JMALS

Journal of Medical and Life Science

https://jmals.journals.ekb.eg/

SPBH

Post-validation and Item Response Analysis of Multiple-Choice Questions in Undergraduate Medical Education: A Psychometric Study

Dr. Jyothsnya S[1], Dr. Manjula MJ[2], Mrs. GokilaVani M[3], Dr Subashini Shanmuganandam[4]

[1]Associate Professor, Department of Pharmacology, St. Peter's Medical College, Hospital and Research Institute, Hosur, Tamilnadu – 635130, India

[2]Assistant Professor, Department of Pharmacology, Rajarajeswari Medical College and Hospital, Affiliated to Dr M.G.R. Educational and Research Institute, Bengaluru 560074, India

[3]Tutor, Department of Pharmacology, St. Peter's Medical College, Hospital and Research Institute, Hosur, Tamilnadu - 635130, India

[4]Assistant Professor, Department of Pharmacology, Nandha Medical College & Hospital, Erode, India

Corresponding author: Dr. Manjula MJ, Assistant Professor, Department of Pharmacology, Rajarajeswari Medical College and Hospital. Email: manjula.mj03@gmail.com

DOI:10.21608/jmals.2025.419704.1069

Abstract

Background: Multiple-choice questions (MCQs) are the most commonly used assessment tool in medical education. However, their effectiveness depends on the quality of item construction. Psychometric evaluation using Classical Test Theory (CTT) and Item Response Theory (IRT) helps in improving assessment validity and reliability. **Objective:** To analyze the quality of MCQs administered to second-year MBBS students using item analysis and IRT parameters, and to identify poorly functioning items for revision. **Methods:** A total of 55 MCQs were administered to 149 second-year MBBS students. The difficulty index, discrimination index, distractor efficiency, and reliability (KR-20) were computed. Items were categorized into high, moderate, or poor based on standard criteria. Data were analyzed using **Xcalibre** and **SPSS**. **Results:** The theta range we obtained was -4.0 to +4.0. The mean difficulty index was 0.55, with 51% of items classified as easy, 31% moderate, and 18% difficult. The average discrimination index was 0.43, with 40% excellent and 22% good items. Distractor efficiency was high, with 91% of items showing 100% efficiency. KR-20 reliability was 0.856. Items with poor or negative discrimination (e.g., Q44, Q52) and non-functional distractors (e.g., Q25, Q26, Q33, Q34, Q39) were identified. **Conclusion:** The MCQ set demonstrated moderate difficulty, high discriminative capacity, and excellent reliability. Periodic item analysis helps in refining flawed items, ensuring the creation of a validated question bank.

Key words: Medical education; Item analysis; MCQs; Reliability; Psychometrics; Assessment validity

Introduction

Assessment in medical education serves not only to evaluate student performance but also to guide learning and curricular development. Among various tools, multiple-choice questions (MCQs) are preferred because they allow objective scoring, cover a wide range of content, and minimize examiner bias (1). Despite these advantages, the utility of MCQs depends on their quality (2). Poorly constructed items may compromise validity, fail to discriminate between high- and low-achieving

Received: August 2, 2025. Accepted: October 9, 2025. Published: October 26, 2025

students, and reduce the reliability of examinations (3,4).

Assessment influences student learning through four key aspects: the content assessed, the format used, the timing of the assessment, and the feedback provided to medical students. Hence, testing for the quality of the mode of assessment is crucial (5). To ensure quality, psychometric analysis of MCQs has become an essential practice. Classical Test Theory (CTT) provides indices such as the difficulty index (proportion of correct responses) and discrimination index (ability of an item to distinguish between strong and weak students). In addition, evaluation of distractor efficiency ensures that incorrect options are plausible and contribute meaningfully to the test (6,7).

Beyond CTT, Item Response Theory (IRT) allows modelling of student ability (theta) and item characteristics, providing a deeper insight into test functioning (8). Together, these approaches can guide the refinement of MCQs, leading to the development of a robust question bank.

Several studies across medical schools in India and abroad have shown that item analysis improves the reliability of assessments and identifies common flaws such as ambiguous stems, implausible distractors, and mis-keyed options. However, systematic reporting of psychometric properties of undergraduate assessments remains limited in Indian settings.

This study was therefore undertaken to evaluate MCQs administered to second MBBS students using CTT and IRT parameters, to assess reliability, and to identify items requiring revision for future use.

