

Hybrid Framework for Pre-Assessing Arabic Multiple-Choice Questions Using ChatGPT and Traditional Item Analysis

Omar Dahham Sarhan 

Al-Qaim College of Education, University of Anbar, Al-Qaim, Iraq

CORRESPONDENCE

Omar Dahham Sarhan
omar.sarhan@uoanbar.edu.iq

ARTICLE INFO

Received: Mar. 30, 2026

Revised: Jul. 11, 2026

Accepted: Jul. 19, 2026

Published: Sep. 30, 2026



© 2026 by the author(s).
Published by Mustansiriyah
University. This article is an
Open Access article distributed
under the terms and condi-
tions of the [Creative Commons Attribution \(CC BY\) li-
cense](https://creativecommons.org/licenses/by/4.0/).

ABSTRACT: Background: Multiple-choice questions (MCQs) are widely used in educational evaluation; nonetheless, their effectiveness depends widely on object quality and psychometric properties. Traditional item analysis methods are typically applied after exam administration, reducing the ability to detect problematic items during the test development phase. **Objective:** This research suggests a hybrid framework that combines AI-based pre-assessment with traditional post-test psychometric analysis to validate Arabic MCQs before and after exam administration. **Methods:** The suggested framework uses ChatGPT-4 to conduct a pre-assessment of MCQ items by estimating item difficulty, discrimination potential, and distractor effectiveness based on linguistic and semantic properties. Traditional psychometric analysis was then performed using actual learner response data collected from 195 individuals to compute item difficulty, discrimination indices, reliability estimates, and distractor effectiveness. **Results:** A moderate positive correlation was observed between AI-based predictions and traditional psychometric results ($r = 0.42$, $p < 0.05$). The AI model demonstrates effectiveness in identifying possible weaknesses in question form and distractor quality. Yet, its performance was limited in estimating discrimination indicators that depend directly on actual learner response behavior. **Conclusions:** The findings suggest that AI-assisted pre-assessment can support the early identification of problematic MCQ items before exam administration. Integrating AI-based linguistic evaluation with traditional psychometric analysis may enhance the quality of Arabic educational assessment.

KEYWORDS: MCQ; Evaluation; ChatGPT; Psychometric Analysis; Arabic assessment; AI assessment

INTRODUCTION

Multiple-choice questions (MCQs) have found wide application in higher education as a standardized evaluation tool for assessing learners' academic performance and the achievement of learning goals. The efficiency of these questions in enabling objective judgment and assessment of large groups of learners has made them an integral component of any testing mechanism in educational settings [1]. In essence, the scholarly literature agrees that MCQs can assess higher-order thinking abilities, such as critical analysis and application, especially when constructed effectively aligned with the intended learning objectives [2]. MCQs are frequently used across many subjects, including the sciences, engineering, and technology, because of their utility in facilitating effective evaluation and standardized administration [3]. However, their effectiveness is highly dependent on the structure of MCQs, which entails item clarity, realistic distractors, and congruence of item difficulty and learning objectives. Poorly designed objects may result in measurement errors and a reduced ability to differentiate between high- and low-performing individuals. They may even compromise the validity of the assessment results [4]. Typical post-test psychometric analyses consist of difficulty indices, discrimination coefficients, and reliability measures such as Cronbach's alpha or KR-20, which are generally used to evaluate the quality of MCQs [1]. Since the detection of faulty questions is only possible after the test administration, these measures provide reliable insights into item performance [4]. In addition, sufficient student samples are required for accurate psychometric analysis, which can

be quite costly and time-consuming in some educational contexts [5]. Recent advances in artificial intelligence, particularly in large language models (LLMs), have significantly improved the ability of automated systems to process, understand, and analyze textual information. It has become feasible to analyze academic content automatically [6]. The educational landscape has witnessed a rapid transformation with the advent of LLMs, which have been highly effective in improving learning efficiency, making learning more engaging, and generating content automatically [7]. In recent times, work using models such as ChatGPT has demonstrated their ability to understand natural language effectively [8]. The AI-based assessment here serves as a preliminary test and a process of checking the questions to detect potential errors before the actual test begins [9]. Several previous studies have shown promising results regarding the use of artificial intelligence models in generating and evaluating questions. However, most of these studies rely either on traditional psychological assessment or on AI-based assessments in isolation [8]. Recent studies indicate that machine learning models can achieve positive results, particularly in content generation and automated assessment, especially within higher education. But there are still concerns about the trustworthiness and consistency of AI use in line with educational goals [10]. One concern is the reliability of AI-based assessment. Furthermore, there are discrepancies between students' actual responses and pre-defined correct answers. This way requires a test trial [11]. Therefore, only a few studies have shown that AI-based assessment can enhance traditional assessment methods [12]. Although AI-based assessment is widely used, few studies have focused on multiple-choice questions in Arabic [13]. This raises concerns that identifying grammatical errors and determining test difficulty is challenging when analyzing only the text. Despite the popularity of AI-based assessment methods, there is virtually no research on MCQs in Arabic [3]. To address the shortcomings mentioned above, this research proposes a hybrid model that integrates pre-assessment with ChatGPT and traditional psychometric analysis of test results. The proposed model attempts to: (a) pre-assess structural or linguistic errors before carrying out the test, (b) estimate the compatibility of predicted AI results with actual data from the performance of students, and (c) determine other essential quality measures, such as difficulty levels or effectiveness of the distractors. By comparing analyses, this research provides recommendations to promote the development of multiple-choice questions in educational testing.

