Advances in Medical Education and Practice

Dovepress
Taylor & Francis Group

Open Access Full Text Article

ORIGINAL RESEARCH

Postgraduate Clinical Residency: The Impact of Multiple-Choice Question Quality on Exam Success Rates

Omer Eladil Abdalla Hamid Mohammed[1,2], Suresh Kumar Srinivasamurthy[3], Raghavendra Bhat[4], Yasir Ahmed Mohamed Alhassan Eltahir[5], Hesham Amin Hamdy Elshamly[6], Fatima Mohammed[7], Bashir Hamad[8]

[1]Department of Neurology & Medical Education, Ras Al Khaimah College of Medical Sciences, Ras Al Khaimah Medical & Health Sciences University, Ras Al Khaimah, United Arab Emirates; [2]Department of Medicine, International University of Africa, Sudan Medical Specialization Board, Khartoum, Sudan; [3]Department of Pharmacology, Ras Al Khaimah College of Medical Sciences, Ras Al Khaimah Medical & Health Sciences University, Ras Al Khaimah, United Arab Emirates; [4]Department of Internal Medicine, Ras Al Khaimah College of Medical Sciences, Ras Al Khaimah Medical & Health Sciences University, Ras Al Khaimah, United Arab Emirates; [5]Goldfarb School of Nursing at Barnes-Jewish College, St. Louis, MO, USA; [6]Department of Surgery, Ras Al Khaimah College of Medical Sciences, Ras Al Khaimah Medical & Health Sciences University, Ras Al Khaimah, United Arab Emirates; [7]Faculty of Arts and Sciences, Qassim University, Buraydah, Kingdom of Saudi Arabia; [8]Faculty of Medicine, International University of Africa, Sudan Medical Specialization Board, Khartoum, Sudan

Correspondence: Omer Eladil Abdalla Hamid Mohammed, Email omereladil@rakmhsu.ac.ae

Objective: To examine the impact of quality parameters in the construction of multiple-choice questions (MCQs) and their associated psychometric analysis for a selected Specialty X (SpX) in the Qualifying Residency Entry Exam (QRE) at a Postgraduate Medical Institute.

Methods: A post-validation cross-sectional analytical study was conducted using a non-probability purposive judgmental sampling technique. The SpX was chosen from one clinical specialities with the lowest exam success rates among the 52 specialities in the 2020–2023 QRE cycles. MCQs were evaluated using standard item analysis parameters: questions were considered acceptable range if they had a difficulty index (DIF) between 0.30–0.70, a discrimination index ≥0.2, and at least two functioning distractors.

Results: Out of 175 candidates who appeared for the QRE, only 19 (10.86%) passed. The exam included 120 A-type MCQs, with just 7 (5.8%) flaw-free Items/Questions. Most questions (98.3%) lacked clinical vignettes, and only 10% used the proper lead-in format. Two-thirds failed the "cover-the-options" test, and 40% showed constructional flaws related to testwiseness or irrelevant difficulty. Psychometric analysis showed a mean difficulty index of 45.9, with 86.7% of Items/Questions in the acceptable range. However, 15% had extremely poor discrimination (mean PBS = 0.17), and the mean distractor efficiency was 66%. A statistically significant relationship ($p < 0.05$) was observed between constructional flaws and DIF, DisI/PBS, the Horst Index, and Bloom's levels. Furthermore, no significant relationship was identified between the exam success rate and the type of MBBS curriculum.

Conclusion: The quality of Items/Questions in the postgraduate residency significantly impacted the QRE. Other potentially influential factors require further multivariate analytical research. This highlights the need for strategic educational initiatives to enhance Exam Bank development, strengthen capacity building, and improve faculty assessment skills.

Keywords: postgraduate assessment, multiple choice questions, constructional flaws, psychometric analysis

Introduction

The World Federation for Medical Education (WFME) has set global standards to improve the quality and uniformity of medical education worldwide. The standards developed for postgraduate medical education serve as a framework for quality improvement that ensures that residency examinations and assessments are conducted with the highest levels of quality. The "Assessment of Trainees" is one of these standards, where the reliability and validity of assessment methods should be documented and evaluated.[1] Additionally, the assessment of any learner is a critical component of instruction

Received: 26 March 2025
Accepted: 24 July 2025
Published: 6 August 2025

and curriculum goal achievements,[2] using different assessment tools, each of which has its pros and cons. Of these, multiple-choice questions (MCQs) are the most popular. The well-constructed MCQ examinations are arguably the most reliable, valid, and cost-effective method among written assessment tools for medical knowledge and psychomotor domains.[3]

This postgraduate study is one of the few international efforts focused on quality in constructed stimulus-response exams. It evaluates the quality and effects of MCQ creation in a postgraduate Specialty MD Qualifying Entry Exam that has one of the lowest exam success rates. The research utilized the Judgmental Sampling Method. Passing this Entry Exam is a prerequisite for candidates before they can begin the necessary training for their specialist degree. Item analysis is an essential, statistically based quality measure used to scientifically evaluate the quality of the MCQs internationally. One of the significant challenges facing exam constructors is the effort to generate highly valid, reliable, acceptable, and feasible exams with educational impacts. Properly constructed MCQs are advocated for building exam banks for both high- and low-stakes exams.

The study's main objectives were to examine the MCQ Items/Questions and test statistics of the chosen entry exam, which includes identifying constructional item flaws and assessing Bloom's taxonomy level of the MCQs. Additionally, the study calculated four primary quality item indices: reliability, the Difficulty Index (DIF), the Discriminating index/ Point Biserial (DisI/PBS), and the mean distractor efficiency (mDE). The study also evaluated the statistical significance of the candidates'/examinees' university curriculum type. For confidentiality, the identity of the selected Programme is referred to as Specialty Programme X (SpX).

Assessment in Medical Education

Assessment is an essential part of candidates' learning. Importantly, it drives learning. Therefore, the assessment tools should be high-quality, valid, reliable, feasible, and acceptable for educational purposes. Those measures crystalize the required competencies in real academic and health practice and reflect various achievement levels.[4] According to Epstein, many postgraduate training programs and licensing bodies have introduced new initiatives to ensure accurate, reliable, and timely assessments of trainee competence.[4]

Assessment Methods

Every assessment method has its own strengths and inherent limitations, whether in written evaluations or assessments conducted by supervising clinicians. Van Der Vleuten identifies five key criteria for evaluating the quality of any assessment method. Reliability refers to the accuracy and reproducibility of the measurement, while validity ensures that the assessment measures what it is intended to assess. The educational impact considers how the assessment influences future learning and clinical practice. Acceptability pertains to how well it is received by both learners and faculty. Lastly, feasibility and cost examine the practicality and financial implications for trainees, institutions, and society as a whole.[5] The assessment tools vary significantly according to the tested domains: knowledge, attitude, or psychomotor. Thus, although each instrument has pros and cons, they overlap, and there is no consensus on the "best" assessment methods.[4]

The standard setting involves determining the necessary level of knowledge and skills for adequate performance and then identifying a corresponding score on the examination scale that reflects this standard.[6] It serves as the concluding phase of the testing process, establishing what acceptable performance looks like. The minimum pass level (MPL) marks the score below which candidates are considered to have failed. The Angoff method and its variations and the Nedelsky method are commonly utilized for criterion-referenced standard setting. A central assessment committee (CAC) should oversee the quality of exams and the standard-setting process to ensure equitable pass-or-fail outcomes. Additionally, the CAC contributes to the creation of an Examination Bank, which supports educators in developing high-quality exam Items/Questions that align with learning outcomes, utilizing insights from candidates, families, and teaching staff.

