Sudan Journal of Medical Sciences
Volume 16, Issue no. 3, DOI 10.18502/sjms.v16i3.9695
Production and Hosting by Knowledge E

Review Article

Item Analysis of Multiple-choice Questions (MCQs): Assessment Tool For Quality Assurance Measures

Amani H. Elgadal[1] and Abdalbasit A. Mariod[2]

[1]Department of Pediatrics and Child Health, Karary University, Omdurman, Sudan
[2]Indigenous Knowledge & Heritage Center, Ghibaish College of Science & Technology, Ghibaish, Sudan

ORCID:
Amani H. Elgadal: https://orcid.org/0000-0003-0934-2755
Abdalbasit A. Mariod: https://orcid.org/0000-0003-3237-7948

Abstract

Background: Integration of assessment with education is vital and ought to be performed regularly to enhance learning. There are many assessment methods like Multiple-choice Questions, Objective Structured Clinical Examination, Objective Structured Practical Examination, etc. The selection of the appropriate method is based on the curricula blueprint and the target competencies. Although MCQs has the capacity to test students' higher cognition, critical appraising, problem-solving, data interpretation, and testing curricular contents in a short time, there are constraints in its analysis. The authors aim to accentuate some consequential points about psychometric analysis displaying its roles, assessing its validity and reliability in discriminating the examinee's performance, and impart some guide to the faculty members when constructing their exam questions bank.

Methods: Databases such as Google Scholar and PubMed were searched for freely accessible English articles published since 2010. Synonyms and keywords were used in the search. First, the abstracts of the articles were viewed and read to select suitable match, then full articles were perused and summarized. Finally, recapitulation of the relevant data was done to the best of the authors' knowledge.

Results: The searched articles showed the capacity of MCQs item analysis in assessing questions' validity, reliability, its capacity in discriminating against the examinee's performance and correct technical flaws for question bank construction.

Conclusion: Item analysis is a statistical tool used to assess students' performance on a test, identify underperformed items, and determine the root causes of this underperformance for improvement to ensure effective and accurate students' competency judgment.

Keywords: assessment, difficulty index, discrimination index, distractors, MCQ item analysis

Corresponding Author:
Amani H. Elgadal;
email:
amanielgaddal@karary.edu.sd
amanielgaddal@gmail.com

Received 27 July 2021
Accepted 02 September 2021
Published 30 September 2021

Production and Hosting by Knowledge E

Editor-in-Chief:
Prof. Mohammad A. M. Ibnouf

1. Introduction

Single or One Best Answer of Multiple-choice Questions (MCQs) is known as an item consisting of a stem with many options, generally three to five, one of them being the right option while the rest distractors. This form of assessment is used in many institutions due to its capability to significantly appraise curricula. It is an efficient and relevant tool to identify the strengths and weaknesses in student knowledge, reflection of educational methods and strategies, however, it needs time, effort, and skill to develop a high-quality one [1].

A well-build MCQ assesses higher cognitive tackles of Bloom's taxonomy like data interpretation, synthesis, and knowledge application more than testing facts recall alone. The stem of the MCQs is a clinical case scenario that can adequately measure core competencies, the intended learning outcome (ILO), evaluating the power of students, give reliable feedback, and reform curricula [1–3]. There are six hierarchically assortments of cognitive scope in Bloom's taxonomy that are arranged in ordered factions: knowledge, comprehension, application, analysis, synthesis, and evaluation. Tarran trivializes Bloom's taxonomy and creates two levels: K1 represents the fundamental knowledge and cognition; K2 embraces analyzing with implementation and analysis [4]. Item analysis is a hokey and avail approach to assess the reliability and validity of test items, performed after the exam. It auditions the effectiveness of stem question and its distractors to enable the examiners to reconstruct/modify or delete questions before the creation of an exam bank for future tests [1–4]. Item analysis shows the questions' difficulty index (DIF-I). Ditto assesses the question's capability to discriminate performance of good or poor students in the test, that is, the discrimination index (DIS-I) [1–5]. Bona MCQs assess perception, effectiveness, and psychomotor scopes better than other assessment methods due to its objectivity covering many subjects, minimizing the assessor's alignment, and its comparative, reliable, conciliated, and easy netting [3–5]. In addition, it is also a relevant method that measures any impairment or strengths of the examinee's knowledge, gaps in teaching methods, or strategies of the institute for better graduate outcomes. It provides a good chance to the staff members to stimulate them in building their MCQ construction skills needed for the clarity of exam questions. [2] The standardization tool characteristics can influence its credibility. MCQ designers ought to pay attention to the examination purpose and its content based on the examinee level, blueprint, and the minimum pass level (MPL). It should fit the purpose and consensus judgment with advantageous implementation. So meticulous evaluation is counseled. Maintaining the standards in medical schools is crucial for high

