Res Dev Med Educ, 2022, 11, 24
doi: 10.34172/rdme.2022.024
https://rdme.tbzmed.ac.ir

Original Research

Assessment of multiple-choice questions by item analysis for medical students' examinations

Marzieh Nojomi[1], Maryam Mahmoudi[2*]

[1]Preventive Medicine and Public Health Research Center, Psychosocial Health Research Institute, Department of Community and Family Medicine, Iran University of Medical Sciences, Tehran, Iran
[2]Department of Community and Family Medicine, Iran University of Medical Sciences, Tehran, Iran

Article info

Article History:
Received: May 24, 2022
Accepted: July 4, 2022
epublished: December 31, 2021

Keywords:
Difficulty index, Discrimination index, Item analysis, Multiple-choice questions

Abstract

Background: Multiple-choice questions (MCQs) are a common assessment method, and it is crucial to design them carefully. Therefore, this study aimed to determine the item analysis of MCQ exams in clerkship tests for general medicine students.

Methods: Following a cross-sectional study, a total of 1202 MCQs designed for fourth-year clerkship medical students in the second semester of 2019 were analyzed. Difficulty and discrimination indices of student scores and taxonomy levels were then computed. Furthermore, the prepared standard structural Millman checklist was utilized.

Results: Of the 1202 MCQs, according to difficulty indices, most questions (666) were considered acceptable (55.39%). In terms of the discrimination index (DI), 530 (44.09%) questions had an average discrimination coefficient. Additionally, 215 (17.88%) had a negative or poor DI and required revision or elimination from the tests bank. Of the 1202 MCQs, 669 (50.7 %) were designed at a lower cognitive level (taxonomy I), 174 (14.5 %) belonged to taxonomy II, and 419 (34.8%) of the questions had taxonomy III. Moreover, according to the structural flaws of the Millman checklist, the most common structural flaw was a lack of negative choices for Stems 1127 (93.8 %), while vertical options 376 (31.3%) were the least common.

Conclusion: Based on the results, it is recommended that easy questions and negative/poor DI of items, a high level of Bloom's taxonomy type I, and questions with unstructured flaws be reviewed and reconstructed to improve the quality of the question banks. Holding training courses on designing test questions could effectively improve the quality of the questions.

Introduction

Evaluation is a crucial element of the education process, and its results reflect both weaknesses and strengths of educational outcomes. The evaluation also enables the development of positive pattern means and improves defects to achieve a positive transformation within an educational system.[1,2]

Evaluation of student performance allows instructors to understand better the effectiveness of administered teaching techniques against specified learning objectives, which helps institutionalize practical teaching techniques or revise ineffective techniques in their pedagogy. Effective assessment not only improves students' motivation but also helps educators in competency assessments.[3]

Multiple choice questions (MCQs) are an efficient method to evaluate basic abilities related to a disease and its signs and symptoms. In addition, MCQs can be used to develop basic diagnostic and treatment strategies. Several advantages have been reported for MCQs in that they allow testing both for knowledge recall and higher cognitive skills as well as lower limitations regarding content specificity, which is based partly on the ease of administering and scoring MCQ tests.[4]

There are numerous examples of using MCQs to evaluate students in various educational streams for objectivity and broad coverage within a shorter duration. MCQs are primarily administered as holistic assessments at the end of academic sessions. In addition, they are widely used to provide feedback on teachers' performance. 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 students' responses to each item, used to assess the quality of those items and evaluate their overall

***Corresponding author:** Maryam Mahmoudi, Email: Mahmoudi.mar@iums.ac.ir

performance to benefit both students and teachers.[5]

According to many studies conducted to assess MCQ exams at universities of medical sciences, it was found that they were most often designed on an unstandardized basis, and it is essential that a curriculum include an appropriate assessment strategy. Therefore, the current study investigated the assessment of item analysis of MCQ exams in clerkship tests of general medicine students in the second semester of 2019 in the fourth year of their clerkship at the Iran University of Medical Sciences (IUMS) in 2019.

