Item Analysis Of Multiple-Choice Questions Of Pre-University Examination Conducted At Nobel Medical College

Joshi B R[1], Rizal S[1]
1. Department of Biochemistry, Nobel Medical College Teaching Hospital, Biratnagar, Nepal

ABSTRACT

Introduction: Multiple-choice questions (MCQs) are useful in assessing student performance, covering a wide range of topics in an objective way. Its reliability and validity depend upon how well it is constructed. Item analysis evaluates the quality of test items using student responses to ensure they measure intended learning outcomes. The objective of the study was to assess the item and test quality of multiple-choice questions and to identify areas for improvement in test construction.

Method: This is a cross-sectional observational study done to analyze 120 MCQs attempted by 100 MBBS students in their 2nd year pre-university exam (30 each from four papers). All the MCQs from 100 students were collected and evaluated using the correct keys. After the evaluation, the marks obtained by the students were entered in MS-excel and data analysis was done in SPSS. Each test paper was assessed for reliability, and individual items were evaluated using the difficulty index (DIF I) and discrimination index (DI).

Results: The test papers demonstrated acceptable reliability, with KR-20 values ranging from 0.78 to 0.89. The mean difficulty index was 53.58 ± 17.84, with 65.84% of items falling within the acceptable range (DIF I 30–70). The mean discrimination index was 0.37 ± 0.18, with 46.66% of items showing excellent discrimination (DI > 0.4). Among the individual subjects, Biochemistry had the highest difficulty index (59.75 ± 21.27), while Physiology and Pathology had the highest discrimination indices (both 0.44). Only 5% of items were classified as defective.

Conclusion: The analysis showed that most of the MCQs had acceptable difficulty and discrimination, indicating effective assessment of student performance. The study highlighted the need to review questions with low discrimination and extreme difficulty levels to improve overall assessment quality.

Keywords: Difficulty index; Discriminating index; Multiple choice questions

INTRODUCTION

Assessment is an ongoing process aimed at understanding and improving student's learning. The goal of assessment in medical sciences is to support learning and to establish the competence of established doctors. It is important because it helps the person being assessed to identify and respond to his/her own learning needs. Besides giving feed-back on performance and suggestions for improvement, grading and certification, it should also be able to differentiate between good and poor candidate and should have high fidelity.[1,2]

Multiple choice questions (MCQs) have become a standard practice in both undergraduate and postgraduate medical assessments due to their practicality, reliability and ease of standardization. The quality of MCQs is important because of its effect on the student's overall competency level during their assessment. When well-constructed, MCQs can assess more than the ability to recall facts – they can also assess higher order cognitive skills including, understanding, application of knowledge, analytical thinking etc. Framing

Corresponding author:
Dr. Bishal Raj Joshi
Department of Biochemistry
Nobel Medical College Teaching Hospital
Biratnagar, Morang, Nepal
Email: drbishaljoshi@gmail.com
ORCID: https://orcid.org/0000-0002-6981-2749
Phone: +977-9852027758

MCQs is a challenging task. A meticulously built MCQ bank after thorough item analysis is a handy tool for any academic institute for conducting assessments.[3] Critics argue that may be limited in evaluating problem-solving and clinical reasoning[4]; however, many research supports that well-developed MCQs can effectively measure learning outcomes across all cognitive levels.[5]

Item analysis is a process of determining the quality of an assessment/test or tool by looking at each individual item or question and determining whether it is performing well. It examines the student responses to individual test items (MCQ) to assess the quality of those items and the test as a whole.[3,6] Key indicators used in item analysis include the difficulty index and discrimination index, which assess the appropriateness and effectiveness of each MCQ. Peer-reviewed, well-analyzed MCQs are associated with its improved reliability and validity of assessments.

