Med & Health Jan 2026; 21(1): 111-127 https://doi.org/10.17576/MH.2026.2101.05

ORIGINAL ARTICLE

The Relationship of Difficulty Index, Discrimination Index and Non-Functioning Distractor in One Best Answer Questions for Undergraduate Medical Program in Universiti Kebangsaan Malaysia

MUHAMMAD NAZIM OTHMAN*, MOHD NASRI AWANG BESAR, MOHAMMAD ARIF KAMARUDIN, MOHAMAD NURMAN YAMAN

Department of Medical Education, Faculty of Medicine, Universiti Kebangsaan Malaysia, 56000 Cheras, Kuala Lumpur, Malaysia

Received: 28 August 2024 / Accepted: 24 June 2025

ABSTRAK

Mencipta soalan satu jawapan terbaik (One Best Answer, OBA) berkualiti tinggi adalah penting untuk penilaian pelajar perubatan yang sah, namun ia menimbulkan cabaran. Isu seperti 'difficulty index' (D) yang tidak sesuai, 'discrimination index' (R) yang lemah, dan 'non-functioning distractor' (NFD) boleh menjejaskan integriti penilaian. Kajian ini bertujuan untuk menyiasat hubungan antara D, R, dan NFD dalam soalan OBA di Universiti Kebangsaan Malaysia (UKM) untuk meningkatkan kualiti item dan keberkesanan penilaian. Menggunakan pendekatan kuantitatif, kami menganalisis 499 soalan OBA dari sesi akademik 2022/2023, termasuk pra-klinikal (Tahun 1 dan 2) (393 soalan) dan klinikal (Tahun 5) (106 soalan), menggunakan 'Smart Question Bank'. Kami mendapati terdapat korelasi dan regresi yang signifikan dalam pra-klinikal, di mana item yang mempunyai lebih banyak NFD menunjukkan korelasi positif yang kuat dengan item yang terlalu mudah ($D > 0.75$) dengan ($r = 0.81$, $p < 0.001$) dan menghasilkan keputusan regresi yang signifikan secara statistik ($F = 279.86$, $p < 0.001$). Manakala, untuk klinikal mempunyai lebih tinggi bilangan item dengan diskriminasi yang lemah ($R < 0.15$) pada 45.28% ($n = 106$), ini menunjukkan item tersebut tidak berupaya untuk membezakan antara pelajar berprestasi tinggi dan rendah, dan tidak menunjukkan korelasi yang signifikan ($r = 0.24$, $p = 0.059$) dan regresi ($F = 3.24$, $p = 0.059$) dengan NFD. Dengan mengenal pasti kelemahan melalui analisis item, kajian ini memberikan pandangan untuk meningkatkan kualiti soalan OBA dan membimbing amalan pembinaan item masa hadapan di UKM dan institusi lain.

Kata kunci: *Distraktor tidak berfungsi; indeks diskriminasi; indeks kesukaran*

ABSTRACT

Creating high-quality one best answer (OBA) questions is crucial for valid medical student assessments, yet it presents a challenge. Issues like inappropriate difficulty index (D), poor discrimination index (R), and non-functional distractors (NFDs) can compromise assessment integrity. This study aimed to investigate the relationship between D, R, and NFD in OBA questions at Universiti Kebangsaan Malaysia (UKM) to improve item quality and assessment effectiveness. Using a quantitative approach, 499 OBA

Correspondence: Muhammad Nazim Othman, Department of Medical Education, Faculty of Medicine, Universiti Kebangsaan Malaysia, Jalan Yaacob Latif, Bandar Tun Razak, 56000, Kuala Lumpur, Malaysia. Tel: +6013 2569072 Email: muhdnazim@ukm.edu.my

questions from the 2022/2023 academic session were analysed, including pre-clinical (Years 1 and 2) (393 questions) and clinical (Year 5) (106 questions), using the Smart Question Bank (SQB). We observed significant correlations and regressions in pre-clinical, items with more NFDs demonstrated a strong positive correlation with excessively easy items (D > 0.75) with ($r = 0.81$, $p < 0.001$) and yielded a statistically significant regression outcome ($F = 279.86$, $p < 0.001$). However, for the clinical had higher number of items with poor discrimination (R < 0.15) at 45.28% (n = 106), this indicated an inability of the item to differentiate between high and low-achieving students, and showed no significant correlation ($r = 0.24$, $p = 0.059$) and regression ($F = 3.24$, $p = 0.059$) with NFD. By identifying weaknesses via item analysis, this study provides insights for improving OBA question quality and guiding future item construction at UKM and other institutions.
Keywords: Difficulty index; discrimination index; non-functioning distractor