Materials and Methods

This cross-sectional study was conducted on a sample of 1202 multiple choice questions covering theoretical courses at the Medical School of the Iran University of Medical Sciences during the fourth year of clerkship of medical students in the second semester of 2019. Their exam scores were obtained from the exams office of the medical school, Deputy of General Medical Education. Since their scores were anonymous, their data could not be linked to personal records. In this study, the inclusion criteria were all MCQ exams in the period, which will be mentioned later, and the exclusion criteria were MCQs deleted according to unstructured questions. The Vice Dean for Academic Affairs at IUMS confirmed this study and allowed the research team to have access to the examination data. Participants' identities were kept anonymous and confidential. This study did not involve human subjects research.

Item Difficulty Index

Item difficulty index (I DIF) is one of the most frequently administered statistics in assessment. It was developed to measure the proportion of examinees answering the item correctly, called the *P* value. Regarding the proportion of examinees who answered an item correctly, the *P* value might more appropriately be called the "item easiness index", rather than the "item difficulty index". It ranges from 0.0 to 1.0, with a higher value indicating a more significant proportion of participants who responded to the item correctly, thus indicating the item is more accessible.

Item Discrimination Index

The item discrimination index (IDI) is developed to evaluate how well an item can differentiate knowledgeable examinees from others, which can also be called masters and non-masters. Various methods exist for achieving item discrimination; however, *point-by serial correlation* is the most commonly used one. IDI investigates the association between an examinee's performance on the provided item (correct or incorrect) and the overall test. For a highly discriminating item, examinees who responded to the item correctly also did well on the overall test, while, in general, those who responded incorrectly also tended to do poorly on the overall test.

The IDI ranges from -1.0 to 1.01; values lower than 0.0 indicate a problem. When an item is negatively discriminated, it can be argued that the most knowledgeable examinees have provided an incorrect answer, and the least knowledgeable examinees have provided a correct answer to the item. Therefore, such items may indicate that the item measures something other than what the other parts of the test measure. More frequently, it may indicate that the item is miskeyed.[6]

In this study, post-valuation was evaluated using item analysis. Scores obtained by all students were arranged in order of merit. The top and bottom 33% were categorized as high and low achievers, respectively. Item analysis was performed according to the following parameters:

- Difficulty index (DIF I) or *P* value using the formula $P = H + L/N \times 100$

H = Number of students who answered the item correctly in the high achieving group

L = Number of students who answered the item correctly in the low achieving group

N = Total number of students in the two groups (including nonresponders)

- DI or d value using the formula, $d = H - L \times 2/N$

Where the symbols H, L, and *N* represent the same values as mentioned earlier

- Distractor effectiveness (DE) or functionality.

Interpretation

Difficulty index (P) if
P < 30% Difficult
P = 30–70% Acceptable
P > 70% Easy
Discrimination index (D) if
D = Negative: Defective item/wrong key
D = 0–0.19: Poor discrimination
D between 0.2 and 0.29: Acceptable discrimination
D between 0.3 and 0.39: Good discrimination
D > 0.4: Excellent discrimination.

The three categories of difficulty index were set as follows: less than 0.30, 0.30 to 0.70, and above 0.70. The DI was classified into five categories: zero, 0.01 to 0.20, 0.21 to 0.40, 0.41 to 0.80, and over 0.81, respectively.[7]

The high and low groups were comprised of 27% of students in each group. The DIF I was computed using the formula H + L/N × 100, where H and L indicate the correct responses in the high and the low groups, and N is the total number of examinees in both study groups. Values between 30% and 70% are considered acceptable, with lower values reflecting greater difficulty and vice versa. DI was computed using the formula DI = H-LX2/N, which expresses the power of the item to differentiate between the high and the low achievers, ranging from 0 to 1.

In this regard, higher values indicate more discrimination power. An item with a DI > 0.35 is considered 'excellent', between 0.25 and 0.34 as 'good', between 0.15 and 0.24 as ‘marginal’, and < 0.15 as ‘poor’. A negative DI (less than zero) indicates that in this case, low achievers answered the item correctly more often than high achievers .[8]

Bloom’s taxonomy levels

All items were categorized into three cognitive levels according to a modified Bloom’s taxonomy: recall, application, and analysis. The first category is comprised of knowledge and cognitive comprehension levels. Answering these items merely requires knowledge or a basic understanding of a fact. Students may be asked to either identify or define factual information. The second category of items may require students to apply knowledge, while “analysis” items require students to analyze, synthesize, or evaluate the obtained data. Such items also require applying knowledge, including calculations or interpretation of data.