Even though MCQs are commonly used in health professional education, there is limited evidence that item analysis is routinely employed to evaluate their effectiveness in Nepal. Few studies have highlighted its importance in medical education in Nepal.[7,8] This study aims to evaluate the quality of multiple-choice questions (MCQs) by analyzing them through established item analysis parameters, specifically the difficulty index and discrimination index and to categorize the items according to difficulty index and discrimination index. The present study will thus be significant because it can serve as a blueprint for an educational tool that can improve educational outcomes within an institution. In addition, this study can also provide future directions for our institution's improvement through sharing its findings.

METHODS

This is a cross-sectional observational study conducted at Nobel Medical College Teaching Hospital among 100 students who appeared for the pre-university examination of MBBS 2nd year, conducted at September 2023, in which the quality of MCQs was assessed using difficulty index and discrimination index for item analysis. Ethical clearance for the study was obtained from IRC, Nobel Medical College Teaching Hospital ref no. 718/2022. A total of four papers each comprising of 30 MCQs as one of the assessment method was taken for analysis. The four papers that students attended in second year pre-university examination were: Paper V: Gastrointestinal and Hepatobiliary system; Paper VI: Renal and Endocrine system; Paper VII: Reproductive system and Paper VIII: Central Nervous System and Special senses.

Each paper was comprised of six subjects namely, Anatomy, Biochemistry, Microbiology, Pathology, Pharmacology and Physiology. The marks distribution of each subject in each paper is according to format determined by Kathmandu University. The pre-university exam assessment test comprised of 30 "single response type" MCQs of 30 marks in each paper. All MCQs had single stem with four responses including, one being correct answer and other three incorrect alternatives (distractor). Each correct response was awarded as 1 mark and incorrect response as 0. Unattempted MCQs was marked as 0. There was no negative marking used for the incorrect answers.

All MBBS 2nd year students (n=100) who appeared in all four pre-university MCQ-based examinations conducted in September 2023 were included in the study. Students who were absent in one or more exams were excluded from our study. Sampling method used was convenience sampling, encompassing the entire cohort of eligible 2nd year MBBS students enrolled at the time of the study. All the MCQs from 100 students were collected and were evaluated using the correct keys. After the evaluation, the marks obtained by the students were entered in MS Excel. Data analysis involved organization of item-wise responses using MS-Excel. Reliability analysis for each paper was performed using the Kuder-Richardson Formula 20 (KR-20) to assess internal consistency. Item analysis was conducted using two classical test theory metrics: Difficulty Index (DIF I) and Discrimination Index (DI).[9] Descriptive statistics including mean, standard deviation, median, and interquartile range (IQR) were computed using SPSS version 16.

Reliability of the MCQs paper was assessed using Kuder- Richardson formula (KR-20). KR-

20 is a measure of internal consistency of items and score consistency, that is, how close related are a set of items as a group. It is a measure of reliability.[10] Formula used was: $K/K\text{-}1(1\text{-}p_iq_i/2)$, where, k: Total number of question, pi: Proportion of individuals who answered question correctly, qi: Proportion of individuals answered question incorrectly and σ2: Variance of scores for all individuals who took the test. Interpretation was done as: 0.9 ≤ KR-20- Excellent, 0.8 ≤ KR-20 < 0.9-Good , 0.7 ≤ KR-20 < 0.8-Acceptable/ Average, 0.6 ≤ KR-20 < 0.7-Questionable, 0.5 ≤ KR-20 < 0.6- Poor, KR-20 < 0.5- Unacceptable.

Difficulty Index (DIF I) measures the proportion of students who answered the item correctly, it determines how difficult or easy was the item. It is calculated as, $DIF\ I = N_c/N_t \times 100\%$ Where, Nc: correct response to a particular item and Nt: Total number of students. Interpretation was done as: Items with DIF I between 30-70% are acceptable, >70% are very easy and < 30% are classified as difficult.[9]