Exam Blueprinting for the Assessment and Alignment

The term "blueprint" is derived from architecture and signifies the need for an assessment process to follow a structured, replicable plan. One of its key objectives is to mitigate two major threats to validity. The first construct, under-representation, occurs when the assessment lacks sufficient coverage of key topics due to undersampling or biased

evaluation. The exam blueprint should be drawn up as a matrix to cover all topics without under-representation. The second is construct-irrelevant variance, where the scores are influenced by the content immaterial to the intended construct. Thus, blueprint matrix emphasizes the importance of the alignment between the three main processes: what is stated in the curriculum/course as expected intended learning outcomes (ILOs), what has been taught/studied and what is to be assessed.[7]

Multiple-Choice Questions (MCQs) as an Assessment Tool

Schuwirth and van der Vleuten (2004) classified written assessment techniques into two categories: (i) based on stimulus format, which refers to what the Items/Questions ask, and (ii) based on response format, which determines how answers are recorded. MCQs are widely used for assessment due to their high content validity, allowing them to cover a broad range of topics. Additionally, they can be administered within a relatively short time frame and efficiently graded using computer-based systems. MCQs primarily assess knowledge recall and comprehension, which are lower levels of Bloom's taxonomy. However, well-constructed MCQs can also evaluate higher-order cognitive skills, including application, analysis, synthesis, evaluation, and creation. Developing a high-quality MCQ-based examination is a complex and time-intensive academic task. Nevertheless, MCQs are often preferred over other assessment tools because they are objective and minimize human bias, as they can be reliably scored either manually or through automated correction systems. The USA National Board of Medical Examiners (NBME), in their book "Constructing Written Test Questions for the Basic and Clinical Sciences", classifies MCQs into two primary types: multiple-response Items/Questions and single-response Items/Questions (one-best-answer format).[2]

Best of Four MCQ Format and Its Prevalence in Postgraduate Medical Assessments

The most commonly used format is the A-type (one-best-answer), which clearly specifies the required number of choices. These questions typically consist of three key components: (i) a stem (often a clinical scenario), (ii) a lead-in question, and (iii) a set of response options, including one correct answer and three to four distractors. Research suggests that using three answer choices may enhance reliability compared to formats with four or five options.[8,9]

In the Membership of the Royal Colleges of Physicians of the United Kingdom (MRCP[UK]), the Royal College of General Practitioners (MRCGP) MCQs are used extensively in Part 1 and Part 2 Written Examinations. Yet, they are the best of Five[10] Similarly, the Australian Medical Programme (AMC)[11] as well UMSLE Step Exams.[12] These organizations have adopted the MCQ-SBA format for its validity, reliability, and practicality, allowing broad coverage of clinical content and high discriminatory power in differentiating between competent and underperforming candidates.

Multiple Choice Questions Flaws

The NBME also identifies two types of common flaws in MCQs, as summarised in Appendix A.

Flaws related to test-wiseness – These allow some candidates to answer correctly based on test-taking strategies rather than actual knowledge. Flaws related to irrelevant difficulty – These make the question unnecessarily challenging for reasons unrelated to the knowledge or skill being assessed.[2] Testwiseness flaws make items easier for strategic guessers, while irrelevant difficulty makes items harder even for knowledgeable students. Both reduce validity, weaken discrimination, and compromise test fairness.

Measures for Quality Item Construction

In addition to being free from the flaws mentioned above, the Item should fulfil the Cover-the-option test: the item is to be answered without looking at the options. The stem should include as much as possible, but not superfluous information, preferably in a scenario format that resembles real life to ensure the application of the Entrustable Professional Activities. Options should be short, grammatically consistent, and logically compatible with the stem. Options are to be listed in logical or alphabetical order. Avoid using negatively phrased Items/Questions (such as "except" or "not") in the lead-in. The lead-in should be in question format.

Item Construction: Bloom's Classification

In 1956, Dr. Benjamin Bloom introduced a framework of learning levels arranged in a hierarchy, ranging from lower-order skills (such as memorization) to higher-order skills (such as evaluation and creation). Bloom and his collaborators outlined six main categories: Knowledge, Comprehension, Application, Analysis, Synthesis, and Evaluation. Bloom's taxonomy serves as a valuable tool in assessment, ensuring alignment between the content being taught, the objectives outlined, and the material included in the exam, all while using the appropriate assessment tools and structure.[13]

Tests Statistics and Analysis of Performance

Item Analysis is a process carried out post-exam to thoroughly evaluate each test item (question) from that exam, assessing the strengths and weaknesses of the Items/Questions used and implementing necessary corrections. This approach serves to evaluate the overall quality of the examination by analysing its components (Items/Questions). It aims to provide insight into how effectively the examination has met its goals. In postgraduate higher education, most examinations primarily function as assessment tools for measuring candidates' achievements, thereby guiding and supporting future learning. There are three major types of item analysis: Classical Test Theory, Item Response Theory, and Rasch Measurement.[14] The classical test theory is commonly used, and its Item Analysis Indicators and Formulas are given in Appendix B.

Difficulty Index (DIF)

The difficulty index, also known as the facility index or p-value, measures the difficulty of an item with a single correct answer worth one point. It represents the percentage of examinees or candidates who answer the item correctly, essentially reflecting the item's mean score.[14,15] The difficulty index ranges from 0 to 100, with higher values indicating an easier item. If an item has alternatives worth more than one point or multiple correct answers, the difficulty index is calculated as the average score divided by the highest possible score for any alternative.

Item difficulty is important for assessing whether candidates have mastered the concept being tested. An accepted difficulty range is between 0.2 and 0.8, although other studies suggest varying acceptable ranges.[14,16,17] For instance, the Hong Kong International Databases for Enhancement of Assessments considers a difficulty level between 50–75 to be optimal.[18]

The classification of difficulty levels is based on value ranges. A Very Difficult (VD) category, with a value below 20, is considered not acceptable The Difficult (D) range, from 20 to 40, represents the acceptable upper level. An Average (AV) score of 41 to 60 is considered excellent. The Easy (ES) category, ranging from 61 to 80, falls within the acceptable lower level. Lastly, a Very Easy (VE) value, above 80, is also not acceptable.[19]

Discrimination Index and Point Biserial

The Discrimination Index (DIS) measures how well an item distinguishes between high- and low-performing examinees. It is calculated using Kelly's method, which compares correct responses from the top and bottom 27% of test-takers:

DIS = (Ru − Rl) / (½T), where Ru is the number of students in the upper group who answered the item correctly, Rl is the number in the lower group who did so, and T is the total number of students in both groups. A DIS value above 0.30 is considered acceptable, with values closer to 1.0 indicating excellent discrimination.[14,16–18,20]

Point-biserial correlation coefficient is often denoted as r_pb or PBS in the literature—both abbreviations refer to the same statistic. Both the terms—PBS, r_pb, are interchangeable and indicate the degree to which performance on a single test item correlates with overall test performance. PBS assesses the relationship between an examinee's response to a single item and their overall test performance.

The Point-biserial Correlation (PBS) represents the Pearson correlation between responses to a specific item and overall test scores. In contrast, the Biserial Correlation simulates item responses to depict the stratification of a normal distribution and computes the correlation based on that model. The Point-Biserial correlation, denoted as rpbis, ranges from +1 to −1. Essentially, PBS connects an examinee's scores on individual Items/Questions with their total raw scores on the test.

A value of ≥ 0.20 is generally acceptable Items with r_pb ≥ 0.40 are considered Very Good (VG), 0.30–0.39 as Good (G), 0.20–0.29 as Below Standard (BS), and ≤ 0.19 as Poor (P).[21]

Distractor Efficiency (DE)

A distractor is an incorrect answer option designed to mislead less experienced test-takers. Effective distractors must be plausible, as their design influences examinee performance. Distractor Efficiency (DE) measures whether distractors effectively divert candidates from the correct answer. There are two types: Functional Distractors (FD), chosen by over 5% of examinees, and Non-Functional Distractors (NFD), selected by less than 5%.[22] DE is determined by the number of Non-Functional Distractors (NFDs)—incorrect options selected by less than 5% of candidates. When an item has three NFDs, its DE is 0%, indicating poor question quality. If there are two NFDs, the DE is 33.3%. A question with one NFD achieves a DE of 66.6%, while an item with no NFDs (all distractors functioning effectively) reaches the highest DE of 100%, signifying an optimally designed question.