Furthermore, items written at the “application” level may ask a student to solve a problem or categorize provided data. The third category of items (“analysis”) may require students to interpret several facts to solve a multistep problem. It is recommended that the student first assess a problem, and then take the necessary implicit steps to come up with a solution.[9]

The Millman checklist

This checklist was used to evaluate the structural quality of these questions based on Millman’s principles for designing stem and choices of questions. This checklist assesses structural errors using 12 items with Yes or No responses. Thereafter, the overall score of each person determines to what extent the designed MCQ questions are standardized for the considered item. It is apparent that a question lacking at least one of the errors mentioned above in the 12 items is viewed as having structural errors, and a question without any structural errors is considered free of structural errors.[10]

Data were analyzed using Microsoft Excel 2010 and SPSS version 23.

Results

In this study, 1202 MCQs obtained from the exam office of the medical school of the deputy of general medical education were analyzed. All students answered their MCQs thoroughly, with no negative scores. The general surgery course had the most MCQs, consisting of 100 questions (8.3%), and the urology course had the minimum number of 30 questions (2.4%) (Table 1).

The DIF I revealed that of 1202 MCQs, 666 (55.39%) had good or acceptable levels of difficulty, whereas 237 (19.71%) of them were highly tricky and 299 (24.9%) were too easy (DIF I > 70%) (Table 2).

Of 1202 MCQs, 53 (4.41%) had a negative DI, 162 (13.48%) items had a poor DI, and 530 (44.09%) items showed an acceptable DI, and 457 (38.02%) items had a good DI, as shown in Table 3.

Table 1. Distribution of courses in MCQs tests in clerkship medical students in the second semester in Iran University Of medical sciences, 2019

Topics	Number (%)
Gynacology1	55 (4.6)
General surgeory1	100 (8.3)
Epidemiology	64 (5.3)
Psychiatry1	70 (5.8)
Orthopedics1	60 (5)
Urology1	30 (2.5)
Neurosurgery 1	40 (3.3)
Emergency Medicine	50 (4.2)
Toxicology	52 (4.3)
Pediatrics 1	89 (7.4)
Neurosurgery 2	40 (3.3)
Infectious diseases medicine	76 (6.3)
Urology 2	30 (2.5)
Orthopedics2	59 (4.9)
Psychiatry2	65 (7.7)
General surgeory2	100 (8.3)
Neurology	74 (6.2)
Gynacology2	55 (4.6)
Pediatrics 2	93 (7.7)
Total	1202 (100)

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

DIF I (P)	Interpretation	Items (%)	Difficulty index (mean ± SD)
< 30	Difficult	237(19.71)	12.47 ± 7.11
30-70	Good/Acceptable	666 (55.39)	35.05 ± 13.03
> 70	Too easy	299 (24.9)	15.73 ± 0.68

Table 3. Classification of questions according to the discrimination (DIS I)

Discrimination index (DI)	Interpretation	Items (%)	Discrimination index (mean ± SD)
< 0	Negative	53 (4.41)	2.78 ± 2.93
0-0.1	Poor	162 (13.47)	8.42 ± 3.96
0.1-0.3	Acceptable	530(44.09)	29.63 ± 12.01
> 0.3	Good	457(38.02)	24.15 ± 7.82

In terms of taxonomy, of a total of 1202 MCQs, 609 (50.7%) were in taxonomy level I, 174 (14.5%) were in taxonomy level II, and 419 (34.8%) were in taxonomy level III. In Table 4, the number and percentage of all items are compared at the taxonomy level in all categories of the tests.

