International Journal of Science Academic Research
Vol. 05, Issue 05, pp.7434-7442, May, 2024
Available online at http://www.scienceijsar.com

ISSN: 2582-6425

Research Article

IMPACT OF MCQS QUALITY ON THE SUCCESS RATE OF A QUALIFYING EXAM FOR A POSTGRADUATE MEDICAL RESIDENCY

[1,*]Omer Eladil A.H.M., [2]Bashir Hamad, [3]Yassir Ahmed Mohammed Alhassan and [4]Fatima Mohammed

[1]Professor of Internal Medicine, Neurology and Medical Education Rak Medical & Health University UAE, International University of Africa IUA, Sudan Medical Specialization Board SMSB
[2]Professor of Medical Education and Community Medicine, Sudan Medical Specialization Board SMSB International University of Africa IU-Sudan
[3]Associate Professor of Anatomy and Medical Education USA
[4]Assistant Professor Faculty of Art and Sciences Qassim University KSA

Received 15th March 2024; **Accepted** 20th April 2024; **Published online** 30th May 2024

Abstract

Objective: To study items' construction and psychometric analysis of the Specialty (SpX) Qualifying Residency Exam (QRE) of the Postgraduate Medical Institute "Y" (PGMY). **Methods**: A post-validation cross-sectional analytical study using a Non-Probability Purposive Judgmental sampling technique. The SpX was selected from the lowest three success rates of the 52 clinical specialties within the 2020-2023 QRE Cycles. **Results:** 175 candidates sat for QRE. The success rate was 10.86% (19). The QRE contained 120 A-type MCQs. Items without any flaws were 7 (5.8%). "Non-Vignette Stems" were 118 (98.3%); the majority of the "Lead-in items" were not in a question format. Two-thirds failed the "Cover-the-options" Test. 48 items (40%) had "Constructional Testwiseness" and "Irrelevant-MCQ" flaws. Furthermore, the mean Difficulty Index (DifI) was 45.9 ± 4.52, where 114 (86.7%) were within the acceptable. The Discrimination Index/Points Biserial (DisI/PBS) mean was 0.17± 0.02. 18 items (15%) had minus values. The Mean Distractor Efficiency (mDE) was 66.0% ± 0.09. Significant associations (p-value <0.05) were found between flaws and, the DifI and DisI/PBSDisI/PBS, HORST Index, and Bloom's levels. Likewise, mDE showed a significant association with DifI but not with DisI/PBS. On the other hand, no significant association between the success rate and the MBBS curriculum style. Unlike the international trend of the same profession, the QRE had a zigzagging low success rate since cycle 2013. Conclusion: The items' quality significantly affected QRE. Other potentially influential factors deserve future multivariate analytical research. This consolidates the PGMY strategic plans for Exam Bank and Health Professions Education.

Keywords: Postgraduate, Psychometric analysis, Educational Assessment, Examination Questions.

INTRODUCTION

The World Federation for Medical Education recommends nine global quality standards for postgraduate medical education, including "Assessment of Trainees," which requires documentation and evaluation of assessment methods. Multiple-choice items (MCQs) are the most popular, reliable, valid, and cost-effective written assessment tools for medical knowledge and psychomotor domains (Tawalare *et al.,* 2020). PGMY Qualifying Degrees in Sudan awards professional medical specialty degrees through an entry qualifying exam consisting of 120 Best-of-Four A-type MCQs. Candidates must pass this exam to enroll in specialty training residency programs, which require 48 months/4 years of specific training competencies (Qureshi, 2020). This postgraduate study was the first of its kind in PGMY and among a few internationally. It analyzed the quality and the impact of MCQ construction in one of the PGMY Specialty MD Qualifying Entry Exams with the lowest success rate in the 2020-2023 QRE Cycle (Doorenweerd *et al.,* 2017). It was chosen by the Judgmental Sampling Method.

Corresponding Author:** ***Omer Eladil A.H.M.,
Professor of Internal Medicine, Neurology and Medical Education Rak Medical & Health University UAE, International University of Africa IUA, Sudan Medical Specialization Board SMSB.

This Entry Exam is a crucial fundamental exam that any candidate must pass before starting the required training to obtain the Clinical MD specialty qualification (Ozair *et al.,* 2023). The study analyzed MCQ items, test statistics, constructor flaws, Bloom's taxonomy level, reliability, Difficulty Index, Discriminating index/Point Biserial, and mean distractor efficiency. It also evaluated test statistics and examination reports. Ethical issues were addressed by apprehending PGMY and keeping the council hidden. The study examines the MD Qualifying Entry Exam of Specialty Council X at Sudan Medical Specialization Board, focusing on test statistics, technical construction issues, and the impact of candidates' university curriculum type and geographical location on exam success rate, as well as calculating the four main quality item analysis parameters.

Theoretical Framework

Item Response Theory (IRT)

The item response theory has suggested that the test-taking success depends on the examiners ability and the item's difficulty, particularly in the MCQs. Although the poorly constructed MCQs might not discriminate among high and low ability, that is potentially affecting the success rate (Edelen & Reeve, 2007).

LITERATURE REVIEW

Assessment in Medical Education

Assessment is an essential part of candidates learning. Importantly, it derives learning. Hence, by assessment, we force them to learn what we want them to learn (Sa-Ngiamsuntorn *et al.,* 2021). Ian Hart's 1998 AMEE Conference emphasized the importance of high-quality, valid, reliable, feasible, and acceptable assessment tools for students to learn and improve their learning outcomes. The innovative assessment practices in legal education in England, emphasizing the need for rigorous analysis and implementation. It provides website resources for readers to follow developments in specific projects (Bone & Maharg, 2019). MCQs are commonly used for assessment because their high content validity encompasses many content areas. Moreover, MCQ items can be administered in a relatively short period and can be graded by computer (Marchant, 2021). MCQs should aim to assess knowledge recollection and comprehension, the low learning levels of Bloom's taxonomy, and masterly measure other high cognitive teaching objectives within that creating a high-quality MCQ examination is a complex academic task requiring designers to understand ILOs and the Blueprint, requiring time and effort (Gordon *et al.,* 2017). MCQs are preferred due to their objectivity and efficiency in scoring, impacting the quality of medical undergraduate and postgraduate institutes, as they can be scored manually or electronically (Singh, 2021).

Quality of assessment and its impact on Accreditation

The World Federation for Medical Education (WFME) global standards embrace all phases of medical education, the basic (undergraduate) medical education, postgraduate medical education, and continuing professional development of medical doctors (Rao, 2024). The World Federation of Medical Education (WFME) global standards cover nine standards for postgraduate medical education quality improvement, including Standard Three, which emphasizes trainee assessment, ensuring knowledge, skills, and attitudes are covered (Schwill *et al.,* 2022). The concept of assessment utility is enhanced by using multiple assessment methods and formats, and establishing standard criteria for passing examinations to evaluate their reliability, validity, and fairness. Higher education is vital for nation development in various aspects, including social, economic, cultural, scientific, and political aspects. In the globalized world, quality education is essential for fostering creativity, talent, adaptability, and research mindset. Accreditation, a quality assurance tool, ensures institutions meet minimum standards (Kumar *et al.,* 2020).

