J Med Edu. 2021 May; 20(2):e116834. doi: 10.5812/jme.116834.
Published online 2021 August 30. **Research Article**



# Multiple Choice Questions with Different Numbers of Options in University Putra Malaysia Undergraduate Medical Program: A Comparative Analysis in 2017 and 2018

Siti Khadijah Adam [1, 2, *], Faridah Idris [1, 3], Puteri Shanaz Jahn Kassim [1, 4], Nor Fadhlina Zakaria [5] and Rafidah Hod [1, 2]

[1] Medical Education Research and Innovation Unit, Faculty of Medicine and Health Sciences, Universiti Putra Malaysia, Serdang, Selangor, Malaysia
[2] Department of Human Anatomy, Faculty of Medicine and Health Sciences, Universiti Putra Malaysia, Serdang, Selangor, Malaysia
[3] Department of Pathology, Faculty of Medicine and Health Sciences, Universiti Putra Malaysia, Serdang, Selangor, Malaysia
[4] Department of Family Medicine, Faculty of Medicine and Health Sciences, Universiti Putra Malaysia, Serdang, Selangor, Malaysia
[5] Department of Medicine, Faculty of Medicine and Health Sciences, Universiti Putra Malaysia, Serdang, Selangor, Malaysia

[*] *Corresponding author*: Medical Education Research and Innovation Unit, and Department of Human Anatomy, Faculty of Medicine and Health Sciences, Universiti Putra Malaysia, Serdang, Selangor, Malaysia. Email: sk.adam@upm.edu.my

**Received** 2021 June 06; **Revised** 2021 August 06; **Accepted** 2021 August 22.

## Abstract

**Background:** Multiple-choice questions (MCQs) are used for measuring the student's progress, and they should be analyzed properly to guarantee the item's appropriateness. The analysis usually determines three indices of an item; difficulty or passing index (PI), discrimination index (DI), and distractor efficiency (DE).
**Objectives:** This study was aimed to analyze the multiple-choice questions in the preclinical and clinical examinations with different numbers of options in medical program of Universiti Putra Malaysia.
**Methods:** This is a cross-sectional study. Forty multiple-choice questions with four options from the preclinical examination and 80 multiple-choice questions with five options from the clinical examination in 2017 and 2018 were analyzed using optical mark recognition machine and Ms. Excel. The parameters included PI, DI, and DE.
**Results:** The average difficulty level of multiple-choice questions for preclinical and clinical phase examinations were similar in 2017 and 2018 that were considered 'acceptable' and 'ideal' ranged from 0.55 to 0.60, respectively. The average DIs were similar in all examinations that were considered 'good' (ranged from 0.25 to 0.31) except in 2018 clinical phase examination that showed 'poor' items (DI = 0.20 $\pm$ 0.11). The questions for preclinical phase showed an increase in the number of 'excellent' and 'good' items in 2018 from 37.5% to 70.0%. There was an increase of 10.0% for preclinical phase, and 6.25% for clinical phase, in the number of items with no non-functioning distractors in 2018. Among all, preclinical multiple-choice questions in 2018 showed the highest mean of DE (71.67%).
**Conclusions:** Our findings suggested that there was an improvement in the questions from preclinical phase while more training on questions preparation and continuous feedback should be given to clinical phase teachers. A higher number of options did not affect the level of difficulty of a question; however, the discrimination power and distractors efficiency might differ.

*Keywords:* Item Analysis, Multiple-Choice Question, Difficulty Index, Discrimination Index, Distractor Efficiency

## 1. Background

The validity of an assessment refers to the evidence presented to support or to refute the meaning or interpretation assigned to the assessment data (1). Validity, therefore, is a degree to which the test measures what it is supposed to measure. This includes test item analysis that is usually done after the assessment has been completed to determine the candidate responses to individual test items, the quality of those items as well as the overall assessment. Difficulty index or passing index (PI), discrimination index (DI), and distractor efficiency (DE) of each item can be obtained from the analyses, which reflect the quality of the test items. The PI of an item is commonly defined as the percentage of students who answered the item correctly. The DI, on the other hand, is defined as the degree to which an item discriminates between students of high and low achievement. The DE is used to assess the credibility of the distractors in an item, whether they are able to distract students from selecting the right answer (2). Any distractor that is selected by less than 5% of the students is consid-

ered to be a non-functional distractor (NFD).