In terms of compliance with Millman’s structural principles for MCQs, a test indicated that the highest frequency of structural format was observed in those choices that lacked negative options for stems (1127; 93.8%). The lowest percentage of compliance with Millman’s principles was also found to be associated with the index of verticality of options (376; 31.3%). The compliance with each of

Table 4. Distribution of taxonomy level of 1202 tests of clerkship for medical students in the second semester in Iran University of Medical Sciences, 2019

Topics	Taxonomy I Number (%)	Taxonomy II Number (%)	Taxonomy III Number (%)	Total
Gynacology1	30 (54.5)	6 (10)	19 (34.5)	55
General surgeory1	40 (40)	80 (80)	52 (52)	100
Epidemiology	55 (86)	6 (9.5)	3 (4.5)	64
Psychiatry1	43 (61)	13(19)	14 (20)	70
Orthopedics	38 (63.3)	3 (5)	19 (31.7)	60
Urology 1	14 (46.7)	4 (13.3)	12(40)	30
Neurosurgery1	28 (70)	5 (12.5)	7 (17.5)	40
Emergency Medicine	14 (28)	10 (20)	26 (52)	50
Toxicology	45 (86.5)	3 (5.7)	4 (8)	52
Pediatrics 1	32 (36)	20 (33)	36 (41)	89
Neurosurgery 2	21 (52.5)	8 (20)	11 (27.5)	40
Infectious diseases medicine	43 (56.5)	11 (14.5)	22 (29)	76
Urology 2	14 (46.5)	3 (10)	13 (43.3)	30
Orthopedics2	37(6.1)	7(12)	16 (27)	59
Psychiatry2	43(66)	(9)	16 (25)	65
General surgeory2	32(32)	17(17)	51 (51)	100
Neurology	22(29.72)	14 (18.92)	38 (51.36)	74
Gynacology2	18(32.74)	8(14.54)	29 (52.72)	55
Pediatrics 2	41(44.08)	21 (22.59)	31 (33.33)	93
Total	609 (50.7)	174 (14.5)	419 (34.8)	1202

Table 5. Compliance with each of Millman's principles according to indices in 1202 MCQs tests of clerkship medical students in the second semester in Iran University of Medical Sciences, 2019

Subjects	Number	Percent
Including the stem the least amount of necessary information	827	68.8
Clearness of stem	739	61.5
Lack of negative options for stem	1127	93.8
Specific option	739.	61.5
Lack of contrastive option	739	61.5
Positive words in the stem	526	43.8
Writing structure of the stem	450	37.5
Lack of duplicate option	901	75
The spelling of stem and option	1051	87.5
Vertically of option	376	31.3
Positively of stem and option	977	81.3
No use of " all or none of the above" phrase in options	826	68.8
Using positive vocabulary used in the stem of the question or if they are negative, determining the negative vocabulary	526	43.8

Millman's principles is shown in Table 5.

Discussion

The item analysis of MCQs is a crucial tool used to identify both the validity and reliability of items.[11] In this study, we have performed an item analysis of 1202 MCQs designed for clerkship medical students to develop a quality MCQ bank. To fulfill this, of 1202 items, the majority, 666 (55.39%) MCQs, had ideal difficulty levels, and 299 (24.9%) MCQs were too easy. In Mehta and Mokhasi's study, they found that a difficulty index of 31 (62%) items was in the acceptable range (P value 30-70%), 16 (32%) items were too easy (P value > 70%), and 3 (6%) items were too tricky (P value < 30%).[7]

It is noteworthy that very easy questions are appropriate for 'warm up' items (initial items) or could be removed completely. In the same vein, an evaluation should be performed for difficult questions to check for any confusion due to language, areas of controversy, or even an incorrect key.[12]

Patel and Mahajan[13] investigated 150 bachelor of medicine, bachelor of surgery (MBBS) students for an MCQ test with 50 items and reported that 10 (20%) of the items were in the unacceptable range (P < 30% or P > 70%), while 40 (80%) were categorized as acceptable (P = 30–70%). Additionally, Mehta and Mokhasi[7] performed an item analysis on 100 MBBS students for the MCQ test with 50 items related to anatomy. According to their findings, the mean DIF I was 63.06 ± 18.95. They also reported that the DIF I of 31 (62%) items was in the acceptable range (P = 30–70%), while 16 (32%) and 3 (6%) items were too easy (P > 70%) and too tricky (P < 30%), respectively. Kolte[14] also reported a mean DIF I of 57.92 ± 19.58. 12. The current study had 237 difficult items, while the number of too easy items was 299, which should be revised and kept for subsequent use along with items within the acceptable range.

Regarding discrimination power, 987 (82.11%) items were in the acceptable range, indicating a good to excellent discrimination/differentiation ability of examinees with higher scores from those with lower scores. Additionally, 53 (4.41%) had negative discriminating power.

