ORIGINAL ARTICLE

Volume 17 Issue 2 2025

DOI: 10.21315/eimj2025.17.2.2

ARTICLE INFO

Submitted: 16-04-2024
Accepted: 07-02-2025
Online: 30-06-2025

The Validity of Multiple-True-False and One-Best-Answer in the Final Professional Undergraduate Medical Examination

Mohd Nasri Awang Besar[1], Karimah Hanim Abd Aziz[2], Heliza Abdul Halim[3]

[1]*Department of Medical Education, Faculty of Medicine, Universiti Kebangsaan Malaysia, Kuala Lumpur, MALAYSIA*

[2]*Community Medicine Department, Kulliyyah of Medicine, International Islamic University of Malaysia, Pahang, MALAYSIA*

[3]*Ophthalmology Department, Kulliyyah of Medicine, International Islamic University of Malaysia, Pahang, MALAYSIA*

To cite this article: Awang Besar MN, Abd Aziz KH, Abdul Halim H. The validity of multiple-true-false and one-best-answer in the final professional undergraduate medical examination. Education in Medicine Journal. 2025;17(2):5–21. https://doi.org/10.21315/eimj2025.17.2.2

To link to this article: https://doi.org/10.21315/eimj2025.17.2.2

ABSTRACT

Multiple-True-False (MTF) and One-Best-Answer (OBA) are commonly used multiple-choice question (MCQ) formats in higher education to evaluate cognitive components. While both evaluate factual knowledge, a well-written OBA question can also stimulate problem-solving and knowledge application in clinical case scenarios. This study aimed to evaluate the convergent and predictive validity of MTF and OBA formats in undergraduate medical education assessments. This cross-sectional study analysed archival records of 143 students sitting for the 2022 final professional undergraduate medical examinations. SPSS version 24.0 was used to enter and analyse the data. Pearson's correlation test was used to assess the convergent validity of MTF and OBA, while a linear regression test was used to evaluate predictive validity. Pearson' s correlation test showed moderate convergent validity (r = 0.25 to 0.6) between the MTF and OBA and other assessment methods. OBA outperformed MTF in predicting key feature question (KFQ) theory assessment ($\beta = 0.40$, $p < 0.01$ vs $\beta = 0.26$, $p < 0.01$), while MTF had stronger predictive validity for clinical components (manned Objective Structured Clinical Examination [OSCE], unmanned OSCE, and modified long case) as compared to OBA ($\beta = 0.43$, $p < 0.01$ vs $\beta = 0.28$, $p < 0.01$). The results are consistent with the literature in that OBA can determine know-how levels compared to MTF. However, a quality improvement exercise must be conducted that focuses on the assessment process of each assessment method, including the assessment blueprint, question structure, examiner calibration, and question vetting. These findings contribute to the enhancement of the quality and validity of assessment practices in medical education.

Keywords: *Multiple-True-False, One-Best-Answer, Assessment, Validity, Malaysia*

CORRESPONDING AUTHOR

Mohd Nasri Awang Besar, Department of Medical Education, Faculty of Medicine, Universiti Kebangsaan Malaysia, Jalan Yaacob Latif, 56000 Cheras, Kuala Lumpur, Malaysia

Email: drmohdnasri@gmail.com

INTRODUCTION

The validity of assessment tools is a critical factor for ensuring the effectiveness of the learning process in the ever-evolving domain of medical education. As Harris et al. (1) have noted, this concept of validity has evolved, shifting from a focus on different types of validity (e.g., content, criterion, and construct) to a more unitary approach that emphasises gathering different forms of evidence to support claims of validity. As Teglasi et al. (2) have described, this contemporary perspective aligns with Dwi Wira's (3) definition of validity, as the extent to which a measurement accurately measures what it intends to measure.

This introduction sets the stage for an in-depth exploration of assessment validity in medical education. By examining the Multiple-True-False (MTF) and One-Best-Answer (OBA) formats through the lens of various validity types, this study aims to provide valuable insights into the field, potentially influencing future assessment strategies in medical education. The findings of this study address a critical gap in theoretical knowledge and performance differences between the MTF and OBA formats and their correlation with student performance outcomes, thereby aiding in the design of more effective assessment strategies.

As Fong et al. (4) highlighted, convergent validity involves examining the correlation between a test and another validated assessment instrument that measures the same construct. This study has demonstrated that an R^2 value of 0.4 aligns with findings that moderate correlations are sufficient to demonstrate meaningful relationships in educational assessments (5). Additionally, multiple linear regression has been applied to determine the predictive validity of individual assessment instruments on overall student performance, using standardised Beta coefficients (β) to measure performance increments and standardised Beta coefficients (β) to compare the dominant roles of assessment tools in predicting overall performance. Predictive validity, on the other hand, as defined by Zach (6), refers to the ability of a test score to predict future scores or outcomes. These forms of validity are instrumental in identifying areas of improvement in students' competencies, directing pedagogical approaches, and assessing the overall efficacy of educational curricula. Moreover, valid assessment instruments are fundamental to establishing both the credibility and reliability of the assessment process.

An effective assessment system in medical education is essential for accurately measuring students' knowledge, skills, competencies, and performance (7). According to Gupta (8), the accuracy of assessments is crucial for ensuring that students are sufficiently equipped to prepare for the demands of the medical profession. To achieve a level of precision and comprehensiveness of assessment, Colbert-Getz (9) have emphasised the importance of integrating a multi-source, multi-method, and multi-purpose approach. Using this method, a unified assessment framework can be established to ensure that a comprehensive assessment of various aspects of a student's educational journey is thoroughly evaluated.

An optimal assessment system is essential to the dynamic landscape of medical education. This study includes a perceptual investigation of the domain of assessment validity. It examines the intricate world of assessment validity, focusing particularly on the formats of OBA and MTF, which are integral components of written assessments. It applies established concepts of validity to assess the efficacy of written examinations, such as the MTF and OBA, towards different assessment methods, such as the manned and unmanned Objective Structured Clinical Examination (OSCE), key feature question (KFQ), and modified long

case (MLC), to measure students' knowledge and clinical reasoning skills. The research question central to this study was as follows: How do the MTF and OBA assessment formats demonstrate different levels of convergent and predictive validity?

This study evaluated the effectiveness of the MTF and OBA assessments in medical education to demonstrate their efficacy in accurately measuring students' knowledge. These assessments fall under the broader category of multiple-choice question (MCQ), which are pivotal in measuring students' comprehension, as highlighted by Puthiaparampil and Rahman (10). A significant finding from Radad et al. (11) has revealed that students scored higher on OBA assessments than on the MTF. Further research supports this viewpoint, because it reduces the impact of guessing, simplifies the scoring process, and evaluates higher-order thinking skills more accurately (12, 13).