INTRODUCTION

Assessment is an important method that aims to determine results in the assessment process, such as pass or fail, in assessing candidates and providing crucial feedback. Each assessment has a particular purpose, targeting either summative outcomes or formative improvement. The one best answer (OBA) question format, a specialised type of multiple-choice question, is crucial for evaluating medical students' ability to apply knowledge and engage in critical thinking (Sam et al. 2019; Wahab et al. 2022). These questions present a clinical scenario, complete with patient history and symptoms. They require students to select the most appropriate response from several options, enhancing their decision-making and clinical reasoning skills (Hassan et al. 2018; Tarrant et al. 2009). OBA questions thus bridge theoretical knowledge and practical application, simulating real-world challenges that healthcare professionals face (Sam et al. 2019). Crafting such questions demands significant expertise and adherence to established guidelines by Balaha et al. (2022), Haladyna & Downing (1989), and Haladyna et al. (2002), which aim to reduce bias and improve clarity, ensuring the fairness and effectiveness of assessments.

Item Analysis

Item analysis describes the validity and reliability of test items, distinguishing between questions that effectively differentiate high-performing students from low-performing ones through metrics such as the difficulty index and discrimination index, as noted by (Considine et al. 2005). There are several components involved in the item analysis process, including the index of difficulty (D), index of discrimination (R), distractor efficiency (DE) and Non-functional distractor (NFD). The item's difficulty index (D) is critical in item analysis because it shows the proportion of students who answered the question correctly. Balance is crucial for accurately assessing student knowledge and adjusting the difficulty level of future tests to meet educational goals better. The accepted normal values of each parameter, which the Faculty of Medicine, Universiti Kebangsaan Malaysia (UKM) endorses, are shown in Table 1.

Difficulty Index (D)

The D serves as a quantitative measure of item difficulty level (too easy or too hard/difficult), reflecting the proportion of test-takers who correctly respond to a given question. Expressed as a decimal ranging from 0.0 to 1.0, or equivalently as a percentage from 0% to 100%, the index provides a direct indication of the item's challenge level. A value of 0.0 signifies that no examinees selected the correct answer, indicating an exceptionally difficult item, whereas a value of 1.0 implies that all participants answered correctly, suggesting an excessively easy question. As postulated by Linn (2008), the optimal range for the D in educational

TABLE 1: Distribution of the value of data according to NFD, D and R

Difficulty Index (D)	Discrimination Index (R)	Non-Functioning Distractor (NFD)
Poor D (Too Easy) (> 0.75)	Good R (> 0.15)	NFD 0 (DE = 100%)
Good D (Acceptable) (0.35-0.75)	Poor R (< 0.15)	NFD 1 (DE = 75%)
Poor D (Too Difficult) (< 0.35)		NFD 2 (DE = 50%)
		NFD 3 (DE = 25%)

Source: Smart Question Bank (SQB)

assessments typically resides between 0.31 and 0.60 (31% to 60%). This range is considered conducive to maintaining a balanced assessment that effectively differentiates student competencies without being overly simplistic or prohibitively complex. Items falling within this range are deemed to provide a suitable challenge, thereby maximising the diagnostic potential of the assessment in gauging student understanding and knowledge application. Deviations from this optimal range may compromise the assessment's ability to accurately reflect student performance, potentially leading to skewed interpretations of student knowledge.

Discrimination Index (R)

The R is a critical psychometric parameter that assesses how an item effectively differentiates between high-achieving and low-achieving test-takers. Ebel (1972) suggests that we typically calculate this index by comparing the performance of the upper and lower 27% percentiles of the test-taking population. The R yields values ranging from -1.0 to +1.0, where positive values indicate that high-achieving students are more likely to answer the item correctly, and negative values suggest the opposite. An ideal R exhibits a positive value, indicating that the item successfully differentiates between students of varying competence levels. A value of +1.0 signifies perfect discrimination, where all high-achieving students answer correctly, and all low-achieving students answer incorrectly. Conversely, a value of -1.0 implies that low-achieving students are more likely to answer correctly, indicating a problematic item. Items with low or negative R are generally considered ineffective, as they fail to reflect student knowledge accurately and may introduce bias into the assessment process. The R thus plays a pivotal role in ensuring the validity and reliability of assessment instruments by verifying their ability to measure student performance accurately.