Item analyses assess the quality of individual test items and the test as a whole by looking at how students respond to them. The advantages of the analysis are to help identify faulty items (3, 4), identify the lower performers and their learning problems such as misconceptions as a guide for remedial actions to be done to students, and as importantly to increase teachers' skills to construct a high quality of test items (5). Test items that do not fulfill the well-designed item criterion can therefore be changed or eliminated, and a viable question bank can be developed (6, 7).

Doctor of Medicine (MD) curriculum in Universiti Putra Malaysia (UPM) consists of two phases; preclinical and clinical phases, which run for two and three years, respectively. Students will sit for a summative assessment at the end of each phase. Students must pass the preclinical examination to proceed to clinical phase. In the fifth year, they need to sit and pass the clinical phase examination before being awarded the degree of medicine. In the written examination, various assessment tools are used, such as multiple-choice questions (MCQs), short answer questions, and modified essay questions. In fact, MCQ is one of the most important well-established written assessment tools widely used for its distinct advantage and ability to evaluate over a broad coverage of concepts in less time. The scoring is also objective and reliable (8). The type A MCQ item (single best response) consists of a 'stem' or 'vignette', followed by a 'lead in' statement and several options. The correct answer in the list of options is called a 'key', and the incorrect options are called 'distractors'.

We analyzed the MCQs given in the preclinical and clinical phase examinations in the years 2017 and 2018 to determine the quality and validity of our test items. A comparison was made between PI, DI, and DE of the items between the examinations for both years as well as between the two phases of the examinations. Good quality items and revised items are going to be stored in the question bank, and faulty items shall be discarded based on the obtained findings.

## 2. Methods

This cross-sectional study was conducted in the Faculty of Medicine and Health Sciences, UPM, during the preclinical and clinical phase examinations in the years 2017 and 2018. In 2017, a total of 84 second-year MD students took the preclinical examination, while 128 final-year students took the clinical phase examination. Meanwhile, in 2018, there were 100 second-year students and 120 final-year students who took the end of preclinical and clinical examination, respectively. The 2-year preclinical phase examinations comprised of 40 MCQs with four options each, while the 2-year clinical phase examination comprised of 80 MCQs with five options each. Each correct response was awarded five marks, and there was no negative marking for the wrong answers. Pre-validation of the items was done by the vetting committee of the faculty.

### 2.1. Item Analysis

Post-validation was done automatically by item analysis using the optical mark recognition (OMR) machine (Scantron iNSIGHT 20 OMR scanner, Minnesota USA). The scores of all students in each examination paper were arranged in order of merit. The upper 27% students were considered 'top' students and lower 27% students as 'poor' students. Each item was analyzed for difficulty and discrimination indices according to Hassan and Hod (5) as well as Abdul Rahim (9):

(1) Difficulty index or Passing Index (PI), using the formula: $PI = (H + L)/N$

(2) Discrimination index (DI), using the formula: $DI = (H - L)/A$

H = number of 'top' students answering the item correctly; L = number of 'poor' students answering the item correctly; N = total number of students in the 'top' and 'poor' groups; A = number of students in 27% of total students.

The interpretation of PI and DI values is presented in Table 1.

(3) Distractor efficiency (DE): Non-functional distractor (NFD) is the option that was selected by less than 5% of students. Based on NFDs in an item, DE ranges from 0% to 100%. If an item with four options contained three or two or one or nil NFDs, then DE would be 0, 33.3%, 66.7%, and 100.0%, respectively. If an item with five options contained four or three or two or one or nil NFDs, then DE would be 0, 25%, 50%, 75%, and 100%, respectively.

## 3. Results

A total of 240 MCQs were analyzed, and the average PI and DI were determined. Overall, it was found that the difficulty level of the questions was similar in both preclinical and clinical phase examinations (Table 2). Interestingly, the values of average PI reduced in 2018, indicating a reduction of an increase of difficulty level of the questions for that particular year in both preclinical and clinical examinations. Nevertheless, the mean DI was similar in all examinations except for a low DI in the 2018 clinical phase examination.

For preclinical phase examination, the numbers of 'difficult' and 'very easy' items were similar in both years (Table 3). Half of the 40 MCQs were 'ideal' and 'acceptable'.