However, this study also considered the challenges in the MTF format. As Puthiaparampil and Rahman (10) have reported, MTF assessments could lead to low scores, failures, and poor performance due to the inclusion of false options. This highlights a potential drawback of the MTF format, as it might complicate the decision-making process of students, leading to confusion and errors. In light of these findings, OBA is often considered a superior alternative to MTF in medical education, with better psychometric performance, more balanced student performance, and wider acceptance by medical students and educators (14, 15). This preference for OBA over MTF indicates a shift towards assessment methods that not only test knowledge but also facilitate a deeper understanding and application of medical concepts.

To explore the effectiveness of MTF and OBA assessments, it is essential to address the inherent challenges of both formats. Complementing this, Simbak et al. (16) research indicates that MTF may provide students with greater opportunities for correct answers than OBA does. This attribution could be due to the structure of the MTF questions, which allows for multiple responses and could lead to confusion and the misinterpretation of the assessment's overall accuracy.

MTF questions have traditionally been associated with evaluating rote memorisation of facts (17). A significant shift in the development of scenario-based assessments and innovative scoring methods that reflect Bloom's taxonomy will promote deeper understanding and critical thinking through real-world scenarios, and Anderson (18) has highlighted that well-crafted MTFs could also measure higher cognitive skills, aligning with the highest level of Bloom's taxonomy.

The purpose of medical education is to assess student knowledge, skills, and competencies. It enables educators to determine whether students have achieved their desired learning outcomes and have been prepared for safe clinical practice. Integrating various assessment methods, such as written assessments (MTF, OBA and KFQ) and practical sessions (manned and unmanned OSCE, MLC), can enhance the overall validity and reliability of student assessments in medical education (19–21). Although written assessments are commonly used to assess cognitive aspects, they often lack structure and fail to test problem-solving abilities (22). These methods offer a comprehensive assessment of cognitive and problem-solving skills, ensuring that students are prepared for the competencies required in their prospective medical careers.

A comprehensive and multifaceted assessment approach is crucial to enhance medical education. This approach integrates various assessment tools and methods to appraise students' academic and clinical competencies. The International Islamic University

of Malaysia (IIUM) has exemplified this approach through the integration of multiple instruments of assessment, including written (MTF, OBA, and KFQ) and clinical assessments (manned and unmanned OSCE and MLC), into the Final Professional Examination. Simultaneously, the International Islamic University of Malaysia (IIUM) also offers scenario-based MTF questions to evaluate clinical reasoning and decision-making skills in applied contexts. These questions present progressive clinical case scenarios, followed by multiple statements related to diagnosis, treatment options, and patient management, challenging students to apply their knowledge to real-world situations. Integrating various assessment formats provides a comprehensive assessment of student competencies aligned with the high standards required in the medical profession.

Integrating different assessment instruments into medical education provides a more comprehensive view of student achievement and enhances the validity and reliability of the assessment process. These instruments assess various aspects of knowledge, skills, and behaviour, ensuring a comprehensive assessment of students' competencies (23). However, attaining assessment validity encompasses more than just psychometric assessment. It requires a comprehensive assessment process, incorporating adequate monitoring and training and applying frameworks such as Messick's unified validity framework (1989), which considers numerous contributing factors (24). The validation process is multifaceted (2).

METHODOLOGY

This study was conducted in the Kulliyyah of Medicine, IIUM. It incorporates a cross-sectional and retrospective record review design. This study focused on a cohort of final-year medical students at IIUM 2020/2021, engaging a sample of 143 students who sat for the Final Professional Examination in 2021. The selection was based on a calculated sample size for multivariate analysis with an initial requirement of 90 cases, which was derived from the recommendation of 10 to 15 cases per variable. This was to ensure statistical robustness for the six variables of interest: MTF, OBA, KFQ, and clinical components (manned and unmanned OSCE and MLC). A total of 143 samples were ultimately incorporated in order to improve the statistical robustness and reliability.

A simple random sampling technique was employed to ensure the equitable representation of the student population, which was adequately represented throughout the five months of data collection from October 2022 to February 2023 among final-year medical students. The study's analytical framework was implemented using IBM SPSS Statistics for Windows, version 24.0, with a significance value of $p < 0.05$. Both univariate (Pearson correlation) and multivariate (multiple linear regression) analyses were performed utilising this statistical analysis. The Pearson correlation test was used to determine the convergent validity of the assessment tools, with correlation coefficients interpreted as follows: less than or equal to 0.20 indicates a weak correlation, more than 0.2 but less than 0.8 indicates a moderate correlation, and equal to or more than 0.8 indicates a strong correlation (4). These analyses aimed to evaluate the convergent and predictive validity of the MTF and OBA questions in relation to the KFQ and its clinical components, offering insight into their effectiveness in medical education assessments. This methodological approach, surpassing the initial sample size predictions and utilising a detailed statistical framework, helps ensure a thorough examination of assessment formats, making a substantial contribution to the literature on assessment practices in medical education.

To ensure the clarity and validity of the assessment process, the study considers each department's established learning outcomes and selects appropriate assessment methods, including MTF, OBA, and KFQ. Each department has prepared an assessment blueprint that encompasses all relevant content areas and competencies. This blueprint was documented in a standardised form, and reviewed and approved during a departmental committee meeting. Subject matter experts constructed 60 MTF questions, 60 OBA questions, and 14 KFQ questions, which then underwent a thorough vetting process at both the departmental and faculty levels to provide a clear framework for constructing assessments that are balanced and representative of the curriculum. Examiner calibration was conducted annually for performance assessments, including four manned OSCE stations, four unmanned OSCE stations, and two MLC, ensuring consistency in scoring. Standard setting procedures were employed to determine cut-off scores based on psychometric principles.

Following the assessments, item analysis was performed to evaluate question performance and identify areas for improvement, culminating in an assessment workshop to review the process, discuss item analysis results, and refine strategies. The theoretical assessments (MTF, OBA, and KFQ) were scheduled on different days from the practical assessments (OSCE and MLC) in order to ensure that students were adequately prepared and to reduce test anxiety (Table 1). While the theoretical framework emphasises a structured and meticulous approach, implementing these rigorous procedures poses practical challenges for academics who balance clinical and academic responsibilities. Therefore, fostering a collaborative environment and providing administrative support are both crucial for alleviating the burden on faculty members. This comprehensive and systematic approach ensures the transparency, rigour, and alignment of the assessment process with educational objectives; addresses the reviewer's concerns; and demonstrates a commitment to high-quality medical education assessments.