Horst Index (HI)

The Horst Index is a statistical measure used to improve the reliability of test Items/Questions. Introduced by Horst in the 1950s, it calculates the difference between the number of candidates selecting the correct answer and those choosing the most frequently selected distractor, relative to the total number of candidates who attempted the item.[23] The resulting values range from −1 to +1. Unlike the discrimination index, the HI evaluates Items/Questions in a unique way, making it useful for identifying potential quality issues. A negative HI signifies that a greater number of candidates preferred a specific incorrect answer (the most popular distractor) over the correct one. This suggests that there may be flaws in the item or question itself (or that the Item is outdated, with the best answer having changed) or it could indicate issues in teaching or understanding. The primary concern arises when a significant proportion of the upper-scoring group selects the most popular distractor rather than the correct option.

Reliability (Internal Consistency)

Reliability indicates the consistency of obtained measurements, with the desired level varying by examination type. For Type A multiple-choice questions (MCQs), a reliability coefficient over 0.90 is ideal, while multiple-mark and short-answer Items/Questions typically have a reliability range of 0.65 to 0.80. Longer essay-type and practical examinations can have lower reliability, around 0.40, without raising issues. Standard formulas for calculating reliability coefficients include Cronbach's alpha and Backhouse's specific alpha for optional Items/Questions, with a reliability level of 0.80 or higher deemed reliable.[14,20]

Quality MCQ

A quality MCQ is one that effectively differentiates between high- and low-performing examinees while including functioning distractors. In this study, item quality was evaluated using standard criteria based on Classical Test Theory. An item was considered acceptable if it had a difficulty index (p) between 30% and 70%, indicating a balanced level of challenge. The discrimination index (DIS) was deemed satisfactory if it was ≥ 0.2, reflecting the item's ability to distinguish between stronger and weaker students. Distractor efficiency (DE) was considered adequate when at least two distractors were selected by ≥5% of examinees, indicating a minimum efficiency of 50%. Additionally, a point-biserial correlation (r_pb) above 0.2 was used as a benchmark for acceptable item discrimination.[24]

Quality Test

Whereas a quality test is characterized by an appropriate balance of item difficulty, sufficient discriminatory power, and strong internal consistency. In this study, the overall test quality was assessed using several key indicators. The mean difficulty index was expected to fall within the range of 0.4 to 0.6, indicating a moderate level of difficulty suitable for most examinees. A mean discrimination index greater than 0.2 was considered indicative of a test capable of effectively differentiating between high and low performers. To evaluate internal consistency and reliability, the Kuder-Richardson

Formula 20 (KR-20) or Cronbach's Alpha (as applicable) was used, with values of ≥0.7 regarded as acceptable benchmarks. The selected test-level indices in Classical Test Theory is given in Appendix C.

This study is grounded in Classical Test Theory (CTT), a widely used psychometric approach for evaluating the quality of test items. Under CTT, key indicators such as difficulty index, discrimination index, distractor efficiency, and point-biserial correlation are used to assess individual question quality and overall test performance. The four specific objectives were established: To examine trends in performance across years of study and academic years in the SpX Entry Qualifying Exam; to assess the structural quality and cognitive level of multiple-choice questions using item-writing standards and Bloom's taxonomy; to evaluate the psychometric properties of exam items using Classical Test Theory (CTT) indices such as difficulty, discrimination, and point-biserial correlation, and to analyze their relationship with item flaws and cognitive levels; to explore the association between candidate performance and curricular models adopted.

Methods

Study Design: This study employed a cross-sectional analytical approach to conduct a post-validation item analysis of the A-Type MCQ exam, emphasizing its construction and psychometric evaluation.

A purposive sampling strategy was employed, targeting candidates who completed the SpX Entry Qualifying Exam during the year 2024. All eligible candidates during this period were included to maximize representativeness within the available cohort.

Study Area and Sampling: The Specialty Qualifying Entry Exam (SpX) was conducted at a postgraduate medical institute using a Non-Probability Purposive Sampling Technique, revealing a low exam success rate of 10.86%. The exam covered three main disciplines: physiology, anatomy, and pathology, in addition to several clinical subjects. It comprised 120 multiple-choice questions (MCQs) with a total of 480 responses, including 120 correct answers and 360 distractors. A group of 175 examinees from over 33 universities took part in the examination.

Data Collection and Techniques: Data from the SpX examination were extracted from the examination system, which recorded the exam consisting of 120 Type-A Items/Questions in a Best of four format, along with accompanying psychometric and statistical reports. Flaw classification and item analysis were performed using an automated, in-built software algorithm based on standardized criteria. Since the process did not involve subjective human judgment, inter-rater reliability assessment was not applicable.

Cognitive Level Classification (Bloom's Taxonomy)

The cognitive domain of each MCQ was classified according to the revised Bloom's taxonomy, which includes six levels: Remember, Understand, Apply, Analyze, Evaluate, and Create. For practical purposes, these were consolidated into three categories: Lower-order thinking skills (LOTS) comprising Remember and Understand; Middle-order (Apply); and Higher-order thinking skills (HOTS) comprising Analyze, Evaluate, and Create. Each item was independently reviewed and categorized by two author faculty members trained in educational assessment. Discrepancies, if any, were resolved by consensus using a predefined rubric based on accepted standards in medical education literature.[24]

Use of Remark Office OMR Software

The Remark Office OMR software was used for optical scanning of answer sheets and automated generation of statistical item metrics such as difficulty index, discrimination index, and distractor selection frequency. However, cognitive level classification based on Bloom's taxonomy was performed manually, and not by the software. This distinction is now made explicit to ensure transparency.

Data Analysis

The collected data were organized into Excel sheets and segmented into 11 subsections to tackle four key objectives: test statistics, MCQ errors, item analysis, and relevant admission characteristics for specialities and candidates. The initially structured data were further processed and analyzed using SPSS version 29.0; results were reported in percentages, mean values, and standard deviations (SDs). Correlations among indices were assessed using Pearson's correlation coefficient

and *t*-tests, with significance established at $p < 0.05$. Prior to analysis, data were cleaned and validated to remove incomplete or duplicate records. MCQs with missing responses or ambiguous keys were excluded.

Ethical Considerations

Approval from the Anonymous Selected Institute For Health Professions Education and Qualification Research Committee Board (22/10/Anonymous Selected Institute For Health Professions Education and Qualification) and approval from the EDC Ethical Committee Board (22/10/Anonymous Selected Institute For Health Professions Education and Qualification/EDC) was obtained. The names of the Specialty Programme and the institute were bound to remain undisclosed.

Results

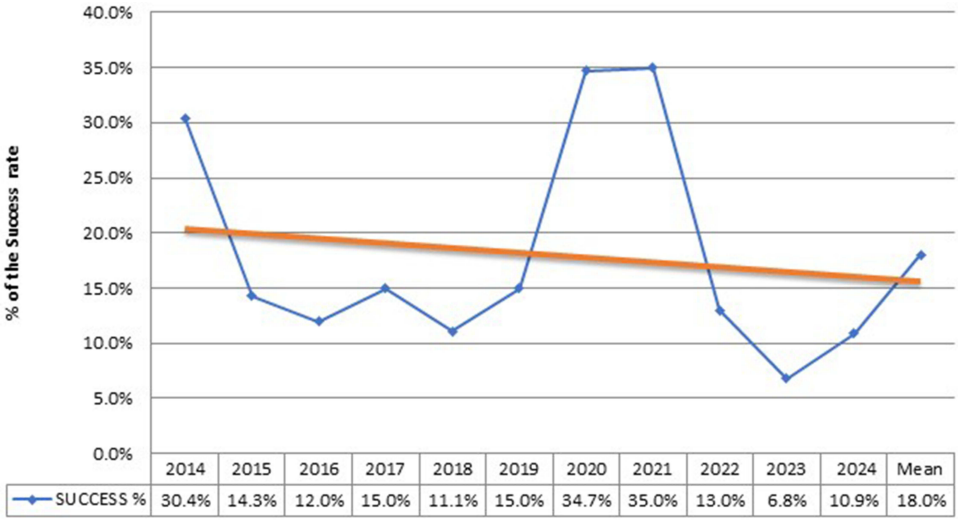
The Trend of the SpX entry Qualifying Exam for the Year of Study and Past Years

Out of 175 candidates who appeared for the SpX Entry Qualifying Exam, only 19 (10.86%) passed. The passing percentage is consistently lower in the past decade. The trend ranged from 35.0% to 6.8% over the ten years from 2014 to 2024 (Figure 1). Figure 2 shows that the candidates' score distribution was fairly concentrated around the 40s–50s, tapering off on both sides. The tallest frequency is the score of 48, with a frequency of 12 candidates.

