Journal of Iranian Medical Council, Volume 4, Issue no. 3
http://dx.doi.org/10.18502/jimc.v4i3.7219

Original Article

Predictability of Discrimination Coefficient and Difficulty Index of Psychiatry Multiple-Choice Questions

Shiva Soraya[1], Amir Shabani[2], Leila Kamalzadeh[2], Fatemeh Kashaninasab[3], Vahid Rashedi[4], Mahdie Saeidi[5], Ruohollah Seddigh[1] and Shabnam Asadi[5*]

1. Spiritual Health Research Center, Iran University of Medical Sciences, Tehran, Iran

2. Mood Disorder Research Group, Mental Health Research Center, Iran University of Medical Sciences, Tehran, Iran

3. Rasoul Akram Hospital, Department of Psychiatry, Faculty of Medicine, Iran University of Medical Sciences, Tehran, Iran

4. Faculty of Behavioral Science and Mental Health (Tehran Institute of Psychiatry), Iran University of Medical Sciences, Tehran, Iran

5. Faculty of Medicine, Iran University of Medical Sciences, Tehran, Iran

Abstract

Background: Multiple-choice questions are among the most common written tests. This study aimed to evaluate the faculty members'ability to determine and predict the level of difficulty and discrimination coefficient of multiple-choice tests at Psychiatry Department.

Methods: All faculty members at Psychiatry Department of Iran University of Medical Sciences participated in this study. The difficulty and discrimination coefficient of all questions (150 questions) of the mid-term exam of psychiatric residents were measured with both software program and formulas by hand. Then, from each group of questions with high, medium, and low difficulty coefficient, 10 questions (30 questions in total) were selected and provided to faculty members for ranking each question in terms of difficulty and discrimination coefficient. Finally, the correlation between faculty members' evaluation and standard results was measured by the Spearman's correlation. To calculate the discrimination coefficient, the number of people who answered a question correctly in the low-score group was subtracted from the high-score group and then the result was divided by the number of people in a group.

Results: Twenty-five faculty members participated in this study. There was a significant negative correlation between difficulty level and discrimination coefficient in the whole group (r=-0.196, p=0.045), but this was not the case in the upper and lower groups (r=-0.063, p=0.733). In addition, the correlation between the discrimination coefficient obtained from the formula and the average discrimination coefficient of faculty members was not significant (r=-0.047, p=0.803).

Conclusion: It seems that the ability of faculty members to predict the discrimination coefficient and difficulty level of questions is not sufficient.

Keywords: Education, Educational measurements, Medical, Psychiatry, Reference standards

*** Corresponding author**

Shabnam Asadi, MD
Faculty of Medicine, Iran University of Medical Sciences, Tehran, Iran
Email: shabnam.asadi23@gmail.com

Received: Mar 16 2021
Accepted: Jun 26 2021

Citation to this article:

Soraya Sh, Shabani A, Kamalzadeh L, Kashaninasab F, Rashedi V, Saeidi M, et al. Predictability of Discrimination Coefficient and Difficulty Index of Psychiatry Multiple-Choice Questions. *J Iran Med Counc*. 2021; 4(3):165-72.

Introduction

In Iran, students enter the medical student course after graduation from high school and passing the comprehensive entrance exam. Medicine is a 7-year course that includes basic sciences, semiology, physiopathology and internships (1).

During the medical course, the training curriculum for students mainly covers theory courses. Students are involved in clinical work and complete two separate one-month training courses in psychiatry rotation at the externship and internship level. In these courses, efforts are made to strengthen the basis of the theory of psychiatry courses and the principles of interviewing psychiatric patients and strengthening clinical skills. After graduating from the general medicine course, following the medical residency exam, the 4-year psychiatric residency course begins with a more profound emphasis on the principles of psychiatry. During the residency course, efforts are made to empower residents in the fields of psychiatric emergencies, child and adolescent psychiatry, geriatric psychiatry, consultation-liaison psychiatry, in-patient and out-patient adult psychiatry, psychotherapy and neurology rotations (2,3). Upon completing the residency course, residents must pass a written examination delivered by the Iranian Board of Psychiatry and Ministry of Health and Medical Education and an oral examination (A 20-minute interview with actual patients and evaluation of two or three examiners) in conjunction with Objective Structured Clinical Examination (OSCE) (4,5).