Table 1: Assessment components and test format

Assessment component	Assessment tools	Number of questions
Theory	MTF	60
	OBA	60
	KFQ	14
Clinical	Manned OSCE	4
	Unmanned OSCE	4
	MLC (observed)	2

RESULTS

The results of this study provide pivotal insights into the performance and correlation patterns of the various test formats. First, the descriptive statistics indicate that the mean mark for the MTF format was 54.5, with a standard deviation (SD) of 6.3, whereas the OBA format scored higher, with an average mark of 66.1 and an SD of 7.8. This indicates higher overall performance in the OBA format than in the MTF format. Delving into the inferential statistical analysis through Pearson's correlation coefficient test has demonstrated the convergent validity of the MTF and OBA in relation to the KFQ and clinical components.

As shown in Table 2, the correlation analysis reveals a moderate correlation between the MTF and OBA assessment tools, with correlation coefficients ranging from 0.25 to 0.60. Specifically, the MTF assessment tool exhibited moderate correlations with the KFQ (r = 0.486), manned OSCE (r = 0.395), and overall clinical components (r = 0.494), suggesting

alignment in measuring similar competencies. The MTF demonstrated a high correlation with the unmanned OSCE (r = 0.600), indicating a strong relationship between the MTF scores and performance in this practical assessment, while showing a weak correlation with the MLC (r = 0.295). The OBA assessment tool also showed a moderate correlation with the KFQ (r = 0.543) and overall clinical components (r = 0.457) but had a low correlation with the manned OSCE (r = 0.307) and the MLC (r = 0.264). Similarly, the OBA scores exhibited a high correlation with the unmanned OSCE scores (r = 0.570). Pearson's correlation test was applied, and the significance level was $p < 0.05$. These findings suggest that the MTF and OBA exhibit varying degrees of correlation with other assessment tools, reflecting their distinct strengths in evaluating specific competencies.

Table 2: Correlation between MTF/OBA with other assessment tools used in undergraduate medical students

	Correlation coefficient® (N = 143)				
Variable	**KFQ**	**Manned OSCE**	**Unmanned OSCE**	**MLC**	**Clinical**
MTF	0.486	0.395	0.600	0.295	0.494
OBA	0.543	0.307	0.570	0.264	0.457

Note: Pearson's correlation test was applied. Significant level $p < 0.05$

Table 3: Predictive values of MTF and OBA towards KFQ, manned OSCE, unmanned OSCE and MLC

Component assessment	Towards another assessment component	B	95% CI (lower, upper)	β	R^2
MTF	KFQ	0.283	0.107, 0.459	0.264	0.343
OBA		0.343	0.200, 0.485	0.394	
MTF	Manned OSCE	0.383	0.1655, 0.601	0.346	0.166
OBA		0.118	-0.058, 0.506	0.295	
MTF	Unmanned OSCE	0.532	0.334, 0.730	0.408	0.439
OBA		0.359	0.199, 0.519	0.340	
MTF	MLC	0.341	0.035, 0.647	0.214	0.101
OBA		0.185	-0.062, 0.433	0.144	

Notes: Multiple linear regression test was performed; CI = Confidence interval; B = Unstandardised Beta-coefficient; β = Standardised Beta-coefficient; $p < 0.05$

Multiple linear regression analysis was employed in this study (Table 3) to evaluate the effects of MTF and OBA on the KFQ and three specific components of the clinical outcome: manned OSCE, unmanned OSCE, and MLC. The results indicated that OBA scores had a stronger association with KFQ outcomes (β = 0.394, $p < 0.05$) than with MTF (β = 0.264, $p < 0.05$). However, MTF demonstrated a higher predictive validity for clinical components, particularly for unmanned OSCE (β = 0.408, $p < 0.05$), compared to OBA (β = 0.340, $p < 0.05$). Additionally, MTF showed significant predictive validity for manned OSCE (β = 0.346, $p < 0.05$) and MLC (β = 0.214, $p < 0.05$), whereas OBA did not show significant predictive validity for these components (β = 0.295 and β = 0.144, respectively, $p > 0.05$). These findings highlight that while OBA is more effective in predicting theoretical knowledge as measured by the KFQ, MTF is more robust in predicting practical competencies and clinical performance.

DISCUSSION

The efficacy of the MTF and OBA examinations as assessment instruments for undergraduate medical students has been demonstrated. This study contributes to the discourse by affirming these assessment instruments and elucidating the intricate aspects of their implementation in the Final Professional Examination. The findings of this study indicate that MTF questions exhibit strong predictive validity for clinical components, as MTF demonstrated higher predictive validity for both unmanned OSCE ($\beta = 0.408$, $p < 0.05$), manned OSCE ($\beta = 0.346$, $p < 0.05$) and MLC ($\beta = 0.214$, $p < 0.05$), compared to OBA, which conversely exhibits higher predictive validity for theoretical knowledge assessment, as evidenced by stronger correlation with KFQ ($\beta = 0.394$, $p < 0.05$). Despite the individual assessments' strengths, neither format comprehensively evaluates clinical and theoretical competencies in medical education. These results are consistent with previous studies, highlighting the need to integrate both assessment formats to complement each other's strengths (25). MTF is highly effective in assessing clinical reasoning and problem-solving skills through structured clinical scenarios, while OBA evaluates theoretical knowledge applications well.

The MTF questions have demonstrated higher predictive validity for clinical assessment supported by the ability to recall knowledge, which aligns with the "know" level of Miller's pyramid of clinical competence (Figure 1). Despite their strengths, the binary true-false structure of MTF questions also increases the likelihood of guessing, thus reducing their reliability. As noted by Puthiaparampil and Rahman (10), the binary format can obscure whether correct responses reflect genuine knowledge or are merely a result of chance. Such limitations necessitate careful consideration in the design and implementation of the MTF questions. However, two main reasons support the strong predictive validity of the MTF questions for the clinical components. First, incorporating clinical scenarios or scenario-based questions offers insights into clinical reasoning and problem-solving abilities. According to Zaidi et al. (26) and Cohen et al. (27), this promotes higher-order thinking. Constructing a progressive clinical case-based question in MTF, followed by multiple statements related to diagnosis, treatment options, and patient management, can induce higher-order thinking and better reflect students' clinical reasoning skills (27). These reasons make MTF indispensable for assessing practical skills within the broader framework of medical education.