Multiple choice items (MCQs) as an assessment tool

Since 2005, EPAs have gained attention, with guidelines and workshops supporting faculty development. The AMEE Guide provides clarity on EPA descriptions, including title, specification, risks, competency domains, knowledge, and supervision levels (Ten Cate & Taylor, 2021). The "AMEE Guide No. 25 2009 (12): The assessment of learning outcomes for the competent and reflective physician" emphasizes assessment as a tool for quality training programs and transitions to Outcome-Based Education. The chosen assessment tools should be valid, reliable, and feasible (Elgadal & Mariod, 2021). A-type items are a widely used multiple-choice format with a stem, lead-in question, and a series of choices with one correct answer and three to four distractors. They are suitable for testing knowledge domains and can be used in summative assessments, national in-service, licensing, and certification examinations (Velou & Ahila, 2020). However, they can be difficult to write, can result in cueing, and may seem artificial. Despite these limitations, they can assess many content areas quickly, have high reliability, and can be graded by computer. However, item editing errors are essential to maintain accuracy (Iqbal *et al.,* 2023). The question's difficulty should be based on the examinee's knowledge on the assessed topic, not their test-taking strategy expertise, as irrelevant difficulty can make the question unrelated to the assessment.

Item Construction; Bloom's Classification

Dr. Benjamin Bloom introduced a hierarchy of learning levels in 1956, categorized into six major categories: Knowledge, Comprehension, Application, Analysis, Synthesis, and Evaluation (Grebin *et al.,* 2020). Knowledge is considered the prerequisite for implementing these skills and abilities, with modifications over time (Betts *et al.,* 2021). Bloom's taxonomy can be helpful in assessment to help the justified alignment between what is written in the objectives, what is taught, and what is to be included in the exam using the appropriate tool and structure (Shelley, 2020).

Standard-setting of MCQs

The study evaluated the Ebel standard-setting method for the 2019 Royal College of Physicians and Surgeons of Canada internal medicine certification exam. It assessed parameters such as inter-rater agreement, correlations between Ebel scores and facility indices, the impact of raters' knowledge of correct answers, and the effects of raters' specialty on inter-rater agreement and Ebel scores. Results showed low correlations between Ebel scores and facility indices, and no significant difference between internists and other specialists (Bourque *et al.,* 2020). Standard-setting is the endpoint of a test, defining the minimum acceptable performance score. A score below the set is considered a failure, and the minimum pass level (MPL) is determined before and after the test, as emphasized by Hamad (Wang *et al.,* 2023). The study compares conventional, norm-referenced, and modified-Angoff dental assessment methods, finding significant differences in pass/fail rates and good inter-rater and test-retest reliability, suggesting potential for improvement (Abd-Rahman *et al.,* 2021). The Smart Standard Set system is a graphical-user interface designed to help higher education instructors remove poor quality items and set appropriate grade boundaries. The system has been evaluated through interviews with teachers and focus groups with 19 students. Both groups found the system to be feasible, accurate, and useful (Brown *et al.,* 2021).

Tests statistics and analysis of performance

Difficulty index (DifI)

The difficulty index (sometimes called facility index or p-value) for items with one correct/bestalternative worth a single point. The item difficulty is simply the percentage of examinees/candidates who answer an item correctly. Hence, it is equal to the item mean.

The difficulty index (DifI) = R/T x 100

Where

R is the number of examinees/candidates who answered the item correctly, and T is the total number of examinees/candidates taking that test (Chidozie & Orluwene, 2021). The item difficulty index ranges from **0** to **100**; the higher the value, the easier the question. When an alternative is worth more than a single point or more than one correct alternative per question, the item difficulty is the average score on that item divided by the highest number of points for any one alternative. Item difficulty is relevant for determining whether examinee/candidates have learned the concept beingtested. The accepted range is 0.2–0.8 (20). Some other researches detailed the acceptable ranges on analysis. Some academic authorities consider the level of difficulty from 50-75 (19) where he referred to The Hong Kong International Databases for Enhancement of Assessments and Learning (Lemke *et al.,* 2004).

Table 1. Different ranges of the Difficulty Index (DifI) (Burud *et al.,* 2019)

Category	Value	Comment
Very Difficult (VD)	< 20	Not acceptable
Difficult (D)	20-40	Acceptable upper level
Average (AV)	41-60	Excellent level
Easy (ES)	61-80	Acceptable lower level
Very Easy (VE)	> 80	Not acceptable

Discrimination index and point Biserial

The DisI is a test indicator that differentiates high and low-achieving examinees/candidates. It ranges from +1.0 to –1.0, with a 1.0 indicating a perfect correlation between correct responses and high marks, and a -1.00 level indicating incorrect answers but high overall scores. A higher DisI value indicates better discrimination between abilities (Sahoo & Singh, 2017). DisI is calculated using Kelly's Method, which involves adding correctly answered items from the upper and lower 27% of examinees/candidates' performance, divided by the total number of both groups.

A. The exam papers are arranged according to the total scores of examinees, from the highest to the lowest.
B. They are then classified into three levels one-third (or 27%) of the highest and one-third (or 27%) of the lowest scores.
C. Take the items one by one and tabulate the number of examinees in the upper third and those in the lower third who selected each distracter as the correct one.
D. The formula used: DisI = (Ru-Rl)/½T x 100

Where

Ru = number of correct responses in the upper 1/3, Rl = number of correct responses in the lower 1/3, and T=total number of examinees in both upper and lower groups (19).

Computerized analyses provide a more accurate assessment of the discrimination power of items because they take into account the responses of all examinees/candidates rather than just high and low-scoring groups (Pradeep Kumar *et al.,* 2021). The item discrimination index is a Pearson Product Moment correlation between examinee/candidate responses to a particular item and total scores on all other items on the test. This index is the equivalent of a Point-Biserial coefficient. It provides an estimate of the degree to which an individual item is measuring the same thing as the rest of the items (Assessment, 2017). The accepted range of DisI is > 0.3 (20). Some other studies say that 35-100% is considered acceptable, while the 'very good' is above or equal to 60%. Some suggested the modified ranges as shown in table 1.3 (Gupta *et al.,* 2022; Sahoo & Singh, 2017). The Point-Biserial Correlation (PBS) is the Pearson correlation between responses to a particular item and scores on the total test. The Biserial Correlation models the responses to the item to represent stratification of a normal distribution and computes the correlation accordingly (Bonett, 2020). Again the point Biserial rpbis ranges are +1 (plus one) to -1 (minus one). The Biserial is always more extreme than the point- Biserial. The Point-Biserial Correlation (PBS) connects an examinee's item scores with their total test scores. A more robust version, the Corrected Point-Biserial Correlation, calculates the relationship after removing the item score, especially important for short tests where one item significantly impacts the total score (Acharya & Tippett, 2022).