educational excellence, patient safety, and total quality management needed for both historic and newly established colleges [5–7].

The authors' aim in this review was to accentuate some consequential points about psychometric analysis displaying its roles in evaluating MCQs, assessing its validity and reliability in discriminating the examinee's performance, and impart some guide to the faculty members especially juniors when constructing their exam questions bank.

2. Materials and Methods

Databases such as Google Scholar and PubMed were searched for freely accessed English articles published since 2010. Synonyms and keywords were also used in the search. The abstracts of the articles were first viewed and read to select suitable matches, and then full-text articles were perused and summarized. Finally, recapitulation of the relevant data was done to the best of authors' knowledge.

3. Results

In any educational institute, assessment is a way to measure supposed mastering of ILOs. It is particularly consequential in clinical college graduates for protected patient care and community needs. Hence, meticulous evaluation and education must be performed. Standardized assessment of students' performances involves measurement aspects that are peculiar of the statistical framework. This process consists of distinct phases, from the definition of the measurement objectives to the development of proper assessment tools, and the analysis of the results in terms of students' achievement [5]. It should match student's ability and items related to specific content domains. The development of a proper assessment method is a rather complex process that starts with the definition of item specifications and ends with the validation of the assessment method itself. It effectively measures the target competencies in a test, its content and format constraints, distractors plausibility, item difficulty, and test consistency. For this purpose, first, a pretest sample is given to an examinee, their responses are then analyzed and validated using psychometric methods before conducting the final exam [6].

4. Discussion

Item analysis is a conciliated and availed method to examine the reliability and validity of the pretested standardized examination items. It is conducted after the exam before banking questions for future tests [5, 6].

4.1. Methods of item analysis

Different methods can be used to investigate the psychometric properties of tests and test items. Descriptive methods based on Classical Test Theory (CTT) and models belonging to modern Item Response Theory (IRT) were reviewed. Regarding the item level, the CTT model is a relatively simple methodology. It is the probative estimate of the examinee's success rate on each item. The CTT appraises reliability, difficulty, DIS-I, and the distractors' efficiency (DE) to check the appropriateness and plausibility of all distractors. The core of this theory is based on the functions of the true test score and the error of random measurement. On the other hand, the Rasch technique of IRT is more grounded to assess the examinee's success at the item level [7]. IRT besides apprizing the test reliability, DI, and DE, assesses the exam global rating similar to Cronbach's alpha. Additionally, it checks the exam invariance that is conclusive for building exam banks with well-calibrated exam questions. Item standardization can be classified as follows [5–8]:

1. Relative approaches (norm-referenced): used for ranking the examinee when a predetermined rating of the examinee is wanted so that there is no fixed MPL and the level fluctuates in accordance with the examinee's overall performance.
2. Absolute approaches (criterion-referenced): judgment based on:
    (a) Exam content: used in high-stake conditions like licensure; e.g., Angoff (1971), Nedelsky (1954), and Ebel (1972) methods where the Standards setter decides the borderline examinee's criteria.
    (b) Compromise: The well-known one is Hofstee, which can be used in a low-resource setting. The designers decide the MPL after consensus.