Distractor Efficiency (DE)

DE evaluates the effectiveness of the incorrect response options, or distractors, in a multiple-choice question. This metric assesses the extent to which distractors function as plausible alternatives, effectively challenging students' understanding and reasoning. According to Tarrant et al. (2009), a distractor is considered non-functional if it is selected by less than 5% of the test-takers, indicating that it does not serve as a viable option for any significant portion of the student population. Kheyami et al. (2018) further suggest that in well-constructed multiple-choice questions (MCQs) with fewer than two non-functional distractors, the overall DE should ideally range between 60% and 90%. This range signifies that the distractors are sufficiently challenging, compelling students to critically evaluate their choices. High DE contributes to the overall validity and fairness of the assessment by ensuring that all response options are meaningful and contribute to the item's ability to differentiate between students of varying competence levels. Conversely, low DE or the presence of numerous non-functional distractors may compromise

the item's effectiveness, potentially reducing the assessment's ability to accurately measure student knowledge.

Non-Functioning Distractor (NFD)

NFDs are incorrect response options in MCQs that are selected by an exceedingly small proportion of test-takers, typically less than 5%, as defined by (Tarrant et al. 2009). The presence of NFDs undermines the psychometric integrity of assessment instruments by reducing the item's discriminatory power and potentially compromising its validity. By failing to serve as plausible alternatives, NFDs diminish the item's capacity to effectively differentiate between students of varying competence levels, thereby reducing the precision of the assessment. The elimination or revision of NFDs is crucial for enhancing the effectiveness of MCQs, ensuring that all response options contribute meaningfully to the assessment process. By minimising the occurrence of NFDs, educators can enhance the reliability and validity of evaluations, thereby ensuring that assessment outcomes accurately reflect student knowledge and understanding. The reduction of NFDs is, therefore, a critical component of item analysis, contributing to the overall quality and effectiveness of assessment practices.

Item Writing Flaws (IWF)

IWF occur when violations against the accepted guidelines for constructing MCQ questions which significantly impact their quality. In this research, we discusses how IWF leads to deficiencies in the quality of questions, subsequently lowering the overall quality of MCQs. Puthiaparampil & Rahman (2020) reported that IWF contributes to deficiencies in question quality, ultimately diminishing the overall quality of MCQs. A poorly constructed questions are revised based on the presence of IWF. By focusing on specific items or questions with IWF, the creator can optimise their time and avoid unnecessary revisions of entire questions. Questions with minimal IWF can be corrected more efficiently, reducing the need for guesswork in identifying which parts require revision. Moreover, numerous studies, including those by Brown and Abdulnabi (2017) and Pham et al. (2018), have consistently proven that IWF negatively affects MCQs.

At UKM, the OBA format constitutes a key component of written assessment, is designed to evaluate medical students' knowledge acquisition, comprehension, analytical reasoning and clinical decision-making abilities. The OBA format at UKM typically presents a clinical scenario in the question stem, followed by four answer options. Among these, one of them are the correct answer, while the remaining three serve as plausible distractors designed to assess the student's ability to distinguish the most appropriate response. Subject matter experts, including medical and university lecturers from various clinical and non-clinical departments develop the questions. Each question undergoes a rigorous validation process, including multiple verification and approval levels. The final stage involves a comprehensive vetting process at both the departmental and faculty levels to ensure the quality and reliability of the assessment.

Creating high-quality OBA questions for assessing medical students at UKM is complex and crucial for valid evaluations. Challenges include setting appropriate difficulty levels and ensuring questions can distinguish between high and low-achieving students. NFD can compromise question integrity, necessitating their removal. A systematic approach involving item analysis, assessing D, R and DE is essential to enhance question quality and exam fairness.
This study includes OBA questions derived from end-of-module assessments and final-year professional examinations, encompassing two distinct cohorts of medical students: the pre-clinical group (Year 1 and Year 2) comprised of 186 and 177 students respectively, and the clinical group (Year 5) comprised of 139 students.