Table 2. Different categories of DisI/PBS and their interpretation

Category	rpbis Value	The suggestion about the item/question
Very good (VG)	≥ 0.40	Retained
Good (G)	0.30-0.39	Reasonable to be retained
Below Standard (BS)	0.20-0.29	Marginal (subject to improvement)
Poor (P)	≤0.19	Rejected if not improved by revision

Distractor Efficiency (DE)

Distractors in MCQs significantly impact test scores, as they must be plausible and close to the key answer. Distractor Efficiency (DE) measures the effectiveness of these distractors, indicating whether they are well-chosen or failed to distract (Ansari *et al.,* 2022). Functional and Non-Functional Distractors are options/distractors selected by over 5% of examinees, respectively, while NFD is the option/distractor selected by less than 5% (Shakurnia *et al.,* 2022). Many factors negatively impact the DE, such as item writing errors (31). Distractor Efficiency (DE) is determined for each item/question based on the number of NFDs. For the Type-A best of four MCQs, it ranges from 0, 33.3%, 66.6%, or 100% (Sajjad *et al.,* 2020).

Horst Index (HI)

The Horst Index (HI) is a statistical measure developed in the 1950s to improve item/question reliability. It measures the difference between correct answer choices and popular distractor choices, divided by total item sats. HI helps identify items with potential quality issues, such as incorrect question answers or outdated teaching. Its central concern is improving item reliability (Farhat *et al.,* 2012). Nineteen anthropometric measures were taken from two racial groups, and a procedure was developed to maximize differentiation. This method is applicable to large-number independent and dependent variables, allowing for rapid estimates of regression weights and multiple correlations at each step (Horst & Smith, 1950). The formula of the Horst Index= [(Frequency of examinees/candidates who have chosen the correct key answer) – (Frequency of examinees/candidates who have selected the most popular distractor)] / the total number of examinees/candidates.

Reliability (internal consistency)

Reliability refers to the consistency of measured data, reducing error variance. Higher reliability results in better test performance and items within it. Two commonly used formulas for calculating reliability coefficients are Cronbach's alpha 2, a generalized Kudor-Richardson 20, and Backhouse's specific alpha coefficient (Henson, 2001). The examinees/candidates' scores of the examination are divided randomly into odd and even numbers in two halves, and the extent of correlation between them is measured. A level of 0.8 or more is considered reliable (Chidozie & Orluwene, 2021). The University of Washington, referring to Psychometric Theory, is ranging it as shown in table 1.5 (Assessment, 2017).

Table 3. Different reliability ranges and its interpretations

Reliability	Interpretation
>0.90	Excellent reliability; at the level of the best-standardized tests
0.80 –0 .90	Very good for a classroom test
0.70 – 0.80	Good and suitable for a classroom test; in the range of most. There are probably a Few items which could be improved.
0.60 – 0.70	Somewhat low. This test needs to be supplemented by other measures (e.g., more Tests) to determine grades. There are probably some items which could be improved. Suggests need for revision of test, unless it is pretty short (ten or fewer items). The
0.50 – 0.60	Test definitely needs to be supplemented by other measures (e.g., more tests) for grading.
≤ 0.50	Questionable reliability. This test should not contribute heavily to the course grade, And it needs revision.

Central Assessment Committee and Examination Bank

The massification of higher education in medicine and other allied health faculties has led to an unprecedented increase in the number of candidates applying to postgraduate studies in PGMY (Pillay, 2022). The Entry exam and the promotion exams include MCQs. Hence, the quality of which is expected to be assessed by medical education experts forming the Central Assessment Committee (CAC) (Nakanishi *et al.,* 2021). CAC has many tasks to perform and evaluate, such as standard-setting as an indispensable mainstay step in making fair decisions on who to pass and who to fail. Moreover, CAC provides structured feedback and training to develop an Examination Bank (Bittner *et al.,* 2020). There is an increasing demand for feedback from the candidates and their relatives. Teaching staff also need feedback on the quality of the items they set. In response to these key teaching and learning issues, examination bank and computer-based analyses are mandatory (Adhikari, 2021).

METHODOLOGY

This analytical cross-sectional study carried out a post-validation item analysis of the MCQ exam, test statistics, and reports of a selected MD Qualifying Entry Exam (first part) at PGMY. The selected MD Qualifying Entry Exam was one of the Specialty Councils Exams Conducted in the E 2020-2023 QRE Cycle. For the purpose of this study, Specialty Council X of PGMY is abbreviated as (SpX).

Study Sample

The study used Non-Probability Purposive Sampling Technique and Judgmental Extreme Case Sampling to select Specialty Council X with the lowest pass rates (10.86%) for a 2020-2023 QRE Cycle consisting of 120 MCQs, 480 options, 120 key answers, and 360 distractors.

Data Collection Methods and Techniques

The SpX exam data were collected from the PGMY Exam Office. These were: (1) The exam paper which contained 120 Type-A, Best of Four, MCQs (2) Condensed Test Report, (3) Detailed Item Analysis Report, (4) Item Analysis Graph Report, (5) Class Frequency Distribution Report, Student response (6).

Data Analysis

The researcher analyzed data from Excel sheets, examining test statistics, MCQ flaws, item analysis, and admission characteristics of candidates and specialty. The analysis revealed findings on performance, MCQ writing flaws, item analysis, specialties, university curriculum, and state location. The Discriminating Index (DifI) ranged from 0% to 100%, based on the difference between high and low achiever scores. The larger the difference between the high- and low-achieving groups, the higher the DisI of an item. The DisI of items ranged from -1 (all and only low group answered correctly) to +1 (all and only high group answered correctly). The categories were adopted as mentioned in table 1.3. The Mean Distractor Efficiency (mDE) is reflected by its reciprocal Non-Functional distractors (NFDs) per item. An NFD is defined as a poor distracter when the MCQ option is selected as an answer (response) by less than 5% of candidates. The categories were adopted as mentioned in table 1.3 and 1.4. The examination's reliability will be assessed using the Kuder-Richardson formula 20 Coefficient (KR20), with higher values indicating better reliability, ranging from <0.3 to ≥ 0.7.

RESULTS

The study analyzed the 2020-2023 QRE Cycle Entry Exam of PGMY Specialty X, focusing on writing language errors, MCQ flaws, item analysis, and test statistics. Results were categorized into four domains: exam success rate, MCQ construction, item analysis, and specialty characteristics.

