cognitive skills. Yaacoub et al. (2025) used machine learning models to evaluate alignment of questions generated by Google's Vertex AI Text Generation Model with lower-order and higher-order cognitive skills as established on Bloom's taxonomy. The authors found that the AI model effectively created questions at lower-order cognitive levels but did not accurately align questions with higher-order cognitive skills. Machine learning and NLP algorithms may be used in combination to produce and then evaluate MCQ stems and answer options to determine alignment between questions and cognitive verbs consistent with Anderson & Krathwohl's (2001) taxonomy, acting as a support tool to improve efficiency and quality of MCQ assessment creation. This remains an area for future research.

An additional possibility for future exploration involves creating discipline-specific, custom AI agents to evaluate MCQ quality as guided by established item-writing guidelines and checklists reviewed in this paper. If they prove to be accurate, such AI agents could improve the efficiency of human MCQ creation. Research is needed to determine whether a well-designed agent could pinpoint MCQ quality issues and suggest remedies that result in items which are as good as expert human-created MCQs.

**Figure 1**

*Checklist for constructing high-quality MCQs*

| **Item Format Selection** |
|---|
| Use effective formats such as single-best-answer or extended matching items. |
| Avoid problematic formats like "all of the above", "none of the above", or complex K-type options. |
| **Content Choice** |
| Ensure each item targets a specific content area and a single cognitive process. |
| Align items with course objectives or learning outcomes. |
| Focus on critical concepts, not trivial details. |
| Emphasize application of knowledge and skills over factual recall. |
| Avoid content that can be easily answered via quick internet searches. |
| **Question Stem Construction** |
| Present items with clear, content-specific vocabulary. |
| Use fact-based language; avoid opinion-driven or misleading phrasing. |
| Eliminate unnecessary wording to reduce cognitive load. |
| Avoid culturally loaded terms, complex sentence structures, or vocabulary that may disadvantage specific learner groups. |
| Embed the central idea within the stem, not in the answer choices. |
| Avoid vague terms like "may," "could be," or "usually." |
| Present only one clearly correct answer. |
| Avoid negatively worded stems unless essential; if used, bold and underline key words like NOT or EXCEPT. |
| Use media only when it serves a clear instructional purpose. |
| Check if the stem alone allows the question to be reasonably answered. |
| **Answer Choice Construction** |
| Include 3–5 response options per item. |
| Use only plausible and relevant distractors. |
| Vary the position of the correct answer to avoid patterns. |
| Arrange choices in logical or numerical order when applicable. |
| Ensure answer randomization does not disrupt formatting or logic. |
| Make all choices mutually exclusive, especially with numerical ranges. |
| Maintain similar content, structure, and length across all options. |
| Avoid formats like "all of the above," "none of the above," or "A and B." |
| Ensure distractors are grammatically consistent and similar in length to the correct answer. |
| Avoid using text from the vignette in distractors. |
| Read each option after the vignette to ensure language fit and readability. |

## Conclusion

Multiple-choice question assessments are widely used due to their efficiency in evaluating a broad range of knowledge and skills within a limited timeframe. Both faculty and students are generally

familiar with the format, and scoring is typically fast and automated. However, as discussed throughout this paper, MCQs are not without limitations. One of the most significant concerns is the risk of inaccurately assessing student performance, particularly when flawed items lead to students being unfairly classified as underperforming.

Despite these limitations, the use of established item-writing guidelines and checklists can support the development of high-quality MCQs that assess higher-order cognitive skills such as analysis, application, and problem-solving. Crafting effective exam items is a skill that must be intentionally developed and continuously refined throughout an instructor's career. To ensure MCQs serve as valid measures of student understanding, rather than simply testing exam-taking strategies, instructors must prioritize alignment between items and learning objectives. Applying item-writing best practices during both the development and review phases of assessment construction helps ensure that MCQs are clear, fair, and capable of accurately evaluating student learning before the exam is administered. With multiple-choice assessments remaining a cornerstone of evaluation for many disciplines, future research into effective faculty training and the utilization of artificial intelligence holds promise for advancing item development practices and strengthening assessment quality at the program level.

