Assessing the quality of multiple-choice questions in allied health science summative exams: A retrospective analysis

Neeti Bhat[1], Satish Kumar Deo[2], Sanyukta Gurung[1*]

[1]Department of Clinical Physiology and Yogic Sciences, [2]Department of Pharmacy and Clinical Pharmacology, Madan Bhandari Academy of Health Sciences, Hetauda, Bagmati, Nepal

ABSTRACT

Introduction: Multiple-choice questions are feasible, reproducible and cost-effective; hence, they are widely embraced in health professions education. The quality of multiple-choice questions is monitored statistically by item analysis. Item analysis confirms whether the questions measure the intended learning outcomes, ensuring fair and equitable assessments. To promote quality assessment, we analyzed the quality standards of multiple-choice questions in a summative examination using item analysis. **Methods:** The multiple choice questions answered by 38 students in the first semester of the allied health science programme of Madan Bhandari Academy of Health Sciences in Bagmati Province were collected, examined, and analyzed. The data gathered were subjected to the performance of each multiple choice question by analysis of difficulty index, discrimination index, and distractor effectiveness. **Results:** Items of most courses were acceptable (30% to 70%) as per the difficulty index with only three courses having (>70%) easy items. Poor discriminatory index (<0.20=poor) was noted in seven courses, particularly discipline-specific courses. A significant proportion of excellent distractors were identified in two courses: 50% each in Medical Terminology & Musculoskeletal and Fundamentals of Pharmacy. Most of the courses had more than 50% of functional distractors. **Conclusions:** The multiple-choice questions in most of the courses had poor discrimination index. These outcomes highlight the need for careful formulation of multiple-choice questions for authentic assessment.

Keywords: Difficulty index, discrimination index, distractor effectiveness, item analysis, mcqs.

***Correspondence:**
Dr. Sanyukta Gurung
Department of Clinical Physiology and Yogic Sciences, Madan Bhandari Academy of Health Sciences, Hetauda, Nepal
Email: sanyukta.gurung@mbahs.edu.np
ORCiD: 0000-0003-2158-735X

Submitted: June 19, 2023
Accepted: December 20, 2023

To cite: Bhat N, Deo SK, Gurung S. Assessing the quality of multiple-choice questions in allied health science summative exams: A retrospective analysis. JGMC Nepal. 2023;16(2):111-7.
DOI: 10.3126/jgmcn.v16i2.55893

INTRODUCTION

An assessment is "any purported and formal action to obtain information about the competence and performance of a candidate".[1] Assessment can be classified into formative which allows feedback during the course or summative which provides a numerical score to the students at the end of the course.[2] Quality assessment tools in health professionals' education are aimed at capturing the ability of assessments to stimulate learning, generate evidence of learner's progress, quantify the efficiency of the learning experience, inform instructors and administrators that programs adhere to their missions, and patient safety. The versatility of multiple-choice questions (MCQs), which can comprehensively evaluate curricula, makes them a popular choice, for assessment, yet, it takes time, effort, and ability to produce a high-quality item. In addition to meticulously drafted stems, good MCQs must suggest misleading but plausible distractor ideas.[3]

A multiple-choice question's quality and accuracy can be improved by the item analysis method which statistically analyzes the quality

standards of MCQ items. Although academic acumen is important in the setting and reviewing of exams, quantitative evidence about assessments would also be crucial, especially when exam results are valued so highly and standardization tools' characteristics can affect their credibility.[4] A test's reliability and validity can be evaluated by item analysis, it may help provide a better pool of questions that can benefit both teachers and students.[5]

Providing appropriate training to faculty members in Nepal presents a logistical challenge due to a lack of training programs. In these settings, item analysis is a highly advantageous practice for faculty members as it provides valuable insights and feedback for improving the quality of Multiple-Choice Questions (MCQs). Additionally, even though MCQs are commonly used in health professionals' education, there is a lack of analysis and feedback on the items constructed highlighted by a lack of studies evaluating the quality of MCQs constructed in Nepal. The assessment process is therefore undervalued, which raises concerns about its effectiveness in the educational system. The present study is 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. Hence, this study was conducted in an academy to assess the quality of MCQ tests administered in its first summative examinations for the first batch of students.