**Table 1.** Interpretation of Item Analysis [a]

| Difficulty @ Passing Index (PI) | Values | Discrimination Index (DI) | Values |
|---|---|---|---|
| **Difficult** | $\leq$ 0.3 | Excellent | $\geq$ 0.35 |
| **Ideal** | 0.31 - 0.59 | Good | 0.25 - 0.34 |
| **Acceptable** | 0.6 - 0.7 | Acceptable | 0.21 - 0.24 |
| **Very easy** | > 0.7 | Poor | $\leq$ 0.2 |

[a] Adapted from Ananthakrishnan (2000) (10)

**Table 2.** Average of Difficulty Index and Discrimination Index of Items in Preclinical Phase Examination (n = 40) and Clinical Phase Examination (n = 80) in 2017 and 2018 [a]

| | Preclinical Phase Examination | | Clinical Phase Examination | |
|---|---|---|---|---|
| | 2017 | 2018 | 2017 | 2018 |
| **Difficulty @ passing index (PI)** | 0.60 $\pm$ 0.24 (Acceptable) | 0.55 $\pm$ 0.23 (Ideal) | 0.60 $\pm$ 0.21 (Acceptable) | 0.56 $\pm$ 0.28 (Ideal) |
| **Discrimination index (DI)** | 0.25 $\pm$ 0.16 (Good) | 0.31 $\pm$ 0.16 (Good) | 0.25 $\pm$ 0.15 (Good) | 0.20 $\pm$ 0.11 (Poor) |

[a] Values are expressed as mean $\pm$ SD.

Clinical phase examination in 2017 had 54% 'ideal' and 'acceptable' items but the value reduced in 2018 to 34% due to the significant increase in the percentage of 'difficult' items in the year 2018.

There was an increase in the total percentages of 'excellent' and 'good' items in preclinical phase examination in 2018, from about 38% to 70% (Table 4). In clinical phase examination of 2017, half of the questions were 'excellent' and 'good'. However, the percentage reduced to 36% in 2018 due to the high percentage of 'poor' questions in the examination (53%). Additionally, there were five questions with zero DI in the paper; one with PI equals one and another with PI equals zero.

The total number of NFD was reduced in 2018 in both examination phases (Table 5). Both examinations in 2018 showed an increase in the number of items with no NFD as compared to the previous year. The number of items with no NFD in preclinical phase examination is higher than clinical phase examination for both years. Similarly, the overall mean DE was increased in 2018 with preclinical phase examination, showing the highest mean DE compared to the rest of examinations.

## 4. Discussion

The end-of-phase examination in UPM MD program is a high-stake summative assessment at the end of preclinical and clinical phase. For preclinical phase examination, the results determine whether the preclinical students are eligible to progress to clinical phase, while the final-year students need to pass the clinical phase examination to graduate. Therefore, valid assessment tools are needed to measure students' knowledge, skills, and attitude in the examination. One of the tools used to test the 'knows' and 'knows how' in Miller's pyramid is with MCQ (11). It is useful in measuring factual recall, but it can also test higher order of thinking skills such as application, analysis, synthesis, and evaluation of knowledge, which are important for medical graduates. Post-validation of test items using item analysis of PI, DI, and DE is a simple yet effective method to assess the validity of the test. In the present study, we analyzed the MCQs from both preclinical and clinical phase examinations taken by two different cohorts of students. Each MCQ in preclinical phase examination has four options, while clinical phase examination has five options.

Based on the findings, the mean PI in both examinations in both years was similar. This indicates that an increased number of options, five versus four options, does not have a significant impact on the difficulty level of the examination. This was supported by the previous study by Schneid et al. who found that there were no significant differences in the difficulty level among MCQs with three, four, or five options (12). On the contrary, Vegada et al. found a slight decreased in the difficulty level when reducing the options from five to four, and the items became much easier when reducing the options to only three (13). They concluded that the items became easier with fewer options due to the increased probability of random guessing to select the correct answer.