## References

Abdulghani, H. M., Irshad, M., Haque, S., Ahmad, T., Sattar, K., & Khalil, M. S. (2017). Effectiveness of longitudinal faculty development programs on MCQs items writing skills: A follow-up study, *PLoS ONE 12*(10), 1–14. https://doi.org/10.1371/journal.pone.0185895

Adesope, O. O., Trevisan, D. A., & Sundararajan, N. (2017). Rethinking the use of tests: A meta-analysis of practice testing. *Review of Educational Research, 87*(3), 659–701. https://doi.org/10.3102/0034654316689306

Anderson, L. W., & Krathwohl, D. R. (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's Taxonomy of educational objectives*. Addison, Wesley, Longman.

Brady, A. M. (2005). Assessment of learning with multiple-choice questions. *Nurse Education in Practice, 5*(4), 238–242. https://doi.org/10.1016/j.nepr.2004.12.005

Breakall, J., Randles, C., & Tasker, R. (2019). Development and use of a multiple-choice item writing flaws evaluation instrument in the context of general chemistry. *Chemistry Education Research and Practice, 2,* 369–382. https://doi.org/10.1039/C8RP00262B

Bridgeman, B. (1992). A comparison of quantitative questions in open-ended and multiple-choice formats. *Journal of Educational Measurement, 29*(3), 253–271. https://doi.org/10.1111/j.1745-3984.1992.tb00377.x

Burton, S. J. (2005). Multiple choice and true/false tests: Myths and misapprehensions. *Assessment and Evaluation in Higher Education, 30*(1), 65-72. https://doi.org/10.1080/0260293042003243904

DiBattista, D., & Kurzawa, L. (2011). Examination of the quality of multiple-choice items on classroom tests. *Canadian Journal for the Scholarship of Teaching and Learning, 2*(2), 1–23. https://doi.org/10.5206/cjsotl-rcacea.2011.2.4

Douglas, M., Wilson, J., & Ennis, S. (2012). Multiple-choice question tests: A convenient, flexible, and effective learning tool? A case study. *Innovations in Education and Teaching International, 49*(2), 111–121. http://dx.doi.org/10.1080/14703297.2012.677596

Downing, S. M. (2002). Threats to the validity of locally developed multiple-choice tests in medical education: Construct-irrelevant variance and construct underrepresentation. *Advances in Health Sciences Education 7*, 235–241. https://doi.org/10.1023/A:1021112514626

Downing, S. M. (2005). The effects of violating standard item writing principles on tests and students: The consequences of using flawed test items on achievement examinations in medical education. *Advances in Health Sciences Education, 10*, 133–143. https://doi.org/10.1007/s10459-004-4019-5

Downing, S. M. (2006). Selected response item formats in test development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 287–301). Routledge.

Fink, L.D. (2013). *Creating significant learning experiences, revised and updated: An integrated approach to designing college courses*. Jossy-Bass.

Flaherty, C. (2020, June 17). *Much ado about class size*. Inside Higher Ed. https://www.insidehighered.com/news/2020/06/18/study-some-things-matter-more-class-size-when-it-comes-student-success

Gupta, P., Meena, P., Khan, A. M., Malhotra, R. K., Singh, T. (2020). Effect of faculty training on quality of multiple-choice questions, *International Journal of Applied Basic Medical Research, 10*(3), 210–214. https://doi.org/10.4103/ijabmr.IJABMR_30_20

Haladyna, T. M. (1997). *Writing test items to evaluate higher order thinking*. Allyn and Bacon.

Haladyna, T. M. (2004). *Developing and validating multiple-choice test items* (3rd ed.). Routledge. https://doi.org/10.4324/9780203825945

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education, 15*(3), 309–333. https://doi.org/10.1207/S15324818AME1503_5

Hansen J. D., & Dexter, L. (1997). Quality multiple-choice test questions: Item-writing guidelines and an analysis of auditing testbanks. *Journal of Education for Business, 73*(2), 94–97. https://doi.org/10.1080/08832329709601623

Krieglstein, F., Beege, M., Rey, G. D., Sanchez-Stockhammer, C., & Schneider, S. (2023). Development and validation of a theory-based questionnaire to measure different types of cognitive load. *Educational Psychology Review*, *35*(9), 1–37. https://doi.org/10.1007/s10648-023-09738-0