All of the above techniques should be executed before conducting the exam [5–8].

There are two types of **Angoff;** the original and the modified methods, both of which are used to decide the cut-off scores for the *exam* items. The original method needs subject experts' panel to decide the probability of a minimally competent student who

can answer each item correctly. Each expert estimates the probability ranging from 0 to 1 for every question and then calculating the average portability as a final cut-off score. The modified Angoff needs test domain expertise and the probabilities choices are eight, e.g., 0.2, 0.3, 0.4, 0. 5, 0.6, 0.7, 0.8, or "do not know" [9].

Angoff method is a predetermined criterion-referenced and test-centered method. The modified-Angoff method allows the panel's setter to discuss the cut-off score and the rating results. For this reason, the modified-Angoff method is used for licensure and professions certification tests. Since the standard-setting is a decision-making process, the criterion setting validity and rating consistency is evaluated by how the process is performed in accordance with the test principle. Evaluation of the standard-setting validity is influenced by internal and external issues. It is consequential to ascertain that all standard-setting activities and measures are done consistently [10].

In the **Nedelsky method**, three Subject Matter Experts (SMEs) are used for the standard-setting of MCQs to assess the probability of a borderline/minimally competent student who will rule out the incorrect options or distractors. The probability is calculated as the reciprocal of the remaining items which the borderline/ minimally competent students are not sure if it is correct or not. For example, a group of experts assess the probability of borderline /minimally qualified students who are expected to rule out two distractors in a four-options item question. The rating will be half (1/2 = 0.50). The cut-off score for the exam is determined by adding up the average Nedelsky values for each item [10, 11].

The Ebel method needs subject experts to judge the difficulty and relevance level of each item in the exam. The panel examines each item to determine its appropriateness, difficulty or simplicity, its relevance, importance, and acceptability. Each item is categorized according to its difficulty and relevance level. Next, the panel experts assess the expected chances of a minimally competent student who can rule out item distractors. Lastly, the number of items in each category is multiplied by the expected probability of correct answers, and the total results are added to calculate the exam cut-off score. Relatively, this method is costly, time-consuming, and needs many standard experts setters. Digital soft wire is important to gather the responses. Backup by the criterion-referenced method is needed like borderline regression. It is widely used in high-stakes exams and if challenged, it can hold up in court [12].

Eclectic Hofstee method was developed in 1983 to address problems that resulted from predictions disagreement between criterion- and norm-referenced items. In this method, the standard setter answers four enquires and presumptions about the candidates who will write the test. Two of these queries are about their apt knowledge level

(referred as k), while the other two are about the failure rate (referred as f); (1) What is the satisfactory maximum cut-off score, even if all of the examinees overreached it? (2) What is the acceptable minimum cut-off score even if all of the examinees do not achieve it? (3) What is the allowed maximum failure rate? (4) What is the minimally accepted failure rate? The first two questions assess the failure rates and range between zeros and a hundred percent; closer to 100% indicate test difficultly and hard for anyone to pass. The last two questions, however, are scored between zero and the total test items numbers, the higher the value, the more difficult the cut-off score [12].

Selection of a suitable psychometric approach is influenced by different factors. It varies depending on the intended goals/objective. In low-resource setting, the CTT psychometric method may be good enough. In a high-stakes exam, IRT and Rasch Measurement Theory must be used, and the final decisions will depend upon the quantitative and qualitative item results. You can select a suitable method according to the psychometric properties you want like the reliability, validity, suitability of item response, scaling assumption, and acceptability [13].