The General Structural Format Flaws Analysis

A structural review of the MCQs revealed several widespread deviations from standard item-writing guidelines. Specifically, the "lead-in" was not framed as a question in 90% of the items (108 out of 120), which may compromise clarity and focus. Additionally, 65.8% of the items (79 out of 120) failed the "cover-the-options" test, indicating that the stem alone did not provide a complete context for answering the question without seeing the options. Furthermore, clinical vignettes were absent in 98.3% of the questions (118 out of 120), limiting the ability to assess application of knowledge in realistic scenarios.

The other flaws in multiple-choice questions (MCQs) were categorized into three groups: irrelevant difficulties, testwiseness, and editing errors. Out of 120 questions, irrelevant difficulties were identified in 24 MCQs, representing 20% of the total. The same percentage was found in the testwiseness category. The most common flaw identified was "long correct answers" at 10%, followed by non-homogeneous options at 11.6% (Table 1).



	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023	2024	Mean
SUCCESS %	30.4%	14.3%	12.0%	15.0%	11.1%	15.0%	34.7%	35.0%	13.0%	6.8%	10.9%	18.0%

Figure 1 The trend of the SpX Entry Qualifying Exam for the Year of Study and Past Years from 2014 to 2024.

https://doi.org/10.2147/AMEP.S525828

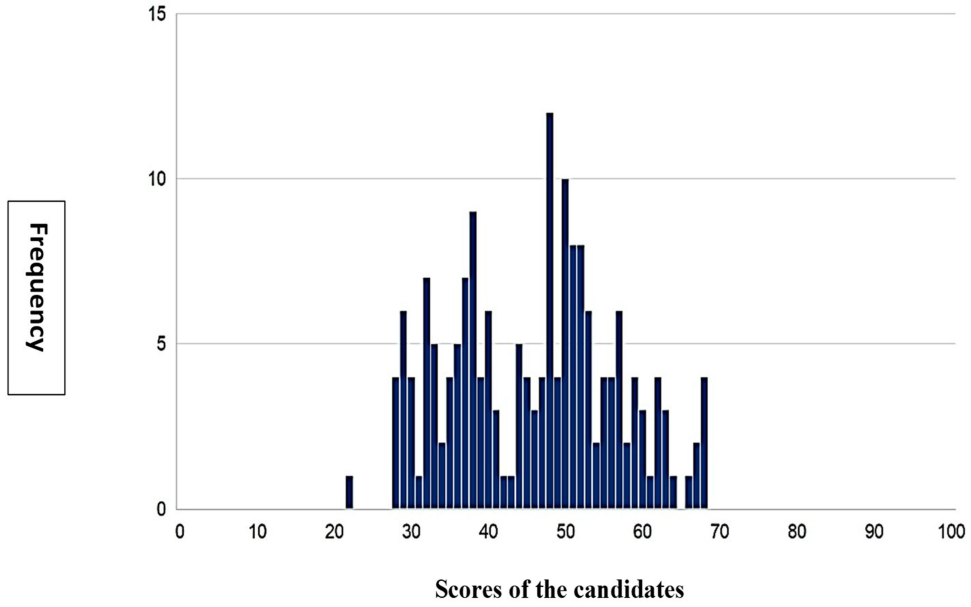


Figure 2 The candidates' scores histogram of the SpX Entry Qualifying Exam.

Bloom's Taxonomy Levels Classification

The SpX Entry Qualifying Exam. N=120 was analyzed based on Bloom's seven levels of cognitive learning. Most questions, 110 (91.67%), fell under the Recall (REC) level. Comprehension (Comp) was assessed in only three questions (2.50%), while the Application (App) level accounted for five questions (4.17%). Higher-order cognitive skills, including Analysis and Scenario-based questions (ANA/SCE), were represented in just two questions (1.67%).

Item Analysis and Statistical Indices

The electronic software provides a comprehensive statistical evaluation of candidate performance, including overall score analysis, exam reliability metrics, and key item indices. These include the Item Difficulty Index (DIF), Point-Biserial Correlation (a measure of item discrimination) (PBS), and a detailed distractor analysis to assess the effectiveness of each response option.

Difficulty Index

The Difficulty Index (DIF) with its five categories are as follows- Very Difficult (VDIF) ≤ 20: was 11 (9.17%), Difficult (DIF) 21–40: was 38 (32.50%), Average (AV) 41–60: was 39 (31.67%), Easy (ES) 61–80: was 27 (22.50%) and Very Easy (VE) > 80: was 5 (4.17%) (Figure 3).

Table 1 Multiple Choice Questions Technical and Editing Flaws of the SpX entry Qualifying Exam (Item N=120)

Flaws	Number and Percentages	Details
Irrelevant Difficulties	24 (20%)	Long, complicated Options" in six Qs (5.0%), "Options are vague" in three Qs (2.5%), and one Q (0.8%) with "Non-logical order". While the "Non-Homogeneous options" were 14 Qs (11.7%)
Testwisenes	24 (20%)	"Long correct answer" was 12 Qs (10%), "Absolute terms" in seven Qs (5.8%), and "Logical cues" in three Questions (1.2%)
Written language errors	59 (49.2%)	Editing errors 38 (52.1%), Punctuation 33 (27.5%), and Grammar 16 (13.3%)

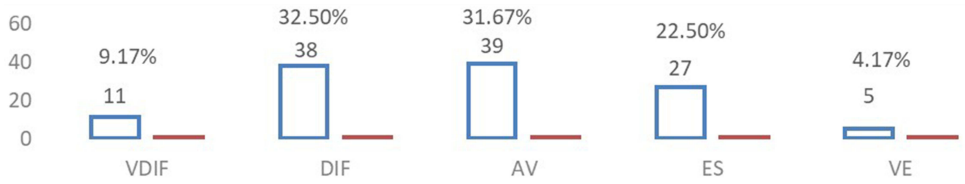


Figure 3 Difficulty Index Categories of the SpX Entry Qualifying Exam. N=120.
Notes: The "y" axis is the frequency: Difficulty levels are categorised as follows: VDIF – Very Difficult, DIF – Difficult, AV – Average, ES – Easy, VE – Very Easy.

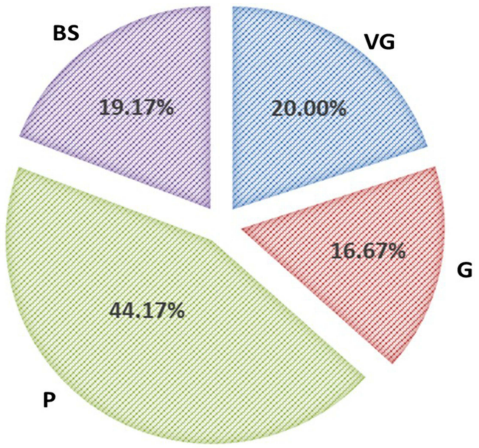


Figure 4 MCQs Point Biserial (PBS) Categories of the SpX Entry Qualifying Exam. Item N=120.
Notes: Point Biserial Correlation are categorized as follows: VG - Very Good, G – Good, P – Poor, BS - Below Standard.

The Point Biserial (PBS) Categories

The Point Biserial (PBS) were categorized into four levels as follows; Very good (VG) ≥ 0.40: 24 (20.00%), Good (G) 0.30–0.39: 20 (16.67%), Below Standard (BS) 0.20–0.29: 23 (19.17%). Poor (P) ≤ 0.19; 53 (44.17%) (Among this category, 18 Items/Questions, 15% of the total were below 0; in minus value) (Figure 4).

Distractor Analysis

The total number of Items/Questions with non-functioning distractors (NFDs) (options selected by less than 5% of examinees) was found to be 27 out of 120 (22.00%), with 6 (5.00%) having two NFDs (Figure 5).