Loepp, E. (2021, June 22). *The benefits of higher-order multiple-choice tests*. Inside Higher Ed. https://www.insidehighered.com/advice/2021/06/23/rethinking-multiple-choice-tests-better-learning-assessment-opinion

Marsh, E. J., Roediger, H. L. III, Bjork, R. A., & Bjork, E. L. (2007). The memorial consequences of multiple-choice testing. *Psychonomic Bulletin & Review, 14*, 194–199. http://dx.doi.org/10.3758/BF03194051

Marzano, R.J., & Kendall, J.S. (Eds.) (2007). *The new taxonomy of educational objectives*. Corwin Press.

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

Mayer, R. E., & Fiorella, L. (2021). Principles for managing essential processing in multimedia learning: Segmenting, pre-training, and modality principles. In R. E. Mayer & L. Fiorella (Eds.), *The Cambridge handbook of multimedia learning* (pp. 243–260). Cambridge University Press.

McCoubrie, P. (2004). Improving the fairness of multiple-choice questions: A literature review. *Medical Teacher, 26*(8), 709-712. https://doi.org/10.1080/01421590400013495

Mehrens, W. A. (1987). Validity issues in teacher licensure tests. *Journal of Personnel Evaluation in Education, 1*(2), 195–229. https://doi.org/10.1007/BF00128894

Moncada, S. M. & Moncada, T. P. (2010). Assessing student learning with conventional multiple-choice exams: Design and implementation considerations for business faculty. *International Journal of Education Research, 5*(2), 15–29. https://go.gale.com/ps/i.do?p=AONE&id=GALE%7CA299759810&v=2.1&it=r&sid=googleScholar&asid=4b9f571e&enforceAuth=true&linkSource=delayedAuthFullText&aty=shibboleth&userGroupName=tamp44898&oweAuth=true

Moreno, R., Martínez, R. J., Muñiz, J. (2006). New guidelines for developing multiple-choice Items. *Methodology European Journal of Research Methods for the Behavioral and Social Sciences, 2*(2), 65–72. https://doi.org/10.1027/1614-2241.2.2.65

National Board of Medical Examiners. (2020). NBME Item-Writing Guide (6th ed.). https://www.nbme.org/educators/item-writing-guide

Ozuru, Y., Briner, S., Kurby, C. A., & McNamara, D. S. (2013). Comparing comprehension measured by multiple-choice and open-ended questions. *Canadian Journal of Experimental Psychology, 67*(3), 215–227. http://dx.doi.org/10.1037/a0032918

Palmer, E. J., & Devitt, P. G. (2007). Assessment of higher order cognitive skills in undergraduate education: modified essay or multiple-choice questions? Research Paper. *BMC Medical Education, 7*(49), 1–7. https://doi.org/10.1186/1472-6920-7-49

Pate, A. & Caldwell, D. J. (2014). Effects of multiple-choice item-writing guideline utilization on item and student performance. *Currents in Pharmacy Teaching and Learning, 6*(1), 130–134. https://doi.org/10.1016/j.cptl.2013.09.003.

Przymuszala, P., Piotrowska, K., Lipski, D., Marciniak, R., & Cerbin-Koczorowska, M. (2020). Guidelines on writing multiple choice questions: A well-received and effective faculty development intervention, *SAGE Open,* 1–12. https://doi.org/10.1177/2158244020947432

Rauf, A. R. & Sultana, S. (2021). Effect of faculty training on quality of Multiple Choice Questions, *Rawal Medical Journal, 46*(2), 430–433. https://www.rmj.org.pk/?mno=133092

Roediger, H. L., & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences, 15*(1), 20–27. https://doi.org/10.1016/j.tics.2010.09.003

Roediger, H. L., & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science, 17*(3), 249–255. https://doi.org/10.1111/j.1467-9280.2006.01693.x

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

Stagnaro-Green A. S. & Downing, S. M. (2006). Use of flawed multiple-choice items by the New England Journal of Medicine for continuing medical education. *Medical Teacher, 28*(6), 566–568. https://doi.org/10.1080/01421590600711153

Suskie, L. (2009). *Assessing student learning: A common sense guide* (2nd ed.). Wiley & Sons.