It means that most examinees from the low achiever category could provide correct answers compared to those from the high achiever category. This negative value could be attributed to either ambiguity of the item or an answer key that was wrongly marked or coded. Moreover, when interpreting the DI of an item, special attention should be paid to the context of the type of test. Items with a wide spectrum of content areas often have lower DI values than more homogeneous tests. Accordingly, items with low DIs mostly have ambiguous wording. Mehta and Mokhasi,[7] in their study, found that the mean of DI was 0.33 ± 0.18. Items with DI > 0.35 were 26 (52%), DI between 0.2 and 0.34 were 9 (18%), and DI 0.35, 42%, with DI between 0.2 and 0.34 as well as 18% with DI < 0.20 were 15(30%).[13] There were no items with negative DIs. However, some studies reported negative DI for some items as in a study one item[6] and in another study two items[12] with negative DI were found. It was observed that items with negative DIs caused the validity of the test to be reduced, indicating the necessity of their removal.

According to the findings, in terms of taxonomy, most levels were in taxonomy level I, 174 (14.5%) were in taxonomy level II, and 419 (34.8%) were in taxonomy level III. In another study, Baig et al[15] reported that 76% of MCQs were about recalling isolated facts, and 24% were about data interpretation skills. They also reported no single MCQ assessing the higher cognitive areas of both administration and analysis. Kowash et al[16] also stated that the majority of MCQs (81.1%) required information recall (level one), while the rest (i.e., 18.9%) needed understanding and interpretation of data (level two).[16] Nevertheless, they reported no higher-order thinking items (third level) to evaluate the application of knowledge. Accordingly, it can be attributed to easier construction of MCQs at the recall level compared to problem-solving MCQ that requires both experience and training, for which higher-level Bloom's taxonomy items can well discriminate among higher- and lower-performing examinees.[16, 17]

Baig et al[15] evaluated 150 undergraduate pharmacology examination MCQs and reported that most items were at cognitive level one (76%), followed by level two (24%). They reported no items at level three. Our research suggests the necessity of enhancing the quality of assessment tools; measuring low cognitive levels can result in decreased validity of the test and compel students to follow surface learning methods that are not appropriate for long-term learning. According to Millman's list, the current study found that of 1202 MCQs, the lowest number of structural flaws were the choices located vertically (31.3%), and the most common structural flaw was avoiding the use of negative choices for negative questions (93.8%). The majority of structural problems were related to the linear order of items (68.7%), followed by the absence of significant data in the item stem (32.2 %), a duplicated option (25%), and heterogeneity of items within terms of length and vocabulary structure (37.5%). Implementing measures to enhance awareness and skills, paying attention to faculty members in this field, and offering suitable training courses can improve the design of MCQs. According to compliance with Millman checklist, there were no structural problems, and educational groups can be organized to help improve such exams by regular assessment of MCQs and feedback. In addition, a review process is crucial in improving the quality of items.

The most critical limitation of the current study is the representativeness of the results. An item analysis of final clerkship exams was merely evaluated. Therefore, the result is not generalizable to students at other stages of medical education courses.

Conclusion

Based on the results, the values of the DIF I and DI indicate that despite the majority of the MCQs' DIF I and DI being in the acceptable range, it is worth noting that there is a need to recommend discarding or revising (reviewing or reconstructing) easy items and negative/poor discrimination indexes by holding workshops for faculty members to improve question banks.

Due to the distribution of taxonomy levels, and the high level of type I items, the cognitive level of the test items needs improving through the use of a test blueprint. Furthermore, it is crucial to encourage and train faculty members to construct MCQs to achieve higher cognitive levels. Reviewing and revising unstructured flaws is necessary according to the Millman checklist.

It is recommended that the item analysis be performed not only for all MCQ-based assessments but also for other types of assessment tools such as patient management problems and key-feature questions. Moreover, it is necessary to analyze Modified Essay Questions exams regularly to determine their validity and reliability. Therefore, comparative analytic methods in other assessment tools of clinical exams for medical students could assess their knowledge, skills, and performance. It is hoped that by executing educational interventions and programs, a standard design of questions will be established based on approved checklist criteria by concentrating on reinforcing merits, removing demerits of previous tests, and giving appropriate feedback to the masters. Standard and structural error-free questions should also be designed for other colleges and disciplines. Psychometric analyses should be conducted for all assessment types, as well as developing a blueprint test to ensure validity. A structured faculty development program is recommended to develop assessment tools. It is also important that similar studies be conducted during other semesters of a medical students' education.

