Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/mjafi

Original Article

Item analysis of multiple choice questions: A quality assurance test for an assessment tool

Dharmendra Kumar [a], *Raksha Jaipurkar* [b,*], *Atul Shekhar* [b], *Gaurav Sikri* [c], *V. Srinivas* [d]

[a] *Assistant Professor, Department of Physiology, Armed Forces Medical College, Pune, India*
[b] *Associate Professor, Department of Physiology, Armed Forces Medical College, Pune, India*
[c] *Professor & Head, Department of Physiology, Armed Forces Medical College, Pune, India*
[d] *Commandant, Military Hospital, Secunderabad, India*

ARTICLE INFO

Article history:
Received 27 October 2020
Accepted 10 November 2020

Keywords:
MCQs
Item analysis
Difficulty index
Discrimination index
Distractor effectiveness

ABSTRACT

Background: The item analysis of multiple choice questions (MCQs) is an essential tool that can provide input on validity and reliability of items. It helps to identify items which can be revised or discarded, thus building a quality MCQ bank.

Methods: The study focussed on item analysis of 90 MCQs of three tests conducted for 150 first year Bachelor of Medicine and Bachelor of Surgery (MBBS) physiology students. The item analysis explored the difficulty index (DIF I) and discrimination index (DI) with distractor effectiveness (DE). Statistical analysis was performed by using MS Excel 2010 and SPSS, version 20.0.

Results: Of total 90 MCQs, the majority, that is, 74 (82%) MCQs had a good/acceptable level of difficulty with a mean DIF I of 55.32 ± 7.4 (mean ± SD), whereas seven (8%) were too difficult and nine (10%) were too easy. A total of 72 (80%) items had an excellent to acceptable DI and 18 (20%) had a poor DI with an overall mean DI of 0.31 ± 0.12. There was significant weak correlation between DIF I and DI ($r = 0.140$, $p < .0001$). The mean DE was 32.35 ± 31.3 with 73% functional distractors in all. The reliability measure of test items by Cronbach alpha was 0.85 and Kuder-Richardson Formula 20 was 0.71, which is good. The standard error of measurement was 1.22.

Conclusion: Our study helped teachers identify good and ideal MCQs which can be part of the question bank for future and those MCQs which needed revision. We recommend that item analysis must be performed for all MCQ-based assessments to determine validity and reliability of the assessment.

* *Corresponding author.*
E-mail address: rakshukarade@gmail.com (R. Jaipurkar).

https://doi.org/10.1016/j.mjafi.2020.11.007

Introduction

Competency-based assessment centres on performance of the students, that is, whether they demonstrate competence.[1] Assessment plays an important role in helping to interpret the magnitude of a student's ability and their own learning progress.[2,3]

The key features of any assessment are content and construct validity (a method assesses what it intends to assess), reliability (how well a score represents an individual's ability) and objectivity (assessment with single correct answer) which can be recalled based on higher-order thinking skills and problem solving. All methods of assessment have some strengths and shortcomings.[4]

Multiple choice questions (MCQs) are commonly used in assessments at undergraduate and postgraduate medical examinations because they are efficient, reliable and can be conveniently standardized. The quality of MCQs is important because of its effect on the students' overall competency level during their assessment. Well-constructed MCQs allow the assessment of higher-order cognitive skills such as interpretation, analytical and critical thinking, application or synthesis in the framework of Bloom's taxonomy and Miller's pyramid.[5] Framing the MCQs is a challenging task. A meticulously built up MCQ question bank after thorough item analysis is a handy tool for any academic institute for conducting the assessments.

The item analysis of an assessment tool provides input about validity and reliability of the item. The MCQ item analysis consists of the difficulty index (DIF I) (percentage of students that correctly answered the item), discrimination index (DI) (distinguish between high achievers and non-achievers), distractor effectiveness (DE) (whether well the items are well constructed) and internal consistency reliability (how well the item are correlated to one another).[6] Each item is evaluated for these indices because if an item is flawed, then it can become a distractor and the assessment can fail.

We have undertaken the project of item analysis of MCQ tests conducted for first-year Bachelor of Medicine and Bachelor of Surgery (MBBS) students in physiology. This item analysis will help to retain quality MCQs, discard or reframe items which have not been well framed. This will help to build a MCQ question bank comprising 'ideal' MCQs. This exercise will also help the faculty to construct good MCQs in future.