RELATED WORKS

Conventional tools for analyzing test item quality require accurate psychometric measures such as the Index of Difficulty, Index of Discrimination, and Reliability. Such tools allow for an understanding of test item quality and of their ability to distinguish between levels within the student base, and are often administered after test administration. This makes it impossible to correct errors before the exam. Such limitations have motivated researchers to explore AI-based pre-assessment techniques that can detect flaws and predict question quality before actual testing [1], [4].

AI-Based Analysis of MCQs Using Large Language Models

Several studies have explored the use of large language models (LLMs) to generate and evaluate MCQs. While these models have presented good performance in enhancing linguistic quality and distractor generation, challenges related to distractor plausibility, ambiguity, and alignment with instructional aims remain unresolved [14]. In their pilot study, Kank (2024) reports that the distractor was effective, but the discrimination and quality of the distractors were inferior to those of instructor-oriented constructed items [15]. This echoes both the hope and shortcomings of AI in item generation. Fardeby *et al.* (2024) evaluated ChatGPT's performance in the knowledge domain. They conclude that the AI-formulated questions are on par with human experts and that the question types and formats are consistent with those of experts, which in turn supports AI item generation [16]. However, most of these efforts focus primarily on the textual characteristics of MCQ items, with limited attention to their psychometric quality and empirical performance, and fail to relate human-predicted congruence test responses and academic achievement to AI-generated predictions. This is the goal of this paper. Although there are clear reasons to be optimistic about LLMs' strengths, they still fall short of accurately addressing advanced reasoning problems in NLP settings [17]. Researchers have also developed a program called 'SAQUET' that detects structural flaws with 91-94% accuracy compared to subject-matter experts [9], [18].

Structural Defect Detection and AI-Assisted Question Evaluation

On another level, we found no difference in the discriminative ability or the level of challenge of the questions produced by ChatGPT compared with those produced by humans [6]. Meanwhile, other researchers found it easier to produce computer-generated questions. For instance, Baudin (2025) examined multiple-choice questions generated by ChatGPT in the context of psychology instruction and found that, although they exhibited acceptable validity and reliability, they were not suitable for assessing higher-level cognitive skills. It is in line with the notion of using hybrid assessment criteria [19]. At the same time, it turns out that the computer-generated questions have discrimination coefficients close to those of the human-written ones, and only human intervention is needed to reveal the defects [1], [4]. Then again, researchers' findings indicated that after the training period, AI's ability to detect problems such as "None of what I just said" was significantly enhanced. So, a study comparing AI-generated and expert-developed MCQs finds that AI-generated questions may be of equivalent quality yet still require human inspection [20].

AI-Based Simulation and Field Testing of MCQs

Moreover, the concept of 'field testing through artificial intelligence' has been proposed, in which deep learning models would essentially serve as human students to answer multiple-choice questions. Some of the latest techniques have sought to enhance the performance of large language models on MCQs by incorporating reasoning that mimics human decision-making [21]. The ChatGPT-4o model and the template-based AIG model achieved very good discrimination scores and could therefore be of great help to teachers in creating multiple-choice questions [22], [23]. A recent example illustrates how these models can accurately model students' behavior and estimate the difficulty and discrimination coefficients, thereby reducing their reliance on human samples. The data support the claim that artificial intelligence can be a reliable partner in increasing the accuracy of educational assessments [3].

Distractor Generation and Evaluation Techniques

One such review has identified three basic methodologies for creating distractors: linguistic methods (including synonymy and grammar-based methods), pattern-matching methods (such as WordNet), and statistical methods (including frequency-based methods). These methods were later complemented by the use of knowledge graphs and text to create high-quality distractors [24]. The target of the distractor-generation task is to produce implausible but incorrect answers to multiple-choice questions. The quality of distractors directly affects the result of multiple-choice items used for evaluation. Recently, studies have focused on evaluation criteria, datasets, and approaches for objective-type questions. The lack of concordance between automatic distractor evaluation approaches and actual test-taking behavior has been highlighted [25].

Research Gap and Motivation

A notable absence in the literature on integrating analytical processes before AI use and subsequent evaluation has been identified. Additionally, a limited degree of alignment has been observed between AI-generated analyses and those from the traditional method. There is a strong need to close the evaluation gap before it is used. In light of the above, it is clear that the previous investigations either focused on the AI-based generation process or on the traditional testing process. The innovative nature of the proposed research lies in integrating both approaches, as reflected in the methodology framework presented in the next section.

Research Questions and Hypotheses

Given the gap in the existing literature, most previous studies have compared manually generated MCQs with those generated by artificial intelligence, with evaluations limited to item analyses based on students' test performance. The aspect of item validation before administration to students was also not highlighted. This gap formed an important motivation for the current study and was addressed through the proposed framework.

(RQ1): Are AI models capable of identifying structural errors and predicting the psychometric quality of multiple-choice questions (MCQ) BEFORE students take them?

(RQ2): How well do the AI predictions relate to the traditional post-test results?

(RQ3): Does the hybrid model of evaluation, consisting of pre-assessment via AI technology and traditional post-testing, constitute a valid framework for assessing the Arabic Multiple-Choice Questions?

Based on the above research questions, the following hypotheses can be proposed:

H1: Artificial intelligence models can identify structurally and linguistically faulty multiple-choice questions before they go live.

H2: AI-calculated difficulty, discrimination, and distractor qualities are significantly related to traditional post-test results.

H3: The combination of the use of AI pre-assessment and the traditional approach of analysis enhances the reliability and validity of the evaluation of MCQs.

