Original Article :

Quality Evaluation of Multiple-Choice Questions as an Assessment Tool for Medical Students.

Moutasem S Aboonq

Department of Medical Physiology, Faculty of Medicine, Taibah University, Al-Madinah Al-Munawwarah, Saudi Arabia

Tel: +966-50-532-7177- E-mail: aboonq@yahoo.co.uk

Received on April 10, 2023, accepted for publication on Oct 23, 2023, dx.doi.org/10.5455/mjhs.2024.01.002

Corresponding Author:

Department of Medical Physiology, Faculty of Medicine, Taibah University, Al-Madinah Al-Munawwarah, Saudi Arabia

Tel: +966-50-532-7177 - E-mail: aboonq@yahoo.co.uk

OPEN ACCESS Distributed under Creative Commons CC BY-NC-ND 4.0

Abstract

Background and Aims:
Assessment is an educational process that measures student ability to accomplish specific objectives, helping to bridge the gap between teaching and learning. This study evaluated the quality of multiple-choice questions (MCQs) used to assess undergraduate medical students' knowledge in the Nutrition & Metabolism course at Taibah University.

Methods:
The study used item analysis to evaluate the MCQs, including item difficulty, item discrimination, reliability, standard error of measurement, and distractor analysis. The study enrolled 171 undergraduate medical students who took a 30-item MCQ exam.

Results:
The results showed that the exam was reliable and consistent in measuring students' knowledge and understanding, but some items had issues that needed revision or deletion from the question bank. The report provides useful information for improving the exam's quality and highlights the importance of ongoing assessment and revision of exam items. The findings suggest that ongoing assessment and revision of exam items are crucial to improving the quality of assessments and creating more effective measures of student learning.

Conclusions:
By leveraging these results, educators can identify areas where students may need further support and improve the design of their assessments.

Keywords:
MCQs, Validity, Reliability, Items analysis.

Background

Assessment is an educational process that measures student ability to accomplish specific objectives, helping to bridge the gap between teaching and learning.[1] A proper assessment provides students with feedback on their performance, identifying areas for improvement and modification of study strategies.[2] For teachers, the, assessment provides evidence for decisions on student progress and learning outcomes.[3] The formative assessment offers feedback on student learning and opportunities to document strengths and areas for improvement, while summative assessment primarily informs student progress decisions.[4] Various types of assessment methods, such as Multiple Choice Questions (MCQs) or Objective Structure Clinical Exam (OSCE), exist to meet different assessment needs.[5]

Guidelines for writing good MCQs have been outlined in the "Item Writing Manual" from the National Board of Medical Examiners.[6],[7] MCQs consist of three parts: the stem, the key, and the distracters. The key is the correct answer, while distracters are incorrect alternatives. The stem is the question or lead-in, which should be designed to be answered by the options provided. To avoid placement bias, correct answers should be distributed throughout the test. Good distractors should be selected by students whose performance is poor, and the key and distractors must be similar in length, style, and grammatical form.[8]

Validity is the degree to which evidence and theory support the interpretations of test scores in relation to the proposed uses of the assessment tool.[9],[10] There are several sources of evidence that support the construct validity of interpretations of test scores, including content-related, response process, internal structure, relations to other variables, and consequences.[11] Content evidence examines the extent to which a test's content represents the construct it is intended to measure, often documented using a table of specifications.[12] Response process is evidence of data integrity that controls or eliminates errors associated with test administration.[13],[14] Relations to other variables examine correlations between scores from different assessment tools [15], while evidence of consequences examines whether preferred results have been met and unintended effects avoided.[16] Internal structure relates to statistical and psychometric characteristics of the assessment tool, such as item discrimination and reliability.[17] Reliability refers to the consistency or repeatability of measurement scores and is an important but insufficient condition for validity.[18]

The aim of the study is to evaluate the

quality of the MCQs in assessing the performance of undergraduate medical students at Taibah University.

Methods

The study employed an observational cross-sectional design and enrolled 171 students who took the first quiz examination for the Nutrition & Metabolism (NM) course during the academic year 2022. The exam consisted of 30 multiple-choice questions (MCQs) with a total mark of 30 and a course total mark of 10. Item analysis was conducted to assess the quality of the MCQs. This analysis involved determining item difficulty, item discrimination, reliability, standard error of measurement, and distractors analysis. Item difficulty was determined by dividing the number of students who answered the question correctly by the total number of students who attempted the question. Item discrimination was calculated by comparing the scores of the top and bottom-performing students, with a point-biserial correlation coefficient calculated for each question. Reliability was measured using Cronbach's alpha coefficient. The standard error of measurement was calculated using a formula that takes into account the standard deviation of scores and the reliability coefficient. Distractor analysis was performed by determining the proportion of students who selected each incorrect answer choice for each question.