Methodology

Study Design and Participants

This was a cross-sectional psychometric analysis of MCQ responses. A total of 149 second-year MBBS students from St. Peter's Medical College and Hospital participated. The test consisted of 55 single-best-answer MCQs, each with four options.

Item Analysis

Responses were analyzed using Classical Test Theory (CTT) parameters:

1. **Difficulty Index (P):** Proportion of students who answered an item correctly. Items were classified as very difficult (<0.20), difficult (0.21–0.40), moderate (0.41–0.60), easy (0.61–0.80), or very easy (>0.80).
2. **Discrimination Index (D):** Calculated by comparing high-achievers (top 50 students) with low-achievers (bottom 50 students). Items were categorized as excellent (>0.40), good (0.30–0.39), fair (0.20–0.29), poor (0.00–0.19), or negative (<0.00).
3. **Distractor Efficiency (DE):** Evaluated based on the percentage of functional distractors. Items were considered high efficiency (100%), moderate (66.6%), or poor (<50%) (9).
4. **Reliability:** Internal consistency reliability was measured using the Kuder-Richardson Formula 20 (KR-20) (9,10).

Statistical Analysis: Item parameters and student ability (theta) were generated using **Xcalibre** software. Correlation analysis between indices was performed using **SPSS**. Graphs and tables were generated to illustrate findings.

Results

The theta range we obtained for 55 items and 149 sample sizes was -4.0 to +4.0. This indicated that the data were suitable for IRT analysis. The obtained analyses are represented as tables and graphs below.

The average difficulty index was 0.55, indicating a moderate overall level. More than half of the items (51%) were classified as easy, while 31% were of moderate/ideal difficulty and 18% were difficult. This distribution shows a slight skew towards easier questions, which may boost student confidence but should be balanced with more moderate items for optimal assessment.

Table 1: Distribution of the items based on the difficulty index

Difficulty Category	**Number of Items**	**Percentage (%)**
Easy (>0.60)	28	51
Moderate (0.40–0.60)	17	31
Difficult (<0.40)	10	18

Table 2: Distribution of the items based on the discrimination index

Discrimination Category	**Number of Items**	**Percentage (%)**
Excellent (>0.40)	22	40
Good (0.30–0.39)	12	22
Fair (0.20–0.29)	8	15
Poor (0.00–0.19)	11	20
Negative (<0.00)	2	4

The mean discrimination index was 0.43, reflecting strong discriminatory capacity. About 62% of items demonstrated excellent to good discrimination, while 15% were fair and 20% were poor. Two items showed negative discrimination, requiring urgent review as they reduce the test's validity.

Table 3: Distraction efficiency of items

Efficiency Category	**Number of Items**	**Percentage (%)**
High (100%)	50	91
Moderate (66.6%)	5	9

Most items (91%) had perfect distractor efficiency, meaning all options were plausible and functional. Only 9% of items had moderate efficiency due to the presence of non-functional distractors, often linked to option 'D'. This highlights the generally strong quality of distractor construction in the test.

Table 4: Reliability of the test: KR 20 calculation analysis

Parameter	**Value**
Number of items (k)	55
Variance of total scores (St^2)	77.8962
Σ (pi × qi) across all items	12.4223
KR-20 Reliability Coefficient	**0.856**

The KR-20 value was 0.856, indicating excellent internal consistency. This suggests that the items reliably measured a common construct and that the test scores were dependable. Such reliability supports the use of this MCQ set for summative evaluation.

Table 5: Correlation matrix

Parameter Pair	Correlation (r)	Interpretation
Distractor Efficiency vs Difficulty	-0.471	Moderate negative correlation (as items get more difficult, distractors become more functional).
Distractor Efficiency vs Discrimination	-0.024	Very weak/near-zero correlation (distractor quality does not strongly influence discrimination).
Difficulty vs Discrimination	+0.594	Moderate positive correlation (moderately difficult items tend to discriminate better).

A moderate positive correlation (r = 0.59) was found between difficulty and discrimination, suggesting moderately difficult items tend to discriminate better. Distractor efficiency showed a weak negative correlation with both difficulty and discrimination, indicating that well-functioning distractors do not necessarily guarantee item discrimination.