Material and methods

Three internal assessments of 150 first-year MBBS students were conducted which included 90 MCQs. Each MCQ consisted of a stem and 04 options with only one correct response and three distractors. The correct response was awarded one mark and the wrong response given zero mark. There was no negative marking. The upper 27% (41) students were considered high achievers (H) and lower 27% (41) as low achievers (L).[7] Each item was analysed for four indices, that is, DIF I, DI, DE and internal consistency reliability.

DIF I/facility value/P value

It is the percentage of students in the high and low achievers' group who answered the item correctly. It ranges between 0% and 100%. The criteria for classification of the DIF I are as follows: DIF I <30 (too difficult), DIF I between 30 and 70% (good/acceptable/average), DIF I >70% (too easy) and DIF I between 50 and 60% (excellent/ideal).[8–10] It was calculated using the formula DIF I or $P = (H + L) \times 100/N$, where H = number of students answering the item correctly in the high achieving group, L = number of students answering the item correctly in the low achieving group and N = total number of students in the two groups (including non-responders).

DI or d value

The DI is the ability of an item to differentiate between students of higher and lower abilities and ranges between 0 and 1. The criteria for classification of the DI are as follows: DI $\leq$ 0.20 (poor), 0.21–0.24 (acceptable), DI 0.25–0.34 (good) and DI $\geq$ 0.35 (excellent).[9] It was calculated using the formula DI = 2 × (H–L)/N, where the symbols H, L and N represent the same values as those mentioned earlier.

Distractor Effectiveness

DE is determined for each item based on the number of non-functional distractors (NFDs) (option selected by <5% of students) in it. The items were categorized on the basis of numbers of NFDs in MCQs, that is, an item having three NFDs, two NFDs, one NFD and zero NFD, the DE would be 0% (poor), 33.3% (moderate), 66.6% (good) and 100% (excellent), respectively.[8–10]

Internal consistency reliability

The SEM is directly related to the reliability of the test. It is an index of the amount of variability in an individual student's performance due to random measurement error. The standard deviation of the distribution is called the SEM and reflects the amount of change in the student's score which could be expected from one test administration to another.[11]

Cronbach alpha is a measure of internal consistency, that is, how closely related a set of items are as a group. It is a measure of reliability. The value of alpha between 0.8 and 0.9 falls in good range.[12] KR-20 is a measure of reliability for a test with binary variables (i.e. answers that are right or wrong). Reliability refers to how consistent the results from the test are or how well the test is actually measuring what you want it to measure. The scores for KR-20 range from 0 to 1, where 0 is no reliability and 1 is perfect reliability.[13] The closer the score is to 1, the more reliable the test. Just what constitutes an 'acceptable' KR-20 score depends on the type of test. In general, a score of more than 0.5 is usually considered reasonable.[14]

Skewness and kurtosis values were measured. Skewness is a measure of lack of symmetry. Kurtosis measures if data are heavy tailed or light tailed relative to normal distribution. The values of skewness and kurtosis between −2 and +2 are

considered as acceptable to prove normal univariate distribution.[15]

Statistical analysis

Statistical analysis was performed by using MS Excel 2010 and SPSS, version 20.0 (IBM, Armonk, NY, United States of America).

Results

The results of study showed that the scores of 150 students ranged from 30 to 91%. The test score of class ranged from 27 to 82 (total score 90) with a mean test score of 47.35 ± 11.6 (mean ± SD). The median was 44, and the interquartile range was 20. The skewness and kurtosis values were 0.59 and 0.69, respectively. The reliability measure of test items by KR-20 was 0.71, with a Cronbach alpha of 0.85 which is in the range of good. The SEM was 1.22 (Table 1).

Results of the DIF I showed that of 90 MCQs, 74 (82%) MCQs had good/acceptable levels of difficulty (DIF I 30–70%), whereas seven (8%) were too difficult (DIF I <30%) and nine (10%) were too easy (DIF I >70%). Among all MCQs, 34 (38%) MCQs had excellent/ideal levels of difficulty (DIF I 50–60%) (Table 2).

Of total 90 MCQs, 18 (20%) MCQ items had a poor DI (≤0.20), 11 (12%) items had an acceptable DI (0.21–0.24), 32 (36%) showed a good DI (0.25–0.34) and 29 (32%) items showed an excellent DI (≥0.35) (Table 3).

Pearson correlation of the DIF I and DI was also analysed for each of the MCQ items. We found a weak correlation between the DIF I and DI ($r = 0.140$, $p < .0001$).

Of the total 270 distractors for 90 items, 198 (73%) were functional distractors (FDs) and 72 (27%) were NFDs. Of all items, two (2%) items had poor DE (0%), eight (9%) had moderate DE (33.3%), 50 (56%) had good DE (66.6%) and remaining 30 (33%) had excellent DE (100%) (Table 4).