Medical students are assessed by various methods including global ranking scales, direct observation by seniors, verbal interviews, written exams, multiple-choice questions and OSCE. Numerous studies have been conducted on how to evaluate medical students in Iran and the world. The results of a survey conducted at Bushehr University of Medical Sciences, Bushehr, Iran showed that the most common method of student assessment is the use of descriptive and multiple-choice midterm exams (6). Multiple-choice tests are the most common type of written tests used in devising functional tests worldwide for about five decades. In Iranian universities of medical sciences, multiple-choice tests are the most common student evaluation tests (7).

Multiple-choice questions can assess a wide range of topic content in a short period. These tests are objective, accurate and can be easily scored, but they also have several limitations. For example, in such trials, students' higher cognitive levels are often not assessed and there is a possibility of choosing the right option based on making a random guess. Sometimes, students have a misinterpretation and misunderstanding of the question that affects their response (8). Given that multiple-choice questions are the cornerstone of evaluation in medical education today, standardization has become very important to judge the appropriateness or quality of multiple-choice questions. Several guidelines have been developed to interpret the difficulty and the power of differentiation of questions (9).

The difficulty of the question is defined by estimating the percentage of the population for whom the test was designed and the percentage of correct answers. In fact, the difficulty is estimated more easily if more individuals ansewr to the question correctly. The difficulty of the question is assessed by relative and absolute approaches. In determining the relative difficulty, the difficulty of the question is ranked compared to the difficulty of other questions on the same test. In contrast, absolute difficulty refers to the actual percentage of people who answered the question correctly (10).

It seems that the ability to predict the difficulty and statistical characteristics of questions with the help of test experts can affect the quality of the test (10). Research shows that sometimes the judgment of test experts reflects problems related to questions or other scales which is not helpful (11); in fact, the results of studies on the ability of test experts to judge the difficulty of test questions are contradictory (12).

For example, in Sherman Tinkelman's study, it was shown that test experts estimated the relative difficulty of questions better than the absolute difficulty of questions. They also overestimated the percentage of people who answered difficult questions and underestimated the percentage of easy questions (13). These discrepancies can be due to test experts' opinions, experiences and thinking processes in estimating the difficulty of test questions (11).

Most studies raise challenging questions and research on discrimination coefficient measures is scarce and insufficient. Also, no study has been conducted so far

on the ability of test experts to determine the difficulty and differentiation power of four-option multiple choice questions in an academic environment in Iran; therefore, due to differences in assessment of test experts judging the quality of multiple-choice questions, conducting a study to evaluate their ability is essential. Therefore, this study examined the skills of faculty members at Psychiatry Department of Iran University of Medical Sciences to determine the difficulty levels and discrimination coefficient measures of multiple-choice tests.

Materials and Methods

The current research is a cross-sectional study in the field of education, which was done in 2017. In this study, all the faculty members of Psychiatry Department of Iran University of Medical Sciences entered the study based on the purposive sampling method. To conduct the research, test questions consisted of mid-term evaluation questions of the Department of Psychiatry at Iran University of Medical Sciences which were given to participants on May 27, 2017. First, the difficulty and discrimination coefficient of all 105 questions were measured by the software. To determine the software's accuracy, the difficulty coefficient of several questions using a statistical formula was measured by hand and the obtained scores by the formula and those by the software were compared. The necessary information for applying the formulas was obtained from the university. Then, from each group of questions with high, medium and low difficulty coefficients, 10 questions (30 questions in total) were selected by simple random sampling using the random selection table. Together, these 30 questions made up the whole questions to determine the degree of difficulty and discrimination coefficient. Faculty members were asked to rate each question in terms of difficulty based on the Likert scale (difficult, medium, easy). Also, to check the discrimination coefficients of the questions, next to each question, columns were added with values ranging from 0 to 100 to distinguish residents' knowledge. Finally, the correlation between faculty members' evaluation and standard results obtained from the software and formula was measured using Spearman's correlation coefficient.