MATERIALS AND METHODS

This article offers a combined approach to assessing Arabic multiple-choice questions (MCQs) by merging AI-assisted pre-assessment with conventional post-test psychometric analysis. It explores the potential of AI-generated results as useful indicators of item quality before actual test administration. Three separate MCQ test forms, each containing 18 items, were administered to 195 students, equally divided into three groups. In the first phase, ChatGPT-4 was utilized to analyze each question and predict item difficulty, discrimination potential, and distractor effectiveness based on the test's linguistic and semantic features. Secondly, a conventional psychometric evaluation using real student response data was conducted to identify indices of item difficulty, discrimination, reliability, and distractor impact. The final approach is to compare AI forecasts with real psychometric data to identify the relationship between the two approaches and to assess the feasibility of using AI-supported pre-assessment in learning measurement in the Arabic context.

Research Design

This study employed an experimental comparative design to assess whether AI-aided pre-assessment helps to predict the psychometric characteristics of Arabic MCQ items before they were administered; that is, during the pre-implementation/ preparation phase. The proposed framework integrates the AI model-based pre-evaluation procedure, implemented via ChatGPT, with classic post-hoc test-based psychometric assessment, using data from real examinees. The experimental procedure was conducted in two consecutive stages. The research adopts an experimental comparative design to investigate the potential of AI-assisted pre-assessment to predict the psychometric characteristics of Arabic MCQ items during the pre-preparation phase. The design first employed the ChatGPT-supported pre-assessment to reexamine the items at the pre-test stage. Then the same items were examined using standard post-test psychometric techniques with the students' responses. Finally, the findings from the two stages were compared to assess the degree of agreement and consistency between the AI results and traditional psychometric parameters, as shown in the methodological approach in Figure 1.

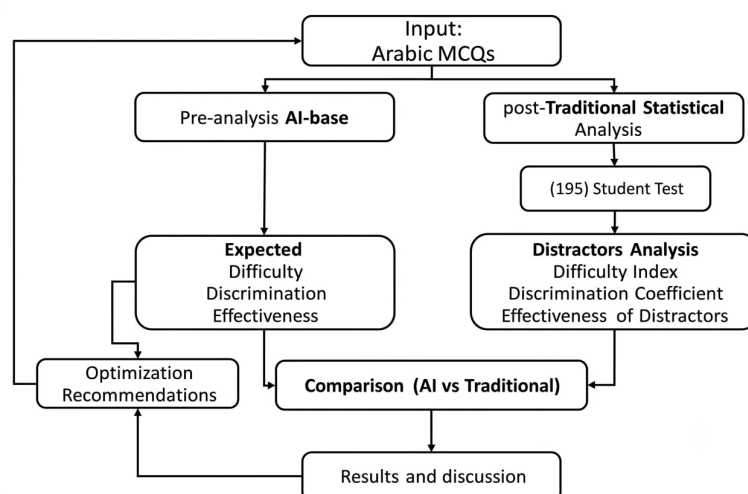


Figure 1. Methodological framework for analyzing MCQs

AI-Based Pre-Assessment Using ChatGPT

Once the MCQs have been entered into the ChatGPT model, it can perform an initial analysis. This first step is performed by relying on linguistic and semantic analysis of the questions, options/chosen answers. ChatGPT writes descriptive predictions of the presumed level of difficulty, the question's correctness, and the effectiveness of the distractors.

Prompting Protocol for ChatGPT Pre-Assessment

For methodological transparency and to allow for re-application in a new study, a pre-assessment by AI was performed using a systematic prompting technique. In this individual study, we applied only the ChatGPT-4 AI model to examine all MCQ items in Arabic before administering the test. To avoid differences due to varying GPT-4 versions and inconsistent overall accuracy, versions were kept consistent throughout the study period. The process of analyzing each Arabic MCQ item and its options was entered into ChatGPT 4 as a standalone input. ChatGPT-4 was asked not only to predict the most popular option among students, but also to predict the item's difficulty estimate while taking into account the language semantics and difficulty factors, such as articulation obscurity, semantic difficulty, and distractor readability, and formulated an estimated difficulty range (for example, 0.7-0.85). Also, it was the model's responsibility to evaluate the distractor's effectiveness based on the accompanying text. Regarding discrimination analysis, the model produced qualitative threshold-based estimates, such as 0.40+, rather than actual psychometric coefficients. This is because such predictions were intended as preliminary indicators before students' actual testing. The same prompting structure was used for all test items to minimize the differences in AI-generated responses. Additionally, Arabic prompts were provided throughout the evaluation process to ensure consistency in the language context of the multiple-choice question (MCQ) items. A sample of the prompting structure employed in this research is available in the Supplementary Material section.

Traditional Post-Test Psychometric Analysis

The questions are applied to a real sample of students, and student response data are collected. Traditional psychometric analysis is conducted after test administration. Based on psychometric principles of traditional measurement, the collected student responses after administering test forms to student groups were scored and analyzed. The analysis was conducted by calculating item difficulty, item discrimination, and distractor effectiveness based on student response patterns. These quantities were measures of item quality and were also used to compare the traditional post-test results with the AI-based pre-assessment outcomes.

The psychometric metrics used in the traditional post-test analysis were computed according to standard item analysis procedures. Item difficulty (P) was determined as the proportion of students who correctly answered an item using Eq. (1):

$$P = \frac{R}{T}, \quad (1)$$

where (R) represents the number of correct responses and (T) represents the total number of students who answered the item. Item discrimination (D) was calculated using the difference in performance between upper and lower student groups using Eq. (2):

$$D = \frac{U - L}{N}, \quad (2)$$

where (U) represents the number of students in the upper group who answered correctly, (L) represents the number of students in the lower group who answered correctly, and (N) represents the number of students in one of the groups (e.g., 27% of the total class size). To make a consistent comparison between AI-based and traditional evaluations, distractor effectiveness was classified according to the pre-set qualitative criteria shown in Table 1, which were derived from well-established principles of item analysis and distractor functioning [26]. These criteria were used as operational definitions in both the AI-based pre-assessment and the traditional post-test evaluation to ensure consistency in the comparative analysis.