4.2. Reliability

The inherent concept is embedded within the CTT, reliability assesses the internal consistency of MCQs items [13, 14]. Reliability and validity are important for defining the result obtained to meet the requirements and measure bias. Reliability shows up to which level the assessments were consistent while validity assesses the assessment accuracy [15]. Reliability-related concepts are internal consistency, stability, equivalence, and precision. Reliability depends both on the standard error of measurement and the standard deviation of the examinee's assessment. Regarding the internal consistency, the estimation depends on the item's average correlation for a test, also it estimates to which degree the MCQs can measure the same knowledge domain characteristics. Typically, internal consistency is obtained by calculating the reliability coefficient. A reliability coefficient estimates the concordance between the observed and true scores of the examinees, it appraises the interlinks between scores obtained by two parallel exams. This estimation explains that an individual's scores are expected to change when retested without alteration in knowledge and perception with the same or any equivalent test [14–16]. Increasing the item numbers in a given exam can augment the reliability but it is expensive, needs time and average correlation effort. Cronbach's alpha of 0.8 or more is needed for high-stakes exams, however, usually, there is a fixed item number in licensure or high-stakes exams; so, you can use other alternatives by increasing

the deployment of the obtained exam scores, for example, test variance. Range of scores/performances as moderately difficult (DI: 0.4–0.8) and sufficient discrimination point biserial correlation (RPB) more or equal to 0.2. It can also increase the standard deviation and the variance of the scores [23–25]. For the assumption that any test can contain score error, SEM is used to estimate the interval within which the true score will be obtained. When the SEM is small, the interval will be narrower and more precise. SEM is inversely related to the reliability coefficient [14–16].

The Kuder-Richardson Formula 20 (Kr-20) measures internal consistency and reliability of an examination. It measures the interior uniformity of the exam with many options. Kr-20 > 0.90 indicates a homogenous test. Kr-20 = 0.8 is acceptable but >0.8 is nonreliable [17].

4.3. Statistical steps of item analysis

The Statistical Analysis System (SAS), Statistical Package for the Social Sciences (SPSS), and similar software are used in data analysis. After conducting the exam, data are gained manually or electronically and then entered into Microsoft Excel sheet, SPSS, or any other statistical methods of your choice. Next, the data are analyzed to get: the mean, standard deviations (SD), unpaired *t*-test, and coefficient of variation, DIF-I and DIS-I, and (DE) [18].

1. *Difficulty index* (DIF-I) is described as the examinee's incapability to reply to the item correctly. To calculate it: rank the examinees in order, then pick one-third of the high or greater achievers (HA) who correctly answered to the item and one-third of the lower achievers (LA) who also choose the correct answer.

   It can be calculated using the following formula:

   DIF-I = [(HA + LA)/*N*] × 100,

   where: *N* is the total number of students in the two groups.

   The DIF-I is expressed as a *P*-value, that is, the proportion of students who correctly answer questions in a given test. The DIF-I can range from zero to a hundred percent. If it is >70%, it is an easy item; 30–70% means average/acceptable difficulty; <30% means a difficult item [18–23].

2. *Discrimination index* (DIS-I) is defined as the ability of an item to differentiate between students with high- and low exam scores. It ranges from –1.00 to +1.00. Those with high value are good discriminator items. Negative DI can be obtained

if the low achievers get more correct answers than the high achievers, and vice versa. DIS-I can be calculated using the formula:

DIS-I = [(HA – LA)/*N*] x 2,

where: HA are the high achievers while LA are the low achievers in the test.

DIS-I can range from 0 to 1; if it is <0.15, it means a poor discriminator; 0.15–<0.25 means good discriminatory items; >0.25 means excellent discriminator [9–12, 20–23].

RPB is another way of measuring item discrimination, defined as the correlation between the item score and the total test score. It is mathematically equivalent to Pearson's correlation. Both DIS-I and biserial correlation are greatly correlated, and a DIS-I or RPB < 0.2 is regarded low [10–16].

3. *Distractor Analysis* aims to determine the capability of item options to distract the examinee when selecting the right answer. Each distractor must be assessed for its frequency of selection by the examinee, it is called DE [18–23].

   DE can be calculated using the formula:

   DE = Frequency of distractor selection ÷ Total no. of item respondent × 100.