The OBA format is a widely used assessment instrument in medical education, as it evaluates both theoretical knowledge and decision-making abilities. Studies by Chauhan et al. (28) confirm the advantages of OBA in promoting a deeper understanding through well-designed questions and eliminating biases associated with guessing. However, despite its strength, OBA has limitations in predicting the performance of clinical components. This study showed that the 95% confidence interval (CI) of the standardised coefficients for manned OSCE and MLC for OBA was zero, demonstrating no statistical significance when using OBA results to predict these assessments directly. This suggests that OBA undermines the efficacy of evaluating the practical assessment outcomes, highlighting its limited usefulness in predicting the performance of clinical components. Other flaws, such as test-wiseness, difficulty index issues, and non-functional distractors (NFD), further reduce robustness and reliability (28). These issues require targeted improvements in OBA design to ensure the accurate assessment of students' competencies. These flaws can make OBA questions less challenging and potentially less effective in assessing true competency, highlighting the need to implement a balanced assessment strategy. Therefore, scenario-based MTF and OBA questions must be integrated into medical education assessments that comprehensively evaluate theoretical and applied knowledge. This balanced approach leverages the strengths

of MTF in clinical reasoning, while retaining OBA's advantages in applied knowledge assessments (26, 27). Furthermore, quality assessment processes, including the proper construction of questions by experts, thorough vetting, and examiner calibration, are essential for alleviating these issues and enhancing the efficacy of OBA.

MTF and OBA question formats are commonly used in medical education assessments. Integrating both assessment formats is essential for a comprehensive evaluation, as these measure student performance in clinical components (29). According to Lahner et al. (30), MTF questions demonstrate superior psychometric results including higher reliability and discrimination indices compared to OBA questions, especially when scored using partial credit methods, such as PS50. This finding aligns with the results of the current study, in which the MTF showed stronger predictive validity for clinical components, highlighting its effectiveness in assessing practical competencies more accurately than the OBA. However, both MTF and OBA formats face common challenges, particularly in terms of question constructions and the limited feedback available to students, which may reduce effectiveness in assessing competencies (10). Despite these challenges, a balanced approach integrating scenario-based MTFs offers superior psychometric results in terms of reliability and discrimination indices, particularly when structured with multiple items per case and combined with OBA, which can effectively alleviate these limitations (31). Thus, the MTF format provides a comprehensive assessment of theoretical knowledge and practical clinical scenarios, whereas the OBA format assesses advanced levels of knowledge essential for clinical practice. The integrated approach of MTF and OBA formats helps to ensure that both theoretical knowledge and clinical performance are evaluated, and at the same time may enhance the psychometric validity of medical education assessments (31).

This study examines the alignment of MTF and OBA in medical education assessments with Miller's pyramid of clinical competence, where MTF is primarily designed to evaluate students at the "knows" level and OBA targets the "knows how" level. For MTF to be valid, the assessment instruments need to demonstrate convergent validity to ensure they correlate with other assessment instruments that evaluate various competency levels, from factual knowledge to practical application. Although MTF and OBA are effective in evaluating lower levels of Miller's pyramid, namely, knows and knows how, they do not adequately assess higher levels of competence, such as "shows how" (performance-based skills) and "does" (real-world application in clinical settings). These higher levels are crucial for evaluating students' ability to perform clinical tasks and interact with patients, requiring assessments like OSCE and MLC designed to evaluate psychomotor and clinical skills. To ensure a comprehensive competency, it is critical to correlate theoretical assessments, such as MTF and OBA, with performance-based assessments, such as OSCE and MLC, as in this study. Furthermore, KFQ assesses the application of knowledge (know how), further enforcing the importance of aligning these formats to evaluate the full spectrum of competencies, from theoretical knowledge to clinical application. By demonstrating the correlation between theoretical assessment (MTF, OBA) and performance-based assessment (OSCE, MLC), this study shows that these formats can measure competencies across all levels of Miller's pyramid. This integrated approach ensures a more comprehensive assessment of theoretical and practical skills, providing a more robust and comprehensive strategy for medical education assessment, as Miller (7) emphasised, encompassing all four levels of competence for a complete assessment of medical students' abilities. By demonstrating the correlation between these assessments and MTF/OBA, this study validates the claim that these formats can accurately measure varying competency levels, thereby establishing convergent validity and enhancing the overall quality and validity of the assessment process.

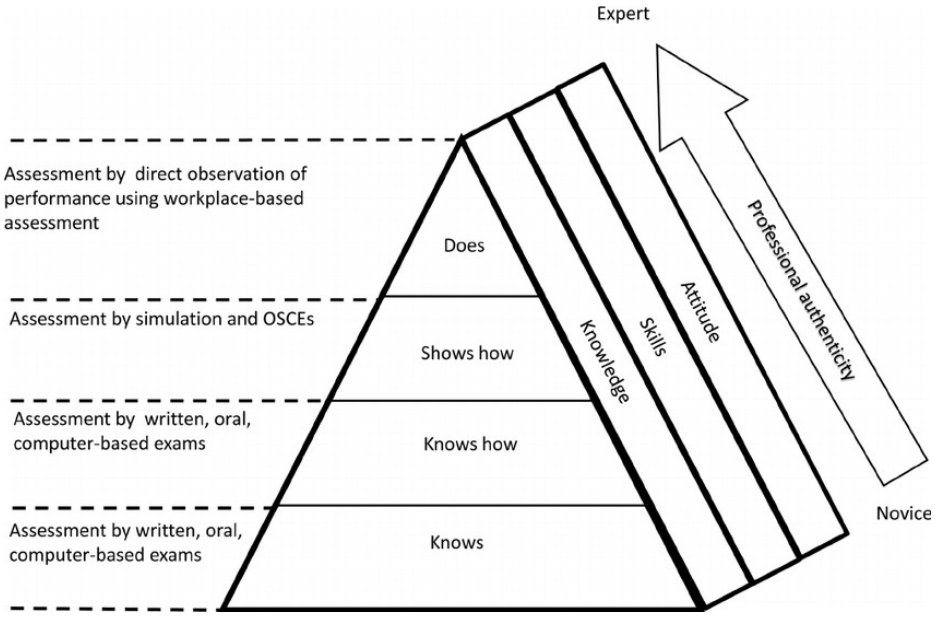

Figure 1: Miller's pyramid of clinical competence. Modified from Miller (7).