Discrimination Index (DI) is the ability of an item to differentiate between students of higher and lower abilities. It helps to know how well an item discriminate between high performers and low performers. It is calculated by using the Kelley's formula as, DI = 2×(H- L/N) Where, H = number of correct responses from high achieving group (33%), L = number of correct responses from low achieving group (33%) and N = total number of students in H and L group. Interpretation was done as: Items with DI between > 0.4 have high discriminating power, 0.30-0.40 have good discriminating power, 0.20-0.30 have low discriminating power and <0.20 have very low discriminating power.[9,11]

RESULTS

In this study, we have analyzed 120 MCQs attempted by 100 students of MBBS second year in their pre-university exam. All papers demonstrated acceptable to good reliability, with KR-20 values ranging from 0.78 to 0.89. Notably, Paper VI had the highest reliability at 0.89, indicating strong internal consistency among its items.

Table 1: Reliability of the MCQs papers

MCQs	KR-20	Interpretation
Paper V	0.80	Good
Paper VI	0.89	Good
Paper VII	0.78	Acceptable
Paper VIII	0.78	Acceptable

Table 2: Categorization of items according to Difficulty index

	Defective item N (%)	Dif<30 N (%)	Dif 30-70 N (%)	Dif> 70 N (%)	Total
Paper V	3 (10%)	2 (6.66%)	17 (56.66%)	8 (26.66%)	30
Paper VI	1 (3.33%)	1 (3.33%)	22 (73.33%)	6 (20%)	30
Paper VII	1 (3.33%)	5 (16.66%)	19 (63.33%)	5 (16.66%)	30
Paper VIII	1 (3.33%)	3 (10%)	21 (70%)	5 (16.66%)	30
Total	6 (5%)	11 (9.16%)	79 (65.84%)	24 (20%)	120 (100%)

Across all papers, 65.84% of the items fell within the acceptable difficulty range (DIF 30–70). Only 20% were relatively easy (DIF >70), showing a favorable distribution for assessing student performance. Paper VI had most MCQs in acceptable difficulty index range (73.33%).

Table 3: Descriptive statistics for difficulty index

	No. of items	Mean (SD)	Median (IQR)
Paper V	27	55.41 (18.16)	55.56 (44.44 – 72.22)
Paper VI	29	55.62 (15.71)	57.41 (45.37 – 67.59)
Paper VII	29	49.30 (18.83)	51.35 (36.11 – 51.85)
Paper VIII	29	54.02 (18.67)	53.70 (37.04 – 64.81)

Table 3 summarizes the descriptive statistics for the difficulty index across each paper. The mean difficulty index across the four papers ranged from 49.30 to 55.62, with Paper VI having the highest mean difficulty score. Collectively, the overall mean difficulty index for all 120 items was 53.58 ± 17.84.

Table 4: Categorization of items according to Discrimination index

	Negative	Poor (0-0.19)	Acceptable (0.2-0.29)	Good (0.3-0.39)	Excellent (>0.4)
V	3 (10%)	4 (13.33%)	4 (13.33%)	4 (13.33%)	15 (50%)
VI	1 (3.33%)	3 (10%)	4 (13.33%)	2 (6.66%)	20 (66.66%)
VII	1 (3.33%)	8 (26.66%)	6 (20%)	5 (16.66%)	10 (33.33%)
VIII	1 (3.33%)	6 (20%)	9 (30%)	3 (10%)	11 (36.66%)
Total	6 (5%)	21 (17.5%)	23 (19.16%)	14 (11.66%)	56 (46.66%)

Table 4 shows the distribution of MCQs by discrimination index, highlighting item quality in distinguishing between high and low performers. Among the 120 items, 46.66% (n=56) demonstrated excellent discrimination (DI > 0.4) where as 17.5 % (n=21) had poor discriminating power. Paper V and VI showed the highest proportions of items with excellent discrimination.

Table 5: Descriptive statistics for discrimination index

	No. of items	Mean (SD)	Median (IQR)
V	27	0.39 (0.18)	0.41 (0.26-0.56)
VI	29	0.46 (0.18)	0.48 (0.31-0.59)
VII	29	0.32 (0.18)	0.33 (0.19-0.44)
VIII	29	0.34 (0.17)	0.30 (0.22-0.46)

The mean discrimination index ranged from 0.32 to 0.46 with overall mean discrimination index of 0.37 ± 0.18.