   DE needs to be assessed in each MCQ to test the presence or absence of NFD. If an MCQ includes 0-NFD, 1-NFD, 2-NFD, or 3-NFD, it means that its capability to act as an efficient distractor is 100, 66.6, 33.3, or 0%, respectively [12–16].

   DE is classed as *Functional Distractor* (FD) when chosen by ≥5% of the examinee and as *Non-functional Distractors* (NFD) if chosen by <5%. NFDs include options other than the right answer chosen by <5% of the examinees. Implausible distractors can be noticed easily, so they ought to be modified or rejected [18–23].

4.4. Item flaws

Faults in item-writing can also influence the overall performance by making questions challenging or too easy.

Example: The use of absolute terms like always, never, or choosing the right option in a lengthy sentence. It is wise to refrain from. using negative words like none of the above OR except.

Grammatical flaws may divert the examinee to the right answer and make the questions easy. Items with many NFDs reduce the DE and DIS-I [24–26].

4.5. The number of item options

Some authors argue that MCQ with three options needs much less time for construction with a greater chance for high reliability and validity than four–five options. Others say that MCQ choices can be three or even two and have the potency to give the same results as 4 or 5 options without affecting the examination quality [27–29].

As cited earlier, evaluation is an essential measure not only for competent graduates but also for college enhancement and quality assurance [30–33]. Valid evaluation techniques aligned with accrediting authorities' requirements are one of the desires for excellence and accreditation. It elevates the importance of the assessment-unit building to lead all evaluation activities within the institute.

Performing collaborative and organized on-job training enhances staff capabilities in the MCQs writing and analyses for higher student success and competence [30–36].

5. Conclusion

Item analysis of MCQs is a statistical tool used to assess students' performance on a test, identify underperformed items, and determine the root causes of this underperformance for improvement in order to ensure effective and accurate students' competency judgment. It is a potent tool to appraise the ILOs in a short time, detect gaps in curriculum contents evident by student's poor performance in a test, and identify strengths and weaknesses in teaching strategies and methods. Exam reliability and validity are important for defining the result obtained to meet the requirements and measure bias. Training and retraining of all faculty members are important to improve their skills in properly standardizing MCQs construction to overcome any assessment challenges.

Acknowledgments

The authors would like to thank the Sudanese Researchers Foundation for their unlimited support.

Ethical Considerations

Not applicable.

Competing Interests

None.

Availability of Data and Material

All materials of this study are available from the corresponding author upon reasonable request.

Funding

None

References

[1] Hingorjo, M. R. and Jaleel, F. (2012). Analysis of one-best MCQs: the difficulty index, discrimination index and distractor efficiency. *Journal of Pakistan Medical Association*, vol. 62, no. 2, pp. 142–147.

[2] Vanderbilt, A. A., Feldman, M., and Wood, I. K. (2013). Assessment in undergraduate medical education: a review of course exams. *Medical Education Online*, vol. 18, no. 1, pp. 1–5.

[3] Gajjar, S., Sharma, R., Kumar, P., et al. (2014). Item and test analysis to identify quality multiple choice questions (MCQs) from an assessment of medical students of Ahmedabad, Gujarat. *Indian Journal of Community Medicine*, vol. 39, no. 1, pp. 17–20.

[4] Abdulghani, H. M., Irshad, M., Haque, S., et al. (2017). Effectiveness of longitudinal faculty development programs on MCQs items writing skills: a follow-up study. *PLoS One*, vol. 12, no. 10, e0185895.

[5] McKinley, D. W. and Norcini, J. J. (2014). How to set standards on performance-based examinations: AMEE Guide No. 85. *Medical Teacher*, vol. 36, no. 2, pp. 97–110.

[6] Ben-David, M. F. (2000). AMEE Guide No. 18: standard setting in student assessment. *Medical Teacher*, vol. 22, no. 2, pp. 120–130.