Preclinical phase examination showed a consistent level of item difficulty for both years. However, half of the questions were 'very easy' and 'difficult'. These questions seem to be unsuitable for assessing students in the high-stake examination as they were unable to discriminate be-

**Table 3.** Summary of the Number of Items in Preclinical Phase Examination (n = 40) and Clinical Phase Examination (n = 80) in 2017 and 2018 Based on Difficulty Index [a]

| Difficulty @ Passing Index (PI) | Preclinical Phase Examination | | Clinical Phase Examination | |
|---|---|---|---|---|
| | 2017 | 2018 | 2017 | 2018 |
| **Difficult** | 7 (17.50) | 7 (17.50) | 7 (8.75) | 21 (26.25) |
| **Ideal** | 12 (30.00) | 15 (37.50) | 29 (36.25) | 18 (22.50) |
| **Acceptable** | 8 (20.00) | 5 (12.50) | 14 (17.50) | 9 (11.25) |
| **Very easy** | 13 (32.50) | 13 (32.50) | 30 (37.50) | 32 (40.00) |

[a] Values are expressed as No. (%).

**Table 4.** Summary of the Number of Items in Preclinical Phase Examination (n = 40) and Clinical Phase Examination (n = 80) in 2017 and 2018 Based on Discrimination Index [a]

| Discrimination Index (DI) | Preclinical Phase Examination | | Clinical Phase Examination | |
|---|---|---|---|---|
| | 2017 | 2018 | 2017 | 2018 |
| **Excellent** | 11 (27.50) | 15 (37.50) | 23 (28.75) | 6 (7.50) |
| **Good** | 4 (10.00) | 13 (32.50) | 17 (21.25) | 23 (28.75) |
| **Acceptable** | 9 (22.50) | 3 (7.50) | 10 (12.50) | 9 (11.25) |
| **Poor** | 16 (40.00) | 9 (22.50) | 30 (37.50) | 42 (52.50) |

[a] Values are expressed as No. (%).

**Table 5.** Distractor Efficiency of Multiple-Choice Questions in Preclinical Phase Examination and Clinical Phase Examination in 2017 and 2018 [a]

| Parameter | Preclinical Phase Examination | | Clinical Phase Examination | |
|---|---|---|---|---|
| | 2017 | 2018 | 2017 | 2018 |
| **Number of items** | 40 | | 80 | |
| **Distractors** | 120 | | 320 | |
| **Functioning distractors** | 83 (56.67) | 86 (71.67) | 155 (48.44) | 177 (55.31) |
| **Non-functioning distractors** | 52 (43.33) | 34 (28.33) | 165 (51.56) | 143 (44.69) |
| **Items with No NFD (DE = 100%)** | 9 (22.50) | 13 (32.50) | 3 (3.75) | 8 (10.00) |
| **Items with 1 NFD (DE = 66.67%)** | 15 (37.50) | 20 (50.00) | 21 (26.25) | 24 (30.00) |
| **Items with 2 NFDs (DE = 33.33%)** | 11 (27.50) | 7 (17.50) | 28 (35.00) | 32 (40.00) |
| **Items with 3 NFDs (DE = 0)** | 5 (12.50) | 0 (0.00) | 24 (30.00) | 9 (11.25) |
| **Items with 4 NFDs (DE = 0)** | - | - | 4 (5.00) | 7 (8.75) |
| **Overall mean DE (%)** | 56.67 | 71.67 | 48.44 | 55.31 |

[a] Values are expressed as mean (SD) unless otherwise indicated.

tween the good and the weak students. Hence, these questions should be revised, by changing either the vignette or the options. All difficult questions should be reviewed for their language and grammar, ambiguity, and controversial statements (5). Some previous studies demonstrated their PI as a percentage, which reflects the percentage of correct answers to the total responses (6, 7, 14). An item with PI percentage between 30% to 70% is considered acceptable, i.e., not too easy and not too difficult (6, 14). Studies by Sim and Rasiah and Rao et al. showed a comparable mean PI similar to our present finding, ranging from 50 to 60% (7, 14). It is evident that proper vetting and training in constructing MCQs led to such desirable findings. It was also suggested that continuous training and feedback should be given to the teachers so that the number of 'difficult' and 'very easy' questions can be reduced in the future.