Table 6: Difficulty and Discrimination index for individual subjects

Subject	Difficulty Index		Discrimination Index	
	Mean (SD)	Median (IQR)	Mean (SD)	Median (IQR)
Anatomy	57.95 (20.55)	60.18 (37.96-74.07)	0.34 (0.18)	0.35 (0.26-0.44)
Physiology	48.89 (13.94)	48.18 (35.64 – 61.97)	0.44 (0.17)	0.43 (0.33-0.56)
Biochemistry	59.75 (21.27)	68.52 (44.44 – 72.22)	0.43 (0.17)	0.52 (0.30-0.56)
Microbiology	49.00 (24.43)	50.00 (32.41 – 74.07)	0.26 (0.17)	0.26 (0.11-0.37)
Pharmacology	47.92 (17.42)	49.07 (38.43 – 60.19)	0.31 (0.22)	0.28 (0.12-0.53)
Pathology	55.27 (9.25)	53.70 (47.69-61.57)	0.44 (0.16)	0.44 (0.30-0.56)

Table 6 compares the difficulty and discrimination indices across individual subjects.

Biochemistry had the highest mean difficulty (59.75 ± 21.27), while Physiology and Pathology had the highest mean discrimination indices (both 0.44). Microbiology items were the most challenging and had the lowest mean discrimination.

DISCUSSION

Evaluation is a crucial element of an education process, and its results reflect both weaknesses and strengths of educational outcomes. Evaluation plays a key role in helping educators recognize what's working well and where improvements are needed, ultimately guiding positive changes in the education system. MCQs are commonly employed across various academic disciplines due to their objectivity and ability to assess a wide range of content in a limited time.[12] MCQs serve as a reliable and objective method for measuring students' learning outcomes.[13] However, their design is both complicated and time-consuming and requires multidisciplinary teams to ensure their high quality, mainly because of rigid standards. Item analysis analyzes student's responses to each item, used to assess the quality of those items and evaluate their overall performance to benefit

both students and teachers.[12] An ideal multiple-choice question (MCQ) is one that has a moderate difficulty level (difficulty index between 30% and 70%), a high ability to distinguish between high and low performers (discrimination index of 0.30 or above), and optimal effectiveness (100% distractor efficiency) with all three distractors functioning properly.[14]

The reliability of the MCQs papers were assessed using Kuder-Richardson formula (KR-20) and all the papers were found to be reliable i.e. it can be used for assessment purpose.

DIF I is used to differentiate between the easy item, acceptable items and the difficult items. In our present study, out of 120 items analyzed for difficulty index (DIF I), 65.84% (n=79) of the items were in the acceptable range which can be preserved for future use where as 20% (n=24) of the item were easy and 9.16% (n=11) of the items were difficult, which has to be revised for future use or discarded. This finding of our study was consistent with the study done by Ingale et.al where 80% of the items were in the acceptable range in terms of DIF I.[13] The mean difficulty index in our study for paper V was 55.41+18.16, for paper VI was 55.62±15.71, for paper VII was 49.30±18.83 and for paper VIII was 54.02± 18.67, which was comparable to the study done by Ingale in 100 students and 30 items where the mean DIF I was 55.10±17.28.[13] The mean DIF I was bit lower in study done by Karelia in 2013 where it was 47.17 ± 19.27.[10] Also, Karelia, showed a range of mean ± SD between 47.17±19.77 to 58.08 ± 19.33 in a study conducted over a period of five years.[15] Patel et al. showed 80% of items in the acceptable range (DIF I 30-70%) and 20% in the unacceptable range (DIF I >70% and <30%).[16] Thus, the findings of our study was in congruence with the previous studies done by many authors. Too difficult items (DIF I ≤ 30%) can lead to deflated scores, while the easy items (DIF I > 70%) may result into the inflated scores and a decline in motivation. Items with high DIF I (>70%) should be placed either at the start of the test as "warm-up" questions to boost the confidence of students or discarded, similarly items with low DIF I (<30%) should be either revised or removed altogether.[17]