[7] Testa, S., Toscano, A., and Rosato, R. (2018). Distractor efficiency in an item pool for a statistics classroom exam: assessing its relationship with item cognitive level classified according to Bloom's taxonomy. *Frontiers in Psychology*, vol. 9, p. 1585.

[8] Kumar, D., Jaipurkar, R., Shekhar, A., et al. (2021). Item analysis of multiple choice questions: a quality assurance test for an assessment tool. *Medical Journal Armed Forces India*, vol. 77, no. 1, pp. S85–S89.

[9] Mubuuke, A. G., Mwesigwa, C., and Kiguli, S. (2017). Implementing the Angoff method of standard setting using postgraduate students: Practical and affordable in resource-limited settings. *African Journal of Health Professions Education*, vol. 9, no. 4, p. 171.

[10] Yim, M. (2018). Comparison of results between modified-Angoff and bookmark methods for estimating cut score of the Korean medical licensing examination. *Korean Journal of Medical Education*, vol. 30, no. 4, pp. 347–357.

[11] Seçil, Ö. M. Ü. R. and Selvi H. (2010). Angoff, Ebel ve Nedelsky yöntemleriyle belirlenen kesme puanlarının sınıflama tutarlılıklarının karşılaştırılması. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, vol. 1, no. 2, pp. 109–113.

[12] Park, J., Ahn, D.-S., Yim, M. K., et al. (2018). Comparison of standard-setting methods for the Korea Radiological technologist Licensing Examination: Angoff, Ebel, Bookmark, and Hofstee. *Journal of Educational Evaluation for Health Professions*, vol. 15, p. 32.

[13] Petrillo, J., Cano, S. J., McLeod, L. D., et al. (2015). Using classical test theory, item response theory, and Rasch measurement theory to evaluate patient-reported outcome measures: a comparison of worked examples. *Value Health*, vol. 18, no. 1, pp. 25–34.

[14] Ali, S. H., Carr, P. A., and Ruit, K. G. (2016). Validity and reliability of scores obtained on multiple-choice questions: why functioning distractors matter. *Journal of the Scholarship of Teaching and Learning*, vol. 16, no. 1, pp. 1–14.

[15] Abdalla, M. E. (2011). What does item analysis tell us? Factors affecting the reliability of multiple-choice questions (MCQs). *Gezira Journal of Health Sciences*, vol. 7, no. 2, pp. 17–25.

[16] Vegada, B. N., Karelia, B. N., Pillai, A., et al. (2014). Reliability of four-response type multiple choice questions of pharmacology summative tests of II. *International Journal of Mathematics and Statistics Invention*. Retrieved from: https://www.semanticscholar.org/paper/{%}22Reliability-of-four-response-type-multiple-choice-Vegada-Karelia/43a896bff1c7b16cee1a5c89643b443f0cd0bf9d#citing-papers

[17] Glen, S. (n.d.). Kuder-Richardson 20 (KR-20) & 21 (KR-21). Retrieved from: https://www.statisticshowto.com/kuder-richardson/

[18] Velou, M. S. and Ahila, E. (2020). Refine the multiple-choice questions tool with item analysis. *IAIM*, vol. 7, no. 8, pp. 80–85.

[19] Coughlin, P. A. and Featherstone, C. R. (2017). How to write a high quality multiple choice question (MCQ): a guide for clinicians. *European Journal of Vascular and Endovascular Surgery*, vol. 54, no. 5, pp. 654–658.

[20] Salih, K. E. M. A., Jibo, A., Ishaq, M., et al. (2020). Psychometric analysis of multiple-choice questions in an innovative curriculum in Kingdom of Saudi Arabia. *International Journal of Family Medicine and Primary Care*, vol. 9, no. 7, pp. 3663–3668.

[21] Harti, S., Mahapatra, A. K., Gupta, S. K., et al. (2021). All India AYUSH post graduate entrance exam (AIAPGET) 2019–AYURVEDA MCQ item analysis. *Journal of Ayurveda and Integrative Medicine*, vol. 12, no. 2, pp. 356–358.