Difficulty coefficient

The difficulty coefficient is by definition the percentage of the total number of test-takers who answered a question correctly and is denoted by the letter P. If all test papers were involved in the analysis of a question, it was sufficient to calculate the difficulty factor of the question by dividing the total number of people who answered the question correctly by the total number of test-takers and the result. Then, the result could be multiplied by 100 and the difficulty factor of the question was reached. In cases where the number of test-takers (Number of sheets) was high and our information was limited to how the upper and lower groups responded, it was necessary to use the following formula:

Difficulty factor=Correct choices of the lower group +Right choices of the upper group *100/Number of people in the top group+Number of people in the bottom group For selecting the top and bottom groups, the following procedure was done; if the number of our respondents was less than 20, after correcting, all the papers were arranged in the order of scores obtained from the first to the last person and were divided into two groups of strong and weak. The top and bottom groups were associated with low scores. If the respondents were between 20 and 40 people, 10 people were selected from the strong group and 10 people from the weak group, respectively and the people with middle scores were excluded And if the number of people was more than 40, 27% of cases for the top group were selected and 27% for the bottom group. The larger the difficulty factor of a question is, the closer to 100, the easier that question would be and the smaller the coefficient was, closer to zero, the more difficult the question would be. The optimal difficulty coefficient for a question, regardless of the type of question, is about 50% (Between 30 and 70%). A difficulty coefficient of less than 30% is considered difficult, between 30 to 70% is considered medium and more than 70% is considered easy (7).

Discrimination coefficient calculation

The discrimination coefficient determines the extent to which the question separates the top and bottom groups. The following formula is used to calculate the discrimination coefficient of a question:

Discrimination coefficient=(Correct choices of the

top group-Correct choices of the bottom group)/ Number of people in a group (Up or down) The larger the discrimination coefficient (Closer to one) is, the greater the discrimination coefficient of the question would be and the smaller value (Closer to zero) corresponds to lower discrimination power. A discrimination coefficient of less than 20% is considered low, 20% to 34% is considered medium and above 35% is considered high (7).

Ideal questions

The combination of two coefficients of difficulty and purity of questions was used to show the ideal questions. Ideal multiple-choice questions have a difficulty coefficient of 30% to 70% and a discrimination coefficient of over 24% (7).

In this study, informed consent was obtained from all participants, and all participants' data was collected confidentially. Final results were reported collectively (Not individually). This project was approved by the ethics committee of Iran University of Medical Sciences (IR.IUMS.REC1396.31760).

Results

Twenty-five faculty members participated in the study. Thirteen questions were selected from 105 questions and provided to faculty members.

A total of 60 psychiatry residents participated in the 105-question test. The mean discrimination coefficient of all questions was 59.58±2.09, of which 52.4% had a medium difficulty coefficient. Also, the mean discrimination coefficient of the questions was 26.04±1.80, of which 16.2% had a medium discrimination coefficient, and 35.2% had a high discrimination coefficient. Calculations showed that only 37 out of 105 questions (35.2%) were ideal, and the rest of the questions did not have this feature (Table 1).The mean of faculty members who correctly predicted the difficulty of the questions was 41.43±3.10% (Figure 1).

Based on the Spearman's correlation coefficient test results, there was a negative and significant correlation between the difficulty coefficient and discrimination coefficient in the whole group (r=-0.196,p=0.045). Still, there was no significant correlation between the difficulty coefficient and discrimination coefficient in the upper and lower groups (r=-0.063,p=0.733). Also, the correlation between the discrimination coefficient obtained from the formula and the average computed coefficient of faculty members was not significant (r=-0.47,p=0.803) (Figure2). There was no significant correlation between the difficulty coefficient calculated by the formula and the one predicted by the faculty members (r:0.208,p:0.269).

Discussion

According to the results of this study, in general, the more complex the questions are, the higher the computing power is, but this is not true among people with very high or very low grades; in other words, difficult questions can not necessarily determine the exact computing power among people with high or low grades. Our study also showed that faculty members were not successful in estimating the discrimination coefficient of questions. In fact, no specific research has been done on evaluating discrimination coefficients, and most studies focus on the coefficient of difficulty. It should be noted that although the difficulty factor of questions in exams that assess a person's knowledge is very important, the discrimination coefficient of questions in competitive exams such as entrance exams can play a significant role in better ranking of people;

Table 1. Frequency distribution of difficulty coefficient and discrimination coefficient of all questions

Percentage	Number	Definition	Difficulty coefficient
14.3	15	Difficult	30>
52.4	55	Medium	30-70
33.3	35	Easy	70<
			Discrimination coefficient
48.6	51	Low	20>
16.2	17	Medium	20-34
35.2	37	High	34<

