International Journal of Medical and Pharmaceutical Research

Online ISSN-2958-3683 | Print ISSN-2958-3675
Frequency: Bi-Monthly
Available online on: https://ijmpr.in/

Research Article

Item Analysis of Multiple-Choice Questions in Undergraduate Medical Assessment: A Cross-Sectional Study in a Medical College in Maharashtra

Dr. Ranjana Zade

[1] *Assistant Professor, Department of Community Medicine, Rajiv Gandhi Medical College and Chhatrapati Shivaji Maharaj Hospital Thane Maharashtra*

OPEN ACCESS

Corresponding Author:

Dr. Ranjana Zade
Assistant Professor, Department of Community Medicine, Rajiv Gandhi Medical College and Chhatrapati Shivaji Maharaj Hospital Thane Maharashtra

Received: 11-09-2025
Accepted: 27-09-2025
Available online: 20-10-2025

ABSTRACT

Background: Multiple Choice Questions (MCQs) provide objective, scalable assessment aligned with competency-based medical education. However, their educational value depends on psychometric quality, not content coverage alone. Item analysis, performed after test administration, yields actionable indices: difficulty index (P) and discrimination index (DI). These indices support data-driven retention, revision or removal of items and promote a curated question bank. **Methodology**: A 30-item, single-best-answer MCQ test (four options per item; one key and three distractors) was administered as per the academic schedule. Answer sheets were scored, and the total marks were used to form performance tertiles (upper, middle, lower) for discrimination calculations. **Results**: A total of 35 sixth-semester MBBS students' scripts were analysed. The mean test score was 15.86 ± 3.33 (median 15, range 7–23) out of 30. This able to discriminate students in good performer and bad performer and also there are significant number of students who require more preparation. There is no ceiling effect or floor effect. Item analysis states that the first item on the test has a difficulty ranking of 30, ie it is the 20th most difficult item on the test, the third item a ranking of 28. The easiest items on the test are items 22, 5 and 26. Clearly the items need to be rearranged so the easiest items are towards the front of the test. **Discussion**: Our test shows a balanced difficulty profile with 50% moderate and 30% easy items, which is pedagogically desirable for separating abilities without overwhelming the cohort.

Keywords: *MCQ, difficulty ranking, ceiling effect, distractors.*

INTRODUCTION

Multiple Choice Questions (MCQs) provide objective, scalable assessment aligned with competency-based medical education. However, their educational value depends on psychometric quality, not content coverage alone. Item analysis, performed after test administration, yields actionable indices: difficulty index (P) and discrimination index (DI). These indices support data-driven retention, revision or removal of items and promote a curated question bank. Prior studies from Indian and international contexts have reported moderate difficulty as desirable, variable discrimination and the importance of plausible distractors in maintaining test validity.

Objectives

Primary Objective: To perform item analysis of an internal assessment MCQ test administered to undergraduate students of Community Medicine.
Secondary Objectives: To determine the difficulty index and discrimination index of each item.

MATERIALS AND METHODS

Study design: Present study was planned and conducted as a small project under "Basic course on Medical Education Technologies" workshop.
Study setting: Department of Community Medicine
Study population: Sixth-semester MBBS students who appeared for the department's internal assessment MCQ examination.
Inclusion criteria: All students who appeared for the exam and consented to use of anonymized data for analysis.

Exclusion criteria: Students absent during the exam or who did not provide consent.
Sample size: Universal inclusion of the appearing cohort (n = 35).
Study method: A 30-item, single-best-answer MCQ test (four options per item; one key and three distractors) was administered as per the academic schedule. Answer sheets were scored, and the total marks were used to form performance tertiles (upper, middle, lower) for discrimination calculations.

Operational definitions:

i. Difficulty index (P): proportion of examinees answering the item correctly; classified as Easy (>70%), Moderate (30–70%), Difficult (<30%).
ii. Discrimination index (DI): difference in the proportion of correct responses between upper and lower quartiles divided by the number per quartile; classified as Excellent (≥0.40), Good (0.30–0.39), Acceptable (0.20–0.29), Poor (<0.20).

Statistical analysis
Descriptive statistics (frequencies, proportions) were computed in Excel to summarise P and DI categories.

Ethical considerations
Participation in item analysis had no bearing on grades. Data were anonymized; confidentiality was maintained.

Results
A total of 35 sixth-semester MBBS students' scripts were analysed. The mean test score was 15.86 ± 3.33 (median 15, range 7–23) out of 30.