Acknowledgments

The authors would like to extend their appreciation to the Deputy of General Medical Education of Iran University Medical sciences and the faculty members who collaborated in implementing the study.

Author Contributions

All authors contributed to the design of the study. MN and MM have conducted the study. MM wrote the manuscript and MN and MM provided detailed comments on the manuscript.

Funding

The Iran University of Medical Sciences funded this research under Grant Number 16400.

Ethical Approval

This study was approved by the Research Ethics Committee, Vice-Chancellor of Research Affairs, Iran University of Medical Sciences, Tehran, Iran (Ethics code: R.IUMS.REC.1399.542).

Competing Interests

The authors declare no conflicts of interests.

References

1. Grauer GF, Forrester SD, Shuman C, Sanderson MW. Comparison of student performance after lecture-based and

case-based/problem-based teaching in a large group. J Vet Med Educ. 2008;35(2):310-7. doi: 10.3138/jvme.35.2.310.

2. Smith-Strøm H, Nortvedt MW. Evaluation of evidence-based methods used to teach nursing students to critically appraise evidence. J Nurs Educ. 2008;47(8):372-5. doi: 10.3928/01484834-20080801-08.
3. Newble DI, Jaeger K. The effect of assessments and examinations on the learning of medical students. Med Educ. 1983;17(3):165-71. doi: 10.1111/j.1365-2923.1983.tb00657.x.
4. Fowell SL, Bligh JG. Recent developments in assessing medical students. Postgrad Med J. 1998;74(867):18-24. doi: 10.1136/pgmj.74.867.18.
5. 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 Educ Psychol Res. 2016;2(4):201-4. doi: 10.4103/2395-2296.189670.
6. D'Sa JL, Visbal-Dionaldo ML. Analysis of multiple choice questions: item difficulty, discrimination index and distractor efficiency. Int J Nurs Educ. 2017;9(3):109. doi: 10.5958/0974-9357.2017.00079.4.
7. 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.
8. Guilbert J. Educational Handbook for Health Personnel. Geneva: WHO; 1981.
9. Tiemeier AM, Stacy ZA, Burke JM. Using multiple choice questions written at various Bloom's Taxonomy levels to evaluate student performance across a therapeutics sequence. Innov Pharm. 2011;2(2):41. doi: 10.24926/iip.v2i2.224.
10. Lockyer J, Carraccio C, Chan MK, Hart D, Smee S, Touchie C, et al. Core principles of assessment in competency-based medical education. Med Teach. 2017;39(6):609-16. doi: 10.1080/0142159x.2017.1315082.
11. 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-S9. doi: 10.1016/j.mjafi.2020.11.007.
12. Hingorjo MR, Jaleel F. Analysis of one-best MCQs: the difficulty index, discrimination index and distractor efficiency. J Pak Med Assoc. 2012;62(2):142-7.
13. Patel KA, Mahajan NR. Itemized analysis of questions of multiple choice question (MCQ) exam. Int J Sci Res. 2013;2(2):279.
14. Kolte V. Item analysis of multiple choice questions in physiology examination. Indian J Basic Appl Med Res. 2015;4(4):320-6.
15. Baig M, Ali SK, Ali S, Huda N. Evaluation of multiple choice and short essay question items in basic medical sciences. Pak J Med Sci. 2014;30(1):3-6. doi: 10.12669/pjms.301.4458.
16. Kowash M, Hussein I, Al Halabi M. Evaluating the quality of multiple choice question in paediatric dentistry postgraduate examinations. Sultan Qaboos Univ Med J. 2019;19(2):e135-e41. doi: 10.18295/squmj.2019.19.02.009.
17. Tarrant M, Ware J. A framework for improving the quality of multiple-choice assessments. Nurse Educ. 2012;37(3):98-104. doi: 10.1097/NNE.0b013e31825041d0.