

AI-Driven System for Automated Generation and Evaluation of Multiple-Choice Questions*

Thoria ALGHAMDI

Department of Information Systems-Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah 21589, Saudi Arabia

Correspondence should be addressed to: Thoria ALGHAMDI, Taalghamdi@kau.edu.sa

*Presented at the 47th IBIMA International Conference, 29-30 June 2026, Madrid, Spain

Abstract

Student assessment is a fundamental aspect of the learning process in higher education, enabling faculty members to measure learning outcomes and obtain feedback on their teaching. Multiple-choice questions MCQs are widely used in higher education assessments due to their low cost, ease of preparation, accuracy, and ability to measure various learning outcomes. However, the effectiveness of this tool in measuring learning outcomes is directly related to the quality of the MCQs. Student responses are used to calculate both the difficulty index (P) and the discrimination index (D), which helps in evaluating and continuously improving the quality of MCQs. However, the process of creating and writing high-quality MCQs and collecting and analyzing student results is time-consuming and labor-intensive, often resulting in low-quality questions or overlooking feedback calculations. This paper presents BankGen, an AI-powered framework that automates the creation, management, and evaluation of MCQs using generative AI. BankGen is developed using the Replit platform, it integrates Google's Gemini engine with a user-friendly interface to generate curriculum-aligned and learning-outcome-related MCQs. BankGen supports the creation of question banks, creation of test, and analysis using difficulty and discrimination indicators. It also allows users to delete or modify low-quality MCQs. The results showed that the BankGen-generated MCQs are curriculum-related, aligned with learning outcomes, and free from spelling and grammatical errors. These results indicate that BankGen can reduce the effort required to evaluate and improve MCQs.

Keywords: Artificial Intelligence, Assessment, Item Analysis, MCQs.

Introduction

Student assessment methods are a fundamental component of academic accreditation in higher education, as they ensure academic quality by measuring student learning outcomes. While various assessment methods exist to measure the mastery of course objectives, multiple-choice questions (MCQs) remain among the most common due to their effectiveness, lower cost, applicability to large classes, accuracy, ease of formulation, and flexibility, as they can measure both recall and higherorder thinking skills (Lui et al 2024). However, these advantages depend heavily on the quality of question formulation; otherwise, the assessment tool loses its credibility (Malau-Aduli and Zimitat, 2012; Parekh and Bahadoor, 2024).

Although there exist educational resources that support the formulation of high-quality multiple-choice questions (Cox, 2019), this is not enough. It is important to review assessment tools and rely on quantitative measures to determine question quality and continuously improve them. The best quality ideally includes MCQs of moderate difficulty that differentiate between high-achieving and low-achieving students. The difficulty index and discrimination index are among the most important measures of such qualities (Malau-Aduli and Zimitat, 2012).

However, the evaluation of question quality often occurs after the test results are received, which limits the possibility of immediate improvement and confines the development process to future sessions. Furthermore, many faculty members rely on a preliminary formal review that focuses on the correctness of the language before the test, while neglecting the statistical analysis of question quality after test administration and grading. This has been indicated by several studies, which have shown weaknesses in assessment systems. This has been attributed to heavy and time-consuming workload of faculty members, which includes scientific research, teaching, assessment, and community service (Iqbal et al, 2024; Brown and Abdalnabi, 2017).

Potter and McLachlan (2025) have recommended careful attention to the design of MCQs. Analytical study involving 90 questions confirmed that an acceptable percentage of questions achieve an appropriate level of difficulty (80%) and a good discrimination index (82%) (Kumar et al, 2021). In another study of 210 items, the discrimination index was (56.7%) and the difficulty index was (50.5%) (Eleragi et al, 2025). However, these results indicate a gap that requires further improvement.

Singer-Freeman and Taylor (2023) mentioned that employing artificial intelligence techniques in the assessment process can generate MCQs of a quality comparable to those produced by humans. The role of these technologies extends beyond simply generating questions; they also improve the fairness of assessment by reducing student confusion resulting from cultural differences, thus contributing to more accurate measurements of their learning level.