Discriminating index (DI) is another important index which differentiates high ability and low ability student. It is obvious that a question which is either too difficult (attempted wrongly by everyone) or too easy (response correctly by everyone) will have nil to poor DI. In our present study the mean discriminating index was 0.39± 0.18 for paper V, 0.46 ± 0.18 for paper VI, 0.32 ± 0.18 for paper VII and 0.34±0.17 for paper VIII respectively. Out of 120 item analyzed for DI, 46.66% (n=56) has high discrimination power where as 11.66% (n= 14) of items had good discriminating power which can be retained for future use. Similarly, 19.16% (n=23) had acceptable discriminating power with 17.5% (n=21) has poor discriminating power which has to be revised for future use or discarded. The findings of our study was comparable with the study done by many authors. Namdeo et al. reported a mean DI of 0.33±0.23 which signifies good discriminative ability of test items.[18] However, Gajjar S, reported the items in his study had a very low DI with mean DI of 0.14 ± 0.19.4.[19] Another study done by Ingale et.al reported quite high mean DI of 0.40 ± 0.33.[13] In an item analysis study by Patil et al., out of total 100 items, 24 had poor DI, 45 had good DI , and 31 had excellent DI which was almost comparable to our study.[20] Lin et al. reported that 28.8% of MCQ items in the basic medical sciences section had a DI of <0.2. Items with poor DIs usually result in low scores due to the use of incorrect answer keys, confusing stems or areas of controversy. Such items should be removed from the question bank as they fail to discriminate between strong and weak academic performances.[21] In our study there were three defective items in paper V, one items each in Paper VI, VII and VIII, which led to negative DI. The wrong keys in those questions were corrected and stored in question bank. A negative discrimination index (DI) suggests that something may be wrong with the question—especially if lower-performing students are the ones answering it correctly, possibly by chance. This can happen for a few reasons, like an incorrect answer key, unclear or confusing wording, or if many students were generally unprepared for the topic. Items with negative DI decrease the validity of the test and should be removed from the collection.[13]

The result for subject-wise item analysis showed that the items from physiology and biochemistry

had excellent discriminating index.

Thus, our experience as educators in Nepal has shown that high stakes testing is a priority at the end of a learning period ('assessment of learning'). Health professional educators must embrace a radical shift in assessment culture in order to incorporate assessment for learning and developing quality assessment tools that will enhance student's learning experiences. If educators are unable to develop valid or reliable assessments because they lack the necessary skills, it is the academic institution's responsibility to train and instruct them. Training can substantially improve the quality of MCQs developed by teaching faculty. Due to the organizational culture and core beliefs, radical changes are met with inherent resistance. It is therefore incumbent on institutions' leaders, or the power structures within their organizations, to spearhead a paradigm shift in assessment culture for learning.[8]

CONCLUSION

Item analysis is thus a useful method to assess the quality of MCQs test papers. Its quality and accuracy can be improved by item analysis helps us to create a valid pool of items and store it in a question bank for future use. Also, it helps us to identify low achievers groups in order to deal with their learning difficulties by counseling or modifying learning methods. The identification of problematic items through the analysis underscores the necessity of ongoing review and revision to maintain the validity and reliability of the assessments. Thus, moving forward, integrating the insights gained from this analysis will be essential in designing more effective and equitable assessment for educational or evaluative process.

Recommendations

Based on above findings, we suggest conducting more and more studies in order to develop a valid question bank and to also identify the students whose score are less. The students should be counselled personally by the staff to identify their difficulties and hence these problems must be dealt either by the modification of the teaching skills or by solving the difficulties the students are facing.