Exam success rate and candidates performance analysis

The PGMY SpX Entry 2020-2023 QRE Cycle full marks were 120; one mark for each MCQ (item). Candidates scores ranged between 81 (67.50%) and 26 (21.67%), the Pass Mark was 72 (60.00%), the mean score was 55.11 (45.83%) and no candidate scored the range of Grade A, B or C as shown in table 4 and table 5.

Table 4. Candidate's score report of the PGMY SpX Entry Exam; Jan 2020-2023 QRE Cycle

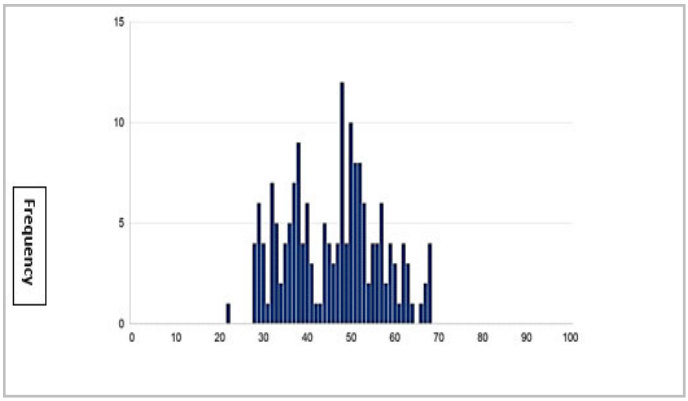
Full marks	Maximum Score	Minimum Score	Pass Mark	Mean Score	Standard Deviation	Median Score
120	81 (67.50%)	26 (21.67%)	72 (60.00%)	55.11 (45.83%)	12.72	57

The score histogram was plotted in Figure 3.1, where the prominent marks were around the 50th percentile.

Table 5. Candidate Grades and Exam success rate of the PGMY SpX Entry Exam; Jan 2020-2023 QRE Cycle

Grade	Percent Score	Raw Score out of 120	Number	%	Decision
D	60.00 - 69.99	72.00 - 83.99	19	10.86%	Pass
F	0.00 - 59.99	0.00 - 71.99	156	89.14%	Fail
Total			175	100%	

Table 6. The candidates' scores histogram of the PGMY SpX Entry Exam; Jan 2020-2023 QRE Cycle



MCQs Construction and Design Results

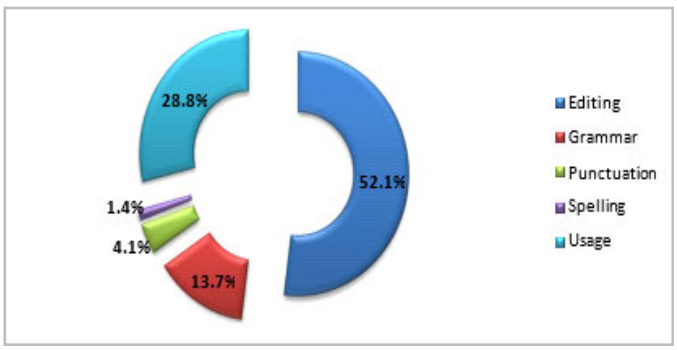
The study identified MCQ flaws in the USA National Board of Medical Examiners (USA NBME), with 7 items without flaws and 42.2% fulfilling the Cover-the-Options test, with non-vignette type and alphabetical options.

Table 7. Construction format; Lead-in, Stem and Options of the PGMY SpX Entry MCQs 2020-2023 QRE Cycle. N=120

The MCQ format issue	Lead-in		Stem
Type	Interrogative required style	Cover-the-options	Vignette
Number/Percentages N=120	12 (10%)	41 (34.2%)	2 (1.7%)

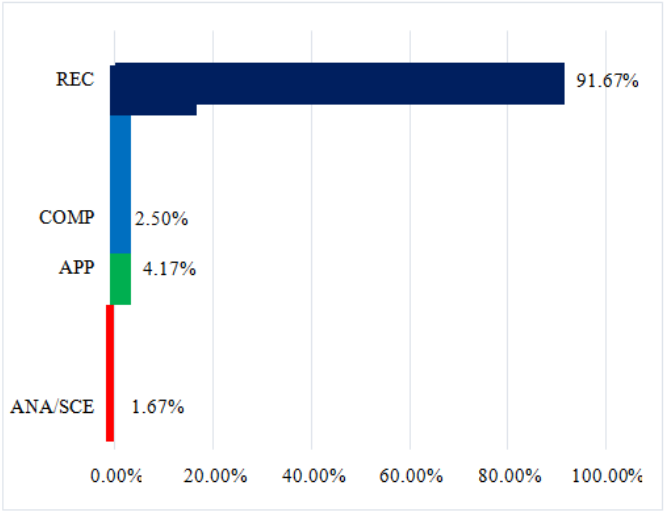
The study identified 24 out of 120 MCQs (20%) with flaws related to Testwiseness and irrelevant difficulties. The flaws included long correct answers (10%), absolute terms (5.8%), and logical cues (1.2%). The flaws related to irrelevant difficulties were found in long, complicated options (5%) and non-homogenous options (11.7%). Out of 120 MCQs, 59 (49.2%) had written language errors, with editing errors being the most common at 52.1%, and spelling errors at 1.4%.

Table 8. The writing language errors of the PGMY SpX Entry MCQs 2020-2023 QRE Cycle. N=120



The Blooms' Seven levels were studied in the SpX exam; Recall (REC) level was the majority 110 Qs (91.67%), Comprehension (Comp) in three Qs (2.50%), Application (App) level was five Qs (4.17%) and Analysis including Scenarios (ANA/SCE) were only found in two Qs (1.67%). It was illustrated in figure 3.6.

Table 9. Bloom's levels of the PGMY SpX Entry MCQs 2020-2023 QRE Cycle. N=120



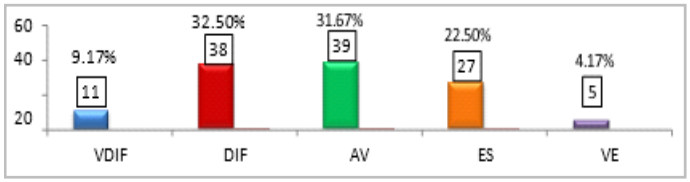
Items Analysis and Indices results

Remark software provides detailed statistical analysis of candidates' scores, exam reliability measurements, Item Difficulty Index (DifI), Point-Biserial Correlation (as a measure of item discrimination) and a detailed distracter analysis.

Difficulty index

The Difficulty Index (DifI) five categories were demonstrated in Figure 3.7.as follows; Very Difficult (VDIF) ≤ 20: was 11 (9.17%), Difficult (DIF) 21-40: was 39 (32.50%), Average (AV) 41-60: was 38 (31.67%), Easy (ES) 61-80: was 27 (22.50%) and Very Easy (VE) > 80: was 5 (4.17%).