METHODS

A quantitative, observational, and cross-sectional study was conducted at Madan Bhandari Academy of Health Sciences (MBAHS) from June 2023 to July 2023. The data was collected retrospectively and was analyzed during this period. Thirty-eight students enrolled in the first batch of Bachelor of Public Health (n=13), Bachelor of Pharmacy (n=23), and Bachelor of Science in Medical Laboratory Technology (n=2) of MBAHS and appeared for the first summative examination conducted by the Office of Examination Controller on July 2022.

Breakdown of Examination Pattern: The students appeared in 14 exams in total, with four exams common to students from all streams and two exams common to Bachelor of Science in Medical Laboratory Technology (BScMLT) and Bachelor of Pharmacy (B.Pharmacy) students. Bachelor of Public Health (BPH) and BScMLT students appeared for eight exams, while B. Pharmacy students appeared for seven. Among the courses examined and the programs they are part of are:

Table 1: Course Code and Respective programs

Code	Name	BPH	BScMLT	B. Pharmacy
CB 111+113	Medical Terminology & Musculoskeletal System			
CB 112+114	General Concepts & Hematopoietic System			
CB 115+116	Respiratory and Cardiovascular System			
CR 111	Research and Biostatistics			
BC111	Cell Biology			
PF 111	Fundamentals of Pharmacy			
PC 111	Pharmaceutical Chemistry I			
HO 111	Occupational Health and Toxicology			
HN 111	Food and Nutrition – I			
HP 111	Primary Health Care – I			
HF 111	Fundamentals of Public Health			
LF 111	Fundamentals of Lab Medicine			
LH 111	General Histology			
LM 111	General Microbiology			

Development of MCQs: MCQs were constructed by faculty members of MBAHS by guidelines from the Office of Examination Controller, which were approved by the Examination Committee. All MCQs had a single stem, one correct answer (key), and three incorrect alternatives (distractors). A moderation team of at least two experts, who worked in a confidential setting, meticulously moderated the questions. The evaluation team focused on the accuracy and validity of the MCQs and aligned them with guidelines from the Office of the Examination Controller. The moderation team moderated three sets from the questions submitted by faculty members in alignment with the guidelines. An anonymous selection process was used to select the MCQs for use in the examination to maintain impartiality and fairness. An examination committee member made a selection from three sets of questions drafted by the moderation team.

Execution of Examination: A total of 30 MCQs were presented in Group A (objective) of the question paper of each course, and each student had 30 minutes to answer questions from Group A. With the questions paper, students received an Optical Mark Recognition (OMR) sheet. The guidelines for filling out the OMR sheets were written on the OMR sheets as well as dictated in the examination hall. A completed OMR answer sheet was submitted by the examination center to the Office of Examination Controller.

Evaluation of the Assessment: The submitted answer sheets were checked by the Office of the Examination Controller. Later, the scrutiny committee reviewed 10% of the OMR sheets of every subject for the accuracy of the marking and grading. They check if the marks are displayed correctly and then hand over the marks to the data entry section for entering the marks. The scrutiny committee is therefore responsible for monitoring the examination process and ensuring its integrity and quality.

Data Entry and Validation: The results were entered into the master sheet. Each MCQ item of 11 examinations (except fundamentals of laboratory medicine, General Histology, and General Microbiology) was included in the study and analyzed for difficulty index, discrimination index, and distractor effectiveness. We did not analyze MCQ items in Fundamentals of Laboratory Medicine, General Histology, and General Microbiology as only two students were enrolled in the BScMLT programme. The data was entered in Microsoft Excel and was checked for duplication. The duplicated data were removed and data validation to cells was applied in Excel where the settings were conditioned as per the variables. Among the 30 MCQs distractors were entered where the setting was limited to string variables a, b, c, and d, and appropriate formulas were applied according to the required variables.

Data Analysis: The difficulty index, discrimination index, and distractor effectiveness were calculated using the formulas below:

Difficulty Index (DIFi): DIFi was calculated using the formula: _Where, H=the top third of the scores obtained by high achieving students, L = the bottom third of the scores obtained by low achieving students, N = the sum of high achieving and low achieving students. The grading was as follows: <30% = too difficult, and 30% to 70% = average, >70% was considered too easy.[6]

Discrimination index (DISi): DISi is an index to calculate the difference between high-achieving students and low-achieving students. It was calculated by the following formula: .[6]

Distractor effectiveness (DE): Distractors which were selected by less than 5% of students were defined as nonfunctional distractors (NFD). NFD was calculated for every item and DE was calculated based on NFDs which had a range from zero to 100% and graded as follows: No NFD, (100%, excellent), 1 NFD, (66.6%, good), 2 NFDs,(33.3% moderate), 3 NFD, (0% poor).[6]

Ethical Approval was obtained from the Institutional Review Committee of the academy IRC-001-079. The difficulty index, discrimination index and distractor effectiveness calculated may help the facilitators in improving the quality of their MCQs which may benefit both the faculty members and the students.