A total of 393 OBA questions were selected from the end-of-module examinations across 16 modules in the pre-clinical cohort (Year 1 and Year 2), comprising four modules per academic

year (Table 2), all of which covered basic medical sciences subjects. For the clinical cohort (Year 5), 106 OBA questions were selected from the final professional examination, encompassing both medical- and surgical-based components (Table 3). All selected OBA questions (Pre-Clinical: 393; Clinical: 106) were analysed using the Smart Question Bank (SQB) system, and the generated item analysis values were extracted as raw data for further statistical analysis in this study.

This study's main objective was to evaluate the relationships between the number of NFDs and the D and R for OBA questions in undergraduate medical students in UKM, specifically for the pre-clinical (Year 1 and Year 2) and clinical (Years 5). This study hypothesises that; (i) There is a statistically significant correlation and regression relationship between NFD and Overall D and Overall R (H_1); (ii) There is a statistically significant correlation and regression relationship

TABLE 2: Number of the value of data according to NFD, D and R on different modules in pre-clinical years

Modules	Total Questions	Good D and Good R with 0 NFD	Good D and Good R with NFD	Poor D (Too Easy) (D>75%, R>0.15)	Poor D (Too Difficult) (D<35%, R>0.15)	Poor R (R<0.15)
Cellular Biomolecules	20	1	7	5	4	3
Tissues of body	19	4	2	9	2	2
Membranes & receptors	30	4	14	5	4	3
Metabolism	20	2	6	9	1	2
Human genetics	19	2	6	5	2	4
Infection & immunity	25	10	7	4	2	2
Mechanism of diseases	26	6	7	7	0	6
Musculoskeletal system	24	8	3	8	1	4
Blood & lymph	28	3	9	8	3	5
Cardiovascular sSystem	30	8	7	5	2	8
Respiratory system	25	9	3	8	0	5
Urinary system	20	4	6	7	0	3
Gastrointestinal & hepatobiliary system	30	8	7	6	1	8
Endocrine system	20	4	5	3	1	7
Neuro sciences	32	9	5	8	7	3
Reproductive system	25	6	3	9	4	3

D: difficulty index; R: discrimination index; NFD: non-functioning distractor. Source: Smart Question Bank (SQB)

TABLE 3: Number of the value of data according to NFD, D and R on the Professional Set 1 (Medical-Based) and Professional Set 2 (Surgical Based) for clinical year

Professional Set	Total Questions	Good D and Good R with 0 NFD	Good D and Good R with NFD	Poor D (Too Easy) (D > 75%, R > 0.15)	Poor D (Too Difficult) (D < 35%, R > 0.15)	Poor R (R < 0.15)
Professional Set 1 (Surgical Based)	58	7	8	6	6	31
Professional Set 2 (Medical-Based)	48	6	7	14	4	17

D: difficulty index; R: discrimination index; NFD: non-functioning distractor. Source: Smart Question Bank (SQB)

between NFD and Good D and Good R (H_2); (iii) There is a statistically significant correlation and regression relationship between NFD and Poor D (Too Difficult) (H_3); (iv) There is a statistically significant correlation and regression relationship between NFD and Poor D (Too Easy) (H_4); and (v) There is a statistically significant correlation and regression relationship between NFD and Poor R (H_5).

Research Significance

This study is crucial for medical education as it provides a focused investigation into OBA question quality within the UKM context, where the application of item analysis is highly context-dependent. By examining the relationships between D, R and NFD, the research offers tailored insights to enhance UKM's assessments. Furthermore, it used item analysis to improve low-performing items, preventing unnecessary removal from the question bank and promoting a sustainable assessment process. The actionable recommendations derived from this study were intended to equip the UKM faculty with a critical view on the quality of OBA items and to provide structured guidance for improving item construction, thereby leading to fairer and more effective assessments. Ultimately, this research extended beyond UKM, offering a valuable benchmark to improve assessment practices in medical education, ensuring competent healthcare professionals and enhancing educational outcomes.

MATERIALS AND METHODS

Study Location

This study was conducted at the Faculty of Medicine, UKM.

Study Design

This was a cross-sectional retrospective study that evaluated the correlation and relationship between parameters involved in item analysis, such as the D, R and NFDs.