Table 6: The items flagged for revision

Issue Identified	Item Numbers	Suggested Action
Negative Discrimination	44, 52	Review keying/clarity of stem and options.
Poor Discrimination (0.00–0.19)	1, 14, 18, 20, 24, 25, 30, 31, 35, 46, 50, 51	Revise the stem and options to better reflect the intended concept.
Non-functional Distractors	13, 25, 26, 33, 34, 39	Reconstruct distractor ‘D’ to make it plausible.

Two items (44 and 52) had negative discrimination, while 11 showed poor discrimination, warranting revision of stems or keys. Additionally, six items had non-functional distractors, requiring modification of incorrect options to improve plausibility. Regular review of these flagged items will strengthen the MCQ bank.

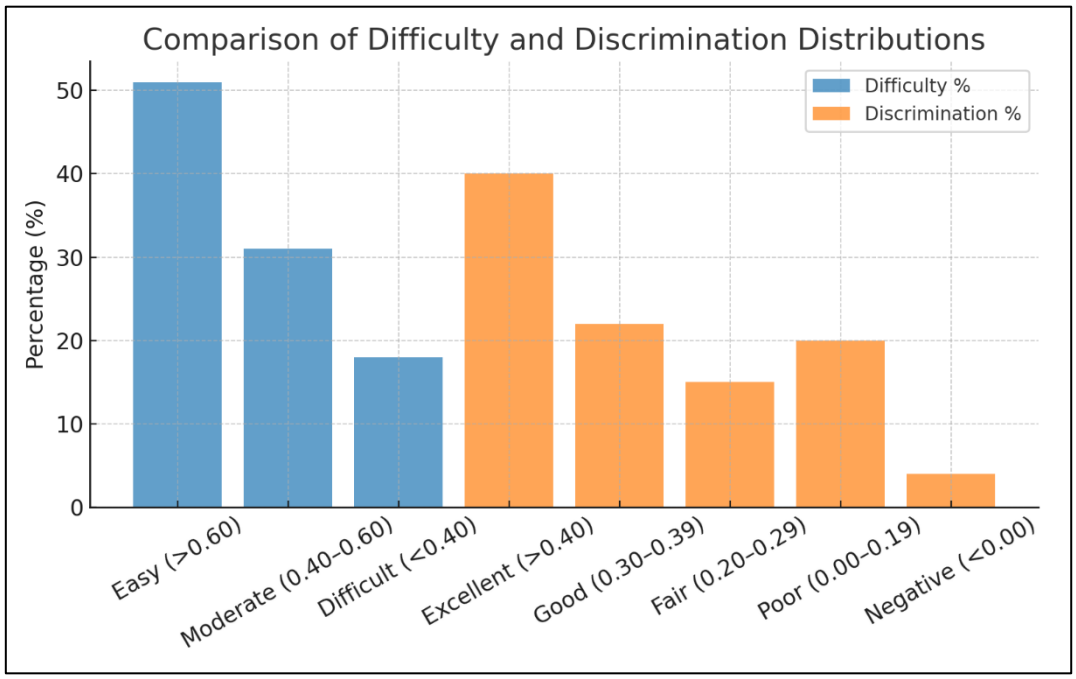

Figure 1: Bar graph depicting the comparison of difficulty and discrimination distributions

In this bar chart, the **x-axis** represents the different categories of item quality: difficulty levels (easy, moderate, difficult) and discrimination levels (excellent, good, fair, poor, negative). The **y-axis** shows the percentage of items in each category. The height of each bar reflects how many questions fell into that category. The use of separate bars for difficulty and discrimination highlights how items were distributed across these two important indices. This graph indicates that although many items were on the easier side, a substantial proportion demonstrated excellent or good discrimination.