RESULTS

The highest mean score was observed in Food and Nutrition-1 and the lowest was observed in Fundamental of Public Health. An overall mean score was higher in discipline-specific courses (Figure 1).

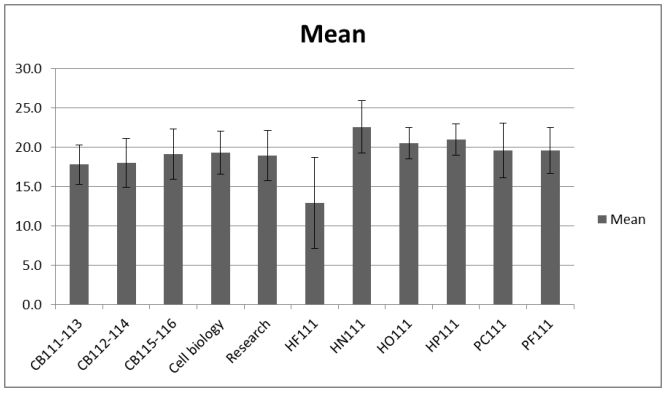

Figure 1: General characteristics of the multiple-choice questions

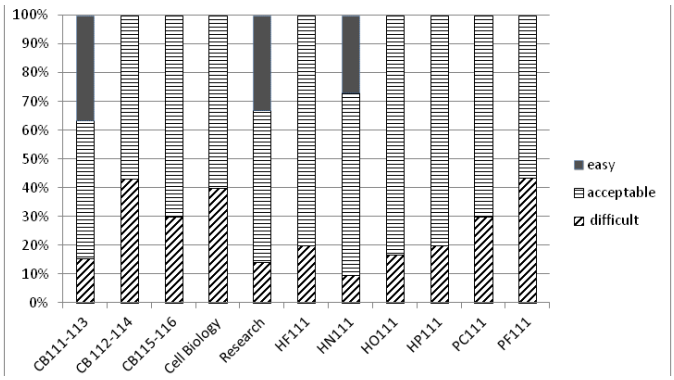

Figure 2: Difficulty index of the multiple-choice questions (DIFi)

Items of most of the courses were acceptable as per DIFi. The difficulty index of CB 112 + 114 (General Concepts and Hematopoietic system) and PF 111 (Fundamentals of Pharmacy) was the highest. Occupational Health and Toxicology (HO 111) and Fundamentals of Public Health (HF 111) had the most number of acceptable items. Easy questions where >70 % answered an item were observed in CB 111+113, Research & Biostatistics (CR 111), and Food & Nutrition (HN 111) (Figure 2).

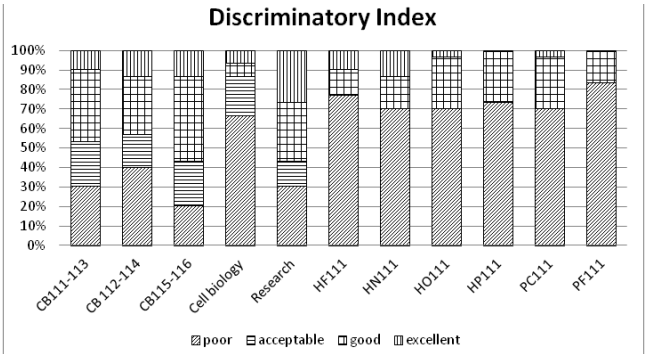

Figure 3: Discrimination index of the multiple-choice questions (DISi)

Most items had a poor discriminatory index. Items of Research and Biostatistics (CR 111) had the maximum number of items with excellent discriminatory index (n=8), but Fundamentals of Pharmacy (PF 111) and Fundamentals of Public Health (HF 111) had the most number of items with poor discriminatory index, with n=25 and 23 respectively (Figure 3).