Despite the availability of theoretical guidelines and modern techniques supporting the construction of MCQs, practical application still lacks integrated tools to help faculty members design and systematically and consistently evaluate highquality questions. This challenge is exacerbated by the multitude of academic responsibilities, including research, teaching, assessment, and service, often leading to a superficial review of questions and neglecting quantitative analysis of their quality. This, in turn, affects the accuracy of learning outcomes measurement and the reliability of assessment.

To address this, our research proposes the development of an intelligent system based on artificial intelligence technologies called BankGen. BankGen is designed to support four main functions: First: to generate MCQs on the given portion of the materials, within the given instructions, alignment with learning outcomes, and according to quality standards. Second: to create a test, a number of questions are drawn from the test bank according to the requirements specified by the faculty member. Third: to evaluate the used generated MCQs in the test by calculating quality indicators such as difficulty and discrimination indices. Fourth: to allow the faculty member to evaluate the quality of the questions in the bank where the question which showed acceptable level of difficulty and achieved good discriminator index is valid otherwise the faculty member can delete or update it.

The remainder of this paper is organized as follows. Section 2 reviews the related studies on utilizing AI to generate and evaluate assessments. Section 3 describes the proposed methodology and implementation details, including the technologies and evaluation procedures. Section 4 explains and discusses the results. Section 5 summarize the conclusion, limitations, and future work.

Previous Studies

The application of artificial intelligence in education has seen significant development, with many studies focusing on large language models (Language Models) and Retrieval-Enhanced Generation (RAG) techniques to enhance intelligent learning environments and address learner queries. Wijaya and Purwarianti (2024) proposed a constructivist-based intelligent learning system that relies on a set language model combined with RAG technology

using the LangChain library. This system aimed to provide direct interaction between students and instructors, and its results demonstrated positive outcomes in terms of answer quality and processing speed. Deng and Yuan (2026) presented a more comprehensive framework that integrates multimedia cultural influences with RAG technology to extract knowledge from videos and textbooks and answer questions with high accuracy, ensuring user satisfaction and system performance.

In the field of question generation using linguistic inferences, several studies have focused on developing graphical algorithms to generate high-quality questions. Gašpar et al (2023) presented a linguistic instruction-based approach to transforming texts into questions, assessing question quality using human criteria. It demonstrated the potential for generating accurate questions with significant accuracy when using paraphrasing. Al Shuraiqi et al (2024) reviewed the latest automated multiple-choice question generation techniques in medical education, emphasizing the role of natural language technologies and scientific learning. It highlighted the need for more adaptive models. Qorich and El Ouazzani, (2026) addressed the organization of question generation based on transformer models (transformers), proposing a hybrid model combining PART and convolutional networks (CNNs). This approach aims to improve question quality while achieving implementation efficiency and computational cost reduction.

Furthermore, some studies have explored question quality assessment and selection optimization using large-scale linguistic analysis. Siro et al (2026) presented a comprehensive framework for automated question submission using large-scale linguistic models, along with CrowdLLM, which simulates human reviewer evaluations. The results demonstrated that the automated-generated questions achieved high quality and contributed to improving information retrieval model tools. A Awalurahman et al (2025) conducted a significant review of methods for creating multiple-choice questions using transformer models and large-scale linguistic models. It concluded that most studies excel at manual evaluation or textual similarity criteria, with limited attention to pedagogical evaluation or statistical analysis of the questions.

Previous studies have largely focused on one of two main aspects: either developing answer questions and engaging learners using large-scale linguistic associations, or developing techniques that do not include multiple-choice questions and their linguistic and semantic quality. While some studies have addressed question quality evaluation, this evaluation is often based on linguistic measures, human assessment, or textual similarity measures, without considering the statistical analysis of the questions after generation.

Despite the progress made in previous studies, significant variation remains due to the absence of an integrated system that combines the generation of multiple-choice questions with their statistical evaluation. Therefore, this research aims to develop an intelligent web-based system that relies on large linguistic inferences to generate MCQs, with the quality of these questions assessed using the Difficulty Index and the Discrimination Index.

