International Journal of Research in Medical Sciences
Patil RP et al. Int J Res Med Sci. 2022 Dec;10(12):2878-2883
www.msjonline.org
pISSN 2320-6071 | eISSN 2320-6012
DOI: https://dx.doi.org/10.18203/2320-6012.ijrms20223091

Original Research Article

Evaluation of validity and reliability of multiple-choice questions in second MBBS competency-based medical education-based pharmacology examination of medical institute of India

Rushikesh P. Patil, Satish E. Bahekar, Madhuri D. Kulkarni, Mirza S. Baig*

Department of Pharmacology, Government Medical College, Aurangabad, Maharashtra, India

Received: 03 October 2022
Revised: 04 November 2022
Accepted: 09 November 2022

***Correspondence:**
Dr. Mirza S. Baig,
E-mail: shirazdoctor@yahoo.com

ABSTRACT

Background: Multiple choice questions (MCQs) are most commonly used assessment tool in undergraduate medical examination. Assessment method must be reliable and valid. To improve quality of MCQs, item analysis was carried out by determining their validity and reliability using parameters like difficulty index, discrimination index, distractor efficiency and Cronbach's alpha value.
Methods: Study was carried out among 193 second year medical students. Each student was given 40 MCQs of 1 mark each. After assessment of MCQs, validity of test was analyzed by using difficulty index, discrimination index and distractor efficiency while reliability was analyzed by using Cronbach's alpha.
Results: Mean ± SD of difficulty index, discrimination index, functioning and non-functioning distractors were 59.80±23.38, 0.25±0.12, 1.98±0.92 and 13.25±13.05 respectively with reliability value of 0.7. About 47.5% items had moderate difficulty index, 22.5% items have excellent discrimination index with 35% items having 100% distractor efficiency. Reliability of test as measured by Cronbach's alpha value was 0.7. There was weak correlation between difficulty index and discrimination index.
Conclusions: It is concluded from study that given MCQs test have reliability but not validity and needs to improve quality of MCQs. Validity of test is improved by improving difficulty index, discrimination index, distractor efficiency of items.

Keywords: Difficulty index, Discrimination index, Distractor efficiency, Cronbach's alpha, Item-analysis

INTRODUCTION

As defined by international competency-based medical education (CBME) collaborator's competency is "An observable ability of a health professional, integrating multiple components such as knowledge, skills, values, and attitudes. Observable competencies can be measured and assessed to ensure their acquisition. CBME education as defined is "an outcomes-based approach to the design, implementation, assessment, and evaluation of medical education programs, using an organizing framework of competencies. Thus, competencies can be assembled like building blocks to facilitate progressive development.[1]

In CBME, assessment of the student is important, and it must be tough enough and multidimensional. Assessment to be effective, it must be continuous, frequent, criterion and work-based must meet a certain minimum standard of quality in terms of validity, reliability, acceptability, educational impact, and cost-effectiveness should have a qualitative approach, drawn according to the wisdom of the group and must involve trainee.[2]

In India, the CBME pattern commenced in 2019 for undergraduate medical students. In CBME, there are 2 types of assessments. One is formative assessment that helps to improve learning and to modify teaching-

learning strategies. Other is a summative assessment conducted by the university and meant for certification or assessment of learning.[3]

MCQs are widely used as an assessment tool in the undergraduate examination. MCQs are simple and easy for scoring, time-efficient, assess higher-order cognitive processing, and tests many skills apart from factual knowledge. Well-constructed MCQs test the application of medical knowledge and should be written according to the difficulty level of candidates.[4]

MCQs provide greater domain sampling having high content validity. They have high validity and reliability. Good objectivity, easiness for analysis, usefulness for banking of items, and transparency are other advantages of MCQs. MCQs (hereafter it is called item) consist of question part called as 'stem' and one correct answer called 'key' and other incorrect options called 'distractors' on which quality of MCQs depends.[5]

Items can be too easy or too difficult. Too easy to guess or do not discriminate positively will give an inappropriate decision about student ability and further decisions like pass/fail and give inappropriate feedback to students and instructors also. So, it is important to evaluate the quality of MCQs in such cases. For that purpose, item analysis is carried out.[6]