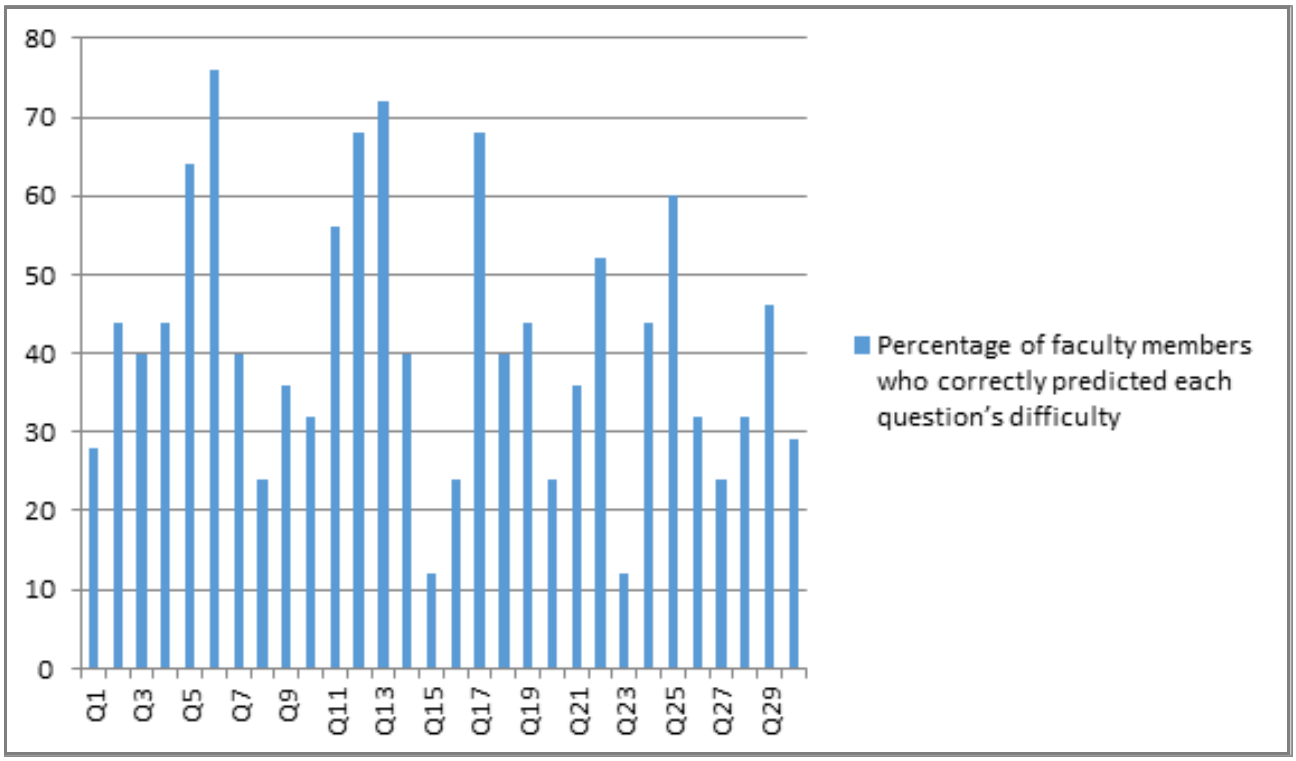

Figure1. Percentage of faculty members who correctly predicted each question's difficulty.