The mean DI for preclinical phase examination in both years and 2017 clinical phase examination ranging from 0.25 to 0.31, which were considered 'good'. An earlier study had shown that a comparable mean of DI proved that the quality of questions has been consistent over the years (5). Nevertheless, the mean DI in clinical phase examination reduced significantly in 2018. This may be due to the high number (66%) of 'very easy' and 'difficult' questions

in the examination. Consequently, about 53% of the questions were considered 'poor' and were not able to discriminate between the good and weak students. A test item ideally should be able to pick out the 'good' students from the 'poor' ones, in which more 'good' students are able to answer the item as compared to the 'poor' students (9). In the present study, some questions were found to have zero and negative DI. Zero DI means that the item was non-discriminating in which either all students were able to answer the item correctly, or an equal number 'good' and 'poor' students were able to answer correctly, or none of the 'good' and 'poor' students managed to correctly answer it. Negative DI indicates that more 'poor' students were able to answer the item correctly. There was also one question with zero DI and zero PI. These demonstrated an extremely low number of students who managed to answer it correctly, and none of them were the 'good' and 'poor' students. We speculate the reasons for these were due to ambiguous framing of the questions and poor preparation of students (7, 15, 16). Another question has zero DI and PI equals to one, demonstrating that all students managed to answer it correctly, probably because it was too easy. Too difficult and too easy questions may contribute to the 'poor' questions based on the dome-shaped correlation between PI and DI (6, 17). These questions were not useful and may reduce the validity of the test, therefore should be eliminated.

Preclinical phase examination in 2018 showed an increase in the number of 'excellent' and 'good' questions, and a decrease in 'poor' questions comparing to 2017. This proves that the preclinical lecturers have shown considerable improvement in constructing MCQs through continuous training and feedback. In contrast, less than 8% of the questions in 2018 clinical phase examination were considered 'excellent', and there was an increased number of 'poor' questions compared to 2017. Several possible reasons for this finding were identified. Some of the clinical lecturers were new and probably were unfamiliar with the test format, while some clinical lecturers had never attended any training on test item construction. There should be a thorough and several levels of vetting by peers who are content experts, and non-content experts are also needed before the questions are used in an examination. With this analysis, feedback should be given to all teachers for them to reflect upon and revise their questions accordingly.

The number of NFDs also affects the discrimination power of an item (14). In this study, more than 20% of the MCQs for preclinical phase had no NFDs in both years. In fact, there was one-third of the 2018 preclinical phase MCQs with three functioning distractors (DE = 100%). We identified that items with a higher number of options in clinical phase examination tend to have higher NFDs in both years. Less than 10% of the MCQs for clinical phase examination had 100% DE. This shows that it was probably difficult for the teachers to develop four equally plausible distractors. Preclinical phase examination in 2018 has no question with 0% DE, and it showed the highest overall mean DE as compared to other examinations. Preclinical lecturers have shown a significant improvement in developing MCQs with plausible distractors and avoiding NFDs between the years. The findings suggested that reducing the number of options may increase the credible of distractors in an item; however, it may reduce its difficulty level.

A meta-analysis by Rodriguez found that having three options in an item is adequate (18). Even though the difficulty level is lowered, but it is more discriminating and more reliable. This is supported by a more recent study which found that questions with even as low as three options would still produce good reliability and less laborious to construct (19). However, this means that students will have a high chance (only 1 in 3) of correctly answering the item with random guessing. Royal and Dorman highlighted that 3-option and 4-option MCQs had similar psychometric properties, which means the former is equally effective as the latter (20). Therefore, the traditional MCQs with four options shall be maintained as more research needs to be done to better understand the effects of 3-option MCQs on guessing strategies and cut score determination decisions to avoid any unintended consequential validity (21).

The present study highlights some interesting findings. First, an increased number of options does not affect the difficulty level of the questions; however, it significantly affects their discrimination power. Questions with a higher number of options tend to have lower DI and a higher number of non-effective distractors. Therefore, it is suggested to standardize the number of options to only four, in both preclinical and clinical phases of examination. Second, teachers from preclinical phase showed considerable improvement in constructing test items with plausible distractors and ideal difficulty level as compared to the clinical phase. Lack of motivation and time constraints may be the possible challenges for the clinical teachers to construct good quality items (22). Despite the availability of the faculty's guidelines for constructing examination questions for reference, training and continuous follow-up and feedback to them are important to decrease items flaws and improve item-writing skills. Institutional support for faculty development programs is crucial to ensure reliable and valid assessment strategies, especially for high-stake examinations.