The use of MTF in medical education has become increasingly controversial and debatable. One significant concern is the 50–50 probability of guessing the correct answer in MTF, which raises doubts about the validity of MTF assessments (32). For this study, with a mean score of 54.5 in MTF, it is critical to determine whether this score represents students' knowledge or is influenced by the probability of correct guessing inherent in the 50–50 format. These issues provide an incomplete or misleading picture of student competencies, particularly in evaluating higher-order cognitive skills such as critical thinking and problem-solving. Traditional MTF with a binary true-false structure will limit its ability to assess practical competencies, and its reliance on knowledge recall may not fully capture the complexity of clinical decision-making (32). Similarly, OBA's design flaws, such as poor distractor quality, can reduce its effectiveness in assessing true competency (29). Future research should focus on refining the design to improve these issues, including enhancing distractor quality in OBA and minimising the impact of guessing in MTF. Additionally, incorporating advanced psychometric methods such as Item Response Theory (IRT) could enhance the precision of these assessment instruments and provide more valid measures of student competencies (33). Moreover, integrating theoretical assessments, MTF, and OBA formats with performance-based assessments, such as OSCE and MLC, would provide a more comprehensive assessment of medical student competencies, addressing both cognitive and clinical performance (34). Finally, continuous quality assessment process measurement would help to ensure that these assessments align with modern educational objectives and psychometric standards.

OBA questions are increasingly recognised as an effective assessment instrument in medical education due to their ability to evaluate higher-order thinking skills, such as critical reasoning, problem-solving, and clinical decision-making (13). OBA offers a more structured approach than MTF formats by minimising guessing and emphasising knowledge application, making them more reliable instruments for assessing theoretical understanding (28). The use of psychometric tools, such as difficulty index and discriminant index, to ensure continuous refining of item quality, thereby enhancing validity and reliability (35, 36). Furthermore, OBAs provide flexibility in integrating complex clinical scenarios, fostering

deeper learning, and aligning assessment with real-world clinical context (37). However, developing high-quality OBA questions is resource-intensive, requiring significant expertise and time (38). Therefore, effectiveness of OBA or MTF assessment instruments depends on the rigorous assessment process.

The Process of Assessment

It is vital to follow a rigorous and appropriate assessment process to ensure the validity and reliability of assessment instruments. Boatright et al. (39) posited that maintaining assessment validity is a meticulous process that necessitates thorough monitoring and training, highlighting the need for a holistic and comprehensive approach in medical education. It demands an ongoing commitment to monitoring and training, underscoring the comprehensive nature of this endeavour in medical education. Although written assessment instruments are generally valid and reliable, this study echoes the notion that they are not standalone measures of students' proficiency. Instead, they are part of a broader pedagogical and assessment framework that requires a thorough and comprehensive approach to maintaining assessment validity (40). This encompasses a spectrum of tasks, beginning with the preliminary formulation of the assessment plan and proceeding to a rigorous item analysis process at the faculty level, as illustrated in Figure 2.

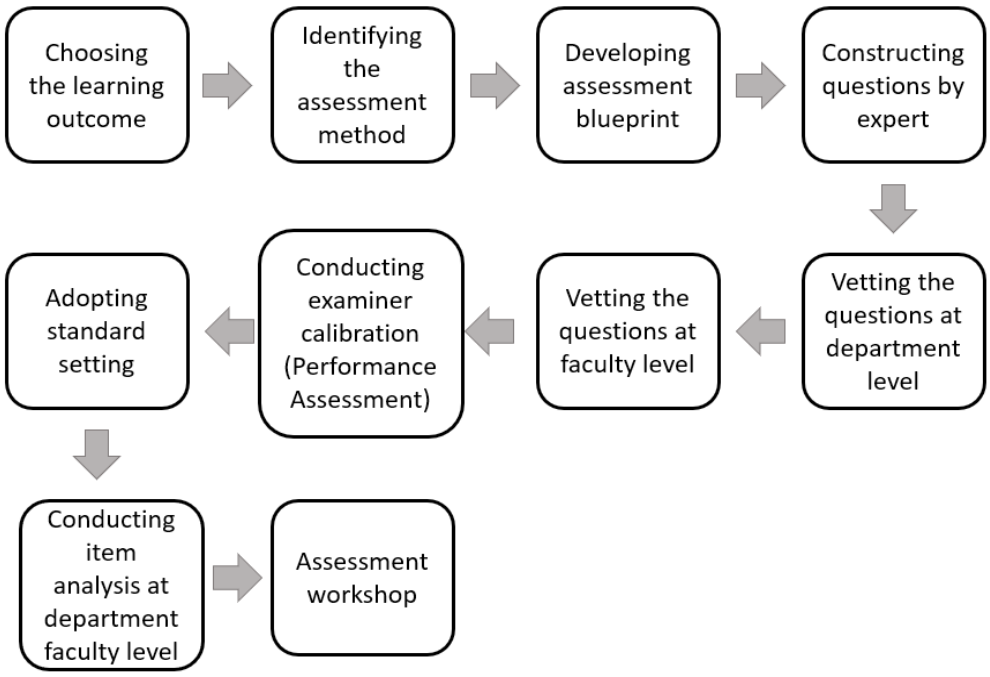

Figure 2: The process of assessment.

The results outlined in this study are consistent with an emerging discourse emphasising the importance of aligning instructional methods with desired learning outcomes, a fundamental principle that guides teaching-learning methodologies, and the development of assessment methods that reflect these outcomes (41). This strategic alignment not only measures student competence but also serves as a diagnostic instrument for curriculum improvement, facilitating precise adjustment of pedagogical methodologies (42). The study outcomes illustrate the necessity of a meticulously planned assessment process

encompassing blueprinting, question construction, examiner calibration, and question vetting in order to maintain the integrity of this alignment.

Various assessment methods have been employed to measure students' progress and to ensure that they achieve their intended learning outcomes. By employing learning taxonomy and Miller's pyramid to categorise learning outcomes, a systematic approach to structuring assessment methods that accurately reflects both cognitive complexity and practical proficiency can be established (43). It is impossible for a single assessment method to effectively evaluate all learning domains equally. MTF and OBA are more suitable for assessing the cognitive domain but are unable to evaluate the psychomotor domain, which is best assessed through performance-based evaluations. Meanwhile, although performance assessments like OSCE can evaluate the cognitive domain, their primary focus should be on assessing students' skills. Therefore, selecting the most suitable assessment method is a crucial early step in the assessment process to ensure it accurately measures what it is intended to assess.

An assessment blueprint or table of specifications is developed to demonstrate that questions are adequately assessed and align with the outcomes. Blueprinting, an essential procedure in assessment design, improves content validity by verifying alignment among curricula, objectives, and topics (44). Each topic in the blueprint should be mapped to outcomes, teaching methodologies, competency levels, learning domain levels, and assessment methods. The selection of topics is based on weightage, which is determined by their importance and may be subjectively influenced by factors such as the frequency or severity of the condition (45). Although a blueprint can be developed by the course coordinator, it should be presented and discussed with others to prevent bias in questions. The assessment blueprint should also represent both theory and performance assessment to ensure adequate sampling across cognitive, psychomotor, and affective domains.