Sampling & Data Collection

A purposive sampling method was employed to select OBA questions from two academic cohorts: the pre-clinical years (Year 1 and Year 2) and the clinical year (Year 5) during the 2022/2023 academic session. We only selected the OBA questions exclusively in Years 1, 2 and 5 of the 2022/2023 academic session specifically because the OBA questions underwent vetting twice at the departmental and faculty levels as compared to Years 3 and 4, where they only underwent vetting at the departmental level. Therefore, the content and construct validity of the OBA questions for Years 1, 2 and 5 were consistent. In the context of OBA question evaluation at UKM, Cronbach's alpha is a standard measure for assessing the reliability of test instruments, however in this study, we did not conduct reliability score calculations for all

499 OBA questions because it was not within the scope of our study. Data collection for this study was done by entering pre-calculated values from the SQB into a Microsoft Excel file. When all OBA question data has been obtained, we collected and reviewed the data. After selecting OBA questions from the end-of-semester exams for the modules involved and two sets of Professional exams, the study obtained 499 OBA questions, 393 from the pre-clinical years and 106 from the clinical year.

Data Analysis

The study results were analysed using Statistical Package for Social Science (SPSS) Version 29 (IBM Corp, Armonk, NY, USA). Data normality tests were conducted by measuring the values of Skewness and Kurtosis (Kim 2013). For the Skewness test, values that fell within the range of -2 to +2 and the Kurtosis test, values that fell within the range of -3 to +3 were accepted as the data showed normal distribution before conducting inferential statistical analysis (Table 4 & Table 5). Pearson correlation coefficient (*r*) was used to analyse each item's value generated from the item analysis parameters presented by SQB. Pearson's correlation was used to explore the relationship between variables (NFD, D, and R) and function to quantify the strength and direction of the linear association between these variables. Pearson correlation coefficient ranged from -1 to +1 indicated a positive or negative correlation. The correlation coefficient (*r*-value) interpretation was referred to (Schober et al. 2018), and a p-value of <0.05 was considered significant. A significant Pearson correlation was followed with a simple linear regression analysis. A simple linear regression was employed to examine the relationships between variables, focusing on a single independent variable (NFD). The choice to use serial regression and exclude multivariate regression was a deliberate decision to maintain focus on the predefined hypotheses concerning NFD as the only independent variable for this study.

TABLE 4: Descriptive statistical analysis of pre-clinical year

	Pre-clinical (n = 393)		
	Difficulty Index (D)	**Discrimination Index (R)**	**Non-Functioning Distractor (NFD)**
Mean	0.65	0.31	1.00
Skewness	-0.52	-0.13	0.58
Kurtosis	-0.60	-0.43	-0.54

D: difficulty index; R: discrimination index; NFD: non-functioning distractor. Source: Smart Question Bank (SQB)

TABLE 5: Descriptive statistical analysis of clinical year

	Clinical (n = 106)		
	Difficulty Index (D)	**Discrimination Index (R)**	**Non-Functioning Distractor (NFD)**
Mean	0.62	0.17	1.31
Skewness	-0.40	0.77	0.20
Kurtosis	-1.12	0.67	-0.95

D: difficulty index; R: discrimination index; NFD: non-functioning distractor. Source: Smart Question Bank (SQB)

RESULTS

Table 2 showed the modules involved in the assessment for the end-of-semester examinations of pre-clinical Year 1 and Year 2. These modules represented the pre-clinical years' wide range of basic medical science topics, with the total number of questions for each module varying from 19 to 32. For the clinical year, there were two sets of professional exams, with a total of 58 for the Set 1 (Medical-Based) and Set 2 (Surgical-Based) OBA questions (Table 3). The pre-clinical years had a higher number of Poor D questions (Too Easy), 106 (26.97%) and the Clinical year had a higher value of Poor R questions, 48 (45.28%) (Table 6).

The Correlation and Regression between NFD and Overall D and Overall R

In the pre-clinical year, there was a significant moderate positive correlation ($r = 0.66$, $p < 0.001$) and a significant regression [$F (1,391) = 312.30$, $p < 0.001$)] between NFD and Overall D. Additionally, there was a significant low-level negative correlation ($r = -0.39$, $p < 0.001$) and regression [$(F (1,391) = 71.53$, $p < 0.001)$] between NFD and the Overall R (Table 7).