Our study showed that 18 (20%) MCQs fulfilled all the three criterias (a good/acceptable level of the DIF I [30–70%] with a high DI [≥0.25] and 100% DE.) which should be present for an ideal MCQ.

The mean values of the DI, DIF I and DE are given in Table 5.

Table 1 – Descriptive results of MCQ test scores.

Parameters	Result
No. of items	90
Percentage of mean test score ± SD	47.35 ± 11.6
Median	44
Range of test scores (%)	30–91
Interquartile range	20
Skewness	0.59
Kurtosis	0.69
Cronbach alpha	0.85
Kuder-Richardson 20 (KR-20)	0.71
Standard error of measurement (SEM)	1.22

MCQ, multiple choice question.

Discussion

Item analysis is a relatively simple, valuable procedure that provides a method for analysing observation, interpretation of the knowledge achieved by the students and information regarding the quality of test items.[9,16] In this study, we have performed item analysis of single best response type MCQ as it is seen as an efficient tool of assessment of the student's level of learning in academics.[8–10] The efficiency of MCQ assessment is solely based upon the quality of test MCQs.

In this study, Skewness and Kurtosis values (0.59 and 0.69) represent the moderately skewed data with tail on the right side of the distribution and data are lighttailed (lack of outliers) i.e distribution of data is approximately normal.[17] In this study, of total 90 items, the majority, that is, 74 (82%) had a good/acceptable level of DIF I (30–70%) with a mean DIF I of 55.32 ± 7.4 (mean ± SD), whereas seven (8%) were too difficult (DIF I <30%) and nine (10%) were too easy (DIF I >70%). Karelia et al.[18] reported a range of DIF I mean ± SD between 47.17 ± 19.79 and 58.8 ± 19.33, which supports our study findings. A similar study by Rao et al.[10] reported that 85% had an acceptable level of DIF I (30–70%) with a mean value of 50.16 ± 16.15, 5% had easy and 10% items were difficult. Another study by Patel et al.[19] showed 80% of items had an acceptable level of DIF I. A similar study showed that 70% of items had an acceptable range of DIF I, 20% were too easy and 10% were too difficult.[9] Mahjabeen et al. reported that 81% of MCQs were in an acceptable group.[20] Another study of item analysis of MCQs showed that 80%, 7% and 13% of MCQs were acceptable, too easy and too difficult, respectively.[21]

In our study, 72 (80%) items had acceptable to excellent discriminating power (DI > 0.20) and 18 (20%) had poor discriminating power (DI ≤ 0.20). The majority of items 61 (68%) had good to excellent DI (≥0.25) which suggests that these items are good to excellent in discriminating or differentiating the ability of the students with higher scores and those with lower scores. The DI values of the present study are comparable with studies on item analysis by Date et al. and others as similar findings with 78% items having acceptable to excellent discriminating power (DI > 0.20), 65–70% having good to excellent (DI ≥ 0.25) and 24% having poor discriminating power (DI ≤ 0.20)[9,22] were reported. Item DIs should be interpreted in the context of the type of test which is being analysed. The values of DI tend to be lower for tests measuring a wide range of content areas than for more homogeneous tests. Items with low DIs are often ambiguously worded and

Table 2 – Classification of questions according to the difficulty index (DIF I).

DIF I (P)	Interpretation	Items (%)	Difficulty index (mean ± SD)
<30	Too difficult	7 (8)	22.65 ± 2.8
30–70	Good/acceptable	74 (82)	55.32 ± 7.4
50–60	Excellent/ideal	34 (38)	55.31 ± 3.0
>70	Too easy	9 (10)	76.96 ± 6.3

DIF I, difficulty index.

Table 3 – Classification of questions by the discrimination index (DI).

Discrimination index (DI)	Interpretation	Items (%)	Discrimination index (mean ± SD)
≤0.20	Poor	18 (20)	0.16 ± 0.03
0.21–0.24	Acceptable	11 (12)	0.24 ± 0.03
0.25–0.34	Good	32 (36)	0.31 ± 0.03
≥0.35	Excellent	29 (32)	0.45 ± 0.06

Table 4 – Distractor analysis and distractor fffectiveness (DE) of MCQ test items.