Item analysis is carried out to review and revise items, to know the quality of items, in turn, test[7]. For the evaluation of the item, two methods are commonly used. One is classical test theory (CTT) and item response theory (IRT). CTT mainly includes traditional statistics such as item difficulty, item discrimination, distractor analysis, item-test correlation while and reliability of the test can be measured by Cronbach's alpha and Kuder-Richardson Formula (KR20) value.[8]

Item difficulty, item discrimination used to summarize, evaluate and compare set of items with regards to difficulty and discrimination and to check for potential that may warrant item revision before item use in test.[7]

The present study was conducted to observe validity and reliability of multiple-choice questions in pharmacology examination conducted at medical institute among undergraduates by using item analysis which is carried out by using classical test theory which includes parameters like difficulty index, discrimination index, and distractor efficiency along with various statistical parameters of reliability like Cronbach's alpha.

METHODS

It was retrospective observational study conducted at the department of pharmacology, government medical college, Aurangabad during November-2021 to January 2021 among 193 second year MBBS students as a part of their internal assessment. Actual study was started after approval of from institutional ethics committee.

Each student was given 40 items (MCQs) of pharmacology and consisting of one key (correct answer) and three distracters (incorrect options). Each item carries one mark. Thus, the total score of the exam was 40. Each correct item was given 1 mark and the incorrect item was given 0 marks. There was no negative marking. If a student did not attempt an item or attempt the same item twice or more, then such item was given 0 marks.

After the assessment, scores of students were analysed and ranked from highest score to lowest score. Students were divided into three groups i.e., high achiever group consists of 33% of students with the highest rank, low achiever group consists of 33 % students from lowest rank, and mid achiever group consists of middle 33% students (remaining students between high achievers and low achievers). Only high achiever and low achiever groups were considered for analysis.

Difficulty index and discrimination index were calculated by following formula:

Difficulty index=(H+L)/N×100
Discrimination index=(H-L)/N×2

Where, H=No. of students answered correctly in the high achiever group, L=No. of students answered correctly in the low achiever group and N=No. of students in high and low achiever group including non-respondent.

Difficulty index/ p is interpreted as shown in Table 1 while discrimination index/ d is interpreted in Table 2.

Table 1: Interpretation of difficulty index/ p value.

Grade	P value (%)
Easy	>70
Moderate	30-70
Difficult	<30

Table 2: Interpretation of discrimination index/ d value.

Grades	Values
Excellent	>0.35
Good	0.25-0.34
Marginal	0.15-0.24
Poor	<0.15

Table 3: Interpretation of distractor efficiency.

Number of NFDs	Distractor efficiency (%)
0	100
1	66.6
2	33.3
3	0

Table 4: The relationship between difficulty index and discrimination index.

P value	D value	Interpretation	Comment
<0.6	<0.15	A difficult item with poor discrimination	Verify answers have been keyed correctly. If no key error, consider removing the item.
<0.6	≥ 0.15	A difficult item with high discrimination	Retain item.
0.6-0.9	≤ 0	Moderate to low difficulty item with negative discrimination	Verify answers have been keyed correctly. If no key error, consider removing an item
0.6-0.9	0<d<0.15	Moderate to low difficulty item with low discrimination	Retain item but consider revising
0.6-0.9	> 0.15	Moderate to low difficulty item with high discrimination.	Retain
>0.9	Disregard	Low difficulty item	Retain item but consider revising.

Distracter efficiency

All the distracters with frequency <5 % were identified i.e., non-functioning distractors (NFDs) and the number of items 0, 1, 2, and 3 NFDs were calculated. Distractor efficiency of an item is determined as shown in the Table 3.[9]

The relationship between difficulty index and discrimination index is shown in Table 1 and interpreted as shown in Table 4.[10]

Statistical analysis

Data obtained was entered in MS-excel and the difficulty index and discrimination index was calculated and expressed as Mean ± SD. The relation between difficulty index and discrimination index was analyzed by using Pearson Correlation Coefficient with the help of SPSS software. The reliability of the test was analyzed by using Cronbach's alpha using SPSS software.

