Quality Assessment Through Learning Engineering: An Evaluation Rubric of LLM-Generated Multiple-Choice Questions

Katerina Christhilf, Linh Huynh, Yu Tian, Shubham Chakraborty, Micah Watanabe, & Danielle S. McNamara

Automated Question Generation | Multiple Choice Questions | Psychometrics

Accurately assessing reading comprehension is essential for monitoring student progress, yet creating high-quality multiple-choice questions is time-consuming for instructors. AI-generated questions offer a promising solution but raise concerns about question quality. To address this gap, we developed a rubric for human raters to evaluate the quality of automatically generated comprehension questions. The rubric incorporates established criteria for question design and documented weaknesses of AI-generated items, including awkward, wordy phrasing and conspicuous correct answers. Through iterative testing with two human raters scoring 200 AI-generated multiple-choice questions, we refined scoring categories and scales, achieving weighted kappa reliability of .79 or higher. The final rubric includes seven categories assessing answer correctness, distractor incorrectness, topic relevance, writing clarity, distractor linguistic features, distractor semantic plausibility, and distractor semantic uniqueness. This tool supports future research comparing question-generation systems and exploring automated question-quality evaluation.

Katerina Christhilf, Linh Huynh, Yu Tian, Shubham Chakraborty, Micah Watanabe, and Danielle S. McNamara

Arizona State University; {kchristh, lthuynh1, ytian126, schak103, mwatana5, dsmcnama}@asu.edu

Abstract.

Keywords: Automated Question Generation, Multiple Choice Questions, Psychometrics.

Challenge

Accurate assessment of reading comprehension is essential for instructors to evaluate students' progress and adjust instruction. Multiple-choice questions (MCQs) are widely used in education since they offer an efficient and objective, and method for large-scale assessments (Haladyna et al., 2002); however, they can be difficult and time-consuming for instructors to create. High-quality MCQs contain unambiguously correct answers, plausible but incorrect distractors, clear and concise wording, alignment with learning objectives, and emphasis on key information rather than minor details (Haladyna et al., 2002). As such, a key educational challenge is how best to support instructors in rapidly creating relevant, appropriately challenging MCQs for learner practice and assessment.

Recent advances in Generative Artificial Intelligence (GenAI) have enabled automatic generation of MCQs at scale, reducing the burden on instructors. Yet, there are concerns regarding the quality of these questions. Automatically generated questions tend to have certain characteristics (Gorgun & Bulut, 2024; Zhou & Li, 2025). Although AI-generated questions are excellent in spelling and grammar, they often overemphasize trivial details, include awkward or overly verbose wording, and produce distractors that are shorter or less developed than the correct answer. These limitations highlight the need for a human-rating rubric that systematically evaluates the quality of AI-generated questions. There have been previous human evaluations of MCQs, but they often focused on surface details or were overly broad (e.g., Cheung et al., 2023; Olney, 2023). The objective of this work is to introduce a domain general question-evaluation rubric designed to assess the quality of MCQs.

Rubric Design

Our team created an initial rubric that was iteratively refined through three rounds of practice scoring sessions. The finalized rubric includes seven categories (see Appendix A): Answer Correctness, Distractor Incorrectness, Topic Relevance, Writing Clarity, Distractor Linguistic Features, Distractor Semantic Plausibility, and Distractor Semantic Uniqueness. Answer Correctness indicates that the "correct answer is clearly correct to those who have comprehended the text", while Distractor Incorrectness indicates that the "distractors are clearly incorrect to those who have comprehended the text." Because GenAI-generated questions rarely include inaccurate answers or accurate distractors, answer correctness and distractor incorrectness have binary rating options: 0 or 1. Other categories capture known challenges for LLM-generated questions, including overemphasis on trivial details, awkward or overly verbose language, and implausible or insufficiently distinct distractors. Therefore, the remaining categories were scored on a 4-point scale to allow nuanced quality judgments while avoiding the midpoint bias common in 3-point scales. Topic Relevance indicates that the question "addresses a central concept in the text". Writing Clarity indicates that the question "stem and response options are concise and unambiguous". While other rubrics often separate clarity into several categories, such as grammar, spelling, and fluency, we condensed clarity into a single category, because artificially generated questions rarely include errors in grammar or spelling. The features of the distractors were split into three categories: linguistic features, semantic plausibility, and semantic uniqueness, to reflect three common yet separate issues with artificially generated questions. Distractor Linguistic Features refers to how similar the language of the distractors is to the correct response, including length, syntactic structure, and style. Distractor Semantic Plausibility refers to the extent to which the meanings of the distractors are plausible, while Distractor Semantic Uniqueness refers to the extent to which each distractor conveys a unique meaning.