Table 1. Distractor effectiveness classification criteria

Class	Description
Good	effectively attract low-performing students and contribute positively to item discrimination.
Accepted	Show partial functionality and provide limited contribution to item differentiation.
Weak	Rarely selected or fail to contribute meaningfully to item discrimination.

Study Sample and Data Collection

The study was conducted using three separate MCQ test forms for the first-year Bachelor of Education computer principles course. Six multiple-choice questions were included in each test form, and each independent test form was administered to a different group of 65 students. As a result, each student in each group was to answer only one test form, rather than all 18 questions. To conduct the element analysis for each item, the analysis was calculated only for students who answered the specific test form group. The 195 students were divided equally across the three independent test forms.

Comparative Evaluation Procedure

The proposed methodology ultimately focuses on comparing the indicators extracted from the AI-powered pre-analysis with the results of the traditional post-analysis. The comparative evaluation phase compared the indicators generated during the AI-based pre-assessment phase with those from the traditional post-test analysis. This comparison was carried out along three facets of the indicators: Discrimination, difficulty and distractor effectiveness. This provided results on the accuracy of the AI-based predictions relative to the actual psychometric measures obtained from student responses.

Research Scope and Limitations

The research scope is limited to only Arabic multiple-choice questions and does not cover other question types or languages. The pre-analysis is based solely on one AI model (ChatGPT), and no comparative analysis with other AI systems is included. The work does not provide a detailed assessment of linguistic complexity or consider how well the questions align with educational goals, since such analysis would require expert human judgment and is beyond the scope of this study. Despite the final sample including 195 students, each item-level analysis included only 65 student responses. As such, results should be described as test cases or preliminary measures within a proof-of-concept mindset rather than as large-scale psychometric standardizations. A proposed direction for future research is to use larger, more diverse sample sizes.

Statistical Analysis

We conducted a statistical analysis to determine item difficulty, item discrimination, and the effectiveness of distractors. We used Pearson correlation analysis to assess the relationship between AI-based predictions and post-test outcomes, with a significance level of $p \leq 0.05$. For correlation analysis, the midpoint of each AI-predicted difficulty interval was used as the representative value for the corresponding item. These center-point values were then compared with the traditional item-difficulty indices obtained from student response data using Pearson's correlation coefficient. We measured the internal consistency of three MCQ test forms using the Kuder-Richardson Formula 20 (KR-20), which is conceptually equivalent to Cronbach's alpha for dichotomously scored items.

We carried out all statistical analyses using the Statistical Package for the Social Sciences (SPSS) version 26.0. SPSS was the tool we used to provide descriptive statistics, perform correlation analysis, and calculate reliability.

RESULTS AND DISCUSSION

Descriptive Results of Traditional Item Analysis

Descriptive statistics for the typically analyzed MCQ items are shown here to provide a general idea of the psychometric properties of the test forms. Three sets of questions, each containing 6 items, were administered and analyzed to 3 classes of 65 students, for a total of 195 participants. Table 2 presents

descriptive statistics for item difficulty, discrimination, and distractor effectiveness. The reported values include the minimum, maximum, mean, and standard deviation for the item difficulty and discrimination indices, as well as the overall distribution of distractor effectiveness classifications.

Table 2. General statistical properties of questions

Indicator	Min	Max	Average	Standard Deviation
Difficulty	0.23	0.92	0.70	0.18
Discrimination	-0.38	0.94	0.73	0.31
Effectiveness	Weak (16.7%)	Good (72.2%)	Accepted (11.1%)	-

Using the classic post-test review, the assessment is shown to have good features with question difficulty, discrimination, and distractor effectiveness. The difficulty index ranged from 0.23 to 0.92 (average: 0.70). Most of the questions were of moderate difficulty level, while a few questions were very easy or very hard. The discrimination index ranged from 0.38 to 0.94 (average: 0.73), suggesting that most questions can distinguish between high- and low-ability students. If some items show negative discrimination, we first have to address their formulation problems or weaknesses. The percentages of good, acceptable, and weak distractors were 72.2%, 11.1%, and 16.7%, respectively. That means the quality of this test is quite good, but some distractors still need to be changed.

Reliability Analysis

A reliability analysis was also performed to assess the internal consistency of the test forms using the Kuder-Richardson Formula 20 (KR-20), which is equivalent to Cronbach's alpha for dichotomously scored MCQ items. The reliability analysis was then performed for each set of independent questions. The KR-20 reliability coefficients obtained for the three MCQ test forms are presented in Table 3.

Table 3. Kuder-Richardson (KR-20) result

Question Group	Question Number	Student Number	KR-20 (Cronbach's Alpha)
1	6	65	0.58
2	6	65	0.46
3	6	65	0.52

The reliability Results from this research provide further insight into the internal consistency of the MCQ items. The reliability scores were tested using the usual guidelines in educational research: a score of less than 0.50 was considered low reliability, a score between 0.50 and 0.80 was considered moderate internal consistency, and a score greater than 0.80 was considered high internal consistency [27]. Given these criteria, the KR-20 values observed across the different test forms appear to indicate varying degrees of reliability, ranging from weak to moderate. These results are not unexpected, given the number of items in each test form (six items each is quite brief) and the fact that brief tests tend to have correspondingly low internal consistency estimates. Weak or negatively discriminating items may also have contributed to the fluctuation in reliability across the test sets.