[22] Namdeo, S. K. and Sahoo, S. (2016). Item analysis of multiple-choice questions from an assessment of medical students in Bhubaneswar, India. *International Journal of Research in Medical Sciences*, vol. 4, no. 5, pp. 1716–1719.

[23] Garg, R., Kumar, V., and Maria, J. (2019). Analysis of multiple-choice questions from a formative assessment of medical students of a medical college in Delhi, India. *International Journal of Research in Medical Sciences*, vol. 7, pp. 174–177.

[24] Tarrant, M. and Ware, J. (2008). Impact of item-writing flaws in multiple-choice questions on student achievement in high-stakes nursing assessments: item-writing flaws and student achievement. *Medical Education*, vol. 42, no. 2, pp. 198–206.

[25] Rao, C., Kishan Prasad, H. L., Sajitha, K., et al. (2016). Item analysis of multiple-choice questions: Assessing an assessment tool in medical students. *International Journal of Educational and Psychological Researches*, vol. 2, no. 4, pp. 201–204.

[26] Kheyami, D., Jaradat, A., Al-Shibani, T., et al. (2018). Item analysis of multiple choice questions at the Department of Paediatrics, Arabian Gulf University, Manama, Bahrain. *Sultan Qaboos University Medical Journal*, vol. 18, no. 1, p. 68.

[27] Vegada, B., Shukla, A., Khilnani, A., et al. (2016). Comparison between three-option, four option and five option multiple choice question tests for quality parameters: a randomized study. *Indian Journal of Pharmacology*, vol. 48, no. 5, pp. 571–575.

[28] Nwadinigwe, P. I. and Naibi, L. (2013). The number of options in a multiple-choice test item and the psychometric characteristics. *Journal of Education and Practice*, vol. 4, pp. 189–196.

[29] Tweed, M. (2019). Adding to the debate on the numbers of options for MCQs: the case for not being limited to MCQs with three, four or five options. *BMC Medical Education*, vol. 19, no. 1, p. 354.

[30] Pawluk, S. A., Shah, K., Minhas, R., et al. (2018). A psychometric analysis of a newly developed summative, multiple choice question assessment adapted from Canada

to a Middle Eastern context. *Currents in Pharmacy Teaching and Learning*, vol. 10, no. 8, pp. 1026–1032.

[31] Gupta, P., Meena, P., Khan, A. M., et al. (2020). Effect of faculty training on quality of multiple-choice questions. *International Journal of Applied and Basic Medical Research*, vol. 10, pp. 210–214.

[32] Ali, R., Sultan, A. S., and Zahid, N. (2021). Evaluating the effectiveness of'MCQ development workshop using cognitive model framework: a pre-post study. *Journal of the Pakistan Medical Association*, vol. 71 1 (A), pp. 119–121.

[33] Alamoudi, A. A., El-Deek, B. S., Park, Y. S., et al. (2017). Evaluating the long-term impact of faculty development programs on MCQ item analysis. *Medical Teacher*, vol. 39, no. 1, pp. S45–S49.

[34] Steinert, Y., Mann, K., Anderson, B., et al. (2016). A systematic review of faculty development initiatives designed to enhance teaching effectiveness: a 10-year update: BEME Guide No. 40. *Medical Teacher*, vol. 38, no. 8, pp. 769–786.

[35] Smeby, S. S., Lillebo, B., Gynnild, V., et al. (2019). Improving assessment quality in professional higher education: could external peer review of items be the answer? *Cogent Medicine*, vol. 6, no. 1, 1659746.

[36] AlKhatib, H. S., Brazeau, G., Akour, A., et al. (2020). Evaluation of the effect of items' format and type on psychometric properties of sixth year pharmacy students' clinical clerkship assessment items. *BMC Medical Education*, vol. 20, no. 1, p. 190.