Parameter	Number (%)
Number of items	90
Total distractors	270
Functional distractors (FDs)	198 (73)
Non-functional distractors (NFDs)	72 (27)
No of items with 3 NFDs/0FD (DE = 0%, poor)	2 (2)
No of items with 2 NFDs/1FD (DE = 33.3%, moderate	8 (9)
No of items with 1 NFD/2FDs (DE = 66.6%, good	50 (56)
No of items with 0 NFD/3FDs (DE = 100%, excellent)	30 (33)

DE, distractor effectiveness; MCQ, multiple choice question.

Table 5 – Mean values of various parameters of MCQ items used in item analysis.

Item analysis parameters	Difficulty index	Discrimination index	Distractor effectiveness
Mean ± SD	54.95 ± 13.4	0.31 ± 0.12	32.35 ± 31.3

MCQ, multiple choice question.

should be reexamined. Higher value of the DI indicates higher efficiency of an MCQ item.[18]

The present study shows weak correlation between the DIF I and DI ($r = 0.140$, $p < .0001$), indicating that with increasing the DIF I, the ability to discriminate between the high and low scorers decreased and fewer easy questions were used in MCQ tests. The weak correlation between the DIF I and DI also indicates that relationship between the DIF I and DI of 90 MCQ items is dome shaped rather than linear. Several studies also showed poor correlation between the DIF I and DI.[3,8]

Our study showed 73% were FDs and 27% were NFDs. Of all items, 2% items had poor DE (DE = 0%), 9% had moderate DE (DE = 33.3%), 56% had good DE (DE = 66.6%) and remaining 33% had excellent DE (DE = 100%). Our study showed that, in 33% MCQs, all three wrong options fully distracted the student's attention. Our findings are in agreement with the study by Date et al who reported that 70% items were FDs, 30% items were NFDs and other studies showed similar findings as 76–83% items were FDs and 17–24% items were NFDs.[9,22] Another study by Mahjabeen et al showed that 72% items were FDs and only 28% items were NFDs.[20] Framing possible distractors and reducing the NFDs is an important aspect for good quality MCQs. More NFD makes MCQ easy and vice versa with more functioning distractors making it difficult.[23] A non-functional distractor can reduce the discrimination power of a MCQ. More training and effort are required for construction of possible options of MCQ items to improve the validity and reliability of the tests. The comparison of various item analysis indices between various studies is shown in 'Table 6'.

For an ideal MCQ, the level of the DIF I should have a good/acceptable level of DIF I (30–70%) with high DI (≥0.25) and 100% DE.[24] The present study showed that a total 20% MCQs fulfilled all the three criteria of an ideal MCQ. Other studies reported that 10–15% MCQs satisfied all the three criteria of ideal MCQs.[24] Our study showed mean and standard deviation for DIF I (good), DI (excellent) and DE (moderate) which suggested that the majority of MCQs were in the category of good MCQs. Our results are comparable to a study by Ingale et al.[21] that analysed 30 MCQs.

In this study, the results showed that the overall validity of the test was good with a Cronbach alpha of 0.85, KR 20 = 0.71 and SEM 1.22. Based on the assumption that any test score contains an error, SEM is used to estimate a band or interval within which a person's true score would fall, that is, the score if there were no error of measurement. The smaller the SEM,

Table 6 – Comparison of various item analysis indices between studies in percentages of items.

Indices	Rao et al.[10] (items in %)	Patel et al.[19] (items in %)	Mahjabeen et al.[20] (items in %)	Date et al.[9] (items in %)	Gomboo et al., Christian et al. [3,8] (items in %)	Our study (items in %)
DIF I 30–70	85	80	81	70	–	82
DI > 0.20	85	–	82	78	–	80
DE (functional distractors)	95	–	72	70	–	73
Correlation between the DIF I and DI	Positive	–	–	–	Weak correlation	Weak correlation

DE, distractor effectiveness; MCQ, multiple choice question; DIF I, difficulty index; DI, discrimination index.

the narrower is the interval. Narrow intervals are more precise, containing less error, than larger intervals.[11]

Conclusion and recommendations

Item analysis is a valuable procedure which should be regularly performed after the assessment providing information regarding the reliability and validity of an item/test by calculating the DIF I, DI and DE and their interrelationship. An ideal single best response item (MCQ) with four options will be the one which has average difficulty (DIF I 30–70%), high discrimination (DI $\geq$ 0.25) and maximum DE (100%) with three FDs.

Items analysed in this study were neither too easy nor too difficult which is acceptable. Therefore, items were acceptably difficult and were good at differentiating students with higher and lower abilities. The values of the DIF I and DI indicated that not too easy and too difficult questions were part of the examination.