AI-based vs Traditional Evaluation Comparison

Here, we seek to compare the degree to which AI-driven pre-assessment outcomes align with those of the traditional post-assessment, which measures students' performance on tests. The pre-assessment was conducted using ChatGPT, an advanced AI model proficient in natural language processing (NLP) and in interpreting educational texts. Because of the moderate correlation between the measures ($r = 0.42$, $p < 0.05$), a notable convergence emerged between traditional and AI-based evaluation.

The proposed methodology evaluates multiple-choice questions and their options generated from Arabic content before the actual test begins, focusing on three main points:

- Preliminary assessment of the difficulty level of the question Predicted Difficulty: by analyzing the texts and formulating the distractors.
- Estimating the level of Discrimination: by predicting the number of students who are attracted to the given question.

- Determine distractors' quality and classify them as (Good, Accepted, or Weak).

At stage two, the test was administered to 195 students (65 per each question model), and their data were collected for analysis using classical statistical methods. The discrimination and difficulty indices were computed for each question, and the effectiveness of the distractors was assessed.

The correlation between the ChatGPT-based pre-assessment and the traditional post-test analysis yielded a moderate and positive ($r = 0.42$, $p < 0.05$). Under typical effect-size interpretations in behavioral research, this represents a moderate relationship, reflecting the magnitude of the association between AI-based prediction and those empirically determined from the psychometric analysis of student data [28]. Such a relationship holds some promise for AI-based prediction in providing preliminary recommendations on item difficulty and distractor effectiveness, which may be helpful before a one-test analysis, but cannot stand as an alternative to post-test psychometric analysis at this juncture. However, discrepancies were observed in several items, where AI-based predictions did not perfectly align with empirical student-derived results. Some items that AI identified as having weak performance showed acceptable to good performance according to traditional methods, whereas others showed surprisingly challenging performance or discrimination levels. The results indicate that mere text analysis may not adequately reflect students' item-behavioral responses.

Difficulty Index (P) and Discrimination factor (D)

Based on student response data, Table 4 presents a direct comparison between the difficulty of the AI-predicted value and the classic item difficulty P-value. With the AI pre-assessment phase, ChatGPT did not give a specific difficulty value; instead, it generated a range of difficulty based on the masking, semantics, and distractor plausibility of the MCQ item. The traditional values use the item difficulty (P) index as explained in the method section.

Table 4. Comparison of AI-predicted and traditional difficulty index for each question

Q. Num	AI-based	Traditional
1	(0.70-0.85)	0.69
2	(0.75-0.90)	0.77
3	(0.85-0.95)	0.85
4	(0.90-0.98)	0.89
5	(0.80-0.90)	0.80
6	(0.80-0.90)	0.92
7	(0.50-0.65)	0.62
8	(0.90-0.98)	0.85
9	(0.75-0.85)	0.74
10	(0.60-0.75)	0.65
11	(0.80-0.90)	0.77
12	(0.55-0.70)	0.23
13	(0.60-0.75)	0.58
14	(0.90-0.98)	0.92
15	(0.45-0.60)	0.69
16	(0.70-0.85)	0.54
17	(0.30-0.50)	0.31
18	(0.50-0.65)	0.62

Figure 2 compares the real item difficulty scores derived from the conventional psychometric method with the difficulty predictions made by the AI model, which were obtained before the test. The area filled with color between the lower and upper AI-predicted limits indicates the difficulty interval estimated by the model for each question. The actual item difficulty values, obtained from the students' responses, are shown by the black line.

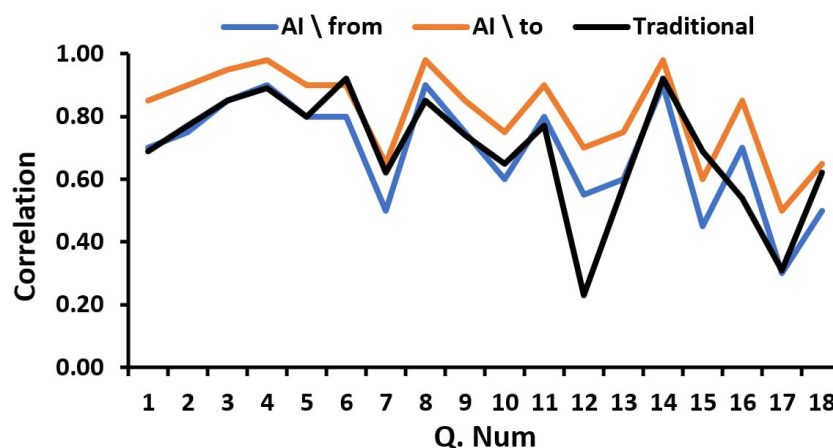


Figure 2. Comparison of AI-predicted difficulty and traditional values

Table 5 compares the initial discrimination estimates from the ChatGPT AI model without administering the test with the actual discrimination values after analyzing student responses to the post-test. Data revealed that a majority of the items had quite high discrimination values. But, two items, viz. Q12 and Q17 appeared to exhibit negative discrimination indices (-0.19 & -0.38, respectively), whereas the other items showed positive discrimination indices. Besides, ChatGPT produced qualitative discrimination forecasts rather than exact numeric psychometric coefficients, which hinder a direct numerical comparison between AI-based and traditional results.