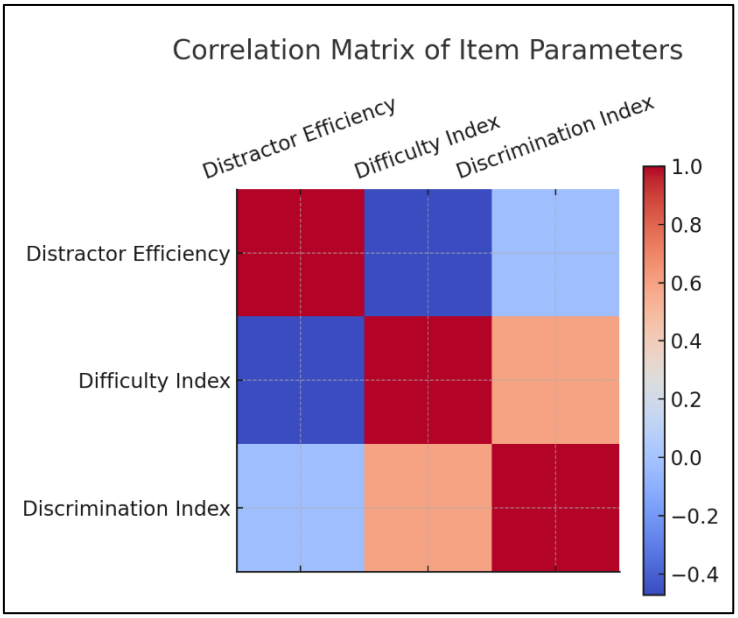

Figure 2: Correlation heat map of item parameters

In the heatmap, both the **x-axis and y-axis** represent the three item parameters: distractor efficiency, difficulty index, and discrimination index. The colours indicate the strength and direction of correlation between these variables, with warm tones (red) showing positive correlation and cool tones (blue) showing negative correlation. The diagonal values (darkest shade) represent perfect correlation of a parameter with itself (r = 1). For example, the blue shading between distractor efficiency and difficulty shows a moderate negative correlation, while the red shading between difficulty and discrimination indicates a moderate positive correlation. This graph visually summarizes how item properties are interrelated in the test.

Discussion

The present study evaluated 55 multiple-choice questions administered to second-year MBBS students through psychometric analysis. The findings suggest that the test was well-balanced, reliable, and capable of differentiating between high- and low-performing students.

The mean difficulty index was 0.55, which falls within the ideal range (0.40–0.60). More than half of the questions (51%) were easy, while 31% were moderately difficult and 18% were difficult. This distribution reflects a slight leaning towards easier items, which can be beneficial for student confidence but may slightly reduce discriminatory power. Previous studies in Indian medical colleges **Patil et al.,** had analyzed the 90 distractors (derived from 3 sets of 30 MCQs). The mean values for the difficulty index, discrimination index, and distractor efficiency in their study were 38.3%, 0.27, and 82.8%, respectively. Among their 30 items, 11 were classified as difficult (DIF I <30%), while 5 were considered easy (DIF I >60%). Overall, 15 items demonstrated a very good discrimination index. Out of the 90 distractors, 16 (17.8%) were identified as non-functional distractors (NFDs), occurring in 13 items (43.3%). Whereas **Gajjar et al.,** also have reported similar findings, emphasizing the importance of maintaining a balanced distribution to ensure fairness and challenge, which were almost similar to our study (11,12).

The mean discrimination index was 0.43, with 62% of items falling into excellent or good categories. This indicates that the majority of items successfully differentiated between high- and low-achieving students. A small proportion (20%) demonstrated poor discrimination, and 4% were negatively discriminating. Negative discrimination is a concerning finding, as it implies that weaker students performed better on those items than stronger students. Possible causes include ambiguous wording, mis-keyed answers, or misleading distractors. Similar issues have been highlighted in previous psychometric studies. **Hingorjo et al** reported, DI > 0.35 was 62%, DI ranging between 0.25 and 0.34 with an incidence of 14%, and DI 0.15 - 0.24 were found to be 12%. Two items each in their study had negative and zero DI (13). Meanwhile, **Tarrant M et al,** underscore the need for continuous item review (14).

Distractor efficiency was generally high, with 91% of items having 100% functional distractors. This reflects careful construction of distractors, ensuring that incorrect options were plausible and contributed to the challenge of the test. Only 5 items had non-functional distractors (primarily option D), which is consistent with international findings where one or two distractors often fail to function effectively. Improving these options by basing them on common misconceptions may further enhance test quality.

Reliability of the test, measured using KR-20, was 0.856, which is considered excellent. This indicates that the test items measured a common construct consistently and provided dependable results. Comparable studies in undergraduate medical education have reported reliability indices ranging from 0.70 to 0.85, placing our findings at the higher end of the spectrum (15,16).