Table 1: Student's total score

Sr no	Score
1	15
2	19
3	17
4	17
5	17
6	21
7	15
8	22
9	17
10	17
11	14
12	21
13	17
14	15
15	15
16	13
17	13
18	15
19	17
20	15
21	7
22	10
23	13
24	16
25	14
26	12
27	17
28	20
29	23
30	13
31	12
32	18
33	15
34	15
35	18

Group statistics:
The whole score table is ranked from lowest to highest obtained score as shown in table 2 and divided into three group as upper third, middle third and lower third as shown in table 3.

Table 2: Group statistics B

Sr no	Score
21	7
22	10
26	12
31	12
16	13
17	13
23	13
30	13
11	14
25	14
1	15
7	15
14	15
15	15
18	15
20	15
33	15
34	15
24	16
3	17
4	17
5	17
9	17
10	17
13	17
19	17
27	17
32	18
35	18
2	19
28	20
6	21
12	21
8	22
29	23

Mean score of group is 15.86.

Table 3: Group statistics B

Group	No of students	Mean score
All	35	15.85
Lower	12	16.31
Middle	12	15.76
Upper	11	19.27

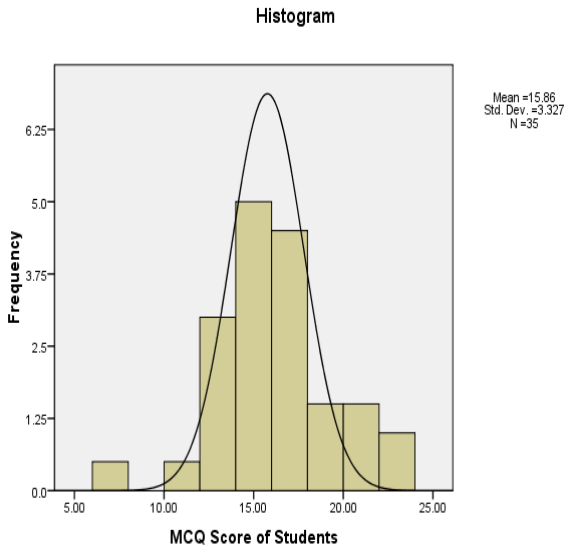
The Distribution of Students' Total Scores
Table 4 shows the Distribution of Students' Total Scores as calculated by cumulative frequency.

Table 4: The Distribution of Students' Total Scores

Score	No of students	Cumulative frequency
0	0	0
1	0	0
2	0	0
3	0	0
4	0	0
5	0	0
6	0	0
7	1	1
8	0	1
9	0	1
10	1	3
11	0	2
12	2	4
13	4	8
14	2	10
15	8	18
16	1	19
17	8	27
18	2	29
19	1	30
20	1	31
21	2	33
22	1	34
23	1	35

Total score frequencies:
Graph 1 shows how is the distribution of score. This shows the distribution which we expect from the test means larger group of students score lies about the mean. It suggests even spread of score about mean. This shows that test was not easier or not even so difficult as average no of student score about mean value. Mean score of all students is 15.86. This able to discriminate students in good performer and bad performer and also there are significant number of students who require more preparation. There is no ceiling effect or floor effect.

Graph1:



Analysis of the items on the test

Difficulty rank and item number
Items are ordered from least to most difficult. Item no 1 was the most difficult one only five students answer it correctly whereas item no 22 was answer correctly by 33 students.

Number of Categories (No. of Cat
Multiple choice tests have two categories of response ie correct or incorrect.

Item Score All: This column shows the number of students in the class who were correct on each of the items.

Item Proportion

all	the proportion of all students who were correct on each item.
lower	the proportion of students in the bottom third of the scores who were correct on each item.
middle	the proportion of students in the middle third of the scores who were correct on each item.
upper	the proportion of students in the upper third of the scores who were correct on each item.

Item difficulty (proportion correct for all students):

- Easy (>70%): 9/30 (30.0%)
- Moderate (30–70%): 15/30 (50.0%)
- Difficult (<30%): 6/30 (20.0%)

Overall, the average difficulty (mean proportion correct) was 0.542 (median 0.525), indicating a test centred near the desirable mid-difficulty range.