Preliminary Outcomes

Two subject-matter experts used the rubric to evaluate a set of 200 AI-generated MCQs. Results of post-discussion weighted kappas reached .79 or higher for every category (see Appendix B), indicating the rubric supported raters in establishing good reliability when scoring these key features of MCQs.

Conclusion

This project demonstrates an application of learning engineering by systematically designing and refining a key assessment artifact, a human-rating rubric for MCQ quality, within a larger effort to evaluate AI-generated questions. Grounded in learning engineering's emphasis on evidence-based design, the rubric was developed through initial research on key question quality considerations, followed by three cycles of use, feedback, and revision; design decisions were informed by rater agreement and scoring discussions. This work aligns with the creation phase of a broader learning engineering cycle by enabling consistent data collection about known strengths and weaknesses of GenAI-produced MCQs. It will be used to support subsequent phases of our larger project evaluating questions generated by different LLMs and multi-agent systems. At the same time, the rubric development itself can be viewed as a smaller, nested learning engineering cycle involving goal definition, artifact design, practical testing, and evidence-based refinement. Looking ahead, we are also exploring the potential

to automate this rubric using LLMs, natural language processing, or machine learning methods, enabling the carefully developed scoring categories from this rubric to be used at a larger scale.

Acknowledgments

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305T240035 to Arizona State University. The opinions expressed are those of the authors and do not represent the views of the Institute or the U.S. Department of Education.

References

Cheung, B. H. H., Lau, G. K. K., Wong, G. T. C., Lee, E. Y. P., Kulkarni, D., Seow, C. S., Wong, R., Co, M. T. H. (2023). ChatGPT versus human in generating medical graduate exam questions – An international prospective study. medRxiv. https://doi.org/10.1101/2023.05.13.23289943

Gorgun, G., & Bulut, O. (2024). Exploring quality criteria and evaluation methods in automated question generation: A comprehensive survey. Education and Information Technologies, 29(18), 24111–24142. https://link.springer.com/article/10.1007/s10639-024-12771-3

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

Olney, A. M. (2023). Generating multiple choice questions from a textbook: LLMs match human performance on most metrics. Proceedings of the Workshop on Empowering Education with LLMs. https://eric.ed.gov/?id=ED630037

Zhou, H., & Li, L. (2025). Qadg: Generating question–answer-distractors pairs for real examination. Neural Computing & Applications, 37, 1157–1170. https://doi.org/10.1007/s00521-024-10658-5

Appendix A

Rubric to Evaluate Multiple-Choice Question Quality

	0- Disagree	1- Agree
Answer Correctness Correct answer is clearly correct to those who have comprehended		
Distractor Incorrectness Distractors are clearly incorrect to those who have comprehended the text.		

	1- Disagree	2- Neutral	3- Agree	4- Strongly Agree

Overall Quality The question and answer choices are clearly worded, concise, and unambiguous. Addresses key concepts central to the text.	Topic Relevance Addresses a central concept in the text	Question contains trivial or irrelevant information	Question contains a key sub-point, but no main ideas	Question contains a key sub-point that connects to main ideas	Entire question focused on a main idea or key concept
	Writing Clarity Stem and options concise & unambiguous	Excessively wordy or unclear	Wordiness or other slightly impacts comprehension, mild ambiguity, or awkward syntax	A few redundancies or minor wordiness	Fully clear and concise
Distractor Plausibility Distractors are plausible to those who have not comprehended the text	Linguistic Features All are similar in linguistic features (e.g., length, number of phrases)	Distractors are visibly different from correct answer (e.g., length by 12+ characters).	Distractors all have small differences, OR at least one distractor has major differences from correct answer.	Distractors have relatively same length and features to correct answer	All distractors are uniform in linguistic features
	Semantic Plausibility All are semantically plausible (e.g., similar topic, no absolute terms, information is not more well known than the correct answer)	All distractors mildly implausible OR two distractors very implausible	Two distractors mildly implausible OR one distractor very implausible	One distractor is mildly plausible	All distractors plausible
	Semantic Uniqueness Each distractor is unique in meaning	All distractors very similar	Multiple distractors slightly similar or two very similar	Two distractors are wrong in similar way	All distractors unique

Appendix B

Post-Discussion Weighted Kappas Per Category

Scoring Category	Weighted Kappa
Answer Correctness	1
Distractor Incorrectness	0.79
Topic Relevance	0.97

Writing Clarity	0.96
Distractor Linguistic Features	0.84
Distractor Semantic Plausibility	0.83
Distractor Semantic Uniqueness	0.89