RESULTS

In our study, a total of 193-second MBBS students were enrolled. Total 40 MCQs were constructed and evaluated among these students. Each item consists of a stem or problem and four options comprising one correct option (key) and three incorrect options (distracters). Thus, a total of 40 correct options and 120 distracters were analyzed.

Test scores ranged from 8 to 35. The Mean±SD score of the test was 24.10±4.75. The range of scores in the high achiever group was from 26 to 35 while that of low achiever group ranged from 12 to 23.

In the present study, the difficulty index ranged from 12.5-94.53. For difficulty index mean ± SD was 59.80±23.38. Out of 40 items, 4 items (10 %) had higher

difficulty while 17 items (42.5%) had an easy difficulty level. 19 items (47.5 %) had a moderate difficulty index (Figure 1).

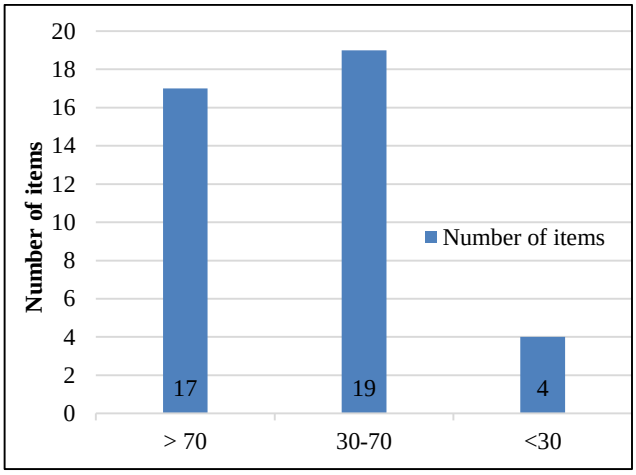

Figure 1: Difficulty index of items versus number of items.

The discrimination index ranged from 0-0.484. Mean±SD for discrimination index was 0.25±0.12. Out of 40 items, 9 items (22.5%) had excellent discrimination index, 13 items (32.5 %) had good discrimination index, 8 (20%) items had marginal discrimination index while 10 items (25%) had poor discrimination index (Figure 2).

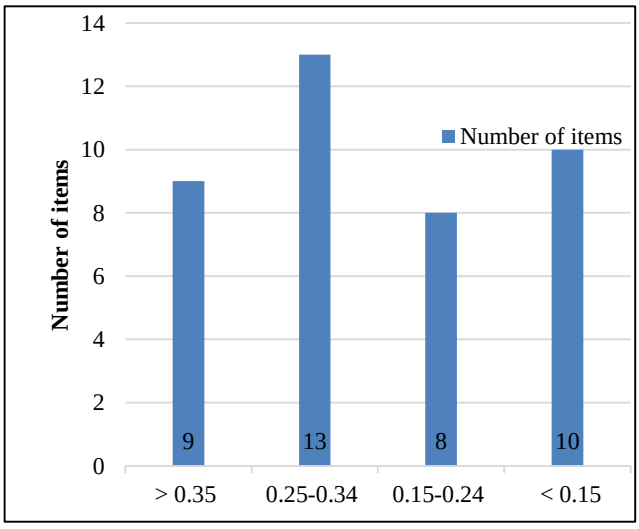

Figure 2: Discrimination Index of items versus number of items.

Present study assessed 40 items and 120 distractors. Out of 120 distractors analyzed, 41 distractors (34.17%) were non-functioning while 79 distractors (65.83 %) were functional. More than half of the distractors were functional. The mean number of functional distractors per item was 1.98±0.92. The mean number of non-functional distractors per item was 13.25±13.05. About 14 items (35 %) have distractor efficiency of 100 % meaning that they did not have non-functional distractors or have three functioning distractors. 2 items (5%) have three non-functioning distractors having low distractor efficiency of 0 %. 2 distractors were not chosen by anyone in the test. The following graph shows several items with 0, 1, 2, and 3 NFDs (Figure 3).