Table 5: Item analysis

Item ranking	Item no	Cat no	Item score	All	Lower	Middle	Upper	DI
1	22	2	33	0.91	0.83	1.00	1.00	0.17
2	5	2	32	0.88	0.91	0.91	0.90	-0.01
3	26	2	31	0.88	0.91	0.83	0.90	-0.01
4	6	2	31	0.88	0.83	0.83	1.00	0.17
5	2	2	31	0.85	0.83	0.91	0.90	0.07
6	25	2	30	0.8	0.83	0.83	0.90	0.07
7	17	2	28	0.8	0.83	0.66	0.90	0.07
8	12	2	28	0.74	0.66	0.83	1	0.34
9	14	2	26	0.71	0.66	0.58	1	0.34
10	30	2	25	0.68	0.5	0.75	0.9	0.4
11	7	2	24	0.68	0.58	0.83	0.63	0.05
12	15	2	24	0.65	0.91	0.58	0.63	-0.28
13	20	2	23	0.65	0.33	0.58	0.9	0.57
14	11	2	23	0.57	0.5	0.5	1	0.5
15	13	2	20	0.54	0.58	0.58	0.54	-0.04
16	24	2	19	0.51	0.16	0.58	0.81	0.65
17	27	2	18	0.48	0.16	0.66	0.72	0.56
18	29	2	17	0.45	0.33	0.41	0.81	0.48
19	28	2	16	0.42	0.08	0.41	0.72	0.64
20	19	2	15	0.4	0.41	0.33	0.63	0.22
21	8	2	14	0.37	0.08	0.41	0.72	0.64
22	21	2	13	0.37	0.25	0.5	0.27	0.02
23	10	2	13	0.37	0.25	0.41	0.45	0.2
24	4	2	13	0.34	0.41	0.12	0.54	0.13
25	16	2	12	0.28	0.25	0.33	0.45	0.2
26	9	2	10	0.25	0.16	0.33	0.45	0.29
27	23	2	9	0.25	0.08	0.12	0.27	0.19
28	3	2	9	0.25	0.08	0.41	0.27	0.19
29	18	2	6	0.17	0.08	0.16	0.36	0.28
30	1	2	5	0.14	0	0.08	0.18	0.18

Table 5 states that the first item on the test has a difficulty ranking of 30, ie it is the 20th most difficult item on the test, the third item a ranking of 28. The easiest items on the test are items 22, 5 and 26. Clearly the items need to be rearranged so the easiest items are towards the front of the test. If the items in the test are arranged in sections, according to how the unit content has been covered, then items can still be ranked according to difficulty within each section.

DI or Discrimination Index

This is calculated by subtracting the proportion of students correct in the lower group from the proportion correct in the upper group. It is assumed that persons in the top third on total scores should have a greater proportion with the item correct than the lower third.

This calculation of the index is an approximation of a correlation between the scores on an item and the total score. Therefore, the DI is a measure of how successfully an item discriminates between students of different abilities on the test as a whole. Any item which did not discriminate between the lower and upper group of students would have a DI=0. An item where the lower group performed better than the upper group would have a negative DI. In general, DI's above +0.30 indicate an item which is working well, but 0.20 is not bad.

The discrimination index is affected by the difficulty of an item, because by definition, if an item is very easy everyone tends to get it right and it does not discriminate. Likewise, if it is very difficult everyone tends to get it wrong. Such items can be important to have in a test because they help define the range of difficulty of concepts assessed. Items should not be discarded just because they do not discriminate.

Items 5,26,13 and 15 have negative DIs this will not help to discriminate the students.
Items 22,2,25,17and 7 have DIs below 0. 10. Item 22 is an easy item and therefore possibly not meant to discriminate between students but an easy item it should not be the last item in the test.

Discrimination index (DI):

- Excellent (≥0.40): 8/30 (26.7%)
- Good (0.30–0.39): 2/30 (6.7%)
- Acceptable (0.20–0.29): 5/30 (16.7%)
- Poor (<0.20): 15/30 (50.0%)

Mean DI was 0.243 (median 0.195). 10/30 (33.3%) items showed DI ≥0.30 (good–excellent), while 15/30 (50.0%) were <0.20 (poor). Overall, it shows that the items on the test discriminate as expected ie students who score well on a particular item tend to score well on all items in the test and students who score poorly on a specific item tend to score poorly across all the items.

Alternative (or Distractor) Analysis

This analysis provides the opportunity to study the responses students make to each alternative on an item. The efficiency of alternatives can be judged by inspecting the tables below. These tables show the number and proportion of students in the lower, middle and upper group who selected the correct answer as well as the number of students choosing each alternative.