Overall, our findings suggest that the majority of items were well-constructed, reliable, and discriminatory. However, items with poor or negative discrimination and those with non-functional distractors must be revised or eliminated before future use. Regular item analysis not only strengthens the validity of assessments but also contributes to the creation of a validated MCQ bank, which is essential for maintaining quality in medical education.

Conclusion

The MCQ set administered to second-year MBBS students showed moderate difficulty, excellent discrimination, and high reliability. A small subset of items with poor discrimination and non-functional distractors requires revision. Incorporating regular psychometric evaluation into assessment practices

will help build a validated question bank, improving the quality and fairness of medical education assessments.

Conflict of interest: NIL
Funding: NIL

REFERENCES

1. Gottlieb M, Bailitz J, Fix M, Shappell E, Wagner MJ. Educator's blueprint: A how-to guide for developing high-quality multiple-choice questions. AEM Educ Train. 2023 Jan 24;7(1):e10836.
2. Rashwan NI, Aref SR, Nayel OA, Rizk MH. Postexamination item analysis of undergraduate pediatric multiple-choice questions exam: implications for developing a validated question Bank. BMC Medical Education (2024) 24:168
3. Al-Faris EA, Alorainy IA, Abdel-Hameed AA, Al-Rukban MO. A practical discussion to avoid common pitfalls when constructing multiple choice questions items. J Family Community Med. 2010 May;17(2):96-102
4. Balaha MH, El-Ibiary MT, El-Dorf AA, El-Shewaikh SL, Balaha HM. Construction and Writing Flaws of the Multiple-Choice Questions in the Published Test Banks of Obstetrics and Gynecology: Adoption, Caution, or Mitigation? Avicenna J Med. 2022 Aug 31;12(3):138-147.
5. CPM VDV. The assessment of professional competence: developments, research and practical implications. Adv Health Sci Educ. 1996;1(1):41–67
6. Salih KEMA, Jibo A, Ishaq M, Khan S, Mohammed OA, Al-Shahrani AM, Abbas M. Psychometric analysis of multiple-choice questions in an innovative curriculum in Kingdom of Saudi Arabia. J Family Med Prim Care. 2020 Jul 30;9(7):3663-3668.
7. Cappelleri JC, Jason Lundy J, Hays RD. Overview of classical test theory and item response theory for the quantitative assessment of items in developing patient-reported outcomes measures. Clin Ther. 2014 May;36(5):648-62.
8. Yang FM, Kao ST. Item response theory for measurement validity. Shanghai Arch Psychiatry. 2014 Jun;26(3):171-7.
9. Rezigalla AA, Eleragi AMESA, Elhussein AB, Alfaifi J, ALGhamdi MA, Al Ameer AY et al. Item analysis: the impact of distractor efficiency on the difficulty index and discrimination power of multiple-choice items. BMC Med Educ. 2024 Apr 24;24(1):445.
10. Kumar D, Jaipurkar R, Shekhar A, Sikri G, Srinivas V. Item analysis of multiple choice questions: A quality assurance test for an assessment tool. Med J Armed Forces India. 2021 Feb;77(Suppl 1):S85-S89.
11. Patil R, Palve SB, Vell K, Boratne AV. Evaluation of multiple choice questions by item analysis in a medical college at Pondicherry, India. Int J Community Med Public Health 2016;3:1612-6.
12. Gajjar S, Sharma R, Kumar P, Rana M. Item and test analysis to identify quality multiple choice questions (MCQs) from an assessment of medical students of Ahmedabad, Gujarat. Indian Journal of Community Medicine. 2014;39(1):17-20.
13. Hingorjo MR, Jaleel F. Analysis of One-Best MCQs: the Difficulty Index, Discrimination Index and Distractor Efficiency. J Pak Med Assoc; Vol. 62, No. 2, February 2012: 142-147.
14. Tarrant M, Ware J. Impact of item-writing flaws in multiple choice questions on student achievement in high-stakes nursing assessments. Med Educ 2008; 42: 198-206.
15. Mukherjee P, Lahiri SK. Analysis of Multiple Choice Questions (MCQs): Item and Test Statistics from an assessment in a medical college of Kolkata, West Bengal. IOSR Journal of Dental and Medical Sciences (IOSR-JDMS). 2015; 14(12): 47 – 52
16. Downing SM. WRITTEN T ESTS: Constructed-Response and Selected-Response Formats. In Assessment in health professions education 2009;1(1): 169-204.