The roles of vetting committee in medical schools have been described in the literature to evaluate the content,

language, and technical aspects of all questions (23, 24). Vetting sessions should be more thorough to ensure the validity of test items by removing any flaws and making them as understandable and clear as possible (25). Additionally, more time, as well as resources, are needed to develop the assessment blueprint to ensure all items are aligned with the learning objectives. A well-developed blueprint also corresponds with the depth of knowledge and level of difficulty of each content area.

Several limitations should be noted in this study. First, this study was confined to one educational setting, limiting its generalization. Any effort to infer the findings to other educational settings needs to be done with caution. Second, several other parameters such as internal consistency and correlation between PI and DI were not measured in this study. Lastly, some variables such as previous training on writing MCQs and students' characteristics were not controlled during the analysis, which might affect the findings of the study.

### *4.1. Conclusions*

The findings suggest standardizing the number of options to only four as it did not much affect the difficulty level of the questions but improve the discrimination degree of the items between high and low achievers. This will also ease the teachers on preparing MCQs with equally plausible distractors. More trainings are required for the teachers, especially from clinical phase, to improve the quality of the items as seen in preclinical phase. Feedback should be given to all teachers after analysis for them to reflect and make improvements. Good quality items have to be stored in the question bank while the poor ones have to be discarded.

## Acknowledgments

The authors wish to thank the Dean and Deputy Dean of Academic of the Faculty of Medicine and Health Sciences, UPM for their support during the running of this study. Furthermore, to the staff in Academic Unit; Ms Siti Zuraida Shahardin, Ms Siti Nor Husaine Husin, Mr. Muhammad Hakimi bin Suhaimi and Mr. Mohd Esham Husain for their valuable assistance in data collection and documentations.

## Footnotes

**Authors' Contribution:** Conceptualization: SKA and RH. Data curation: SKA. Formal analysis: SKA, FI, PSJK, and RH. Methodology: SKA and RH. Writing the original draft: SKA. Writing, review, and editing: SKA, FI, PSJK, NFZ, and RH.

**Conflict of Interests:** The authors declare there is no conflict of interest.

**Funding/Support:** This study did not receive any funding.

## References

1. Messick S. Linn R, editor. *Educational measurement*. Washington, DC: American Council on Education; 1989. p. 13–103.
2. Linn R, Gronlund N. *Measurement and assessment in teaching*. 8th ed. New Jersey: Prentice Hall; 2000.
3. Kaur M, Singla S, Mahajan R. Item analysis of in use multiple choice questions in pharmacology. *Int J Appl Basic Med Res*. 2016;**6**(3):170–3. doi: 10.4103/2229-516X.186965. [PubMed: 27563581]. [PubMed Central: PMC4979297].
4. Caldwell DJ, Pate AN. Effects of question formats on student and item performance. *Am J Pharm Educ*. 2013;**77**(4):71. doi: 10.5688/ajpe77471. [PubMed: 23716739]. [PubMed Central: PMC3663625].
5. Hassan S, Hod R. Use of item analysis to improve the quality of single best answer multiple choice question in summative assessment of undergraduate medical students in Malaysia. *Edu Med J*. 2017;**9**(3):33–43. doi: 10.21315/eimj2017.9.3.4.
6. Kheyami D, Jaradat A, Al-Shibani T, Ali FA. Item analysis of multiple choice questions at the department of paediatrics, Arabian Gulf University, Manama, Bahrain. *Sultan Qaboos Univ Med J*. 2018;**18**(1):e68–74. doi: 10.18295/squmj.2018.18.01.011. [PubMed: 29666684]. [PubMed Central: PMC5892816].
7. Sim SM, Rasiah RI. Relationship between item difficulty and discrimination indices in true/false-type multiple choice questions of a para-clinical multidisciplinary paper. *Ann Acad Med Singap*. 2006;**35**(2):67–71. [PubMed: 16565756].
8. Kar S, Lakshminarayanan S, Mahalakshmy T. Basic principles of constructing multiple choice questions. *Indian J Community Med*. 2015;**1**(2). doi: 10.4103/2395-2113.251640.
9. Abdul Rahim A. *What those numbers mean: A guide to item analysis*. Kota Bahru, Kelantan: KKMED Publications; 2010.
10. Ananthakrishnan N. Item analysis-validation and banking of MCQs. In: Sethuraman K, Kumar S, editors. *Medical Education Principles and Practice*. Pondicherry: JIPMER; 2000. p. 131–7.
11. Miller GE. The assessment of clinical skills/competence/performance. *Acad Med*. 1990;**65**(9 Suppl):S63–7. doi: 10.1097/00001888-199009000-00045. [PubMed: 2400509].
12. Schneid SD, Armour C, Park YS, Yudkowsky R, Bordage G. Reducing the number of options on multiple-choice questions: response time, psychometrics and standard setting. *Med Educ*. 2014;**48**(10):1020–7. doi: 10.1111/medu.12525. [PubMed: 25200022].
13. Vegada B, Shukla A, Khilnani A, Charan J, Desai C. Comparison between three option, four option and five option multiple choice question tests for quality parameters: A randomized study. *Indian J Pharmacol*. 2016;**48**(5):571–5. doi: 10.4103/0253-7613.190757. [PubMed: 27721545]. [PubMed Central: PMC5051253].
14. Rao C, Kishan Prasad HL, Sajitha K, Permi H, Shetty J. Item analysis of multiple choice questions: Assessing an assessment tool in medical students. *International Journal of Educational and Psychological Researches*. 2016;**2**(4). doi: 10.4103/2395-2296.189670.
15. Hassan S, Amin RM, Bt. Mohd Amin Rebuan H, Thwe Aung MM. Item analysis, reliability statistics and standard error of measurement to improve the quality and impact of multiple choice questions in undergraduate medical education in faculty of medicine at UNISZA. *Malaysian J Public Health Med*. 2016;**16**(3):7–15.
16. Quaigrain K, Arhin AK, King Fai Hui S. Using reliability and item analysis to evaluate a teacher-developed test in educational