Table 5. AI Prognosis vs real discrimination index results

Q. Num	AI	Traditional
1	(0.40+)	0.56
2	(0.50+)	0.81
3	(0.40+)	0.81
4	(0.35+)	0.88
5	(0.45+)	0.81
6	(0.40+)	0.88
7	(0.55+)	0.81
8	(0.40+)	0.94
9	(0.40+)	0.94
10	(0.45+)	0.94
11	(0.50+)	0.94
12	(0.40+)	- 0.19
13	(0.40+)	0.88
14	(0.30+)	0.88
15	(0.40+)	0.81
16	(0.40+)	0.81
17	(-) Negative	- 0.38
18	(0.45+)	0.81

The report on the volume of difficulties suggested that the vast majority of observed levels were within the margins estimated by the AI tool. So, it indicates the possibility of using AI as a first approximation tool for assessment. Still, there were some points of departure, especially in cases where the actual level of difficulty is very different from the range. As for discrimination, ChatGPT only provided a qualitative estimate rather than a quantitative measure. As for discrimination, ChatGPT only provided a qualitative estimate rather than a quantitative measure. Yet, the model identified

questions that potentially discriminate against problems related to discrimination; two items (Q12, Q17) display negative discrimination pointers (-0.19 and -0.38, respectively). Testing of the response distributions revealed different answer-selection patterns. In Q12, more participants (32 out of 65) left the question unanswered. At the same time, Q17 showed a substantial concentration of responses on distractor A (22 participants) compared with only 6 participants choosing the correct answer. These patterns suggest possible issues related to item wording, distractor construction, or answer-key alignment that may have confused participant and affected their performance. The AI-based pre-assessment had already identified these items as potentially problematic before the exam was administered. These results highlight the practical value of AI-assisted pre-assessment as an early diagnostic technique to identify weak questions before they are used in real exams.

Distractors' Effectiveness

Table 6 compares the outcomes of the standard post-test analysis based on students' actual performance with the ChatGPT AI model's pre-assessment of distractor quality before the test.

Table 6. Comparison of AI-based and traditional evaluation of distractor effectiveness

AI/Traditional	Good	Accepted	Weak
Good	9	2	0
Accepted	1	1	1
Weak	0	1	3

A confusion matrix was developed to visually illustrate the agreement between AI-based distractor classifications and the traditional method. This matrix allows visualization of whether items received the same classification consistently or inconsistently. Table 6 shows the level of agreement and disagreement between the AI-based pre-assessment and conventional post-test distractor classification for each of the analyzed pre-assessment MCQ items. The diagonal entries indicate agreement, and the off-diagonal entries indicate disagreement. The obtained values are presented in Table 7, which also shows that agreement between AI-based and traditional distractor classifications was evaluated using Cohen's kappa coefficient.

Table 7. Agreement between two evaluation types

Measure	Value
Agreement	77.8%
Cohen's Kappa (k)	0.584
Description	Moderate agreement

Finally, for distractor assessment, the agreement test showed that the AI's marking decisions and those of a human agreed with 77.8% accuracy across the items analyzed. The confusion matrix of the agreement test that the bulk of the classification was consistent and formed the main diagonal, while the disagreements were mostly due to borderline cases between neighboring categories. In addition, Cohen's kappa statistic revealed moderate agreement ($k = 0.584$) after accounting for chance agreement, as described in the established interpretive guideline [29]. Really, ChatGPT was mainly effective at detecting harmful distractors, mostly those involving very few students. But ChatGPT was quite a bit less reliable in borderline cases, where the assignment of distractors depended far more on the genuine distributions of student responses. In these situations, which arise from live human interaction, textual analysis is insufficient for classification. Because of this, these findings support the use of AI-assisted assessment as an aid to, not a replacement for, standard psychometric analysis.

These experimental results lend credence to the theory that using artificial intelligence-based pre-assessment as a first diagnostic tool will provide valuable support for traditional psychometric testing and that combining the two methods will yield higher-quality multiple-choice test items that have yet to be used. It will also provide a feasible, practical assessment measure that could spark interest in Arab classrooms, especially in resource-limited settings or classes with large student populations.

CONCLUSION

This research paper suggests a new hybrid setup that integrates ChatGPT-based upfront assessment with conventional psychometric post-test analysis for Arabic multiple-choice questions (MCQs). This approach fuses AI-assisted linguistic evaluation with empirically based evaluation of actual learner response data to enable early detection of item-quality problems, even before the examination is administered. The findings indicate that AI-based pre-assessment can identify unmet needs in question and distractor development and show a moderate correlation with standard psychometric results ($r = 0.42$). And the findings suggest the central role of AI as an immediate performance tool for inspecting test items even before the examination starts. Still, AI predictions were less precise for the discrimination parameter, which is highly dependent on actual student response samples. This pattern is consistent with using AI-assisted evaluation as an addition rather than a replacement. Though the outcomes are encouraging, the research was done on a limited dataset of only 18 MCQs, distributed over three separate test formats. Therefore, the results should be viewed as initial verification and may not correctly represent many content sectors or difficulty levels. Future studies could leverage larger, more diverse item groups that include additional linguistic and semantic characteristics and compare many AI models to increase the effectiveness and generalizability of AI-based assessment models.

SUPPLEMENTARY MATERIAL

The following supporting information can be downloaded at: <https://mjs.uomustansiriyah.edu.iq/index.php/MJS/article/view/1857/843>

AUTHOR CONTRIBUTIONS

Omar Dahham Sarhan: Conceptualization, Methodology, Software, Data curation, Formal analysis, Investigation, Validation, Visualization, Writing – original draft, Writing – review & editing.