Methodology

To develop BankGen, AI based system for generating and evaluating MCQs, we used The Replit platform (Replit, n.d) and integrate it with the AI generative tool Gemini (Gemini, n.d). The following steps are followed as the research methodology:

Requirements

To ensure BankGen is used effectively and replaces traditional question creation and evaluation methods, we designed it to provide the following features:

- Ease of MCQs generation, evaluation, and modification.
- Fast in MCQs generation, calculation of difficulty and discrimination indices, and creation of question quality reports.
- Reliability of generated MCQs in terms of their relevance to the curriculum, compliance with instructions, ability to measure specified learning outcomes, and free from spelling and grammatical errors.

- Ease of labeling high-quality MCQs and modifying or deleting questions that do not meet the required indicators.

To evaluate MCQs quality, BankGen calculates difficulty and discrimination indices.

Difficulty Index

The difficulty index measures the proportion of students who answered a question correctly. It is calculated by using equation 1. Where P is the difficulty index, R is the number of students who answered correctly, and T is the total number of students.

$0.3 \leq P \leq 0.7$ indicates acceptable questions, otherwise the question either very easy or very difficult.

$$P = \frac{R}{T} \quad \text{Equation 1}$$

Discrimination Index

The discrimination index measures the ability of a question to differentiate between high-performing and low-performing students. It is calculated by using equation 2. Where D is the discrimination index, U is the number of correct responses in the upper group, L is the number of correct responses in the lower group, and N is the number of students in one group. $0.2 \leq D \leq 0.29$ indicates acceptable discrimination where $D < 0.2$ indicates weak discrimination. Questions with weak statistical indicators are flagged for revision or removal from the question bank.

$$D = \frac{U-L}{N} \quad \text{Equation 2}$$

Design

We designed BankGen to support four main functions: generating bank questions, generating exam using all or some of the generated questions in the bank questions, analyzing the questions of the exam by calculating difficulty and discrimination indices, and adjusting the bank questions based on the analysis report by labeling the good quality question as valid and delete, or replace others.

BankGen begins with two main options on the homepage: creating a course or browsing existing courses, as shown in figure 1.

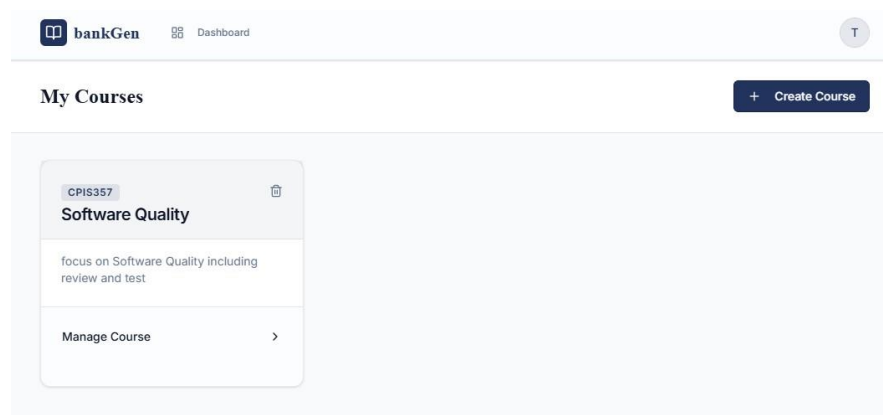
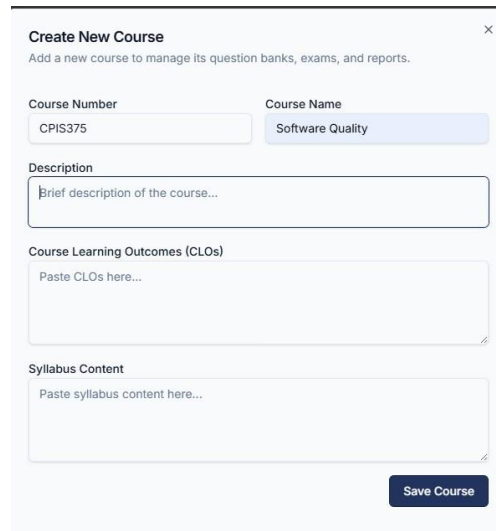


Fig 1. BankGen home page

By clicking on “create course”, a faculty member can create a page for his course. He needs to enter the course code, the course name, a brief description, and the learning outcomes, as shown in Figure 2.