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science,* 12, 257–285. https://doi.org/10.1207/s15516709cog1202_4

Sweller, J., Van Merriënboer, J. J., & Paas, F. (2019). Cognitive architecture and instructional design: 20 years later. *Educational Psychology Review,* 31, 261–292. https://doi.org/10.1007/s10648-019-09465-5

Tarrant M., Knierim A., Hayes S. K., & Ware J. (2006). The frequency of item writing flaws in multiple-choice questions used in high-stakes nursing assessments. *Nurse Education Today, 26(*8), 662–671. https://doi.org/10.1016/j.nedt.2006.07.006

Tarrant, M., & Ware, J. (2008). Impact of item-writing flaws in multiple-choice questions on student achievement in high-stakes nursing assessments. *Medical Education, 42*(2), 198–206. https://doi.org/10.1111/j.1365-2923.2007.02957.x

Thissen, D., Mislevy, R. J. (2000). Testing algorithms. In H. Wainer (Ed.) *Computerized adaptive testing: A primer* (pp. 101–134). Routledge.

Towns, M. H. (2014). Guide to developing high-quality, reliable, and valid multiple-choice assessments. *Journal of Chemical Education. 91*(9), 1426–1431. https://doi.org/10.1021/ed500076x

Wu, H., Zerner, T., Lee, D., Court-Kowalski, S., Devitt, P., & Palmer, E. (2025). GPT-4 versus human authors in clinically complex MCQ creation: A blinded analysis of item quality. *Medical Teacher,* 1–14. https://doi.org/10.1080/0142159X.2025.2505122

Xu, X., Kauer, S., and Tupy, S. (2016). Multiple-choice questions: Tips for optimizing assessment in-seat and online. *Scholarship of Teaching and Learning in Psychology, 2*(2), 147–158. https://doi.org/10.1037/stl0000062

Yaacoub, A., Da-Rugna, J., & Assaghir, Z. (2025). Assessing AI-Generated Questions' Alignment with Cognitive Frameworks in Educational Assessment. *International Journal of Computer Theory and Engineering, 17*(3), 114–125. https://doi.org/10.48550/arXiv.2504.14232

Zaidi, N. B., Hwang, C., Scott, S., Stallard, S., Purkiss, J., Hortsch, M. (2017). Climbing Bloom's taxonomy pyramid: Lessons from a graduate histology course. *Anatomy Science Education, 10*(5), 456–464. https://doi.org/10.1002/ase.1685

Zheng, A. Y., Lawhorn, J. K., Lumley, T., & Freeman, S. (2008). Application of Bloom's taxonomy debunks the "MCAT myth". *Science, 319*(5862), 414–415. https://doi.org/10.1126/science.1147852

Zimmaro, D. M. (2016, December 1). *Writing good multiple-choice exams*. The University of Texas at Austin Faculty Innovation Center. https://ctl.utexas.edu/sites/default/files/writing-good-multiple-choice-exams-fic-120116.pdf

## About the Authors

René Schmauder, Ph.D., M.B.A., Director of Undergrad Assessment, Clemson University, aschmau@clemson.edu

Kimberly Daugherty, Pharm.D., Ph.D., Vice President and Provost, Sullivan University, kdaugherty@sullivan.edu

Michael Rudolph, Ph.D., Associate Dean and Associate Professor, College for Healthy Communities, A.T. Still University, mjrudolp@gmail.com

Bryant Hutson, Ph.D., Director of Assessment, University of North Carolina at Chapel Hill, bryant.hutson@gmail.com

Kelly McCarthy, Ph.D., M.B.A., Associate Director for Institutional Effectiveness and Research, Florida Polytechnic University, kmccarthy@floridapoly.edu

Sarah McCauley, M.A., Assistant Director of Assessment, College of Medicine, Department of Medical Education, University of South Florida, samccaul@usf.edu

Dina Eagle, M.A., Academic Director, Assessment, Strayer University, dina.eagle@strayer.edu