Alternative analysis for item 9:

Answer key= B				
Observations				
alt	lower	middle	Upper	All
A	2	2	3	7
B	**2**	**4**	**5**	**11**
C	2	1	1	4
D	6	5	2	13
total	12	12	11	35
Proportion				
alt	lower	middle	upper	All
A	0.16	0.16	0.25	0.2
B	**0.18**	**0.33**	**0.41**	**0.31**
C	0.16	0.08	0.08	0.11
D	0.5	0.41	0.16	0.37
Mean score	0.55	0.98	0.9	0.99

This is the difficult item in the test only 31% of students have passed this item. Alternative C does not appear to serving any function as only 4 students have selected it. Alternative A and D are appear to be good distractor. But this item should be placed later in the test.

Alternative analysis for item no 5

Answer key= A				
Observations				
alt	lower	middle	upper	All
A	**11**	**11**	**10**	**32**
B	1	0	1	2
C	0	1	0	1
D	0	0	0	0

total	12	12	11	35
Proportion				
alt	lower	middle	upper	All
A	**0.91**	**0.91**	**0.83**	**0.91**
B	0.08	0	0.09	0.05
C	0	0.08	0	0.02
D	0	0	0	0
Mean score	0.99	0.99	0.92	0.69

This is the easiest item in the test as 91% of the students have passed this item. Alternatives C and D does not appear to be serving any purpose as only 1 and no student selected it respectively. Alternative C is clearly discriminating in the right direction but with most students making the right choice. This item could be the first on the test.

DISCUSSION

Overall calibration (difficulty):

Our test shows a balanced difficulty profile with 50% moderate and 30% easy items, and an average p-value around 0.54, which is pedagogically desirable for separating abilities without overwhelming the cohort. This compares favourably with Bhattacherjee et al. (2022), who reported that their items were "mostly easy" in an online internal assessment of 6th-semester students; a predominantly easy paper tends to compress score spread and blunt discrimination. Our distribution suggests better calibration than a mostly-easy paper.

Discrimination:

About one-third (33.3%) of our items attained good-excellent discrimination (DI ≥0.30), whereas half (50%) fell into the poor band. Bhattacherjee et al. (2022) also found "most of the items were of poor discrimination," so our profile echoes a common issue in undergraduate settings, especially when items have ambiguous stems/keys or implausible distractors. In contrast, Shahat et al. (2024) (Postgraduates of Paediatrics) reported stronger option performance and a high DE (~81.4%), which typically correlates with better discrimination because plausible distractors attract lower performers. The difference likely reflects cohort level and content focus (postgraduates vs. undergraduates), item-writing rigour and pre-review practices.

CONCLUSION

This item analysis of a 30-item internal assessment in Community Medicine demonstrated that the paper was broadly well-calibrated on difficulty, with half the items (50.0%) in the moderate band and an overall mean proportion correct of 0.542. However, discrimination performance was heterogeneous: one-third of the items achieved good to excellent discrimination (DI ≥0.30), while half (50.0%) showed poor discrimination (DI <0.20). In practical terms, the test already contains a substantial core of bankable items, those combining moderate difficulty with good/excellent DI, yet a meaningful minority require redrafting to improve stem clarity and the plausibility of distractors. These results mirror patterns reported in comparable undergraduate cohorts and reinforce the value of routine post-test analysis to incrementally lift assessment validity. Institutionalizing an edit–retest cycle, strengthening blueprinting and pre-administration peer review, and capturing complete option-wise response data will together improve both the fairness and interpretability of future assessments.

Strengths and Limitations

Strengths: A principal strength of this work is its authentic, real-world context: a complete cohort (n=35) was analyzed exactly as the assessment was delivered, ensuring that conclusions are directly actionable for the department's assessment practice. The study used metrics such as item difficulty, discrimination via performance quartiles and qualitative inspection of distractors, offering clear thresholds for retaining, revising or retiring items. All computations were performed on anonymized tests, with consistent scoring, which enhances internal validity and reproducibility. Another strength is the explicit benchmarking against published literature in similar settings, this situates the findings, highlights what is already working (balanced difficulty) and pinpoints where focused effort will yield maximum gains (items with DI <0.20).

REFERENCES

1. Bhattacherjee S, Mukherjee A, Bhandari K, Rout AJ. Evaluation of multiple-choice questions by item analysis, from an online internal assessment of 6th semester medical students in a rural medical college, West Bengal. Indian J Community Med. 2022;47(1):92–95.
2. Nojomi M, Mahmoudi M. Assessment of multiple-choice questions by item analysis for medical students' examinations. Res Dev Med Educ. 2022;11:24.
3. Shahat KA. Item Analysis of Multiple-Choice Question (MCQ)-Based Exam Efficiency Among Postgraduate Pediatric Medical Students: An Observational, Cross-Sectional Study From Saudi Arabia. Cureus. 2024;16(9):e69151.