Table 10. Difficulty Index Categories of the PGMY SpX Entry MCQs Exam; Jan 2020-2023 QRE Cycle. N=120



The Point Biserial (PBS) & the Discrimination Index (DisI) Categories

The Point Biserial/Discrimination Index (PBS) were categorized into four levels in figure 3.8 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) ≤ 0.19-0:53 (44.17%) (Among this category, 18 items 15% of the total were below 0; in minus value). No significant difference was found when comparing the Point Biserial and the discrimination Index; the p-value is 0.96 ($p > 0.05$).

Reliability Coefficient

The PGMY SPX Entry 2020-2023 QRE Cycle was reliable with a Reliability Coefficient of 0.85. 175 candidates from 35 universities passed the exam, but the classification based on the curriculum type was insignificant (p-value 0.946).

Table 11. MCQs Point Biserial (PBS) Categories of the PGMY SpX Entry MCQs Exam; Jan 2020- 2023 QRE Cycle. N=120

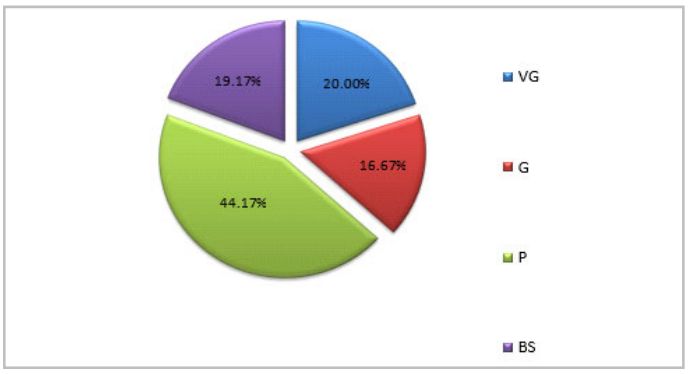


Table 12. the curved non-linear relationship between DifI and the DisI/PBS of the PGMY SpX Entry Exam; Jan 2020-2023 QRE Cycle

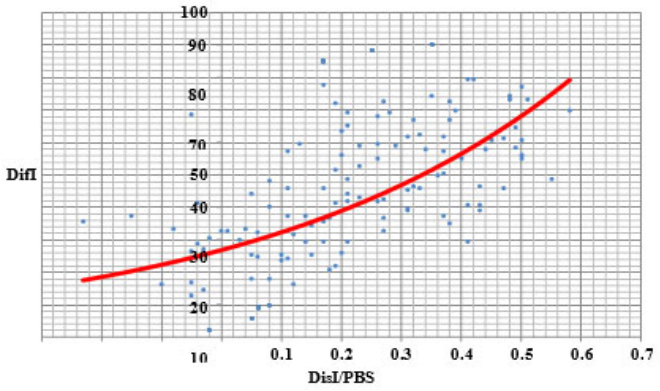


Table 13. Number of candidates who passed the PGMY SPX Entry 2020-2023 QRE Cycle versus type of the curriculum

Type of Curriculum	Number of all candidates per Curriculum Type	Number of successful candidates	% Success Rate
Traditional	36	7	19.44%
New "SPICES"	82	9	10.98%
Hybrid	53	3	5.66%
Unspecified	4	0	0.00%
Total Number	175	19	10.86%

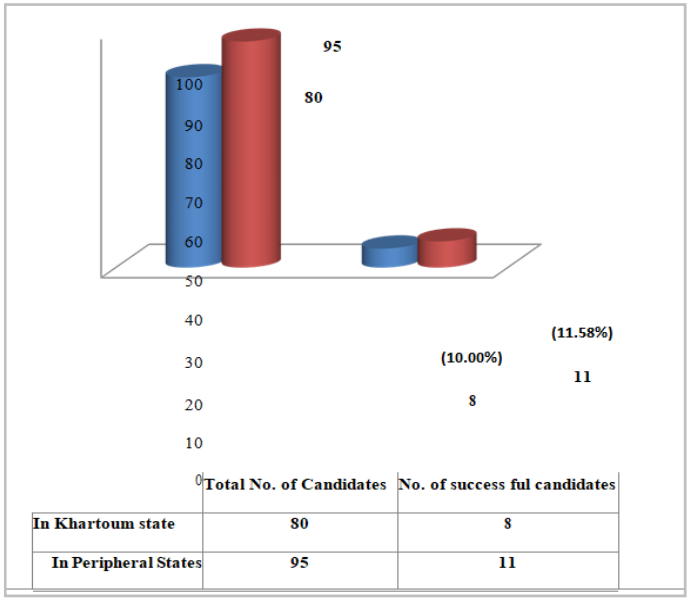
Table 14. The P-value (0.946) > 0.05; insignificant association

Table 3.9 classifies candidates/examinees who passed the PGMY SpX Entry Exam from 2016 to 2020-2023 QRE Cycle based on university location, with no significant difference found between Khartoum State and Peripheral States universities.

The JAN 2020-2023 QRE Cycle Exam candidates/ examinees have been classified based on their university location

The university's location from which those passed 19 candidates/examinees were graduated has no statistical difference between the Khartoum State universities and the Peripheral States universities, as shown in figure 3.12.

Table 15. The Number and percentages of successful candidates/examinees in the PGMY SpX Entry Exam; Jan 2020-2023 QRE Cycle per the location of their universities in Khartoum or peripheral states



	Total No. of Candidates	No. of success ful candidates
In Khartoum state	80	8
In Peripheral States	95	11

Table 3. 10: The Number and percentages of successful candidates/examinees in the PGMY SpX Entry Exam; for the last five years, 2016 to 2020-2023 QRE Cycle, per the location of their universities in Khartoum state or peripheral states

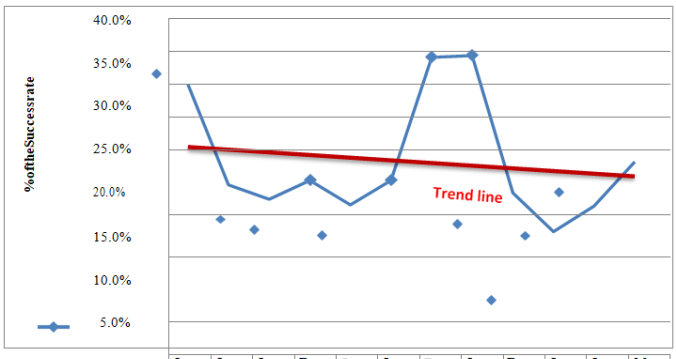
Location of the Candidates/examinee's University	No. of all candidates	No. of pass	Percent%
In the Khartoum State	266	57	21.43%
In the Peripheral States	313	73	23.32%
Total	579	130	22.45%

The p-value is (0.664).) > 0.05; insignificant association.