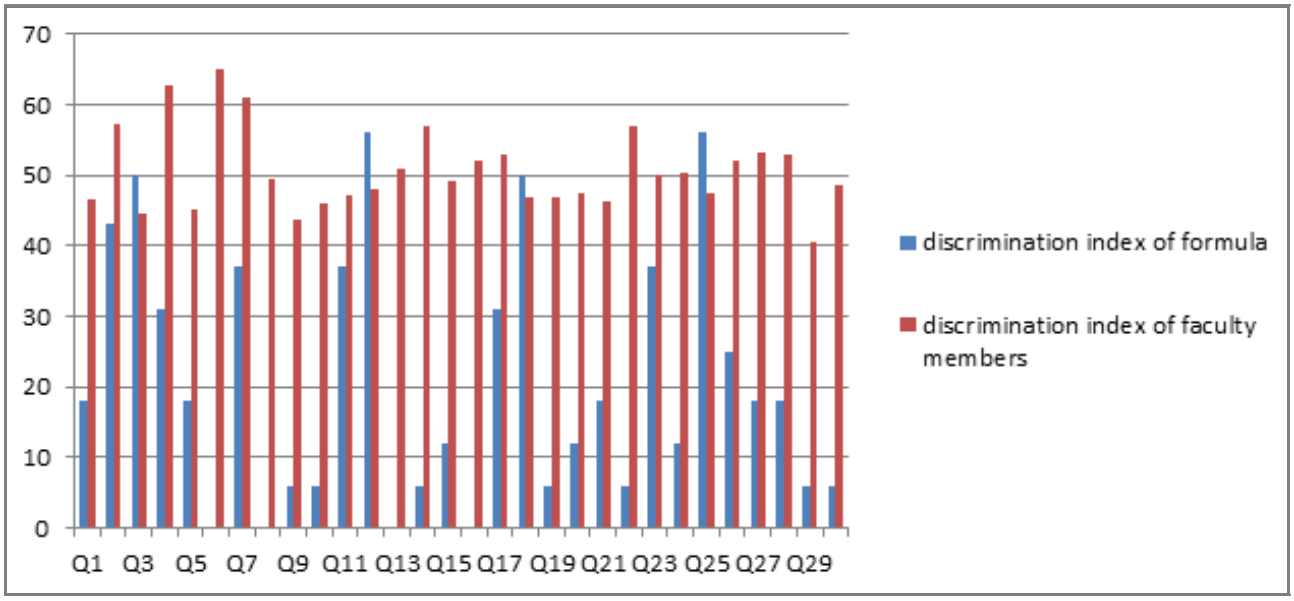

Figure 2. Discrimination coefficient based on faculty members' opinion and calculation with formula.

so conducting studies in this field and reviewing the factors affecting the ability of question designers to predict the discrimination coefficient of questions is necessary.

In this study, 52.4% of the questions had a medium difficulty coefficient, 51.4% had a medium and high discrimination coefficient, and only 35.2% included ideal questions. In the study of Shakurnia *et al*, these values were 46.2% for the medium difficulty coefficient, 57.3% for the medium and high discrimination coefficient, and 30.7% for the ideal questions (7). Also, most similar studies have shown that about 50% of the questions have the medium difficulty coefficient, all of which are similar to the results of our study (14,15). In their research, Shakurnia *et al* introduced an strategy to improve the quality of questions by forming a question bank in educational groups and eliminating inappropriate questions by periodic analysis (7).

Examination of quantitative indices of multiple-choice questions in Qazvin University of Medical Sciences has shown that more than half of the designed questions did not have medium discrimination coefficient (14). In this study, 48.6% of the questions had an inappropriate discrimination coefficient, which indicates the weakness of the questions in distinguishing between strong and weak students.

In Mehta and Mokhasi's study, 24% of the multiple-choice questions were ideal, which is almost consistent with the findings of this study (16). But in a study in Pakistan, the frequency of ideal questions was reported to be 64% (17), which is higher than our study and probably indicates a higher ability of question designers.

Bejar (18), Melican *et al* (19) and Cross *et al* (20), in their studies, evaluated the ability of judges to estimate the difficulty level of questions. In all of these studies, there was a weak correlation between the grading of the difficulty of the questions by the judges and the actual difficulty of the questions.

In the study by Kibble and Johnson, eight physiology professors graded several questions before the test in terms of difficulty. Data analysis generally showed a significant but relatively low correlation between the estimated difficulty and actual student scores (21), while Attali *et al* believed that the reason for the poor results in previous studies was that the judges evaluated questions independently and without comparison with other questions, and if the questions were comparatively assessed, the results could be more accurate and precise.

In their study, 7 judges evaluated math questions in terms of difficulty. The questions were compared with each other and relatively ranked. The results showed that the judges were successful in this task, and there was a high correlation between actual and estimated difficulty (12). However, predicting the difficulty of math questions may differ from medical questions, and generalizing the results of this study to exams in other fields does not seem reasonable.

In our study, the number of evaluators was more than the last two studies, but the number of questions in Kibble and Johnson's study was more than our study (272 questions). There was no significant correlation between the actual difficulty coefficient of the questions and the difficulty coefficient predicted by the faculty members in our study, which was consistent with other studies.

One of the limitations of the present study was the availability of only one educational group, so it is recommended to repeat the experiment in other educational groups. Another limitation is the lack of standard questions to assess the discrimination coefficient in various fields of psychiatry and to evaluate the faculty members' ability as well. Due to small sample size in our study, similar studies with larger samples are needed to achieve generalizable results.