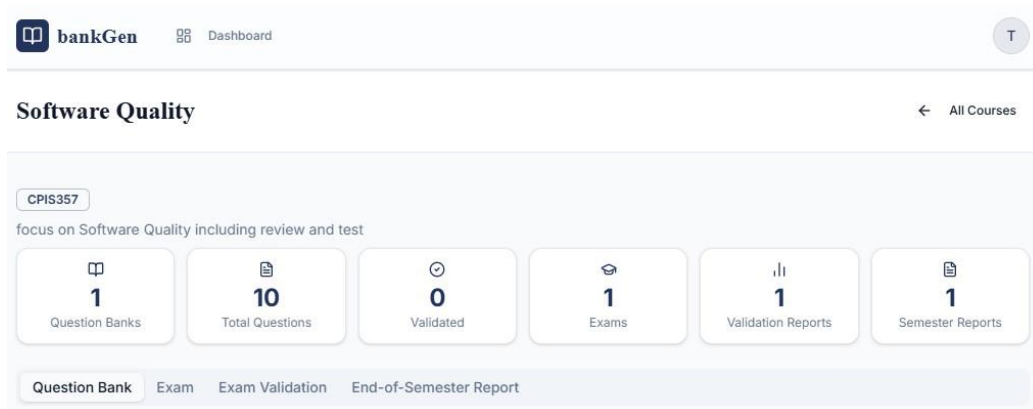


The 'Create New Course' form is a modal window with a close button (X) in the top right corner. It contains the following fields and sections:

- Course Number:** A text input field containing 'CPIS375'.
- Course Name:** A text input field containing 'Software Quality'.
- Description:** A text area with a placeholder 'Brief description of the course...'.
- Course Learning Outcomes (CLOs):** A text area with a placeholder 'Paste CLOs here...'.
- Syllabus Content:** A text area with a placeholder 'Paste syllabus content here...'.
- Save Course:** A blue button at the bottom right.

Fig 2. Create course page

As shown in Figure 3, once the course page is created, all the main functions appear in the menu, including the Question Bank, Exam, Exam Validation, and End -of-Semester Report.



The 'Software Quality' course page shows the following details:

- Course ID:** CPIS357
- Focus:** focus on Software Quality including review and test
- Statistics:**
 - Question Banks: 1
 - Total Questions: 10
 - Validated: 0
 - Exams: 1
 - Validation Reports: 1
 - Semester Reports: 1
- Menu:** Question Bank, Exam, Exam Validation, End-of-Semester Report

Fig 3. Course page

Clicking on the “Question Bank” provides two options: creating a bank or viewing existing ones. A faculty member can create a question bank by entering a title, uploading the curriculum file or part of it as PDF file, entering the required learning outcomes, and instructions specifying the number of generated questions, the questions format, difficulty level, and other requirements for Gemini to generate the questions. Using structured prompt engineering, Gemini generates curriculum-aligned MCQs with answer choices, correct answers, and CLO mapping. The generated MCQs are stored in the question bank and can later be selected automatically for exam generation. This provides high flexibility in generating questions to meet the faculty member's needs, Figure 4 shows the creating bank question page.

Create Question Bank

×

Provide material and instructions. AI will generate MCQ questions automatically.

Bank Name

e.g. Chapter 3 MCQs

Course Material (PDF)

↑

Upload a PDF with course material, lecture notes, or textbook content

Click to browse or drag & drop

CLOs (Course Learning Outcomes)

Paste CLOs here...

Additional Instructions

e.g. Focus on theoretical concepts, difficulty level medium...