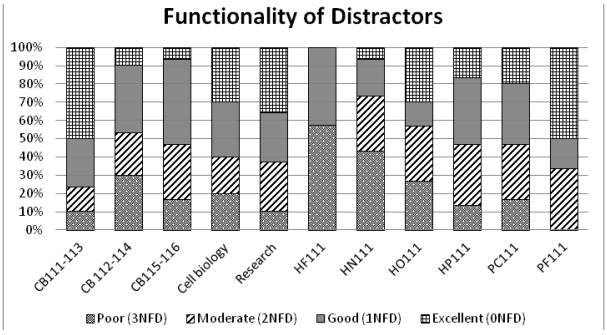

Figure 4: The functionality of the distractors in the multiple-choice questions (Distractor efficiency)

Figure 4 shows that CB 111+113 (Medical terminology and musculoskeletal system) and PF 111 (Fundamentals of Pharmacy) had the highest number of items with an excellent functional distractor (i.e. zero non- functional distractor). Fundamentals of Public Health had the majority of poor functionality of distractors (i.e. two non-functional distractors).

Summary of item analysis of all courses reveals that CB 111+ 113 and PC 111 had the highest number of functional distractors (72.2% each) and the highest number of items with excellent functionality of distractors (n=15 each). HN 111 (70%), HO 111(52%), and CB 115+ 116 (52%) had the highest number of non-functional distractors. CR 111 had the highest (n=8) with an excellent discriminatory index. Although most items fell within acceptable ranges based on the difficulty index, the item analysis showed overall poor performance.

DISCUSSION

We identified deficiencies in the quality of the summative exam administered to allied health science undergraduate students in the first semester. Ideally, there should be a DIFi between 30 and 70%, a DISi of at least 0.30, and a maximum DE of three functional distractors in a MCQs.[7] While a majority of the courses had a fair number of items that were acceptable according to the difficulty index, other indices, such as the discriminatory index and functionality of distractors, were not upto standards.

Discriminatory Index: Seven out of eleven courses had (>50%) majority of items with a poor discriminatory index ≤0.20. The discrimination index for item PF 111 was among the poorest. Despite this, there were 72.2% of high Functional Distractors (FDs). FDs with negative DISi are not FDs, since they distract top performers more than low-performers. Similar findings were supported by Puthiaparamil et al.[7] that four FD or 100% DE did not seem to be confined to high-class items. In contrast to our finding for PF 111, DIFi showed a significant negative correlation with FDs/item in their study, while DISi showed a significant positive correlation with FDs/item. In exam items, discrimination refers to the ability of an item to differentiate statistically between groups of examinees in a desired manner. As assessment is developed to measure the learning outcomes. We can use discriminatory indexes to analyze how well the test can distinguish between students who met their learning outcomes and those who did not. We use a scoring system to identify those who score higher and those who score lower on the test to create groups that more or less meet the learning outcomes we are trying to measure. It is based on the assumption that the entire test is a reasonably valid measure of the learning outcomes.[8]

Functionality of Distractors: Distractor efficiency was found to be low among items constructed. Puthiaparampil et al.[9] recommend using three-option MCQs instead of NFDs due to their difficulty in constructing plausible questions. The psychometric properties of the three-option test are not significantly different from those of the four- and five-option tests, based on a meta-analysis by Vyas et al.[10] As per meta-analysis by Rodriguez et al.,[11] item discrimination increases when options are fewer. Item writing errors are more prevalent in MCQs created at lower cognitive abilities, according to Tarrant et al.[12] The relationship between the distractor's performance and the complexity of the cognitive processes involved in selecting the right response has repeatedly been studied using Bloom's taxonomy, item difficulty, and discrimination index.[11,12] In a study by Kim et al.,[13] application and synthesis questions that demand critical thinking when compared to knowledge or comprehension questions significantly improved the discrimination index ($p<0.05$) using post-hoc analyses to the univariate ANOVA, however, did not reduce the proportion of right responses.

Difficulty Index: All courses had a majority of items that were acceptable in terms of difficulty index. Only three assessments had around one-third too easy MCQ items. The assessment of the correct response fractions showed that test questions involving analysis and synthesis/ evaluation, which required multiple areas of knowledge,

were significantly more difficult than questions associated with a single concept. When a test item appears to be very difficult (i.e., P is very small), it may be that the topic tested is inappropriate at this stage of students' training, or that it is not taught well in this particular academic session. Other possible reasons for poor performance on the items (i.e., D is very small) include ambiguity in the wording, areas of controversy, and perhaps, even that the wrong key was given. It is possible that a "good" student might not risk attempting a "difficult" MCQ item for fear of losing hard-earned marks on the other items of the same question. However, a "weak" student might take the risk to guess as he knows so little on the topic that he has nothing much to lose, and the least he can obtain for the whole question is zero marks. This could then result in a negative discrimination index.