Results

The item analysis of the Nutrition & Metabolism (NM) multiple-choice question (MCQ) exam revealed important findings regarding the exam's validity, reliability, and individual item performance. As shown in table 1, the exam demonstrated acceptable reliability with a KR20 score of 0.79. This indicates that the exam scores were consistent and accurately reflected students' knowledge. The exam also demonstrated good discrimination overall, as all 30 questions had a positive discrimination index. This suggests the exam adequately differentiated between high- and low-performing students. However, 13 items did have discrimination indices below 0.3, indicating room for improvement in their ability to distinguish levels of student understanding. Regarding difficulty, seven items were found to have difficulty indices above 0.8, meaning they were too easy. Item difficulty ranged from 0.38 to 0.92 across the exam, demonstrating variation in question difficulty levels. As shown in Table 2, item analyses were conducted for each question. This provided insight into individual item properties like discrimination, difficulty, and the effectiveness of distractors. Some questions

performed better than others according to these metrics. Table 3 shows the item analysis of distractors for each question. This revealed that some distractors were more effective than others at challenging students and disguising the right answer. Distractors that were selected by a very high or low percentage of students may need improvement.

Table 1: Item analysis report.

Parameter	Results
Exam Median score	21.00
Exam Average score (Mean)	20.23
Standard Error of Measurement	2.69
Standard Deviation (SD)	5.87
Minimum Score	0.00
Maximum Score	30.00
Total Difficulty	68.20%
Reliability Index (KR20)	0.79
Items with Difficulty Index <30%	0
Items with Difficulty Index >80%	7
Items with –ve discrimination	0
Items with zero discrimination	0
Items with +ve discrimination Index <0.3	13
Items with +ve discrimination Index>=0.3	17

Table 2: Psychometric analysis of each question.

Question	Difficulty Index	Discrimination Index	No. of Distractor not used
1	0.54	0.24	0
2	0.8	0.33	0
3	0.81	0.36	0
4	0.76	0.48	0
5	0.84	0.18	0
6	0.78	0.27	0
7	0.5	0.36	0
8	0.59	0.36	0
9	0.83	0.27	1
10	0.6	0.57	0
11	0.67	0.42	0
12	0.57	0.57	0
13	0.92	0.12	0
14	0.74	0.36	0
15	0.67	0.27	0
16	0.7	0.48	0
17	0.89	0.18	0

Question	Difficulty Index	Discrimination Index	No. of Distractor not used
18	0.7	0.42	0
19	0.39	0.03	0
20	0.38	0.42	0
21	0.8	0.24	0
22	0.91	0.24	0
23	0.49	0.27	0
24	0.85	0.21	0
25	0.4	0.36	0
26	0.39	0.42	0
27	0.77	0.42	0
28	0.64	0.57	0
29	0.76	0.27	0
30	0.77	0.42	0

Table 3: Distractors analysis.

Question	A	% of student chose A Distractor	B	% of student chose B Distractor	C	% of student chose C Distractor	D	% of student chose D Distractor	Key Answer
1	23	13.50%	94	55.00%	12	7.00%	42	24.60%	B
2	6	3.50%	8	4.70%	20	11.70%	137	80.10%	D
3	9	5.30%	139	81.30%	10	5.80%	13	7.60%	B
4	130	76.00%	24	14.00%	13	7.60%	4	2.30%	A
5	1	0.60%	145	84.80%	6	3.50%	19	11.10%	B
6	7	4.10%	13	7.60%	17	9.90%	134	78.40%	D
7	87	50.90%	57	33.30%	10	5.80%	17	9.90%	A
8	7	4.10%	11	6.40%	101	59.10%	52	30.40%	C
9	143	83.60%	24	14.00%	4	2.30%	0	0.00%	A
10	37	21.60%	17	9.90%	14	8.20%	103	60.20%	D
11	116	67.80%	28	16.40%	15	8.80%	12	7.00%	A
12	19	11.10%	99	57.90%	30	17.50%	23	13.50%	B
13	4	2.30%	5	2.90%	159	93.00%	3	1.80%	C
14	18	10.60%	17	10.00%	8	4.70%	127	74.70%	D
15	115	67.30%	13	7.60%	41	24.00%	2	1.20%	A
16	19	11.10%	27	15.80%	120	70.20%	5	2.90%	C
17	9	5.30%	153	89.50%	4	2.30%	5	2.90%	B
18	121	70.80%	6	3.50%	15	8.80%	29	17.00%	A
19	18	10.60%	67	39.40%	65	38.20%	20	11.80%	B
20	9	5.30%	8	4.70%	89	52.00%	65	38.00%	D
21	10	5.80%	14	8.20%	138	80.70%	9	5.30%	C
22	5	2.90%	156	91.20%	7	4.10%	3	1.80%	B
23	84	49.10%	46	26.90%	29	17.00%	12	7.00%	A