Generate Question Bank

Fig 4.Creating bank question page

Once the "Generate" button is clicked, the question bank appears on the course page under the "Question Bank" item. The created Question Bank can contain the generated questions, each with four answer choices, the correct answer, and the mapped CLO. Each question has three icons: a check button to mark the question as valid, a second icon to edit the question if it needs minor adjustments, and a third icon to delete the question if it does not meet the required quality standards. Figure 5 shows the generated bank questions.

bankGen

Dashboard

T

Chapter 1

← Back to Course

Print

Download PDF

10 questions

0 / 10 validated

Q1

○

✎

🗑

A logistics company evaluates a new fleet management system. The manager states, 'This software is high quality because it allows our drivers to plan routes 20% faster than before, directly meeting our operational efficiency goals.' Which of David Garvin's quality views is being expressed?

CLO 1

A. The Transcendental view

B. The User view

C. The Manufacturer's view

D. The Product view

Fig 5. The generated bank questions

After creating the question bank, the faculty member can create an exam by entering the exam title, selecting a question bank from the available and previously created banks, and specifying the number of questions. As shown in figure 6.

Create Exam

Randomly select questions from a question bank to generate an exam.

Exam Name

Question Bank

Select a question bank

Number of Questions

OPTIONAL — PRINT TEMPLATES

Coversheet (optional)


Upload exam coversheet PDF — shown as first page when printing
Click to browse or drag & drop

Exam Template (optional)


Upload exam format/layout PDF — used to style the printed exam
Click to browse or drag & drop

Generate Exam

Fig 6. Create an exam

The created exam appears on the course page under “Exam” item, and the faculty member can easily view, print, or download it, as shown in figure 7.

bankGen Dashboard

Midterm Exam Back Print Exam Questions Only Download PDF Delete

5 Questions Created April 12, 2026

Print-ready exam. Use Print Exam to print with coversheet · Questions Only to skip the cover · Download PDF then choose "Save as PDF" in the print dialog.

Midterm Exam
April 12, 2026

Student Name: _____ Student ID: _____
Section: _____ Date: _____

Total Questions: 5 Time Allowed: _____ minutes Total Marks: _____

1. A logistics company evaluates a new fleet management system. The manager states, 'This software is high quality because it allows our drivers to plan routes 20% faster than before, directly meeting our operational efficiency goals.' Which of David Garvin's quality views is being expressed?

(A) The Transcendental view (B) The User view
(C) The Manufacturer's view (D) The Product view

[CLO: CLO 1]

Fig 7. The generated exam

After completing the exam, correcting it, and obtaining the grades, the faculty member can analyze the exam questions and obtain the validation report by entering a report name, the questions used in the quiz, and the scores. Clicking "Analyze" displays the report, which includes values of difficulty level and discriminator factor for each question, with an indication of whether the question was weak or not. As shown in figure 8.

Exam Validation Reports

medterm exam Apr 12, 2026

2 weak questions identified

q1
Difficulty: 100% Discrimination: 0% Weak

q2
Difficulty: 0% Discrimination: 0% Weak

q3
Difficulty: 58% Discrimination: 80%

q4

The assessment contains 5 questions, 2 of which are classified as weak. Question 1 was answered correctly by all students, and Question 2 was answered correctly by none; both fail to provide useful psychometric data. Questions 3 and 4 are excellent items with strong discrimination. Question 5 is acceptable but could be refined to improve its discriminatory power. Recommendation: Revise or remove q1 and q2.

Fig 8. The validation report

All created components, including courses, banks, exams, and reports, can be deleted.

Implementation

To implement BankGen, we use Replit platform. Where we inserted the prompt which encompassed all the requirements and we kept improving the Replit output including the interface and functions. BankGen uses a database created by Replit to store and manage data related to courses, banks, exams, and reports. And utilizes Gemini to generate the MCQs and reports.

Software Testing

Functional tests were conducted on the system to ensure the integrity of all its functions and the absence of software errors. These tests included verifying the functionality of the user interface, uploading files, generating MCQs, displaying results, calculating question quality indicators, and testing the handling of errors and incorrect inputs to ensure the system's stability and performance.