A typical finding was found in HF 111 (Fundamentals of Public Health) where the mean score was lowest among all courses assessed yet almost 80% of items were in acceptable range as per difficulty index i.e. >30 to 70% of students could attempt 24 out of 30 MCQ items. No questions were too easy for the students, i.e. no items were answered correctly by more than 70% of students. Interestingly the discrimination index is however found to be poor with maximum percentage of non-functional distractors among all courses. Accordingly, a low mean score does not necessarily mean that most of the questions were difficult. The widespread of item discrimination values for similar-level questions might reflect some level of guessing. It is essential to perform a quality analysis of the items at this stage in order to evaluate a finding of this nature. In this way, it provides departments with effective feedback regarding their educational activities.[14]

Despite the majority of difficult items, CB112+114 (Applied Integrated Basic Sciences: General Concepts and Hematopoietic System), BC111 (Cell Biology) had poor distractor efficiency. This implies that DE and DIFi are not related in a predictable way. These findings are supported by some studies.[9,15] We thus recommend faculty members to construct MCQ items as per student's readiness. In a study by Pande et al.,[16] the discrimination index correlated positively with the difficulty index ($r = 0.191$, $p=0.003<0.01$) using Pearson correlation test.

Poor performance on the items i.e. low distractor index can be caused by ambiguities in the wording, controversy, or even giving the wrong key. MCQ items that are "difficult" may not be attempted by a "good" student as he or she might be afraid of losing marks on other questions in the same question. However, "weak" students may guess as they know so little about the topic and have nothing to lose since zero marks is the worst they can hope for the whole question. A negative discrimination index could result from this.[14] The study by Licona-Chavez et al.[17] analyzed DIFi, DISi, DE and Cronbach alpha to evaluate 20 MCQ items, but found there to be no parallel performance across all the metrics.

It is first important to flag poorly performing questions based on statistics. We recommend reviewing every flagged question because its statistical results might have an explanation. The feedback from students might also be helpful at this stage. Feedback may explain statistics that weren't obvious to the person reviewing the question. However, good statistics don't guarantee quality items, as we discovered in CR 111, where despite relatively better indexes, the items performed poorly on content analysis.

Our experience as educators in Nepal has shown that high stakes testing is a priority at the end of a learning period (called '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 students' learning experiences. If educators are unable to develop valid or reliable assessments because they lack the necessary skills, it is the academic institutions' responsibility to train and instruct them. Training can substantially improve the quality of MCQs developed by teaching faculty.[12] 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.[18]

Our study has few limitations. First, the data analysis for internal evaluation was conducted by the Examination Controller which we retrieved later. Thus, the study is limited due its retrospective nature which is why various other analyses to test quality of assessment couldn't be performed. Institutional restriction further did not allow access to the questions limiting our findings. The nature of the study does not allow us to establish cause and effect. Randomization was not done which may have led to selection bias. The sample size of study was small (n=38), which could have resulted in erroneous findings. A study with an improved methodology and larger sample size is recommended. We remain supportive of the study as such, since Nepal's educators of health professionals are unaware of construction of statistically assessed MCQ items. We strongly recommend other universities and autonomous higher education institutions of Nepal to

conduct similar studies with improved methodology and larger sample size including multiple centers to analyze their MCQ items constructed for formative and summative examinations. We also specifically request item analysis of items used in entrance exams such as Common Entrance Exam for Health Science, Kathmandu University Common Admission Test and Institute of Engineering Entrance Exam for standardized assessment.

CONCLUSIONS

Our institution had an overall low standard of MCQ items in most of the courses. Difficult items do not necessarily guarantee better discriminatory performance. Educators should be encouraged to reflect on the findings and take proactive steps to fix deficiencies in assessment tools. The study findings highlight the need for careful and rigorous formulation of MCQs to evaluate learner's competence. Similar studies ought to be conducted at all medical institutions in Nepal in order to promote quality assessment.

ACKNOWLEDGMENTS:

The authors would like to express our deepest gratitude to Dr. Alina Karna and Dr. Sansar Babu Tiwari for their valuable help.

CONFLICTS OF INTEREST: None declared

SOURCE OF FUNDING: None

AUTHORS CONTRIBUTION

SKD conceptualized the research, NB prepared the manuscript, SG analyzed the data, NB and SG collected the data. All authors read and approved the manuscript.