Horst Analysis

The Horst Index (item-level) was used to evaluate all questions and categorize them into three groups based on response patterns. Items with a Horst Index of 9 or higher were considered accepted and well-constructed, comprising 67 items (55.83%). Those with scores between 0 and 8 were identified as potentially miskeyed, indicating possible ambiguity or

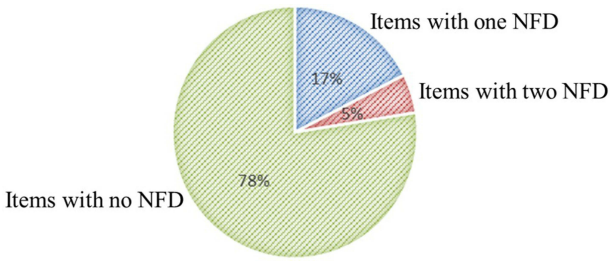


Figure 5 Distractor Efficiency of the SpX Entry Qualifying Exam. Item N=120.
Abbreviation: FTD, Non-Functional Distractor.

https://doi.org/10.2147/AMEP.S525828

Table 2 The Mean of the Item Indices of the SpX entry Qualifying Exam (Item N=120)

Index	Mean ± SD
Difficulty index	45.9 ± 4.52
Discrimination index/Point Biserial	0.17± 0.02
Mean Distractor Efficiency	0.66 ± 0.09
Horst Index (test level)	0.23
Reliability coefficient (using Coefficients Cronbach's alpha2)	0.85

the presence of two functioning options, accounting for 18 items (15.00%). Items with a Horst Index below zero suggested serious flaws or incorrect keying, such as multiple correct responses, and totaled 35 items (29.17%).

The Mean of Item Indices

These mean ± statistical deviation of Difficulty Index was 45.9 ± 4.52 and the mean Point Biserial was 0.17± 0.02. and Horst index was 0.23 ± 0.02 (Table 2).

The Relationships Between Item Indices and Flaws in the SpX Entry Qualifying Exam

The relationships between item indices and flaws in the SpX Entry Qualifying Exam showed significant results ($P < 0.05$) between the DIF categories vs All MCQS flaws; DIF categories and Mean Functional Distractibility; DisI/PBS categories and HORST Index; DisI/PBS categories and All MCQS flaws (Table 3).

Table 3 Test of Significance Among Flaws and Item Indices of the SpX entry Qualifying Exam (Item N=120)

Items/Questions Cross Tabulation	P-value	Interpretation
DIF categories * MCQs Testwiseness flaws	(0.569)	Insignificant
DIF categories * MCQs Irrelevant flaws	(0.084)	Insignificant
DIF Categories * Editing Errors	(0.653)	Insignificant
DIF categories * All MCQS flaws	(0.047)	Significant
DIF categories * Mean Functional Distractibility	(0.0035)	Significant
PBS categories * MCQs Testwiseness flaws	(0.074)	Insignificant
PBS categories * MCQs Irrelevant flaws	(0.063)	Insignificant
PBS Categories * Editing Errors	(0.044)	Significant
PBS categories * All MCQS flaws	(0.043)	Significant
PBS categories * Mean Functional Distractibility	(0.418)	Insignificant
PBS categories * Horst Index	<0.00001	Significant
Editing Errors * Horst Index	(0.253)	Insignificant

Notes: Mean Functional Distractibility = Average number of functioning distractors (selected by ≥5% of examinees); Testwiseness Flaws = Structural flaws that aid guessing rather than knowledge. Irrelevant Flaws = Unnecessary difficulty not related to content; Editing Errors = Language, formatting, or structural issues; Horst Index = Item-level metric used to detect miskeyed or ambiguous questions.
Abbreviations: MCQs, Multiple Choice Questions; DIF, Difficulty Index; PBS, Point-Biserial Correlation.

Table 4 Test of Significance Between Bloom's Taxonomy and Item Indices of SpX entry Qualifying Exam (Item N=120)

Items/Questions Cross Tabulation	P value	Interpretation
Bloom's level * PBS Categories	(0.946)	Insignificant
Bloom's level * DIF Categories	(0.619)	Insignificant
Bloom's level * Horst Index	(0.047)	Significant

Notes: Bloom's Level = Cognitive classification of items based on Bloom's Taxonomy (eg, recall, application, analysis); Horst Index = Item-level index used to detect miskeyed or poorly performing questions.
Abbreviations: PBS, Point-Biserial Correlation; DIF, Difficulty Index.

The relationships between Bloom's taxonomy and item indices in the SpX Entry Qualifying Exam showed significant results ($P < 0.05$) between BLOOMS level and HORST Index; Editing Errors and DisI/PBS Categories (Table 4). The relationships between DisI/PBS cat versus DIFCat in the SpX Entry Qualifying Exam showed significant results ($P < 0.05$) (Table 5).

Classifications of the Candidates/Examinees Who Passed the Exam According to the Type of Curriculum

Those 175 candidates who sat the exam graduated from 35 different universities, classified in Table 6. The classification of the exam successful candidates versus the type of the adopted curriculum was insignificant; p-value (0.946).

Discussion

This study distinguishes itself from numerous comparable research initiatives through three fundamental considerations: evaluating constructional flaws prior to calculating psychometric indices, exploring the correlation between these constructional flaws and the derived psychometric indices, and examining its association with a postgraduate residency program.

The three main potential factors that may influence candidates' exam success rates are: (i) the type and implementation of the curriculum, (ii) the quality of the assessment tools used, and (iii) the teaching and learning environment, including overall wellbeing. The latter factor goes beyond the scope of this study. Regarding the first factor—the type of curriculum—the study analyzed various models of the basic medical curriculum that described by Flexner.[25] The study

Table 5 PBS Categories Versus DIF Categories of the SpX entry Qualifying Exam (Item N=120)

Cross Tabulation		DIF CAT					Total
		VDIF	DIF	AV	ES	VE	
DIS/PBS CAT	VG	0 (0.0%)	2 (8.3%)	10 (41.7%)	10 (41.7%)	2 (8.3%)	**24 (100.0%)**
	G	0 (0.0%)	4 (20.0%)	7 (35.0%)	8 (40.0%)	1 (5.0%)	**20 (100.0%)**
	BS	0 (0.0%)	4 (17.4%)	12 (52.2%)	6 (26.1%)	1 (4.3%)	**23 (100.0%)**
	P	11 (20.8%)	29 (54.7%)	9 (17.0%)	3 (5.7%)	1 (1.9%)	**53 (100.0%)**
Total		**11 9.2%**	**39 32.5%**	**38 31.7%**	**27 22.5%**	**5 4.2%**	**120 100.0%**

Notes: A significant association was observed between DIF categories and DIS/PBS categories ($p < 0.0001$).
Abbreviations: DIF CAT, Difficulty Index Categories; VDIF, Very Difficult; DIF, Difficult; AV, Average; ES, Easy; VE, Very Easy; DIS/PBS CAT, Discrimination Index/Point-Biserial Correlation Categories; VG, Very Good; G, Good; BS, Below Standard; P, Poor.