The Trend of the Pgmy Spx Entry Exam for the Past Years

The success rates from the year 2013 to the year 2020-2023 QRE Cycle demonstrated the zig-zag plot points. The highest success rate in June 2018 reached 35.0%, to the lowest 6.8% in June 2019.

Figure 3.11. The success rate graph and trend line of the PGMY SpX Entry Exam from 2013 to 2023-2020 QRE Cycle



	Jun-13	Jun-14	Jun-15	Dec-15	Jun-16	Jun-17	Dec-17	Jun-18	Dec-18	Jun-19	Jan-20	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%

DISCUSSION

This study is the first of its kind in Sudan for published postgraduate and undergraduate studies, using purposive judgmental sampling to analyze extreme deviant exam results (Mahees *et al.,* 2021). The exam was the Specialty Council X entry qualifying exam at PGMY, chosen from the lowest three success rates in the 2020-2023 QRE Cycle (Koubrak, 2023). The PGMY Specialty Council X uses MCQ-type A for written exams, consisting of 120 Best of Four items with 480 options (Dwivedi, 2019). The examinations are secure, with top security measures taken by the PGMY, Academic Affairs Secretariat, and Central Examination Department. The marking system is electronic and computer-based, with a pass rate of 10.85% (37). The success rate of candidates depends on exam quality, candidate academic and social issues, curriculum implementation, and surrounding environment (Calderon & Nagy, 2020). This study analyzes the SpX 2020-2023 QRE Cycle exam, focusing on Sudanese candidates with similar backgrounds. It evaluates test performance, technical item flaws, item quality parameters, and examines factors like candidate background, exam committee, and success rate since 2013 (DIMA et al.). The study examines SpX 2020-2023 QRE Cycle MCQ construction and identifies flaws in the USA National Board of Medical Examiners' book (Moretti *et al.,* 2021). The study found that only 7.8% of items without flaws were in the technical MCQs, highlighting the importance of proper construction for effectiveness. A study found 48 flaws in 63 Qs out of 120 items (52.5%), with some items having multiple flaws. The main flaws were long correct answers (10%), absolute terms (5.8%), and logical cues (1.2%).

The remaining 24 (20%) were "irrelevant difficulty" flaws, including non-homogenous options (11.7%), long, complicated options (5%), and vague options (2%). Kwoash's study on Paediatric Dentistry Postgraduate Examinations revealed common flaws in lead-in, stems, and convergence strategy usage, with 17.7% and 13.3% errors respectively (Grindrod *et al.,* 2020). Factors attributed include designer experience, training, and time. Short-duration faculty training is insufficient for correcting flaws in Multiple Choice Items writing. The committee member with a postgraduate degree in Medical Education suggests that flaws in the SpX 2020-2023 QRE Cycle exam may have negatively affected the success rate, potentially affecting the MPL (Mean Percentage of Solved Questions) of 60% (≥ 72 Qs out of 120). This assumption is based on the known university cutoff score for higher postgraduate studies. Downing's analysis of four medical student examinations supports this claim (49). Tarrant and Ware found flaws in high-stakes medical exams penalizing high-performing examinees more than average ones, and "Construct Irrelevance Variance" in pathology MCQs hindering meaningful interpretation of scores.

The study analyzes the SpX 2020-2023 QRE Cycle exam's reliability, difficulty, discrimination/point Biserial indices, and distractor efficiency post-validation. It evaluates correlations between these indices and technical flaws, using Pearson chi-squared test ($\chi 2$ test) for suitability over Fisher's exact test. The SpX 2020-2023 QRE Cycle exam's Difficulty Index showed that 91.7% of items were in the middle range, with a mean of 45.9 ± 4.52. This is consistent with other studies, such as the 2nd Year Nursing Qualifying Exam in Cavite, Philippines, and the undergraduate Physiology Department exam at the University of Khartoum. Some other studies have a wide range of difficult indices (22) who studied 12 Pre-clinical Semester Multidisciplinary Summative Tests from 2003 to 2006 in International Medical University-Malaysia. Their mean difficulty index scores of the individual tests ranged from 64% to 89%. On the other hand, some studies revealed DifI as low as (38.34±2.25). All of these factors were not well elaborated on in those studies. Even though 96 Qs (80%) of the "Recall" level were within the "Acceptable DifI category," Chi-Square Tests failed to show statistical significance between Bloom's level and DifI. Nevertheless, there was a significant statistical association between item constructional collective flaws and DifI at $\alpha = 0.05$. The findings suggest that each MCQ quality measure is a standalone assessment tool, particularly for high-stakes postgraduate exams, with medical educators recommending mean value based on average difficulty index trends (Pugh *et al.,* 2020).

The study analyzed Discrimination Index (DisI) and Point Biserial Correlation (PBS) in four categories based on rpbis value, a term often used interchangeably to describe an item's ability to differentiate high and low scorers. The study found that the mean DisI/PBS was 0.17 ± 0.02, with 36.7% in upper categories and 63.2% in lower categories. 15% of items in the P category were below 0 with minus values, suggesting lower ability candidates answered more correctly. The wide range of DisI/PBS indicates inconsistency in the Entry Exam planning, which was not based on a curriculum map or well-structured blueprint. This suggests a need for future review and rejection of negative values. The DifI and DisI/PBS indices were positively correlated, but not linear or pyramidal. Extreme DifI categories were predominantly in the poor discriminating category. The PGMY SpX Entry Exam's distractibility was measured using Distractors Efficiency (DE) and HORST Index (HI). The HI was 0.23 ± 0.02, with negative HI in 23.4%) items (Tomasevich *et al.,* 2022). HI was significantly related to Bloom's level, possibly due to question errors or incorrect teaching. The study by Ismail Burud found that the mean distractor efficiency was 66.0% ± 0.09, indicating that distractors should be reviewed, despite the low variation in studies in this area of DE (Burud *et al.,* 2020). The DE study found constructional issues in MCQs, suggesting the Best of Three MCQs is superior. No significant relationship was found between DE and Dis/PBS, but a statistically significant co-relation was found between DifI categories and FD. The PGMY SpX Entry Exam Reliability (KR20) was estimated at around 0.85, acceptable for type-A exams (Ntumi *et al.,* 2023).

Conclusion

The study analyzes the MCQs items indices and test statistics of a postgraduate exam, the Entry MD Qualifying Exam of Specialty Council X at Sudan Medical Specialization Board. The exam had the lowest success rate of the Jan 2020-2023 QRE Cycle. The study found editing errors in 60% of the exam items, with 40% being constructional Testwiseness and irrelevant MCQs flaws. The mean Distractor Efficiency was 66.0% ± 0.09, with one and two NFDs. The study found significant associations between MCQS flaws and DifI and DisI/PBS, with some indices having statistical significance. Despite these low markers, the exam showed reasonable internal consistency with 0.85 reliability. No significant difference was found between success rate and university curriculum types, and Khartoum State universities had no superiority over peripheral states.