LIMITATIONS

Only one set of the test items were analyzed thus cannot be generalized for all the assessment of our institution.This study can be considered a pilot study, providing preliminary data and insight into the quality of MCQs, and laying the groundwork for larger-scale, longitudinal analyses across multiple exam sessions in the future.

Acknowledgement:
The authors express their sincere thanks to the Examination Section of Nobel Medical College Teaching Hospital for providing the MCQ test papers

Conflict of interest: None
Source of funding: None

REFERENCES

1. Olayemi E. Multiple choice questiones as a tool for assessment in medical education. Ann Biomed Sci. 2013; 12(1): 15-23
2. Epstein RM. Assessment in medical education. New England journal of medicine. 2007; 356(4):387-96.
3. 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; 77 (Suppl 1):S85–9.
4. van Wijk EV, Janse RJ, Ruijter BN, Rohling JH, van der Kraan J, Crobach S, et al. Use of very short answer questions compared to multiple choice questions in undergraduate medical students: an external validation study. PLoS One. 2023;18(7):e0288558.
5. Javaeed A. Assessment of higher ordered thinking in medical education: multiple choice questions and modified essay questions. MedEdPublish. 2018;7(2):128.
6. Rao C, Kishan Prasad HL, Sajitha K, Permi H, Shetty J. Item analysis of multiple choice questions: Assessing an assessment tool in medical students. Int J Educ Psychol Res. 2016; 2(4):201-4.
7. Lama CP, Kharbuja R, Karki D, Dhungel S. Study on item analysis of multiple–choice

questions amongst the undergraduate dental students. Nepal Medical College Journal. 2023 ;25(4):301-5.
8. Bhat N, Deo SK, Gurung S. Assessing the quality of multiple-choice questions in allied health science summative exams: A retrospective analysis. Journal of Gandaki Medical College-Nepal. 2023;16(2):111-7.
9. Kuder GF, Richardson MW. The Theory of the Estimation of Test Reliability. Psychometrika. 1937; 2(3):151–60.
10. Kelley TL. The selection of upper and lower groups for the validation of test items. J Educ Psychol. 1939; 30(1):17–24.
11. Nojomi M, Mahmoudi M. Assessment of multiple-choice questions by item analysis for medical students' examinations. Res Dev Med Educ. 2022; 11:24.
12. Ingale AS, A. Giri P, Doibale MK. Study on item and test analysis of multiple choice questions amongst undergraduate medical students. Int J Community Med Public Health. 2017; 4(5):1562.
13. Sahoo DP, Singh R. Item and distracter analysis of multiple choice questions (MCQs) from a preliminary examination of undergraduate medical students. Int J Res Med Sci. 2017; 5(12):5351.
14. Karelia BN, Pillai A, Vegada BN. The levels of difficulty and discrimination indices and relationship between them in four-response type multiple choice questions of pharmacology summative tests of year II MBBS students. IeJSME. 2013;7(2):41–6.
15. Patel Dr KA, Mahajan DrNR. Itemized Analysis of Questions of Multiple Choice Question (MCQ) Exam. Int J Sci Res. 2012; 2(2):279–80.
16. Kaur M, Singla S, Mahajan R. Item analysis of in use multiple choice questions in pharmacology. Int J Appl Basic Med Res. 2016; 6(3):170.
17. Namdeo SK, Sahoo B. Item analysis of multiple choice questions from an assessment of medical students in Bhubaneswar, India. Int J Res Med Sci 2016; 4(5): 1716-19.
18. 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 J Community Med. 2014;39(1):17–20.
19. Patil R, Palve S, Vell K, Boratne A. Evaluation of multiple choice questions by item analysis in a medical college at Pondicherry, India. Int J Community Med Public Health. 2016; 3(6):1612–6.
20. Kheyami D, Jaradat A, Al-Shibani T, Ali FA. Item Analysis of Multiple Choice Questions at the Department of Paediatrics, Arabian Gulf University, Manama, Bahrain. Sultan Qaboos Univ Med J SQUMJ. 2018;18(1):68.