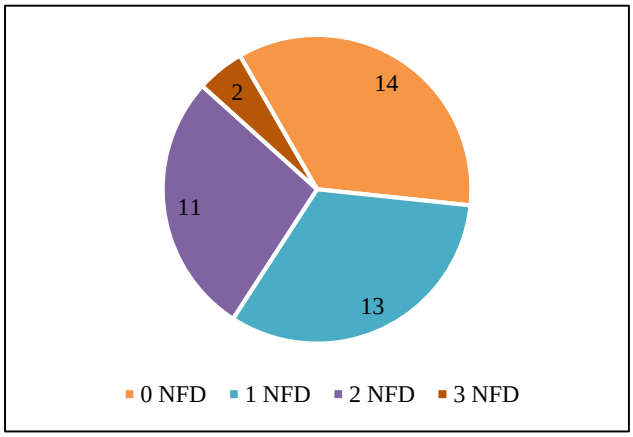

Figure 3: Number of items with NFDs.

The correlation of the difficulty index with the discrimination index is parabolic with an R2 of 0.52 signifying 52% of the variability in the discrimination index could be explained by the difficulty index. This significant correlation ($p<0.01$) implies that the too low and too high difficulty index of the question leads to a poor discrimination index. The small sample size is the limitation of the study else this R2 value could have been high of close to 0.7 (Figure 4).

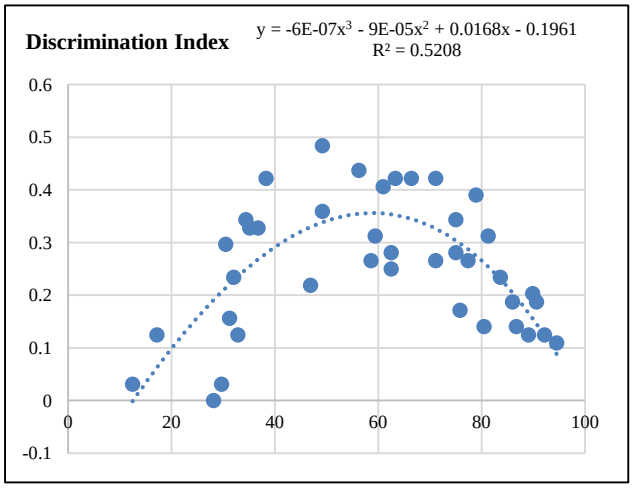

Figure 4: Difficulty index with discrimination index.

The Cronbach's alpha value of the test is 0.7. This means that the given test is reliable but not valid.

DISCUSSION

Assessment is an important tool in medical education for further educational opportunities, maintenance of standards, encouragement of learning, to provide feedback to students for their performance as well as to teachers to improve and modify their teaching, and lastly for preparation of students for real life. There are five attributes for the evaluation of the assessment method-reliability, validity, educational impact, feasibility, and acceptability.[11]

MCQs is one of the assessment methods in academics. MCQ evaluation is a practically possible method as it can test many subjects at a time, do not have a negative impact on student performance, and provide an objective evaluation. Apart from that, it is a reliable method for covering wide knowledge and used for assessing higher cognitive functions like application and analysis if framed correctly. Although reliable, its' validity is questionable and for that item analysis techniques are used for improving the quality of MCQs.[12]

Percentage of people who answer an item correctly is noted as Item difficulty. It is denoted by 'p' and depends upon the sample tested. P value ranges from 0 to 100% and in the four option MCQs test, a p25% is considered as the lowest value. Item discrimination index measures how well an item differentiates between low- and high-performing students. It is denoted by d. D value ranges from -1 to +1. A minimum value of d should be 0.15 for course-based examination while for standardized examination d value should be at least 0.3.[10]

In the present study, there are 19 items (47.5 %) had moderate difficulty index and 4 items had higher difficulty index. In our study, 30 items have a discrimination index >0.15 having marginal to excellent discrimination index while 10 items have a poor discrimination index. Thus, 30 items discriminate between low-performing as well as high-performing students.

Many studies analyze MCQs for their quality. In our study, the difficulty index and discrimination index were 59.80±23.38 and 0.25±0.12 respectively. This is in correlation with the difficulty index and discrimination index observed by Rao et al where they reported difficulty index and discrimination index of 50.16±16 and 0.34±0.17 respectively.[13] A similar finding was also found in a study conducted by Namdeo et al having difficulty index and discrimination index 65.92±22.2 and 0.33±0.23 respectively.[14] In the same study, 56% of items have acceptable difficulty and 10% of items have the high difficulty. This is consistent with the result of our study.