Question	A	% of student chose A Distractor	B	% of student chose B Distractor	C	% of student chose C Distractor	D	% of student chose D Distractor	Key Answer
24	3	1.80%	14	8.30%	146	86.40%	6	3.60%	C
25	69	40.60%	30	17.60%	44	25.90%	27	15.90%	A
26	28	16.40%	26	15.20%	50	29.20%	67	39.20%	D
27	10	5.80%	7	4.10%	22	12.90%	132	77.20%	D
28	26	15.20%	111	64.90%	3	1.80%	31	18.10%	B
29	130	76.00%	23	13.50%	3	1.80%	15	8.80%	A
30	7	4.10%	7	4.10%	133	77.80%	24	14.00%	C
1	23	13.50%	94	55.00%	12	7.00%	42	24.60%	B
2	6	3.50%	8	4.70%	20	11.70%	137	80.10%	D
3	9	5.30%	139	81.30%	10	5.80%	13	7.60%	B
4	130	76.00%	24	14.00%	13	7.60%	4	2.30%	A
5	1	0.60%	145	84.80%	6	3.50%	19	11.10%	B
6	7	4.10%	13	7.60%	17	9.90%	134	78.40%	D
7	87	50.90%	57	33.30%	10	5.80%	17	9.90%	A
8	7	4.10%	11	6.40%	101	59.10%	52	30.40%	C
9	143	83.60%	24	14.00%	4	2.30%	0	0.00%	A
10	37	21.60%	17	9.90%	14	8.20%	103	60.20%	D
11	116	67.80%	28	16.40%	15	8.80%	12	7.00%	A
12	19	11.10%	99	57.90%	30	17.50%	23	13.50%	B
13	4	2.30%	5	2.90%	159	93.00%	3	1.80%	C
14	18	10.60%	17	10.00%	8	4.70%	127	74.70%	D
15	115	67.30%	13	7.60%	41	24.00%	2	1.20%	A
16	19	11.10%	27	15.80%	120	70.20%	5	2.90%	C
17	9	5.30%	153	89.50%	4	2.30%	5	2.90%	B
18	121	70.80%	6	3.50%	15	8.80%	29	17.00%	A
19	18	10.60%	67	39.40%	65	38.20%	20	11.80%	B
20	9	5.30%	8	4.70%	89	52.00%	65	38.00%	D
21	10	5.80%	14	8.20%	138	80.70%	9	5.30%	C
22	5	2.90%	156	91.20%	7	4.10%	3	1.80%	B
23	84	49.10%	46	26.90%	29	17.00%	12	7.00%	A
24	3	1.80%	14	8.30%	146	86.40%	6	3.60%	C
25	69	40.60%	30	17.60%	44	25.90%	27	15.90%	A
26	28	16.40%	26	15.20%	50	29.20%	67	39.20%	D
27	10	5.80%	7	4.10%	22	12.90%	132	77.20%	D
28	26	15.20%	111	64.90%	3	1.80%	31	18.10%	B
29	130	76.00%	23	13.50%	3	1.80%	15	8.80%	A
30	7	4.10%	7	4.10%	133	77.80%	24	14.00%	C

Discussion

Content validity, a critical concept in testing, determines the degree to which a test accurately measures the intended construct, especially in written assessments like MCQs. The NM course underwent rigorous content validation through a test blueprint, ensuring a direct link between test questions and learning objectives. By doing so, we guarantee the MCQs utilized in the test are pertinent and aptly representative of the construct under evaluation. In addition, the internal structure of the test provides valuable evidence to determine its validity. A meticulous analysis of the test items and computation of item discrimination and difficulty indices is essential to either corroborate or discredit the internal structure of the test. Furthermore, it is crucial to assess the test's reliability to establish the soundness of its internal structure.[18]

After each exam a special assessment committee prepares a comprehensive report about the item analysis, table 1. As emphasized in table 1, the item analysis report for the exam shows a median score of 21.00 and an average score (mean) of 20.23. The exam had a good reliability index (KR20 = 0.79) and a standard error of measurement of 2.69, indicating that the exam scores were consistent and accurately reflected students' knowledge. The exam also had accepted discrimination, as all 30 questions had a positive discrimination index, indicating that the exam discriminated well between high and low performers.