Validation Testing

After ensuring the correctness of the systems, its suitability was validated by engaging faculty members from various disciplines. They were asked to use the system to generate MCQs based on their courses and then evaluate the quality of the resulting questions using a questionnaire that included several criteria. These criteria included the relevance of the questions to the course content, clarity of wording, linguistic accuracy, alignment with learning outcomes, and the quality of the answer choices and the correct answer. The questionnaire results were then used to assess the system's suitability for use in the educational environment.

Results

The following results present a comprehensive evaluation of the quality of questions generated by BankGen. Overall, the findings indicate a high level of satisfaction among faculty members who participated in the testing phase.

All participants (100%) reported that the questions generated by BankGen were free from spelling and grammatical errors. In addition, they unanimously agreed that the content of the questions was highly relevant to the prescribed curriculum, reflecting strong alignment with course objectives and subject matter requirements.

For the clarity, ease of understanding, and appropriateness of difficulty level for students, participants' responses were consistently positive. The majority of faculty members indicated either agreement or strong agreement, suggesting that the generated questions are well-structured, clearly phrased, and suitable for the intended student level.

Similarly, in evaluating the effectiveness of the questions in measuring the intended learning outcomes, participants' opinions again ranged between agreement and strong agreement. This indicates that BankGen is capable of producing assessment items that meaningfully evaluate students' knowledge and skills in relation to the specified learning objectives.

However, when considering the comprehensiveness of topic coverage, responses were slightly more varied. While 66% of participants agreed or strongly agreed that the generated questions adequately covered all aspects of the topic, the remaining participants expressed a neutral stance. This suggests that, although the system performs well overall, there may be opportunities to further enhance the breadth and depth of content coverage.

The overall evaluation of BankGen-generated questions was highly positive, with responses consistently falling between "agree" and "strongly agree" across all measured criteria. These findings demonstrate the system's effectiveness in generating high-quality, relevant, and pedagogically sound assessment questions, while also highlighting a potential area for improvement in ensuring more comprehensive topic coverage.

Conclusion

The results of this study demonstrate that BankGen's AI-powered tool provides a robust and efficient solution for supporting faculty in the development of high-quality multiple-choice questions that align with the curriculum. Further, BankGen allows the calculation of key psychometric indicators such as difficulty levels and discrimination indices via an accessible userfriendly interface.

By providing immediate, data-driven feedback on assessment quality, BankGen allows faculty members to refine questions and ensure they are valid, reliable, and aligned with established academic quality standards. This reduces the time and effort required for manual evaluation, allowing instructors to focus more on pedagogical design and student learning outcomes.

To further minimize teacher's administrative workload, BankGen can be further developed in the future to generate a wider range of assessment methods other than multiple-choice questions and automatically generate end-of-semester reports that summarize performance metrics and assessment insights.

In conclusion, the overall findings highlight BankGen's potential as an effective assistant in educational environments, where artificial intelligence is used to enhance both efficiency and quality in assessment practices. By streamlining routine tasks and supporting evidence-based decision-making, BankGen provides a more structured and consistent evaluation approach. This makes it a promising step toward the adoption of intelligent assessment tools that offer practical benefits and opportunities for continuous improvement in educational assessment systems.