Recommendations

For the PGMY leadership

- Prioritise the establishment of the central assessment committee to ensure the transparency, validity, and quality assurance of examinations.
- Infest in the development of and exam software, leveraging the international expertise and partnerships with the authenticated exam bodies and through exam soft companies.

For the Specialty Council

Implement the structured residency program model in offering the aspiring medical professionals with hands-on training and mentorship opportunities.

For the Candidates/Examinees:

- Take an informed approach to specialty selection by considering individual strengths, interests, and career aspirations, aligning with an intelligent career pathway approach.
- Prepare for the exam diligently by focusing on the prescribed syllabus and blueprint, ensuring comprehensive coverage of all relevant topics and competencies.

Acknowledgment: We acknowledge that this research was conducted without any external funding.

Statement of Competing Interests: The authors have no competing interests to declare.

List of Abbreviations

SpX - Specialty
QRE - Qualifying Residency Exam
PGMY - Postgraduate Medical Institute "Y"
MCQs - Multiple-Choice Questions
DifI - Difficulty Index

REFERENCES

Abd-Rahman, A. N., Baharuddin, I. H., Abu-Hassan, M. I., & Davies, S. J. (2021). A comparison of different standard-setting methods for professional qualifying dental examination. *Journal of dental education*, *85*(7), 1210-1216.

Acharya, N., & Tippett, M. K. (2022). Point-biserial correlation-based skill scores for probabilistic forecasts. *Authorea Preprints*.

Adhikari, C. L. (2021). Curriculum for Masters in General Practice–Bhutan. *Journal of Family Medicine and Primary Care*, *10*(6), 2061.

Ansari, M., Sadaf, R., Akbar, A., Rehman, S., Chaudhry, Z. R., & Shakir, S. (2022). Assessment of distractor efficiency of MCQS in item analysis. *The Professional Medical Journal*, *29*(05), 730-734.

Assessment, O. (2017). Understanding Item Analysis. *University of Washington*.

Betts, A., Thai, K.-P., & Gunderia, S. (2021). Personalized Mastery Learning Ecosystems: Using Bloom's Four Objects of Change to Drive Learning in Adaptive Instructional Systems. International Conference on Human-Computer Interaction,

Bittner, D. O., Mayrhofer, T., Budoff, M., Szilveszter, B., Foldyna, B., Hallett, T. R., Ivanov, A., Janjua, S., Meyersohn, N. M., & Staziaki, P. V. (2020). Prognostic value of coronary CTA in stable chest pain: CAD-RADS, CAC, and cardiovascular events in PROMISE. *Cardiovascular Imaging*, *13*(7), 1534-1545.

Bone, A., & Maharg, P. (2019). Introduction: Legal education assessment in England. *Critical Perspectives on the Scholarship of Assessment and Learning in Law*.

Bonett, D. G. (2020). Point-biserial correlation: Interval estimation, hypothesis testing, meta-analysis, and sample size determination. *British Journal of Mathematical and Statistical Psychology*, *73*, 113-144.

Bourque, J., Skinner, H., Dupré, J., Bacchus, M., Ainslie, M., Ma, I. W., & Cole, G. (2020). Performance of the Ebel standard-setting method for the spring 2019 Royal College of Physicians and Surgeons of Canada internal medicine certification examination consisting of multiple-choice questions. *Journal of Educational Evaluation for Health Professions*, *17*.

Brown, G. T., Denny, P., San Jose, D. L., & Li, E. (2021). Setting standards with multiple-choice tests: A preliminary intended-user evaluation of SmartStandardSet. Frontiers in Education,

Burud, I. A. S., Alsagof, S. M. I., Ganesin, R., Selvam, S. T., Zakaria, N. A. B., & Tata, M. D. (2020). Correlation of ultrasonography and surgical outcome in patients with testicular torsion. *Pan African Medical Journal*, *36*(1).

Calderon, T. G., & Nagy, A. L. (2020). A closer look at research on CPA exam success. *Advances in Accounting Education: Teaching and Curriculum Innovations*, *24*, 165-178.

Chidozie, E. O., & Orluwene, G. (2021). Discrimination and Difficulty Indices of Senior Secondary Certificate Examination Multiple Choice Physics Questions from 2016–2018 in Rivers State. *Glob Acad J Humanit Soc Sci*, *3*.

DIMA, B., DIMA, S. M., & IOAN, R. The 'Short-Run'Impact of Expectations' Past Volatility on the Current Predictions of Investors. The Case of VIX. *The Case of VIX*.

Doorenweerd, C., Van Nieukerken, E. J., & Hoare, R. J. (2017). Phylogeny, classification and divergence times of pygmy leaf-mining moths (L epidoptera: N epticulidae): the earliest lepidopteran radiation on A ngiosperms? *Systematic Entomology*, *42*(1), 267-287.

Dwivedi, C. (2019). A study of selected-response type assessment (MCQ) and essay type assessment methods for engineering students. *Journal of Engineering Education Transformations*, *32*(3), 91-95.

Edelen, M. O., & Reeve, B. B. (2007). Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. *Quality of life research*, *16*, 5-18.

Elgadal, A. H., & Mariod, A. A. (2021). Item Analysis of Multiple-choice Questions (MCQs): Assessment Tool For Quality Assurance Measures. *Sudan Journal of Medical Sciences*, *16*(3), 334-346.

Gordon, C., Hughes, J., & McKenna, C. (2017). Assessment Toolkit II: Time-constrained examinations.

Grebin, N., Grabovska, S., Karkovska, R., & Vovk, A. (2020). Applying Benjamin Bloom's Taxonomy Ideas in Adult Learning. *Journal of Education Culture and Society*, *11*(1), 61-72.

Grindrod, M., Barry, S., Albadri, S., & Nazzal, H. (2020). How is paediatric dentistry taught? A survey to evaluate undergraduate dental teaching in dental schools in the United Kingdom. *European Journal of Dental Education*, *24*(4), 715-723.

Gupta, S., Hellwing, W. A., Bilicki, M., & García-Farieta, J. E. (2022). Universality of the halo mass function in modified gravity cosmologies. *Physical Review D*, *105*(4), 043538.

Henson, R. K. (2001). Understanding internal consistency reliability estimates: A conceptual primer on coefficient alpha. *Measurement and evaluation in counseling and development*, *34*(3), 177-189.

Iqbal, Z., Saleem, K., & Arshad, H. M. (2023). Measuring teachers' knowledge of student assessment: Development and validation of an MCQ test. *Educational Studies*, *49*(1), 166-183.