However, the analysis also reveals that some items had issues. Six items had multiple defects, and one item had a distractor that was chosen more than the key answer, indicating that it was not functioning well as a distractor. Moreover, [13] items had a positive discrimination index below 0.3, indicating that they were not good discriminators. Seven items had a difficulty index above 80%, indicating that they were too easy. The report suggests that these items should be revised or deleted from the bank.

For further evaluation, each item analysis was conducted as seen in table 2. The first column in the table represents the question number, followed by the difficulty index, the discrimination index, and the number of distractors not used. The difficulty index is a measure of how difficult the question is for students to answer correctly. The values range from 0 to 1, with 1 being the easiest question and 0 being the most difficult.19,20 In this set of questions, the difficulty index ranges from 0.38 to 0.92, indicating that the questions vary in their level of difficulty.

The discrimination index is a measure of how well a question differentiates be-

tween students who are performing well and those who are not.[21],[22] Discrimination indices range from -1 to 1, with values closer to 1 indicating a better discriminatory power.[21],[22] In this set of questions, the discrimination index ranges from 0.03 to 0.57, indicating that some questions are better at differentiating between high and low performing students than others.

The final column of the provided table highlights the number of unused distractors for each of the 30 MCQs. Distractors are designed to challenge students and make the correct answer less apparent, thereby testing their knowledge and understanding of a topic.[23],[24] Interestingly, all of the questions in the set had at least one distractor, except for question 9, which had one distractor that was not utilized.

Furthermore, the provided results, table 3, show the item analysis of the distractors for a set of the 30 MCQs. Each row portrays a single question, and the columns depict the percentage of students who opted for each answer choice (labeled A, B, C, D) as a distractor. For instance, for question 1, a measly 23 scholars (13.5%) picked answer choice A as a distractor, a whopping 94 students (55.0%) selected B, a puny 12 students (7.0%) selected C, and a moderate 42 students (24.6%) chose D. It is essential to note that distractors are supposed to answer choices that students are likely to pick instead of the correct answer. Thus, the aim of item analysis is to scrutinize how potent the distractors are in accomplishing this purpose. Perusing the findings, it is manifest that some distractors were more potent than others. For example, for question 4, a massive 76.0% of students opted for the distractor labeled A, indicating that it was an exceedingly effective distractor. Correspondingly, for question 9, a staggering 83.6% of students selected the distractor labeled A, denoting that it was also an extremely effective distractor. Conversely, some distractors were way less effective. For example, for question 13, only a negligible 2.3% of students selected the distractor labeled C, implying that it was not a very effective distractor. Likewise, for question 28, only a paltry 1.8% of students chose the distractor labeled C, signifying that it was also not a very effective distractor.

These findings suggest that some questions are more effective than others in measuring students' knowledge and understanding. The questions with higher difficulty and discrimination indices are generally considered better assessment items. Furthermore, the presence of unused distractors may indicate that the question needs revision to make it more challenging and less transparent.

In summary, the Nutrition & Metabo-

lism MCQ exam was reliable and discriminated well, but certain items need revision. The report provides useful information for improving the exam's quality and highlights the importance of ongoing assessment and revision of exam items.

By leveraging these results, educators can improve the design of their assessments and create more effective measures of student learning and identify any areas where students may need further support.

Conclusion

In conclusion, the item analysis report for the Nutrition & Metabolism MCQ exam showed that the exam was reliable and consistent in measuring students' knowledge and understanding. The exam had good discrimination and a low standard error of measurement, indicating that the scores accurately reflected students' performance. However, the analysis revealed that some items had issues, including multiple defects, ineffective distractors, and poor discrimination and difficulty indices. The report suggests that these items should be revised or deleted from the exam bank. Overall, the results highlight the importance of ongoing assessment and revision of exam items to improve the quality of assessments and create more effective measures of student learning. By leveraging these results, educators can identify areas where students may need further support and improve the design of their assessments.