Conclusion

The ability to predict the discrimination coefficient and difficulty of questions can lead to designing tests with higher standards and thus a better assessment of the achievement of educational goals. However, due to the impossibility of predicting the discrimination coefficient and difficulty of four-option multiple choice questions based on the findings of this study and similar studies, the use of various tests, including descriptive and ASCII tests is necessary to evaluate assistants/students, as well as to determine the difficulty of each test before giving them to participants; moreover, applying the opinions of newly graduated teachers is a fundamental help which of course requires further research.

In general, the ability of faculty members and question designers in this area seems insufficient, so

it is necessary to find influential factors and ways to strengthen them.

Acknowledgements

The project was approved by the ethcs committee of Iran University of Medical Sciences (IR.IUMS.REC 1396.31760).

Authors' contribution

All the authors met the standards of authorship based on the recommendations of the Journal of Iranian Medical Council (JIMC).

Conflict of Interest

None declared.

References

1. Tavakol M, Murphy R, Torabi S. A needs assessment for a communication skills curriculum in Iran. Teach Learn Med 2005;17(1):36-41.

2. Azizi F. Medical Education in the Islamic Republic of Iran: three decades of success. Iran J Public Health 2009;38(1):19-26.

3. Sadeghi M, Mirsepassi Gh. Psychiatry in Iran. Int Psychiatry 2005;2(10):10-2.

4. Eissazade N, Shalbafan M, Eftekhar Ardebili M, Pinto da Costa M. Psychotherapy training in Iran: A survey of Iranian early career psychiatrists and psychiatric trainees. Asia Pac Psychiatry 2021;13(1):e12434.

5. Sadeghi M, Taghva A, Mirsepassi G, Hassanzadeh M. How do examiners and examinees think about role-playing of standardized patients in an OSCE setting? Acad Psychiatry 2007;31(5):358-62.

6. Komeili Gh, Rezai Gh. Methods of student assessment used by faculty members of Basic Medical Sciences in Medical University of Zahedan. Iranian J Medical Education 2001;1(4):52-7.

7. Shakurnia A, Ghafourian M, Khodadadi A, Ghadiri A. Analytical study of quantitative indices of multiple-choice questions of immunology department in Ahvaz Jundishapur University of Medical Sciences. JundiShapur Educational Development 2018;9(2):72-83.

8. Farley JK. The multiple-choice test: writing the questions. Nurse Educ 1989;14(6):10-2,39.

9. Young M, Cummings B-A, St-Onge C. Ensuring the quality of multiple-choice exams administered to small cohorts: A cautionary tale. Perspect Med Educ 2017;6(1):21-8.

10. Lorge I, Kruglov L. A suggested technique for the improvement of difficulty prediction of test items. Educational and Psychological Measurement 1952;12(4):554-61.

11. Hambleton RK, Sireci SG, Swaminathan H, Xing D, Rizavi S. Anchor-based methods for judgmentally estimating item difficulty parameters. LSAC Research Report Series 2003.

12. Attali Y, Saldivia L, Jackson C, Schuppan F, Wanamaker W. Estimating item difficulty with comparative judgments. ETS Research Report Series 2014;2014(2):1-8.

13. Tinkelman S. Difficulty prediction of test items. Teachers College Contributions to Education. 1947.

14. Derakhshan F, Ahmady S, Allami A. Quantitative and qualitative indicators evaluation of residency exams in Qazvin University of Medical Sciences (2012-13). J Med Educ Dev 2016;8(20):29-38.

15. HosseiniTeshnizi S, Zare S, Solati S. Quality analysis of multiple choice questions (MCQs) examinations of noncontinuous undergraduate medical records. Hormozgan Med J 2010;14(3):177-83.

16. 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.

17. 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.

18. Bejar II. Subject matter experts assessment of item statistics. ETS Research Report Series 1981;1981(2):i-47.

19. Melican GJ, Mills CN, Plake BS. Accuracy of item performance predictions based on the Nedelsky standard setting method. Educational and Psychological Measurement 1989;49(2):467-78.

20. Cross LH, Impara JC, Frary RB, Jaeger RM. A comparison of three methods for establishing minimum standards on the National Teacher Examinations. J Educational Measurement 1984;21(2):113-29.

21. Kibble JD, Johnson T. Are faculty predictions or item taxonomies useful for estimating the outcome of multiple-choice examinations? Advances in Physiology Education 2011;35(4):396-401.