FUNDING

This research received no external funding.

DATA AVAILABILITY STATEMENT

The main data used to reach the conclusions of this research are published in the paper and supplementary material. To further transparency and reproducibility, additional response matrices have been shared, which display the frequency distribution of student responses for each MCQ item.

ACKNOWLEDGMENTS

The authors would like to thank the University of Anbar for its support in the present work.

CONFLICTS OF INTEREST

The author declares no conflicts of interest.

DECLARATION OF GENERATIVE AI USE

Generative AI technology (ChatGPT) was used both as the research topic and the study's analytical instrument in line with the study's experimental design. The AI system assisted with the preliminary assessment of Arabic multiple-choice questions (MCQs) using the methods described in the paper. The authors did not use any generative AI-based systems to automatically write the manuscript, explain the results, or draw scientific conclusions.

REFERENCES

- [1] E. Emekli and B. N. Karahan, "AI in radiography education: Evaluating multiple-choice questions difficulty and discrimination," *Journal of Medical Imaging and Radiation Sciences*, vol. 56, no. 4, Art no. 101896, 2025, doi: [10.1016/j.jmir.2025.101896](https://doi.org/10.1016/j.jmir.2025.101896).
- [2] Q. Zhang, Z. Huang, Y. Huang, G. Wang, R. Zhang, J. Yang, Y. Cheng, B. Chen, H. Wang, K. Qiu, et al., "Generative AI in medical education: Feasibility and educational value of LLM-generated clinical cases with MCQs," *BMC Medical Education*, vol. 25, no. 1, Art no. 1502, 2025, doi: [10.1186/s12909-025-08085-8](https://doi.org/10.1186/s12909-025-08085-8).
- [3] C. Grévisse, M. A. S. Pavlou, and J. G. Schneider, "Docimological quality analysis of LLM-generated multiple choice questions in computer science and medicine," *SN Computer Science*, vol. 5, no. 5, Art no. 636, 2024, doi: [10.1007/s42979-024-02963-6](https://doi.org/10.1007/s42979-024-02963-6).
- [4] M. Kaya, E. Sonmez, A. Halici, H. Yildirim, and A. Coskun, "Comparison of AI-generated and clinician-designed multiple-choice questions in emergency medicine exam: A psychometric analysis," *BMC Medical Education*, vol. 25, no. 1, Art no. 949, 2025, doi: [10.1186/s12909-025-07528-6](https://doi.org/10.1186/s12909-025-07528-6).
- [5] A. Strugatski and G. Alexandron, "Applying IRT to distinguish between human and generative AI responses to multiple-choice assessments," in *Proceedings of the 15th International Learning Analytics and Knowledge Conference*, ACM, Mar. 2025, pp. 817–823, doi: [10.1145/3706468.3706490](https://doi.org/10.1145/3706468.3706490).
- [6] M. Sabqat, R. A. Khan, M. Jawaid, and M. Sajjad, "Artificial intelligence meets item analysis (AI meets IA): A study of chatbot training and performance in detecting and correcting MCQ flaws," *Pakistan Journal of Medical Sciences*, vol. 41, no. 3, pp. 652–656, 2025, doi: [10.12669/pjms.41.3.11224](https://doi.org/10.12669/pjms.41.3.11224).
- [7] Y. Shi, K. Yu, Y. Dong, and F. Chen, "Large language models in education: A systematic review of empirical applications, benefits, and challenges," *Computers and Education: Artificial Intelligence*, vol. 10, Art no. 100529, Jun. 2026, doi: [10.1016/j.caeai.2025.100529](https://doi.org/10.1016/j.caeai.2025.100529).
- [8] S. Moore, H. A. Nguyen, T. Chen, and J. Stamper, "Assessing the quality of multiple-choice questions using GPT-4 and rule-based methods," in *Responsive and Sustainable Educational Futures*. Springer Nature Switzerland, 2023, vol. 14200, pp. 229–245, doi: [10.1007/978-3-031-42682-7_16](https://doi.org/10.1007/978-3-031-42682-7_16).
- [9] S. Moore, E. Costello, H. A. Nguyen, and J. Stamper, "An automatic question usability evaluation toolkit," in *Artificial Intelligence in Education*. Springer Nature Switzerland, 2024, vol. 14830, pp. 31–46, doi: [10.1007/978-3-031-64299-9_3](https://doi.org/10.1007/978-3-031-64299-9_3).
- [10] S. Guizani, T. Mazhar, T. Shahzad, W. Ahmad, A. Bibi, and H. Hamam, "A systematic literature review to implement large language model in higher education: issues and solutions," *Discover Education*, vol. 4, no. 1, Art no. 35, 2025, doi: [10.1007/s44217-025-00424-7](https://doi.org/10.1007/s44217-025-00424-7).
- [11] T. A. May, Y. K. Fan, G. E. Stone, K. L. K. Koskey, C. J. Sondergeld, T. D. Folger, J. N. Archer, K. Provinzano, and C. C. Johnson, "An effectiveness study of generative artificial intelligence tools used to develop multiple-choice test items," *Education Sciences*, vol. 15, no. 2, Art no. 144, 2025, doi: [10.3390/educsci15020144](https://doi.org/10.3390/educsci15020144).
- [12] C. Isley, J. Gilbert, E. Kassos, M. Kocher, A. Nie, E. Brunskill, B. Domingue, J. Hofman, J. Legewie, T. Svoronos, et al., "Assessing the quality of AI-generated exams: A large-scale field study," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 40, no. 45, pp. 38 626–38 634, 2026, doi: [10.1609/aaai.v40i45.41205](https://doi.org/10.1609/aaai.v40i45.41205).
- [13] R. B. Jabr and A. M. Azmi, "Knowledge-Aware Arabic question generation: A transformer-based framework," *Mathematics*, vol. 13, no. 18, Art no. 2975, 2025, doi: [10.3390/math13182975](https://doi.org/10.3390/math13182975).
- [14] Y. Artsi, V. Sorin, E. Konen, B. S. Glicksberg, G. Nadkarni, and E. Klang, "Large language models for generating medical examinations: Systematic review," *BMC Medical Education*, vol. 24, no. 1, Art no. 354, 2024, doi: [10.1186/s12909-024-05239-y](https://doi.org/10.1186/s12909-024-05239-y).
- [15] M. Kanik, "The use of ChatGPT in assessment," *International Journal of Assessment Tools in Education*, vol. 11, no. 3, pp. 608–621, 2024, doi: [10.21449/ijate.1379647](https://doi.org/10.21449/ijate.1379647).
- [16] S. Al Faraby, A. Romadhony, and Adiwijaya, "Analysis of LLMs for educational question classification and generation," *Computers and Education: Artificial Intelligence*, vol. 7, Art no. 100298, Dec. 2024, doi: [10.1016/j.caeai.2024.100298](https://doi.org/10.1016/j.caeai.2024.100298).
- [17] L. Goorts, R. Hollevoet, V. Xia, F. Cammaerts, and A. Güngör, "How do LLMs perform in the context of MCQs across different levels of thinking skills in a business education course at higher education? A comparison of ChatGPT, Gemini, and Copilot," *Computers and Education: Artificial Intelligence*, vol. 9, Art no. 100475, Dec. 2025, doi: [10.1016/j.caeai.2025.100475](https://doi.org/10.1016/j.caeai.2025.100475).
- [18] J. Doughty, Z. Wan, A. Bompelli, J. Qayum, T. Wang, J. Zhang, Y. Zheng, A. Doyle, P. Sridhar, A. Agarwal, et al., "A comparative study of AI-generated (GPT-4) and human-crafted MCQs in programming education," in *Proceedings of the 26th Australasian Computing Education Conference*, ser. ACE 2024, ACM, Jan. 2024, pp. 114–123, doi: [10.1145/3636243.3636256](https://doi.org/10.1145/3636243.3636256).