References

- Al Shuraiqi, S., Aal Abdulsalam, A., Masters, K., Zidoum, H. & AlZaabi, A. 2024, "Automatic Generation of Medical Case-Based Multiple-Choice Questions (MCQs): A Review of Methodologies, Applications, Evaluation, and Future Directions", *Big data and cognitive computing*, vol. 8, no. 10, pp. 139.
- Awalurahman, H.W., Fathoni Aji, R. & Budi, I. 2025, "Transformer and Large Language Models for Automatic MultipleChoice Question Generation: A Systematic Literature Review", *IEEE access*, vol. 13, pp. 127100-127112.
- Brown, G.T.L. & Abdulnabi, H.H.A. 2017, "Evaluating the Quality of Higher Education Instructor-Constructed MultipleChoice Tests: Impact on Student Grades", *Frontiers in Education*, vol. 2.
- Cox, C.W. 2019, "Best Practice Tips for the Assessment of Learning of Undergraduate Nursing Students via MultipleChoice Questions", *Nursing education perspectives*, vol. 40, no. 4, pp. 228-230.
- Deng, C. & Yuan, B. 2026, "Research on an intelligent tutoring system based on automatic construction of multimodal knowledge graphs and retrieval-augmented generation", *Frontiers in computer science (Lausanne)*, vol. 8.
- Eleragi, A.M.S., Miskeen, E., Hussein, K., Rezigalla, A.A., Adam, M.I.E., Al-Faifi, J.A., Alhalafi, A., Al Ameer, A.Y. & Mohammed, O.A. 2025, "Evaluating the multiple-choice questions quality at the College of Medicine, University of Bisha, Saudi Arabia: a three-year experience", *BMC medical education*, vol. 25, no. 1, pp. 233-8.
- Gašpar, A., Grubišić, A. & Šarić-Grgić, I. 2023, "Evaluation of a rule-based approach to automatic factual question generation using syntactic and semantic analysis", *Language resources and evaluation*, vol. 57, no. 4, pp. 1431-1461.
- Gemini. [Online], [Retrieved March 1, 2026], <https://gemini.google.com/app>
- Iqbal, S., Taib, C.A.B. & Razalli, M.R. 2024, "The effect of accreditation on higher education performance through quality culture mediation: the perceptions of administrative and quality managers", *TQM journal*, vol. 36, no. 2, pp. 572-592.
- Kumar, D., Jaipurkar, R., Shekhar, A., Sikri, G. & Srinivas, V. 2021, "Item analysis of multiple choice questions: A quality assurance test for an assessment tool", *Medical journal. Armed Forces India*, vol. 77, no. Suppl 1, pp. S85-S89.

- Liu, Q., Wald, N., Daskon, C. & Harland, T. 2024, "Multiple-choice questions (MCQs) for higher-order cognition: Perspectives of university teachers", *Innovations in education and teaching international*, vol. 61, no. 4, pp. 802-814.
- Malau-Aduli, B.S. & Zimitat, C. 2012, "Peer review improves the quality of MCQ examinations", *Assessment and evaluation in higher education*, vol. 37, no. 8, pp. 919-931
- Parekh, P. & Bahadoor, V. 2024, "The Utility of Multiple-Choice Assessment in Current Medical Education: A Critical Review", *Curēus (Palo Alto, CA)*, vol. 16, no. 5, pp. e59778.
- Potter, H.G. & McLachlan, J.C. 2025, "Assessing medical knowledge: A 3-year comparative study of very short answer vs. multiple choice questions", *Medical teacher*, vol. 47, no. 10, pp. 1669-1677.
- Qorich, M. & El Ouazzani, R. 2026, "Context-based automatic question generation for educational assessment via lightweight transformer augmentation", *Engineering applications of artificial intelligence*, vol. 177, pp. 114953.
- Replit – Build apps and sites with AI. [Online], [Retrieved March 1, 2026], <https://replit.com/>
- Singer-Freeman, K., Janio, J. & Taylor, J. 2023, "Making a Grand Contribution: Using ChatGPT to Improve Assessment in Higher Education", *Assessment update*, vol. 35, no. 5, pp. 10-11.
- Siro, C., Yuan, Y., Aliannejadi, M. & de Rijke, M. 2026, "AGENT-CQ: Automatic Generation and Evaluation of
- Clarifying Questions for Conversational Search with Large Language Models", *ACM transactions on information systems*.
- Wijaya, O.C. & Purwarianti, A. 2024, "An Interactive Question-Answering System Using Large Language Model and Retrieval-Augmented Generation in an Intelligent Tutoring System on the Programming Domain", *IEEE*, , pp. 1.
- Wu, H., Lee, D., Zerner, T., Court-Kowalski, S., Devitt, P. & Palmer, E. 2026, "A comparison of the psychometric properties of GPT-4 versus human novice and expert authors of clinically complex MCQs in a mock examination of Australian medical students", *Medical teacher*, vol. 48, no. 1, pp. 74-84.