measurement and evaluation. *Cogent Education*. 2017;**4**(1). doi: 10.1080/2331186x.2017.1301013.

17. Mitra NK, Nagaraja HS, Ponnudurai G, Judson JP. The levels of difficulty and discrimination indices in type A multiple choice questions of pre-clinical semester 1 multidisciplinary summative tests. *Int eJournal Sci Med Educ*. 2009;**3**(1):2–7.
18. Rodriguez MC. Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. *Educ Meas*. 2005;**24**(2):3–13. doi: 10.1111/j.1745-3992.2005.00006.x.
19. Loudon C, Macias-Munoz A. Item statistics derived from three-option versions of multiple-choice questions are usually as robust as four- or five-option versions: implications for exam design. *Adv Physiol Educ*. 2018;**42**(4):565–75. doi: 10.1152/advan.00186.2016. [PubMed: 30192185].
20. Royal K, Dorman D. Comparing item performance on three- versus four-option multiple choice questions in a veterinary toxicology course. *Vet Sci*. 2018;**5**(2):55. doi: 10.3390/vetsci5020055. [PubMed: 29890727]. [PubMed Central: PMC6024797].
21. Royal KD, Stockdale MR. The impact of 3-option responses to multiple-choice questions on guessing strategies and cut score determinations. *J Adv Med Educ Prof*. 2017;**5**(2):84–9. [PubMed: 28367465]. [PubMed Central: PMC5346173].
22. Karthikeyan S, O'Connor E, Hu W. Barriers and facilitators to writing quality items for medical school assessments - a scoping review. *BMC Med Educ*. 2019;**19**(1):123. doi: 10.1186/s12909-019-1544-8. [PubMed: 31046744]. [PubMed Central: PMC6498649].
23. Gopalakrishnan S, Udayshankar PM. Question vetting: The process to ensure quality in assessment of medical students. *J Clin Diagn Res*. 2014;**8**(9):XM01–3. doi: 10.7860/JCDR/2014/9914.4793. [PubMed: 25386509]. [PubMed Central: PMC4225961].
24. Hassan S, Simbak N, Yussof H. Structured vetting procedure of examination questions in medical education in faculty of medicine at Universiti Sultan Zainal Abidin Malaysia. *J Public Health Med*. 2016;**16**:29–37.
25. Wadi MM. Question vetting: Theory and practice. *Educ Med J*. 2012;**4**(1). doi: 10.5959/eimj.v4i1.29.