References

1. Chowdhury RR, Kalu G. Learning to give feedback in medical education. Obstet Gynaecol. 2004 Oct;6(4):243–7.
2. Takada T, Isaji S, Yoshida M, Horiguchi A, Ando H, Miyakawa S, et al. Modified Socratic Method (planned and executed by Takada) for medical education: Grade II Acute Cholecystitis of Tokyo Guidelines 2018 as an example case. J Hepato-Biliary-Pancreatic Sci. 2022;29(5):505–20.
3. Arjoon JA, Xu X, Lewis JE. Understanding the state of the art for measurement in chemistry education research: Examining the psychometric evidence. J Chem Educ. 2013;90(5):536–45.
4. Nicol D, Macfarlane-Dick D. Rethinking formative assessment in HE: a theoretical model and seven principles of good feedback practice. C Juwah Macfarlane-Dick B Matthew Nicol Smith B2004 Enhancing Stud Learn Eff Form Feedback York High Educ Acad. 2004;
5. Vincent SC, Arulappan J, Amirtharaj A, Matua GA, Al Hashmi I. Objective structured clinical examination vs traditional clinical examination to evaluate

students' clinical competence: A systematic review of nursing faculty and students' perceptions and experiences. Nurse Educ Today. 2022;108:105170.

6. Bhattacharya S, Mondal S, Mitra K. Performance of Undergraduate Medical Students in Formative and Summative Evaluations in Community Medicine in a Medical College in India. J Med Educ. 2021;20(4).
7. Al-Wardy NM. Assessment methods in undergraduate medical education. Sultan Qaboos Univ Med J. 2010;10(2):203.
8. Adnan S, Sarfaraz S, Nisar MK, Jouhar R. Faculty perceptions on one-best MCQ development. Clin Teach. 2022;e13529.
9. Woo SE, LeBreton JM, Keith MG, Tay L. Bias, fairness, and validity in graduate-school admissions: A psychometric perspective. Perspect Psychol Sci. 2023;18(1):3–31.
10. Zhao K. Measuring the non-existent: validity before measurement. 2023;
11. Ramazanzadeh N, Ghahramanian A, Zamanzadeh V, Valizadeh L, Ghaffarifar S. Development and psychometric testing of a clinical reasoning rubric based on the nursing process. BMC Med Educ. 2023;23(1):1–13.
12. Memon MA, Joughin GR, Memon B. Oral assessment and postgraduate medical examinations: establishing conditions for validity, reliability and fairness. Adv Health Sci Educ. 2010;15:277–89.
13. Lane S, Raymond MR, Haladyna TM, Downing SM. Test development process. In: Handbook of test development. Routledge; 2015. p. 19–34.
14. Downing SM, Juul D, Park YS. Statistics of testing. In: Assessment in health professions education. Routledge; 2019. p. 70–85.
15. Cook DA, Zendejas B, Hamstra SJ, Hatala R, Brydges R. What counts as validity evidence? Examples and prevalence in a systematic review of simulation-based assessment. Adv Health Sci Educ. 2014;19(2):233–50.
16. AlFallay IS. Test specifications and blueprints: Reality and expectations. Int J Instr. 2017;11(1):195–210.
17. Cook DA, Beckman TJ. Current concepts in validity and reliability for psychometric instruments: theory and application. Am J Med. 2006;119(2):166-e7.
18. Stevens SP, Palocsay SW, Novoa LJ. Practical Guidance for Writing Multiple-Choice Test Questions in Introduc-

tory Analytics Courses. Inf Trans Educ. 2022;

19. Iqbal Z, Saleem K, Arshad HM. Measuring teachers' knowledge of student assessment: Development and validation of an MCQ test. Educ Stud. 2023;49(1):166–83.

20. Miller MD, Linn RL, Gronlund NE. Measurement and Assessment in Teaching. Merrill/Pearson; 2009. 551 p.

21. Baldwin BA. The role of difficulty and discrimination in constructing multiple-choice examinations: With guidelines for practical application. J Account Educ. 1984 Mar 1;2(1):19–28.

22. Scheim AI, Bauer GR. The Intersectional Discrimination Index: Development and validation of measures of self-reported enacted and anticipated discrimination for intercategorical analysis. Soc Sci Med. 2019;226:225–35.

23. 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;7(1):e10836.

24. Hussein A, Gasmalla HEE. Introduction to the Psychometric Analysis. In: Written Assessment in Medical Education. Springer; 2023. p. 111–35.