Table 6 Number of Candidates Who Passed the SpX entry Qualifying Exam Versus Type of the Curriculum (Item N=120)

Type of Curriculum	Number of All Candidates per Curriculum Type	Number of Exam Successful Candidates	% Exam Success Rate
System/Discipline-Based	36	7	19.44%
SPICES-Based	82	9	10.98%
Hybrid	53	3	5.66%
Unspecified	4	0	0.00%
Total Number	175	19	10.86%

Notes: The p-value =0.946 (> 0.05); insignificant association; SPICES (Student-Centred, Problem-Based, Integrated, Community-Based, Elective Driven and Systematic).

found no significant performance differences among these curriculum models: the classical discipline-based model, the system/organ-based model, the newer SPICES model, or hybrid models. As well, no significance in multivariate analysis. After an extensive review of the available literature, there appears to be a lack of studies directly comparing written examination performance based on formative versus transformative types of Curriculum. With regard to the second factor, the study posits that the quality of the Items/Questions employed in the SpX Entry Qualifying Exam is the primary contributor to the observed low exam success rate. The selected Anonymous Specialty Programme X employs the MCQ-type A tool for its written entry examination, which is conducted electronically via computers. The issue does not lie with the chosen assessment tool; rather, it is associated with its implementation. The A-type MCQs assessment tool is favoured due to its superior "Utility" (Validity X Reliability X Educational Impact X Acceptability X Feasibility)[5] when compared with other written assessment tools.[3–5] It is acceptable in such exams, and MCQs are commonly used in similar exams.[10–12]

In the evaluation of objective one, a total of 175 candidates participated in the SpX Entry Qualifying Exam, of which only 19 candidates, representing 10.86%, exam successfully passed. This outcome reflects the lowest exam success rate recorded in that year for which it was selected for study. When studying the scores of the candidates in the selected exam (the SpX Entry Qualifying Exam), the distribution of candidates' scores reveals a concentration predominantly in the 40s to 50s range, tapering off on either end. Notably, the most frequently achieved score was 48/120, with a count of 12 candidates attaining this score (Figure 2). The analysis of exam results spanning from 2014 to 2024 indicates that the exam success rates have exhibited fluctuations and inconsistencies, deviating from the overall trend line. For the issue of inappropriate preparedness and the nature of such a profession, the literature search highlighted that the candidates of Specialty X in the American Board Written Qualifying Examination for example, have one of the highest board exam pass rates of any medical speciality.[26,27] Also, it reached 94.3%, one of the highest pass rates for Canadian Medical Degree and Post Graduate degree, taking the exam for the first time candidates over the last three years, 2022–2024.[28] In the analysis of these two big qualifying institutes, we realize the critical importance of the prior preparedness and the exam Blueprint which was not the case in the studied SpX Entry Qualifying Exam.

In the evaluation of objective two, strikingly only two questions with a clinical vignette. Clinical vignette-based MCQs are often more effective because they assess the application of knowledge to real patient scenarios and present relevant clinical challenges.[29] Candidates also prefer them,[30] however overemphasis on factual recall in assessments risks under-evaluating higher-order competencies such as clinical reasoning, potentially leading to graduates who are less prepared for complex decision-making in real-world clinical settings.

MCQ flaws are classified based on the criteria from the USA National Board of Medical Examiners (NBME).[2] In a crucial summary, the study found that only seven out of 120 Items/Questions (5.8%) were perfect. The Table 1 documented 24 were categorized under "Testwiseness", reflecting specific issues that could compromise the validity of the assessment. The most common flaw within this category was the "Long correct answer" type, observed in 12 Items/Questions (10%), where the best option stood out due to its excessive length compared to the other three distractors. This was followed by the presence of "Absolute terms" in seven Items/Questions (5.8%), such as using definitive words like

"always" or "never", which may inadvertently guide candidates toward or away from certain options based on this flawed item rather than their knowledge. Additionally, "Logical cues" were identified in three Items/Questions (1.2%), where the structure or wording of the options provided unintended hints toward the correct answer. Notably, there were no instances of other commonly observed flaws. The Second group of the "Irrelevant difficulty" type of flaws found in 24 Items/Questions (20%), mainly involved the "Non-Homogenous options" flaw in 14 Items/Questions (11.7%), "Long, complicated options" in six Items/Questions (5.0%), and vague options in three Items/Questions (2.5%), as well as one Item (0.8%) with a "Non-logical order". No detection of "None/All of the above", inconsistent numeric data. These easily preventable flaws disrupted the candidate/examinee's thought processes and affected their performance, as illustrated in the cross-tabulation of the Items/Questions psychometric indices. Downing SM noted, "10–15% of those classified as 'failures' would have passed if flawed Items/Questions were excluded".[31] Moreover, flawed Items/Questions often penalize high-performing examinees more than their average-performing peers.[32] The persistence of flawed MCQs may stem from systemic issues, including the absence of standardized item-writing training and overreliance on legacy question banks that lack quality control mechanisms.

In evaluating objectives three and four, the findings from the post-validation analysis of MCQ Items/Questions for the SpX Entry Qualifying Exam included analysis of reliability, difficulty, point biserial indices, and distractor efficiency. The study examined the significance of correlations among the indices and their relation to identified constructional flaws. It employed the Pearson chi-squared test ($\chi 2$ test). Inter-rater reliability was not assessed, as flaw classification was carried out using an automated software-based algorithm. Given the absence of subjective human judgment, the potential for rater-related bias was minimized. The Difficulty Index (Feasibility Index) (DIF) of the SpaceX Entry Qualifying Exam reveals the distribution of item difficulty. Items classified as "Very Difficult (VDIF)" (DIF $\leq$ 20) accounted for 11 items, or 9.17%, while "Very Easy (VE)" items (DIF > 80) comprised 5 items, or 4.17%. The remaining 114 items, representing 86.67%, fell within the mid-range of 0.2–0.8. The average DIF was calculated to be 45.9 ± 4.52 Table 3, indicating a wide range of acceptable difficulty levels.[14,20] A wide range of values emerged compared with these findings, for example some studies reported high difficulty indices in postgraduate research.[33] Moreover, a broad range of difficulty indices was observed in a study by 29 researchers conducted from 2003 to 2006 at the International Medical University-Malaysia, where mean difficulty index scores of individual tests ranged from 64% to 89%.[16] Conversely, some studies indicated a significantly lower DIF, such as (38.34±2.25) documented by Patil et al study.[34] In critically evaluating these studies, there was no clear single factor to explain the discrepancies. However, it is essential to consider item construction, the assessment blueprint, and the intended learning outcomes (ILOs), which were not adequately discussed in these referenced studies. Although 96 Items/Questions (80%) at the Recall level fell within the Acceptable DIF category, Chi-Square tests did not demonstrate a statistically significant relationship between Bloom's levels and DIF. However, a significant association was observed between collective item construction flaws and DIF at $\alpha = 0.05$. The most plausible explanation for these findings is that each MCQ quality measure operates independently in determining whether an exam is a valid assessment tool, particularly in high-stakes postgraduate examinations. Medical educators emphasize the importance of the mean difficulty index, as it provides insight into the overall trend of item difficulty.

Notably, several studies use these two terms, DIS and PBS interchangeably to characterize an item's ability to differentiate between high and low scorers, which ranges from −1.00 to +1.00.[8,14,21] In our study mean PBS was found to be 0.17 ± 0.02 (within the Poor Category). Here, the higher PBS categories (VG and G) constituted 36.7%, while the majority fell into the two lower categories (BS and P) at 63.2%. Furthermore, within the P category, 18 Items/Questions (15% of the total) scored below zero, indicating that lower-ability candidates answered more correctly than their higher-ability counterparts. This may result from guessing correctly without proper understanding, while more capable candidates may be wary of the complex and misleading nature of the item.[4,20,35] Hingorjo and Jaleel reported a mean discrimination index (DI) of 0.356 ± 0.17, while Musa et al found a mean point-biserial value of 0.37 ± 0.13.[36,37] Rao et al categorized 60% of items as having excellent discrimination (DI > 0.40), though no overall mean was reported.[38] Thus there is varied reporting of DIS in literature.

In our study mean PBS reflects the inconsistency in critical, careful planning for this Entry Qualifying Exam, which may have not based on a curriculum map or a well-structured blueprint. The implication of these findings will entail

a mandatory future review of many of the used Items/Questions and oblige a rejection of many others, especially those of a negative value PBS.

The DIF and PBS indices, when taken together in Table 6, were significantly correlated (P-value =0.000<0.05) in a positive manner. The extreme DIF categories Very Difficult (VD), Difficult (D), Easy (ES), and Very Easy (VE) were the majority in the poor discriminating category rpbis <0.20 [44 (83%) out of the total 53 Items/Questions]. On the other hand, only three (6.8%) out of the 44 Items/Questions of the VG & G categories of the PBS lay in the extreme DIF categories, the noted Very Difficult (VD) and Very Easy (VE). Many studies support this trend of the relationship between DisI/PBS and DIF.