REFERENCES

1. Swanwick T, Forrest KAT, O'Brien BC. Understanding Medical Education: Evidence, Theory, and Practice. 3rd ed. Hoboken, NJ: Wiley-Blackwell, 2018. 600 p. DOI: 10.1002/9781119373780
2. Ismail SM, Rahul DR, Patra I, Rezvani E. Formative vs. summative assessment: impacts on academic motivation, attitude toward learning, test anxiety, and self-regulation skill. Lang Test Asia. 2022;12(1):40. DOI: 10.1186/s40468-022-00191-4
3. McCoubrie P. Improving the fairness of multiple-choice questions: A literature review. Medical Teacher. 2004;26(8):709–12. DOI: 10.1080/01421590400013495
4. Chiavaroli N, Familari M. When majority doesn't rule: The use of discrimination indices to improve the quality of MCQs. Bioscience Education.2011;17(1):1–7. DOI: 10.3108/beej.17.8
5. Kaur M, Singla S, Mahajan R. Item analysis of in the use multiple choice questions in pharmacology. Int J Appl Basic Med Res. 2016;6(3):170-3. DOI: 10.4103/2229-516X.186965 PMID: 27563581.
6. Bhat SK, Prasad KHL. Item analysis and optimizing multiple-choice questions for a viable question bank in ophthalmology: A cross-sectional study. Indian J Ophthalmol. 2021;69(2):343-6. DOI: 10.4103/ijo.IJO_1610_20 PMID: 33463588.
7. Sahoo DP, Singh R. Item and distractor analysis of multiple choice questions (MCQs) from a preliminary examination of undergraduate medical students. International Journal of Research in Medical Sciences. 2017;5(12):5351. DOI: 10.18203/2320-6012.ijrms20175453.
8. Hogan TP. Psychological Testing: A Practical Introduction, 4th ed. New York: Wiley;2018.
9. Puthiaparampil T, Rahman M. How important is distractor efficiency for grading Best Answer Questions?. BMC Med Educ 21, 29 (2021). DOI: 10.1186/s12909-020-02463-0
10. Vyas R, Supe A. Multiple choice questions: a literature review on the optimal number of options - PubMed. The National medical journal of India. 2008;21(3): 103-133. PMID: 19004145.
11. Rodriguez MC. Three Options Are Optimal for Multiple-Choice Items: A Meta-Analysis of 80 Years of Research. Educational Measurement: Issues and Practice. 2005;24(2):3–13. DOI:10.1111/j.1745-3992.2005.00006.x
12. Tarrant M, Ware J. Impact of item-writing flaws in multiple-choice questions on student achievement in high-stakes nursing assessments - PubMed. Medical education. 2008;42(2):198-206. DOI: 10.1111/j.1365-2923.2007.02957.x PMID: 18230093.
13. Kim M-K, Patel RA, Uchizono JA, Beck L. Incorporation of Bloom's Taxonomy into Multiple-Choice Examination Questions for a Pharmacotherapeutics Course. American Journal of Pharmaceutical Education. 2012;76(6):114. DOI: 10.5688/ajpe766114 PMID: 22919090.

14. Sim S-M, Rasiah RI. Relationship between item difficulty and discrimination indices in true/false-type multiple choice questions of a para-clinical multidisciplinary paper - PubMed. Annals of the Academy of Medicine, Singapore. 2006;35(2):67-71. PMID: 16565756.

15. Mahjabeen W, Alam S, Hassan U, Zafar T, Butt R, Konain S, et al. Difficulty Index, Discrimination Index and Distractor Efficiency in Multiple Choice Questions. Annals of PIMS-Shaheed Zulfiqar Ali Bhutto Medical University. 2017;13(4):310–5. DOI: 10.48036/apims.v13i4.9

16. Pande SS, Pande SR, Parate VR, Nikam AP, Agrekar SH. Correlation between difficulty & ; discrimination indices of MCQs in formative exam in Physiology. South-East Asian Journal of Medical Education. 2013;7(1):45. DOI: 10.4038/seajme.v7i1.149

17. Licona-Chávez AL, Montiel Boehringer PK, Velázquez-Liaño LR. Quality assessment of a multiple-choice test through psychometric properties. MedEdPublish. 2020;9:91. DOI: 10.15694/mep.2020.000091.1

18. Harrison, Könings, Schuwirth, Wass, Vleuten van der. Changing the culture of assessment: the dominance of the summative assessment paradigm. BMC Medical Education. 2017;17(1):1–14. DOI: 10.1186/s12909-017-0912-5