Constructing questions by content experts is another essential aspect of the process, in which the focus is on creating items that accurately measure intended knowledge and skills (36) which improve construct validity. The items of a written test such as MTF and OBA should be constructed or selected based on the pre-determined blueprint. The question author may either constructs a new question, adapts high-quality questions from the question bank, or reconstructs poor-quality questions.

Vetting is the process of reviewing assessment items, including questions, answer schemes, marking forms, and rubrics, to ensure their quality and appropriateness. This includes evaluating the questions' technical accuracy, content level, and language (46). Another important role of question vetting is to ensure alignment with course learning outcomes and teaching and learning experiences, assess the relevance of the assessment method, and maintain an appropriate difficulty level. It is considered good practice for the vetting session to take place at both departmental and faculty levels, creating a dual-layered vetting process. Incorporating a broader educational perspective, including interdisciplinary integration during faculty-level vetting, serves as a valuable cross-check mechanism, especially for ensuring that questions are appropriate for undergraduate or postgraduate levels. Faculty vetting, involving non-experts, plays a crucial role in maintaining assessment integrity. Additionally, too easy or too difficult questions due to question error can introduce construct-irrelevant variance, which could compromise the construct validity of the assessment (47).

Although this article focuses on MTF and OBA, in the context of performance assessment, which requires the examiner to observe students' performance, such as the OSCE, short case, and long case, examination calibration is crucial to improve inter-rater reliability

(48). Examiner judgement can be attributed to variations in candidate marks, and assessing borderline candidates can be highly cognitively demanding (49). Examiner calibration plays a crucial role in aligning expectations across examiners and minimising subjectivity in assessments based on the rubric or marking form. A drawback of assessments involving multiple assessors is the potential for low inter-rater reliability if assessors are not adequately trained (50).

During the calibration process, examiners score a video of students performing a psychomotor task, identify and discuss any scoring discrepancies, particularly outliers, and then rescore the students' performance. This process not only ensures consistency in grading but also allows examiners to familiarise themselves with the assessment criteria, rubrics, and examination procedures.

The role of standard setting exercise is to determine the cut-off score, or pass mark, of an item or a test. It is essential to employ a systematic methodology grounded in validity instead of relying on arbitrary cut-off scores (51). A minor adjustment to cut-off scores can lead to considerable differences in the number of students who pass or fail. There is no universally ideal standard-setting method for all tests, but each test in a specific context likely has a method that is most suitable.

Item analysis was conducted to evaluate the performance of individual assessment items in assessing the quality and effectiveness of the assessment (52). Most theoretical assessments focus on analysing the difficulty index and discrimination index. In contrast, for OBA, analysing non-functioning distractors plays a crucial role in helping question authors refine poorly constructed questions. The author can enhance the quality of assessments by identifying problematic questions and implementing the required adjustments by examining the statistics of these items (53). Another important role of item analysis is to detect potential issues in teaching, learning, and student performance, especially when no flaws are found in the question itself.

Assessment workshops are important to the ongoing enhancement of educational quality. They fulfill dual purposes: they provide an introductory setting for newly appointed faculty members and a refresher for experienced lecturers. These workshops focused not only on providing orientation but also on identifying areas for improvement through item analysis and other forms of critical examination preparation. Research has demonstrated that faculty-level assessment workshops lead to improvements in the quality and content of assessment methods, particularly in terms of developing valid and reliable assessment questions (54). By facilitating a deep dive into assessment items, this workshop empowers faculty members to identify and rectify flaws, comprehend the diverse cognitive levels required by questions, and enhance the assessment quality.

The integration of different assessment methods into medical education is a comprehensive and dynamic process. This involves not only various assessment methods but also careful planning and appropriate assessment processes. This thorough assessment process ensures that the assessments are valid, reliable, and effectively aligned with the educational objectives and learning outcomes of the medical programmes.

In conclusion, this study demonstrates that the MTF and OBA exams are effective assessment instruments in the Final Professional Examination, with the MTF exhibiting better predictive validity. However, assessment validity requires a rigorous assessment process with adequate monitoring and training, rather than relying solely on psychometric measurements.

CONCLUSION

This study revealed that OBA has stronger predictive validity for theoretical knowledge, particularly for KFQ assessments, whereas MTF exhibits stronger predictive validity for clinical components. These results can be attributed to two main factors. First, scenario-based MTF can enhance clinical reasoning skills. Second, OBA may be easier for students due to item flaws, such as test-wiseness, indicating the necessity for further training to construct more robust OBA items. Integrating more assessment methods is crucial to overcoming their weaknesses. This highlights the need for continuous improvement and rigorous construction of assessment tools to optimise their psychometric properties. Further research is suggested to explore long-term impacts and refine question formats to enhance their effectiveness.

Take Home Messages

To enhance the validity and effectiveness of the MTF and OBA in undergraduate medical examinations, we consider the following key points:

a. Balanced evaluation: Both formats assess theoretical knowledge and clinical skills.
b. Question design: Enhance MTF question construction to minimise guessing, ensure each format assesses different competency levels, and incorporate clinical scenarios to evaluate clinical reasoning skills.
c. Continuous improvement: Regularly implement quality improvements, including examiner calibration and question vetting.
d. Faculty training: Offers ongoing training on innovative assessment techniques and the use of MTF and OBA.
e. Advanced analysis: Use advanced item analysis to ensure question effectiveness and reliability.

ACKNOWLEDGEMENTS

The researchers would like to express heartfelt gratitude to the Medical Education Department, Universiti Kebangsaan Malaysia (UKM) and Medical Education Unit, International Islamic University of Malaysia (IIUM), for their invaluable support. This study was conducted without any external funding or grant references.

ETHICAL APPROVAL

This research was approved by the ethics committee at UKM under the project code FF-2022-404. To obtain examination data from IIUM, this research was approved by the ethics committee of IIUM with research ID 914.

REFERENCES

1. Harris P, Bhanji F, Topps M, Ross S, Lieberman S, Frank JR, et al. Evolving concepts of assessment in a competency-based world. Med Teach. 2017;39(6):603–8. https://doi.org/10.1080/0142159X.2017.1315071

2. Teglasi H, Nebbergall AJ, Newman D. Construct validity and case validity in assessment. Psychol Assess. 2012;24(2):464–75. https://doi.apa.org/doi/10.1037/a0026012

3. Yuniahans DWG, Parlika R, Arhinza RS, Majid VF, Alifian MG. Uji validitas aplikasi si-book menggunakan SPSS dengan kombinasi metode R-tabel dan Cohen's Kappa. JTI. 2022;16(2):121–33. https://doi.org/10.47111/jti.v16i2.5001

4. Fong CL, Sidi H, Mahady ZA, Mohamad N, Saini SM, Saiful M, et al. The validity of the assessment methods in an undergraduate psychiatry module examination among a sample of fourth-year Malaysian medical students. Int Med J. 2012;19(4):347–51.