Distractors Efficiency (DE), and Horst Index (HI) were the two statistical poles used to measure the functional Items/Questions distractibility of the SpX Entry Qualifying Exam (Table 3). The HI was 0.23 ± 0.02, while the negative HI was found in 41 Items/Questions (23.4%), and they have statistical significance with PBS, which is understandable based on the calculation proximity. HI was significantly related to Bloom's level, which could be from an error in the Item itself or even incorrect teaching and knowledge perception. The DE findings reflected MCQ constructional problems when the designers had difficulty developing plausible distractors on the one hand, and the implausible distractors were used only as fillers. The Figure 5 shows that the number of Items/Questions with one NFD < 5.0% was 21 (17.50), two NFDs were 6 (5.00%), and no three NFDs. These findings surpass those of numerous studies, which conflicts with the other detected psychometric indices. Additionally, the DE demonstrates significant variation. These results bolster the notion that the Best of Three MCQs are superior to the Four or Five A-Types.[9,39–41] Despite the known importance of the DE, the study found no significant relation between DE and PBS, similar to the extensive research conducted by Rush et al study.[42] One possible explanation is that flawed distractors may not impact high-performing examinees, who are more likely to identify the correct response despite suboptimal distractor design.

The present study estimated the SpX Entry Qualifying Exam Reliability (KR20) at around 0.85. (Table 2) This result is acceptable for such type-A exams.[14,20] However, the study acknowledge considerable discussions among researchers regarding reliability limitations.

This indicates that the exam was created without a proper blueprint, affecting its alignment, congruence, and overall validity. The results highlight various shortcomings related to the SpX Entry Qualifying Exam. Attendance records for the training workshops at the anonymous institute showed that only 20% of the Examination Committee members who were involved in constructing the exam had undergone structured training in constructing Quality MCQs. Jozefowicz et al illustrate the considerable effect of training, revealing a significant difference ($p < 0.01$) between those who received training by NBME.[43] While previous studies address factors influencing item quality by faculty, research on the specific barriers and facilitators faced by individual item writers is lacking. Understanding these challenges could lead to targeted interventions to improve the quality and quantity of assessment Items/Questions.[3]

The findings suggest that the deficiencies in the SpX Entry Qualifying Exam have negatively impacted its exam success rate. Many examinees labelled as "failures" might have passed if these issues had not affected the cut-off score, which was set to be 60% per the rules and regulations of the anonymous institute. No Standard-Setting methods were adopted to calculate the Minimum Passing index and Level (MPI/MPL). On the other hand certain flaws, such as the presence of a "long correct answer", may unintentionally cue test-takers, thus inflating correct response rates regardless of actual knowledge. For example, in our dataset, a specific item had a significantly higher correct response rate despite low discrimination, likely due to this flaw—a pattern that illustrates how such cues can undermine the validity of assessment outcomes.

The recommendations for Academic Institutes, Leadership and Educators are as follows: (1) The Central Assessment Committee (CAC) is critical for exam oversight and quality evaluation. As Candidates can escape the effects of poor teaching, they cannot escape the effects of poor assessment.[44] (2) Create an Exam Bank and software for managing and automating testing processes in collaboration with recognized exam bodies. (3) The examination committees should comprise trained experts in health profession education and assessment. (4) Promote research on factors affecting exams and graduate outcomes. (5) The syllabus and blueprint for the Entry Qualifying Exam should be shared with the candidates. New approaches to teaching and learning with more candidate involvement will pave the way for the best results.[45] (6) Perform post-validation item analyses for exams. (7) Participate in workshops by the Education

Development Center. For Candidates/Examinees: (1) Choose specialities based on a strategic career pathway. (2) Prepare using the syllabus and blueprint. (3) Prepare well for the exam.

This study employed purposive (non-random) sampling, which may limit the generalizability of findings to broader populations. Future studies using random or stratified sampling across multiple institutions may enhance external validity. Further, to establish causal relationships between item-writing quality and exam performance, future studies could adopt experimental designs—such as pre- and post-training interventions for item writers—which may provide stronger evidence of impact.

Conclusion

This study examines the question indices and test statistics of the Entry MD Qualifying Exam of Specialty Programme X, selected for its low exam success rate. Comprising 120 Best of Four Items/Questions, the exam revealed editing errors in 60% of questions and 40% contained issues related to Constructional Testwiseness and irrelevant MCQs. Common flaws included non-homogeneous options, lengthy correct answers, and complicated choices. Only 10% of lead-ins were in interrogative format, and over two-thirds failed the "cover-the-options" test. The mean DIF was 45.9 ± 4.52, with 86.7% falling within the acceptable range. The mean PBS was 0.17 ± 0.02, indicating a poor performance, with 15% of Items/Questions scoring below zero. There were significant associations between MCQ flaws and both DIF and PBS. Despite low markers, the exam showed reasonable internal consistency (KR20 = 0.85). No correlation between distractor efficiency and discrimination, may reflect the possibility that ineffective distractors primarily affect borderline or low-performing students, while high performers are less influenced by distractor quality.

This study highlights the critical role of MCQ quality in influencing candidate performance in postgraduate entry examinations. By applying Classical Test Theory metrics—such as difficulty index, discrimination index, and point-biserial correlation—and classifying items using Bloom's taxonomy, we identified specific structural and cognitive flaws that compromise the validity of assessments.

A quality MCQ was defined in this study as one with acceptable difficulty (30–70%), good discrimination (DI ≥ 0.2), and distractor efficiency (≥50% functional distractors). Test-level quality was assessed through the mean of these indices and internal consistency measures. Despite most items falling within acceptable statistical thresholds, several items exhibited structural flaws (eg, poor lead-ins, lack of vignettes) and low cognitive demand, limiting their ability to assess higher-order thinking.

The study also found a limited but notable association between item flaws and poor psychometric performance, underscoring the need for structured item writing training and peer review. The application of Bloom's taxonomy—though manually classified—helped reveal a dominance of lower-order questions, calling for deliberate reform in question design to foster deeper learning.

We acknowledge that the absence of inter-rater reliability for cognitive classification and the use of purposive sampling are limitations that may affect generalizability. Nonetheless, the comprehensive inclusion of all eligible candidates and consistent item analysis strengthen the internal validity of the findings.

Future studies should employ experimental designs, include inter-rater validation, and explore longitudinal impacts of faculty development on item quality.

Disclosure

The researchers assert that there are no current or potential conflicts of interest.

References

1. WFME. WFME standards for postgraduate medical education. 2023. Available from: https://wfme.org/wp-content/uploads/2017/pdf. Accessed July 28, 2025.
2. Teacher section NBME constructing written test questions for the basic and clinical sciences.pdf. Available from: https://www.ohsu.edu/sites/default/files/2019-03/Teacher%20section%20NBME%20Constructing%20Written%20Test%20questions%20for%20the%20basic%20and%20clinical%20sciences.pdf. Accessed January 22, 2025.
3. Norcini JJ, McKinley DW. Assessment methods in medical education. *Teach Teach Educ*. 2007;23(3):239–250. doi:10.1016/j.tate.2006.12.021
4. Epstein RM. Assessment in medical education. *N Engl J Med*. 2007;356(4):387–396. doi:10.1056/NEJMra054784