Difficulty index and discrimination index have a reciprocal relationship and items with high difficulty index are said to be poor discriminators while items with low discrimination index have low difficulty index. But it is not true always.[15] In our study, it was observed that items with difficulty index <30 i.e., difficult items have low discrimination index while most of the items with high difficulty index have high discrimination index that is d>0.15, concluding that their relationship is not always reciprocal. Apart from that most of the items with an acceptable range of p, have higher d value.[13] In our study, about 19 items have an acceptable level of difficulty with d value ranges between 0.125-0.484 with only one item having a d value less than 0.15.

In our study, 5 items were having a p<60% and d<0.15, so these items need to be removed. About 13 and 16 items have p<0.6 and 60-90% respectively and d≥0.15 and >0.15 respectively, so these items should remain in the test. There are no items with a negative d value. Many items with a p=60% to 90% and d value between 0 and 0.15 are 3 and these items should retain in the test but requires revision. Similarly, 3 items having a p>90 % also require revision.

In our study, there were 41 (34.17%) distractors which are selected by <5% examinees and should be revised, replaced, or removed. Two distractors were not selected by anyone and so these distractors should be removed from the test. Items with 0% DE should be removed and those with >0% but < 5% are revised or replaced with better choices. Distractors have an impact on total test scores and on which student performance depends.[16] It is observed that items with one NFD have the excellent discriminative ability as compared to items without NFD.[13] This finding is observed in our study. About 30.8% of items with 1 NFD had an excellent discriminative index (d>0.35) as compared to 28.6% of items without NFD.

In a study conducted by Kolte, there were 19 (47.5%), 16 (40%) and 3 (7.5%) items have 0, 1 and 3 NFDs respectively which is consistent with our result showing 14 (35%), 13 (32.5%) and 2 (5%) items with 0, 1 and 3 NFDs respectively.[17] D'Sa et al also had finding consistent with our study with 13 (27.08%), 9 (18.75%) and 2 (4.17%) items having 1, 2 and 3 NFDs respectively.[18] In our study items with 0% distractor efficiency was 5%.

While constructing the better MCQs, it is important to choose the good distractor that is plausible to correct the answer which is selected by the students that don't know the correct answer. It is important to select a plausible distractor otherwise it is unable to discriminate between good and poor scoring students.[19]

In our study, Pearson's correlation coefficient was 0.7. This significant correlation implies that the too low and too high difficulty index of the question leads to a poor discrimination index. It is concluded that there is a weak relationship between difficulty index and discrimination index and the relationship between them is not linear throughout the range of their values. Reliability refers to the reproducibility or consistency of assessment results over time.[11] It is used for evaluating the extent to which items in the test relate to each other and is measured by using Cronbach's alpha value which ranges from 0 meaning no reliability to 1 meaning a high degree of reliability. For most of the course-based exams, Cronbach's alpha value ranges from 0.6 to 0.8, and in general, the value should be at least 0.5.[10] In our study, the reliability of the test was 0.77 which means the given MCQs test is reliable.

It is concluded that MCQs asked in the examination have low validity depending upon the value of difficulty index (59.80±23.38 %), discrimination index (0.25±0.12), and distractor efficiency. Given MCQs were reliable based upon the value of Cronbach's alpha 0.77. It is important not to pick up MCQs directly from the MCQs question bank and take time to make effective MCQs identify strengths and weaknesses in students' understanding, for higher-order cognitive processing, to encounter real situations in practice, and to conjure complex thought processes of students.[20] Although reliable and valid, MCQs have some limitations too. MCQs are difficult to construct requiring a long time, difficult to find plausible distractors, non-recognition of partial knowledge of students, ineffective in assessing problem-solving approach, and maybe a double-cut weapon if taken individually.[21]

Limitations

Limitations of present study are less number of MCQs were analysed in small number of students and this cannot be generalised for all MCQs based assessment examination. Item analysis should be carried out after each MCQs based on the examination to improve their quality.