- [19] J. S. P. Baudin, "Assessing the psychometric properties of AI-generated multiple-choice exams in a psychology subject," *Journal of Pedagogical Sociology and Psychology*, vol. 7, no. 3, pp. 18–34, 2025, doi: [10.33902/jpsp.202536891](https://doi.org/10.33902/jpsp.202536891).
- [20] M. Elzayyat, J. N. Mohammad, and S. Zaqout, "Assessing LLM-generated vs. expert-created clinical anatomy MCQs: A student perception-based comparative study in medical education," *Medical Education Online*, vol. 30, no. 1, Art no. 2554678, 2025, doi: [10.1080/10872981.2025.2554678](https://doi.org/10.1080/10872981.2025.2554678).
- [21] Q. Zhao and M. Zhang, "Elimination-based reasoning with LLM for multiple-choice educational question answering," *Journal of King Saud University Computer and Information Sciences*, vol. 37, no. 7, Art no. 204, 2025, doi: [10.1007/s44443-025-00122-2](https://doi.org/10.1007/s44443-025-00122-2).
- [22] H. Maeda, "Field-Testing multiple-choice questions with AI examinees: English grammar items," *Educational and Psychological Measurement*, vol. 85, no. 2, pp. 221–244, 2024, doi: [10.1177/00131644241281053](https://doi.org/10.1177/00131644241281053).
- [23] H. Fawareh, "Evaluation of cloud computing for advancement LMS through different environments," *International Journal of Advances in Soft Computing and its Applications*, vol. 16, no. 3, pp. 125–148, 2024, doi: [10.15849/ijasca.241130.08](https://doi.org/10.15849/ijasca.241130.08).
- [24] H. W. Awalurahman and I. Budi, "Automatic distractor generation in multiple-choice questions: A systematic literature review," *PeerJ Computer Science*, vol. 10, Art no. e2441, Nov. 2024, doi: [10.7717/peerj-cs.2441](https://doi.org/10.7717/peerj-cs.2441).
- [25] L. Benedetto, S. Taslimipoor, and P. Buttery, "A survey on automated distractor evaluation in multiple-choice tasks," in *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, Association for Computational Linguistics, Jul. 2025, pp. 55–69, doi: [10.18653/v1/2025.bea-1.5](https://doi.org/10.18653/v1/2025.bea-1.5).
- [26] T. M. Haladyna and S. M. Downing, "A taxonomy of multiple-choice item-writing rules," *Applied Measurement in Education*, vol. 2, no. 1, pp. 37–50, 1989, doi: [10.1207/s15324818ame0201_3](https://doi.org/10.1207/s15324818ame0201_3).
- [27] M. Tavakol and R. Dennick, "Making sense of Cronbach's alpha," *International Journal of Medical Education*, vol. 2, pp. 53–55, Jun. 2011, doi: [10.5116/ijme.4dfb.8dfd](https://doi.org/10.5116/ijme.4dfb.8dfd).
- [28] J. Cohen, *Statistical power analysis for the behavioral sciences*. Routledge, May 2013, doi: [10.4324/9780203771587](https://doi.org/10.4324/9780203771587).
- [29] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, vol. 33, no. 1, pp. 159–174, 1977, doi: [10.2307/2529310](https://doi.org/10.2307/2529310).