5. van der Vleuten CPM, Schuwirth LWT. Assessing professional competence: from methods to programmes. *Med Educ*. 2005;39(3):309–317. doi:10.1111/j.1365-2929.2005.02094.x
6. How to set standards on performance-based examinations: AMEE guide no. 85 - PubMed. Available from: https://pubmed.ncbi.nlm.nih.gov/24256050/. Accessed January 22, 2025.
7. Blueprinting for the assessment of health care professionals - Hamdy - 2006 - the clinical teacher - Wiley online library. Available from: https://asmepublications.onlinelibrary.wiley.com/doi/abs/10.1111/j.1743-498X.2006.00101.x. Accessed January 22, 2025.
8. Burud I, Nagandla K, Agarwal P. Impact of distractors in item analysis of multiple choice questions. *Int J Res Med Sci*. 2019;7(4):1136–1139. doi:10.18203/2320-6012.ijrms20191313
9. Al-lawama M, Kumwenda B. Decreasing the options' number in multiple choice questions in the assessment of senior medical students and its effect on exam psychometrics and distractors' function. *BMC Med Educ*. 2023;23(1):212. doi:10.1186/s12909-023-04206-3
10. Part 1 | the federation. Available from: https://www.thefederation.uk/examinations/part-1. Accessed June 20, 2025.
11. AMC computer adaptive test (CAT) multiple choice question (MCQ) examination. Available from: https://www.amc.org.au/pathways/standard-pathway/amc-assessments/mcq-examination/. Accessed June 4, 2025.
12. Step exams | USMLE. Available from: https://www.usmle.org/step-exams. Accessed June 4, 2025.
13. Adams NE. Bloom's taxonomy of cognitive learning objectives. *J Med Libr Assoc JMLA*. 2015;103(3):152–153. doi:10.3163/1536-5050.103.3.010
14. McAlpine M. *A Summary of Methods of Item Analysis*. Bluepaper HEFCE; 2002. https://www.academia.edu.
15. Miller MD, Linn RL, Gronlund NE. *Measurement and Assessment in Teaching*. 10th ed. Upper Saddle River, N J: Prentice Hall; 2013.
16. Mitra NK, Nagaraja HS, Ponnudurai G, Judson JP. The levels of difficulty and discrimination indices in type A multiple choice questions of pre-clinical semester 1 multidisciplinary summative tests. *Int E-J Sci Med Educ*. 2009;3(1):2–7. doi:10.56026/imu.3.1.2
17. Puthiaparampil T. Assessment analysis: how it is done. 2017. doi:10.15694/mep.2017.000142
18. IDEAL Consortium. Available from: https://www.idealmed.org/. Accessed January 22, 2025.
19. Haladyna TM, Rodriguez MC. Using full-information item analysis to improve item quality. *Educ Assess*. 2021;26(3):198–211. doi:10.1080/10627197.2021.1946390
20. Understanding item analyses | office of educational assessment. Available from: https://www.washington.edu/assessment/scanning-scoring/scoring/reports/item-analysis/. Accessed January 22, 2025.
21. Thorndike RM. *Psychometric Theory*. 3rd ed. McGraw-Hill; 1995.
22. School of Medicine, Ahvaz Jundishapur University of Medical Sciences; Ahvaz I, Shakurnia A, Ghafourian M, et al. 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. doi:10.21315/eimj2022.14.4.5
23. ANRKCT. Measuring item reliability - horst. Maxinity. 2019. Available from: https://maxinity.co.uk/blog/horst/. Accessed January 22, 2025.
24. Haladyna TM, Downing SM, Rodriguez MC. A review of multiple-choice item-writing guidelines for classroom assessment. *Appl Meas Educ*. 2002;15(3):309–333. doi:10.1207/S15324818AME1503_5
25. Papa FJ, Harasym PH. Medical curriculum reform in North America, 1765 to the present: a cognitive science perspective. *Acad Med J Assoc Am Med Coll*. 1999;74(2):154–164. doi:10.1097/00001888-199902000-00015
26. Puscas L. Otolaryngology resident in-service examination scores predict passage of the written board examination. *Otolaryngol Neck Surg*. 2012;147(2):256–260. doi:10.1177/0194599812444386
27. Puscas L. Junior otolaryngology resident in-service exams predict written board exam passage. *Laryngoscope*. 2019;129(1):124–128. doi:10.1002/lary.27515
28. Average pass rates by specialty or subspecialty. Available from: https://www.royalcollege.ca/en/eligibility-and-exams/exam-results/average-pass-rates. Accessed June 8, 2025.
29. Case SM, Swanson DB, Becker DF. Verbosity, window dressing, and red herrings: do they make a better test item? *Acad Med J Assoc Am Med Coll*. 1996;71(10 Suppl):S28–30. doi:10.1097/00001888-199610000-00035
30. Do accompanying clinical vignettes improve student scores on multiple choice questions (MCQs) testing factual knowledge? International association of medical science educators - IAMSE. 2011. Available from: https://www.iamse.org/mse-article/do-accompanying-clinical-vignettes-improve-student-scores-on-multiple-choice-questions-mcqs-testing-factual-knowledge-2/. Accessed June 8, 2025.
31. Downing SM. The effects of violating standard item writing principles on tests and students: the consequences of using flawed test items on achievement examinations in medical education. *Adv Health Sci Educ Theory Pract*. 2005;10(2):133–143. doi:10.1007/s10459-004-4019-5
32. Tarrant M, Ware J. Impact of item-writing flaws in multiple-choice questions on student achievement in high-stakes nursing assessments. *Med Educ*. 2008;42(2):198–206. doi:10.1111/j.1365-2923.2007.02957.x
33. Kowash M, Hussein I, Al Halabi M. Evaluating the quality of multiple choice question in paediatric dentistry postgraduate examinations. *Sultan Qaboos Univ Med J*. 2019;19(2):e135–e141. doi:10.18295/squmj.2019.19.02.009
34. Patil R, Palve SB, Vell K, Boratne AV. Evaluation of multiple choice questions by item analysis in a medical college at Pondicherry, India. *Int J Community Med Public Health*. 2016;3(6):1612–1616. doi:10.18203/2394-6040.ijcmph20161638
35. Item discrimination indices. Available from: https://www.rasch.org/rmt/rmt163a.htm. Accessed January 28, 2025.
36. Hingorjo MR, Jaleel F. Analysis of one-best MCQs: the difficulty index, discrimination index and distractor efficiency. *JPMA J Pak Med Assoc*. 2012;62(2):142–147.
37. Musa A, Shaheen S, Elmardi A, Ahmed AAE. Item difficulty & item discrimination as quality indicators of physiology MCQ examinations at the faculty of medicine Khartoum university. *Khartoum Med J*. 2018. Available from: https://www.semanticscholar.org/paper/Item-difficulty-%26-item-discrimination-as-quality-of-Musa-Shaheen/e0235af6e80708863665da26d4bde1ad9796196e. Accessed February 9, 2025.
38. Rao C, Kishan Prasad H, Sajitha K, Permi H, Shetty J. Item analysis of multiple choice questions: assessing an assessment tool in medical students. *Int J Educ Psychol Res*. 2016;2(4):201. doi:10.4103/2395-2296.189670
39. Fozzard N, Pearson A, du Toit E, Naug H, Wen W, Peak IR. Analysis of MCQ and distractor use in a large first year health faculty foundation program: assessing the effects of changing from five to four options. *BMC Med Educ*. 2018;18(1):252. doi:10.1186/s12909-018-1346-4
40. Vegada B, Shukla A, Khilnani A, Charan J, Desai C. Comparison between three option, four option and five option multiple choice question tests for quality parameters: a randomized study. *Indian J Pharmacol*. 2016;48(5):571–575. doi:10.4103/0253-7613.190757

41. Rahma NAA, Shamad MMA, Idris MEA, Elfaki OA, Elfakey WEM, Salih KMA. Comparison in the quality of distractors in three and four options type of multiple choice questions. *Adv Med Educ Pract*. 2017;8:287–291. doi:10.2147/AMEP.S128318
42. Rush BR, Rankin DC, White BJ. The impact of item-writing flaws and item complexity on examination item difficulty and discrimination value. *BMC Med Educ*. 2016;16(1):250. doi:10.1186/s12909-016-0773-3
43. Jozefowicz RF, Koeppen BM, Case S, Galbraith R, Swanson D, Glew RH. The quality of in-house medical school examinations. *Acad Med J Assoc Am Med Coll*. 2002;77(2):156–161. doi:10.1097/00001888-200202000-00016
44. Boud D. Assessment and learning: contradictory or complementary? 1995. Available from: https://www.semanticscholar.org/paper/Assessment-and-learning%3A-contradictory-or-Boud/0a219014d65c5e5c1ce81b92eb3e1cdb5c768237. Accessed February 9, 2025.
45. Srinivasamurthy SK, Bhat R, Eladil AHMO. The tale of designing a clinical-cases manual for rotations and mixed methods analysis of students' participatory experience in co-creation. *Adv Med Educ Pract*. 2024;15:875–882. doi:10.2147/AMEP.S472544