5. Malapati A, Murthy NLB. Performance of students across assessment methods and courses using correlation analysis. In: 2013 IEEE International Conference in MOOC, Innovation and Technology in Education (MITE). Jaipur, India: IEEE; 2013. p. 325–8. https://doi.org/10.1109/MITE.2013.6756359

6. Zach S. Predictive validity in educational and psychological research. 4th ed. US: John Wiley & Sons; 2022.

7. Miller GE. The assessment of clinical skills/competence/performance. Acad Med. 1990;65(9 Suppl):S63–7. https://doi.org/10.1097/00001888-199009000-00045

8. Gupta S. Authentic assessment in medicine. J Postgrad Med Educ Res. 2019;53(1):42–4. https://doi.org/10.5005/jp-journals-10028-1311

9. Colbert-Getz JM, Shea JA. Three key issues for determining competence in a system of assessment. Med Teach. 2021;3(7):853–5. https://doi.org/10.1080/0142159X.2020.1804540

10. Puthiaparampil T, Rahman M. Very short answer questions: a viable alternative to multiple choice questions. BMC Med Educ. 2020;20(1):1–8. https://doi.org/10.1186/s12909-020-02057-w

11. Radad K, Taha M, Rausch WD. Multiple choice questions versus very short answered questions in the evaluation of students of veterinary pathology. Rev Esp Edu Med. 2022;4(1). https://doi.org/10.6018/edumed.548861

12. Brassil CE, Couch BA. Multiple-true-false questions reveal more thoroughly the complexity of student thinking than multiple-choice questions: a Bayesian item response model comparison. IJ STEM Ed. 2019;6(1):16. https://doi.org/10.1186/s40594-019-0169-0

13. Abdul Rahim AF, Simok AA, Abdull Wahab SF. A guide for writing single best answer questions to assess higher-order thinking skills based on learning outcomes. Educ Med J. 2022;14(2):111–24. https://doi.org/10.21315/eimj2022.14.2.9

14. Sam AH, Field SM, Collares CF, Van Der Vleuten CPM, Wass VJ, Melville C, et al. Very-short-answer questions: reliability, discrimination and acceptability. Med Educ. 2018;52(4):447–55. https://doi.org/10.1111/medu.13504

15. Van Wijk EV, Janse RJ, Ruijter BN, Rohling JHT, Van Der Kraan J, Crobach S, et al. Use of very short answer questions compared to multiple choice questions in undergraduate medical students: an external validation study. PLoS ONE. 2023;18(7):e0288558. https://doi.org/10.1371/journal.pone.0288558

16. Simbak N, Aung MMT, Ismail S, Jusoh NM, Ali T, Yassin WA, et al. Comparative study of different formats of mcqs: multiple true-false and single best answer test formats, in a new medical school of Malaysia. 2015 April. https://doi.org/10.13140/RG.2.1.2228.6881

17. Ingale AS, Giri PA, Doibale MK. Study on item and test analysis of multiple-choice questions amongst undergraduate medical students. Int J Community Med Public Health. 2017;4(5):1562. https://doi.org/10.18203/2394-6040.ijcmph20171764

18. Anderson J. Medical teacher 25th anniversary series multiple-choice questions revisited. Med Teach. 2004;26(2):110–3. https://doi.org/10.1080/0142159042000196141

19. Becker A, Nekrasova-Beker T. Investigating the effect of different selected-response item formats for reading comprehension. Educ Assess. 2018;23(4):296–317. https://doi.org/10.1080/10627197.2018.1517023

20. Somannavar M. Constructed response items as an assessment method for undergraduate medical course: improving the validity. J Sci Soc. 2019;46(1):8. https://doi.org/10.4103/jss.JSS_36_13

21. Özkaya G, Aydin MO, Alper Z. Distance education perception scale for medical students: a validity and reliability study. BMC Med Educ. 2021;21(1):400. https://doi.org/10.1186/s12909-021-02839-w

22. Schiekirka S, Raupach T. A systematic review of factors influencing student ratings in undergraduate medical education course evaluations. BMC Med Educ. 2015;15(1):30. https://doi.org/10.1186/s12909-015-0311-8

23. Wan Mohamad Akil WFH, Mohd Matore ME@ E. The 21st century assessment strategies in medical laboratory education: sharing experiences between two higher institutions in Malaysia. IJARBSS. 2023;13(7):107–30. https://doi.org/10.6007/IJARBSS/v13-i7/17196

24. Messick S. Validity. 3rd ed. New York: Macmillan Publishing Co.; 1989.

25. Schuwirth LWT, Van Der Vleuten CPM. Programmatic assessment: from assessment of learning to assessment for learning. Med Teach. 2011;33(6):478–85. https://doi.org/10.3109/0142159X.2011.565828

26. Zaidi NLB, Grob KL, Monrad SM, Kurtz JB, Tai A, Ahmed AZ, et al. Pushing critical thinking skills with multiple-choice questions: does Bloom's taxonomy work? Acad Med. 2018;93(6):856–9. https://doi.org/10.1097/ACM.0000000000002087

27. Cohen Aubart F, Lhote R, Hertig A, Noel N, Costedoat-Chalumeau N, Cariou A, et al. Progressive clinical case-based multiple-choice questions: an innovative way to evaluate and rank undergraduate medical students. La Revue de Médecine Interne. 2021;42(5):302–9. https://doi.org/10.1016/j.revmed.2020.11.006

28. Chauhan GR, Chauhan BR, Vaza JV, Chauhan PR. Relations of the number of functioning distractors with the item difficulty index and the item discrimination power in the multiple choice questions. Cureus. 2023. https://doi.org/10.7759/cureus.42492

29. Oki O, Naqvi Z, Jordan W, Guilliames C, Archer-Dyer H, Santos MT. Evaluating student clerkship performance using multiple assessment components. PRiMER. 2024;8. https://doi.org/10.22454/PRiMER.2024.160111

30. Lahner FM, Nouns ZM, Sören Huwendiek, Huwendiek S. Multiple true–false items: a comparison of scoring algorithms. Adv Health Sci Educ. 2018; 23:455–63. https://doi.org/10.1007/S10459-017-9805-Y