CONCLUSION

It is concluded from the study that our MCQs were reliable but not valid and need to improve validity in terms of difficulty index, discrimination index, and distractor efficiency. Relationship between the difficulty index and the discrimination index is not linear. It is not possible to comment on the validity of the MCQs test, but the test was reliable. Although non-valid and reliable, it is possible to improve the validity of the test by improving difficulty level, discrimination score and improving and managing plausible distractors.

Funding: No funding sources
Conflict of interest: None declared
Ethical approval: The study was approved by the Institutional Ethics Committee

REFERENCES

1. Frank JR, Snell LS, Cate OT, Holmboe ES, Carraccio C, Swing SR et al. Competency-based medical education: theory to practice. Med Teacher. 2010;32(8):638-45.
2. Shah N, Desai C, Jorwekar G, Badyal D, Singh T. Competency-based medical education: An overview and application in pharmacology. Ind J Pharmacol. 2016;48(1):S5.
3. Medical Council of India. Assessment Module for Undergraduate Medical Education Training Program. 2019;1-29.
4. Al-Rukban MO. Guidelines for the construction of multiple choice questions tests. J Family Community Med. 2006;13(3):125.
5. Amin Z, Khoo HE. Basics in medical education. World Scientific. 2003.
6. Brown GT, Abdulnabi HH. Evaluating the quality of higher education instructor-constructed multiple-choice tests: Impact on student grades. In Frontiers in Education. 2017;2:24.
7. Moses T. A review of developments and applications in item analysis. Advancing Human Assessment. 2017;19-46.
8. Bichi AA. Classical Test Theory: an introduction to linear modeling approach to test and item analysis. Int J Social Studies. 2016;2(9):27-33.
9. Mehta G, Mokhasi V. Item analysis of multiple choice questions-an assessment of the assessment tool. Int J Health Sci Res. 2014;4(7):197-202.
10. Rudolph MJ, Daugherty KK, Ray ME, Shuford VP, Lebovitz L, DiVall MV. Best practices related to examination item construction and post-hoc review. Am J Pharmaceutical Educ. 2019;1;83(7).
11. Sood R, Singh T. Assessment in medical education: Evolving perspectives and contemporary trends. Natl Med J India. 2012;25(6):357-64.
12. Brady AM. Assessment of learning with multiple-choice questions. Nurse Educ Pract. 2005;5(4):238-42.
13. Rao C, Prasad HK, Sajitha K, Permi H, Shetty J. Item analysis of multiple choice questions: Assessing an assessment tool in medical students. Int J Educational Psychological Res. 2016;2(4):201.
14. 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):1716-9.
15. Hingorjo MR, Jaleel F. Analysis of one-best MCQs: the difficulty index, discrimination index and distractor efficiency. J Pak Med Asso. 2012;62(2):142.
16. Burud I, Nagandla K, Agarwal P. Impact of distractors in item analysis of multiple choice questions. Int J Res Med Sci. 2019;7(4):1136-9.
17. Kolte V. Item analysis of multiple choice questions in physiology examination. Indian J Basic Applied Med Res. 2015;4(4):320-6.
18. D'Sa JL, Visbal-Dionaldo ML. Analysis of Multiple Choice Questions: Item Difficulty, Discrimination Index and Distractor Efficiency. Int J Nursing Educ. 2017;9(3).
19. Alsubait T, Parsia B, Sattler U. A similarity-based theory of controlling MCQ difficulty. In 2013 second international conference on e-learning and e-technologies in education (ICEEE). 2013;283-8.
20. Velou MS, Ahila E. Refine the multiple-choice questions tool with item analysis. Int Arch Integrated Med. 2020;7(8):80-5.
21. Tangianu F, Mazzone A, Berti F, Pinna G, Bortolotti I, Colombo F et al. Are multiple-choice questions a good tool for the assessment of clinical competence in Internal Medicine? Italian J Med. 2018;20;12(2):88-96.

Cite this article as: Patil RP, Bahekar SE, Kulkarni MD, Baig MS. Evaluation of validity and reliability of multiple-choice questions in second MBBS competency-based medical education-based pharmacology examination of medical institute of India. Int J Res Med Sci 2022;10:2878-83.