Our study helped teachers identify good and ideal MCQs which can be part of the question bank for future and those MCQs which needed revision. We recommend that item analysis should be performed for all MCQ-based assessments to determine validity and reliability of the assessment.

Disclosure of competing interest

The authors have none to declare.

REFERENCES

1. Davis MH, Harden RM. Competency-based assessment: making it a reality. *Med Teach*. 2003;25(6):565–568.
2. Epstein RM. Assessment in medical education. *N Engl J Med*. 2007:387–396.
3. Gomboo A, Gombo B, Munkhgerel T, Nyamjav S, Badamdorj O. Item analysis of multiple choice questions in medical licensing examination. *Cent Asian J Med Sci*. 2019;5(2):141–148.
4. Singh T. Student assessment: moving over to programmatic assessment. *Int J Appl basic Med Res*. 2016;6(3):149–150.
5. McCoubrie P. Improving the fairness of multiple-choice questions: a literature review. *Med Teach*. 2004;26(8):709–712.
6. Kiat JE, Ong AR, Ganesan A. The influence of distractor strength and response order on MCQ responding. *Educ Psychol*. 2018;38(3):368–380.
7. Kelley TL. The selection of upper and lower groups for the validation of test items. *J Educ Psychol*. 1939;30(1):17–24.
8. Christian DS, Prajapati AC, Rana BM, et al. Evaluation of multiple choice questions using item analysis tool: a study from a medical institute of Ahmedabad, Gujarat. *Int J Commun Med Publ Health*. 2016;4(6):1876–1881.
9. Date AP, Borkar AS, Badwaik, et al. Item analysis as tool to validate multiple choice question bank in pharmacology. *Int J Basic Clin Pharmacol*. 2019;8:1999–2003.
10. Rao C, Kishan Prasad HL, Sajitha K, Permi HSJ. Item analysis of multiple choice questions: assessing an assessment tool in medical students. *Int J Educ Psychol Res*. 2016;2:201–204.
11. McManus IC. The misinterpretation of the standard error of measurement in medical education: a primer on the problems, pitfalls and peculiarities of the three different standard errors of measurement. *Med Teach*. 2012;34(7):569–576.
12. Cronbach LJ. Coefficient alpha and the internal structure of tests. *Psychometrika*. 1951;16(3):297–334.
13. Kuder GF, Richardson MW. The theory of the estimation of test reliability. *Psychometrika*. 1937;2(3):151–160.
14. Feldt LS. The approximate sampling distribution of Kuder-Richardson reliability coefficient twenty. *Psychometrika*. 1965;30(3):357–370.
15. Cain MK, Zhang Z, Yuan K-H. Univariate and multivariate skewness and kurtosis for measuring nonnormality: prevalence, influence and estimation. *Behav Res Methods*. 2017;49(5):1716–1735.
16. Pande SS, Pande SR, Parate VR, Nikam AP, Agrekar SH. Correlation between difficulty and discrimination indices of MCQ's in formative exam in physiology. *South East Asian J Med Educ*. 2013;7:45–50.
17. Mishra P, Pandey CM, Singh U, Gupta A, Sahu C, Keshri A. Descriptive statistics and normality tests for statistical data. *Ann Card Anaesth*. 2019;22:67–72.
18. Karelia BN, Pillai AVB. 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 M.B.B.S students. *IeJSME*. 2013;7:41–46.
19. Patel KA, Mahajan NR. Itemized analysis of questions of multiple choice question (MCQ) exam. *Int J Sci Res*. 2013;2(2):279.
20. Mahjabeen W, Alam S, Hassan U, et al. Difficulty index, discrimination index and distractor efficiency in multiple choice questions. *Ann Pak Inst Med Sci*. 2017;310–5.
21. Ingale AS, Giri PADM. Study on item and test analysis of multiple choice questions amongst undergraduate medical students. *Int J Commun Med Publ Health*. 2017;4(5):1562–1565.
22. Patil VC, Patil HV. Item analysis of medicine multiple choice questions (MCQs) for undergraduate (3rd year MBBS) students. *Res J Pharmaceut Biol Chem Sci*. 2015;6:1242–1251.
23. Tarrant M, Ware J, Mohammed AM. An assessment of functioning and non-functioning distractors in multiple-choice questions: a descriptive analysis. *BMC Med Educ*. 2009;9(1):40.
24. Ananthkrishnan N. Item analysis-validation and banking of MCQs. In: Ananthkrishnan N, Sethuraman KR, Kumar S, eds. *Medical Education Principles and Practice Pondichery: JIPMER*. 2000:131–137.