Koubrak, O. (2023). Effectiveness of Marine Species at Risk Conservation within the UNEP Regional Seas Programme: Taking Stock and Charting Future Courses.

Kumar, P., Shukla, B., & Passey, D. (2020). Impact of accreditation on quality and excellence of higher education institutions. *Revista Investigacion Operacional*, *41*(2), 151-167.

Lemke, M., Sen, A., Pahlke, E., Partelow, L., Miller, D., Williams, T., Kastberg, D., & Jocelyn, L. (2004). International Outcomes of Learning in Mathematics Literacy and Problem Solving: PISA 2003 Results From the US Perspective. Highlights. NCES 2005-003. *US Department of Education*.

Mahees, M., Amarasinghe, H. K., Usgodaarachchi, U., Ratnayake, N., Tilakaratne, W., Shanmuganathen, S., Ranaweera, S., & Abeykoon, P. (2021). A sociological analysis and exploration of factors associated with commercial preparations of smokeless tobacco use in Sri Lanka. *Asian Pacific journal of cancer prevention: APJCP*, *22*(6), 1753.

Marchant, J. R. (2021). *Assessing the validity of multiple-choice questions, using them to undertake comparative analysis on student cohort performance, and evaluating the methodologies used*

Moretti, M., Malhotra, A., Visonà, S. D., Finley, S. J., Osculati, A. M. M., & Javan, G. T. (2021). The roles of medical examiners in the COVID-19 era: a comparison between the United States and Italy. *Forensic Science, Medicine and Pathology*, *17*, 262-270.

Nakanishi, R., Slomka, P. J., Rios, R., Betancur, J., Blaha, M. J., Nasir, K., Miedema, M. D., Rumberger, J. A., Gransar, H., & Shaw, L. J. (2021). Machine learning adds to clinical and CAC assessments in predicting 10-year CHD and CVD deaths. *Cardiovascular Imaging*, *14*(3), 615-625.

Ntumi, S., Agbenyo, S., & Bulala, T. (2023). Estimating the Psychometric Properties (" Item Difficulty, Discrimination and Reliability Indices") of Test Items Using Kuder-Richardson Approach (KR-20). *Shanlax International Journal of Education*, *11*(3), 18-28.

Ozair, A., Bhat, V., & Detchou, D. K. (2023). The US Residency Selection Process After the United States Medical Licensing Examination Step 1 Pass/Fail Change: Overview for Applicants and Educators. *JMIR Medical Education*, *9*(1), e37069.

Pillay, D. (2022). Ngugi Wa Thiong'o. In *Encyclopedia of African Religions and Philosophy* (pp. 508-509). Springer.

PradeepKumar, A. R., Shemesh, H., Nivedhitha, M. S., Hashir, M. M. J., Arockiam, S., Maheswari, T. N. U., & Natanasabapathy, V. (2021). Diagnosis of vertical root fractures by cone-beam computed tomography in root-filled teeth with confirmation by direct visualization: A systematic review and meta-analysis. *Journal of endodontics*, *47*(8), 1198-1214.

Pugh, D., De Champlain, A., Gierl, M., Lai, H., & Touchie, C. (2020). Can automated item generation be used to develop high quality MCQs that assess application of knowledge? *Research and Practice in Technology Enhanced Learning*, *15*, 1-13.

Qureshi, M. (2020). *The impact of different teaching methods on the performance of physician assistants: a survey response from Canadian PAs on the perception of their clinical skills*

Rao, N. R. (2024). Globalization and Medical Education in a Post-Pandemic World. *The Mental Health of Medical Students: Supporting Wellbeing in Medical Education*, 12.

Sa-Ngiamsuntorn, K., Suksatu, A., Pewkliang, Y., Thongsri, P., Kanjanasirirat, P., Manopwisedjaroen, S., Charoensutthivarakul, S., Wongtrakoongate, P., Pitiporn, S., & Chaopreecha, J. (2021). Anti-SARS-CoV-2 activity of Andrographis paniculata extract and its major component andrographolide in human lung epithelial cells and cytotoxicity evaluation in major organ cell representatives. *Journal of natural products*, *84*(4), 1261-1270.

Sahoo, D. P., & Singh, R. (2017). Item and distracter analysis of multiple choice questions (MCQs) from a preliminary examination of undergraduate medical students. *International Journal of Research in Medical Sciences*, *5*(12), 5351-5355.

Sajjad, M., Iltaf, S., & Khan, R. A. (2020). Nonfunctional distractor analysis: An indicator for quality of Multiple choice questions. *Pakistan Journal of Medical Sciences*, *36*(5), 982.

Schwill, S., Kadmon, M., Hahn, E. G., Kunisch, R., Berberat, P. O., Fehr, F., & Hennel, E. (2022). The WFME global standards for quality improvement of postgraduate medical education: Which standards are also applicable in Germany? Recommendations for physicians with a license for postgraduate training and training agents. *GMS Journal for Medical Education*, *39*(4).

Shakurnia, A., Ghafourian, M., Khodadadi, A., Ghadiri, A., Amari, A., & Shariffat, M. (2022). Evaluating Functional and Non-Functional Distractors and Their Relationship with Difficulty and Discrimination Indices in Four-Option Multiple-Choice Questions. *Education in Medicine Journal*, *14*(4).

Shelley, A. W. (2020). Reverse Bloom: A new hybrid approach to experiential learning for a new world. *J Edu Innovat Commun*, *2*, 30-45.

Singh, T. (2021). *Principles of assessment in medical education*. Jaypee Brothers Medical Publishers.

Tawalare, K., Pawar, J., Tawalare, K., & Karade, R. (2020). Need of multiple choice questions (MCQs) in assessment criteria of BAMS curriculum. *Journal of Education Technology in Health Sciences*, *7*(2), 54-57.

Ten Cate, O., & Taylor, D. R. (2021). The recommended description of an entrustable professional activity: AMEE Guide No. 140. *Medical teacher*, *43*(10), 1106-1114.

Tomasevich, K., Ohlsen, S., O, D., Featherall, J., Aoki, S., & Mortensen, A. (2022). Poster 193: Increased hip distractability in the revision hip arthroscopy setting: A comparison between revision and native contralateral hips with an intra-operative axial stress exam under anesthesia. *Orthopaedic Journal of Sports Medicine*, *10*(7_suppl5), 2325967121S2325900754.

Velou, M. S., & Ahila, E. (2020). Refine the multiple-choice questions tool with item analysis. *International Archives of Integrated Medicine*, *7*(8), 80-85.

Wang, Y., Zeng, D., Li, J., & Wen, J. (2023). Does engagement in international alliances affect a firm's influence on domestic standard-setting? The moderating role of government–market relations. *Technology Analysis & Strategic Management*, 1-15.

****