31. O'Reilly T, Sabatini J, Wang Z. Using scenario-based assessments to measure deep learning. In: Deep comprehension. 1st ed. Routledge: Taylor and Francis; 2018. https://doi.org/10.4324/9781315109503-16

32. Puthiaparampil T, Gudum HR, Rahman MM, Saimon R, Lim IF. True-false analysis reveals inherent flaws in multiple true-false tests. Int J Community Med Public Health. 2019;6(10):4204. https://doi.org/10.18203/2394-6040.ijcmph20194476

33. Lahner FM, Schauber S, Lörwald AC, Kropf R, Guttormsen S, Fischer MR, et al. Measurement precision at the cut score in medical multiple-choice exams: theory matters. Perspect Med Educ. 2020;9(4):220–8. https://doi.org/10.1007/S40037-020-00586-0

34. Wilkinson TJ, Frampton CM. Comprehensive undergraduate medical assessments improve the prediction of clinical performance. Med Educ. 2004;38(10):1111–6. https://doi.org/10.1111/j.1365-2929.2004.01962.x

35. Shakurnia A, Ghafourian M, Khodadadi A, Ghadiri A, Amari A, Shariffat M. Evaluating functional and non-functional distractors and their relationship with difficulty and discrimination indices in four-option multiple-choice questions. Educ Med J. 2022;14(4):55–62. https://doi.org/10.21315/eimj2022.14.4.5

36. Murias Quintana E, Rodríguez Castro J, Sánchez Lasheras F, Vega Villar J, Curbelo García JJ, Cadenas Rodríguez M, et al. Improving the ability to discriminate medical multiple-choice questions through the analysis of the competitive examination to assign residency positions in Spain. BMC Med Educ. 2024;24(1):367. https://doi.org/10.1186/s12909-024-05324-2

37. Gardner NP, Gormley GJ, Kearney GP. Is there ever a single best answer (SBA): assessment driving certainty in the uncertain world of GP? Educ Prim Care. 2023;34(4):180–3. https://doi.org/10.1080/14739879.2023.2243447

38. Walsh J, Harris B, Tayyaba S, Harris D, Smith P. Student-written single-best answer questions predict performance in finals. Clin Teach. 2016;13(5):352–6. https://doi.org/10.1111/tct.12445

39. Boatright D, Edje L, Gruppen LD, Hauer KE, Humphrey HJ, Marcotte K. Ensuring fairness in medical education assessment. Acad Med. 2023;98(8S):S1–2. https://doi.org/10.1097/ACM.0000000000005244

40. Mohamad Tahir AI, Shariffuddin FY, Lichyn L, Ng LY, Ling W, Nagandla K, et al. Validity of medical students self-assessment of proficiency in clinical long case examination. Malays J Med Health Sci. 2022;18(5):41–6. https://doi.org/10.47836/mjmhs.18.5.7

41. Naeem SS, Roy V. Use of students' learning outcomes as a tool for changing teaching content and methodology: assessment of impact. MAMC J Med Sci. 2022;8(1):26. https://doi.org/10.4103/mamcjms.mamcjms_116_21

42. Kushari B, Septiadi L. A learning outcome assessment information system to facilitate outcome-based education (OBE) implementation. J Pendidik Teknol Kejuru. 2022;28(2):238–50. https://doi.org/10.21831/jptk.v28i2.42339

43. Modi RN, Kavya PK, Poddar R, Natarajan S. Question classification based on cognitive skills of Bloom's taxonomy using TFPOS-IDF and GloVe. In: Noor A, Saroha K, Pricop E, Sen A, Trivedi G, editors. Proceedings of emerging trends and technologies on intelligent systems. Singapore: Springer Nature; 2023. p. 25–37. https://doi.org/10.1007/978-981-19-4182-5_3

44. Schuwirth LWT, Van Der Vleuten CPM. How to design a useful test: the principles of assessment. In: Swanwick T, editor. Understanding medical education: evidence, theory and practice. 1st ed. UK: Wiley-Blackwell; The Association for the Study of Medical Education (ASME); 2010. p. 195–207. https://doi.org/10.1002/9781444320282.ch14

45. Kenwright DN, Wilkinson T. Quality in medical education. In: Swanwick T, Forrest K, O'Brien BC, editors. Understanding medical education: evidence, theory, and practice. 3rd ed. Hoboken, NJ: Wiley-Blackwell; 2018. p. 101–10. https://doi.org/10.1002/9781119373780.ch7

46. Gopalakrishnan S. Question vetting: the process to ensure quality in assessment of medical students. J Clin Diagn Res. 2014;8(9):1–3. https://doi.org/10.7860/JCDR/2014/9914.4793

47. Downing SM, Haladyna TM. Validity threats: overcoming interference with proposed interpretations of assessment data. Med Educ. 2004;38(3):327–33. https://doi.org/10.1046/j.1365-2923.2004.01777.x

48. Hefti AF, Preshaw PM. Examiner alignment and assessment in clinical periodontal research. Periodontol 2000. 2012;59(1):41–60. https://doi.org/10.1111/j.1600-0757.2011.00436.x

49. Malau-Aduli BS, Hays RB, D'Souza K, Smith AM, Jones K, Turner R, et al. Examiners' decision-making processes in observation-based clinical examinations. Med Educ. 2021;55(3):344–53. https://doi.org/10.1111/medu.14357

50. Patterson F, Ferguson E, Zibarras L. Selection into medical education and training. In: Swanwick T, Forrest K, O'Brien BC, editors. Understanding medical education: evidence, theory, and practice. 3rd ed. Hoboken, NJ: Wiley-Blackwell; 2018. p. 375–88. https://doi.org/10.1002/9781119373780.ch26

51. Alshawwa L. Standard setting: a review of methods. Asian J Educ Soc Stud. 2023;42(2):1–7. https://doi.org/10.9734/ajess/2023/v42i2909

52. Nojomi M, Mahmoudi M. Assessment of multiple-choice questions by item analysis for medical students' examinations. Res Dev Med Educ. 2022;11(1):24. https://doi.org/10.34172/rdme.2022.024

53. Yahia AIO. Post-validation item analysis to assess the validity and reliability of multiple-choice questions at a medical college with an innovative curriculum. Natl Med J India. 2021;34(6):359–62. https://doi.org/10.25259/NMJI_414_20

54. Rahim MF, Qassim Bham S, Khan S, Ansari T, Ahmed M. Improving the quality of MCQs by enhancing cognitive level and using psychometric analysis. Pak J Health Sci. 2023;115–21. https://doi.org/10.54393/pjhs.v4i04.700