In the Clinical year, similar patterns observed with a medium-level positive correlation ($r = 0.63$, $p < 0.001$) and a significant regression [$(F (1,104) = 68.88$, $p < 0.001)$] between NFD and Overall D. Furthermore, there was a significant low-level negative correlation ($r = -0.36$, $p < 0.001$) and regression [$(F (1,104) = 15.92$, $p < 0.001)$] between NFD and Overall R (Table 8).

The Correlation and Regression between NFD and Good D and Good R

Pre-clinical year, there was a significant low-level positive correlation ($r = 0.262$, $p < 0.001$). There was a significant regression [$(F (1,201) = 14.79$, $p < 0.001)$] between NFD and Good D. There was also a very low negative correlation significance ($r = -0.24$, $p < 0.001$) and regression significance [$(F (1,332) = 21.15$, $p < 0.001)$] between NFD and Good R (Table 7).

In the clinical year, there was no significant correlation ($r = -0.05$, $p = 0.76$) and no significant regression [$(F (1,35) = 0.09$, $p = 0.763)$] between NFD and Good D. There was also no significant correlation [$(r = -0.14$, $p = 0.266)$] and no significant regression [$(F (1,56) = 1.26$, $p = 0.266)$] between NFD and Good R (Table 8).

The Correlation and Regression between NFD and Poor D (Too Difficult)

Pre-clinical year, there was no significant correlation ($r = 0.17$, $p = 0.277$) and no significant regression [$(F (1,41) = 1.21$, $p = 0.277)$] between NFD and Poor D (Too Difficult) (Table 7) and for the clinical year, there was a significant negative correlation ($r = -0.73$, $p < 0.001$), and there was a

TABLE 6: The percentage of different categories of D and R for pre-clinical and clinical years

	Pre-clinical (n = 393)	Clinical (n = 106)
	Number	Number
Good D, Good R without NFD	88 (22.39%)	13 (12.26%)
Good D, Good R with NFD	97 (24.68%)	15 (14.15%)
Poor D (Too Easy)	106 (26.97%)	20 (18.86%)
Poor D (Too Difficult)	34 (8.66%)	10 (9.43%)
Poor R	68 (17.30%)	48 (45.28%)
Total	393 (100%)	106 (100%)

D: difficulty index; R: discrimination index; NFD: non-functioning distractor. Source: Smart Question Bank (SQB)

TABLE 7: Pearson's Correlation and Linear Regression analysis of NDF, D and R values for pre-clinical years

Hypothesis	Category	Pearson's Correlation		Linear Regression			Conclusion
		r	p-value	R^2	F	p-value	
H_1	NFD with Overall D	0.66	<0.001	0.44	312.30	<0.001	Supported
	NFD with Overall R	-0.39	<0.001	0.15	71.53	<0.001	Supported
H_2	NFD with Good D	0.26	<0.001	0.06	14.79	<0.001	Supported
	NFD with Good R	-0.24	<0.001	0.06	21.15	<0.001	Supported
H_3	NFD with Poor D (Too Difficult)	0.17	0.277	0.02	1.215	0.277	Not supported
H_4	NFD with Poor D (Too Easy)	0.81	<0.001	0.65	279.86	<0.001	Supported
H_5	NFD with Poor R	0.24	0.059	0.06	3.72	0.059	Not supported

D: difficulty index; R: discrimination index; NFD: non-functioning distractor. Source: Smart Question Bank (SQB)

TABLE 8: Pearson's Correlation and Linear Regression analysis of NDF, D and R values for clinical years

Hypothesis	Category	Pearson's Correlation		Linear Regression			Conclusion
		r	p-value	R^2	F	p-value	
H_1	NFD with Overall D	0.63	<0.001	0.39	68.88	<0.001	Supported
	NFD with Overall R	-0.36	<0.001	0.13	15.92	<0.001	Supported
H_2	NFD with Good D	-0.05	0.763	0.03	0.09	0.763	Not Supported
	NFD with Good R	-0.14	0.266	0.02	1.26	0.266	Not Supported
H_3	NFD with Poor D (Too Difficult)	-0.73	<0.001	0.54	24.819	<0.001	Supported
H_4	NFD with Poor D (Too Easy)	0.75	<0.001	0.56	56.56	<0.001	Supported
H_5	NFD with Poor R	-0.15	0.308	0.02	1.06	0.308	Not Supported

D: difficulty index; R: discrimination index; NFD: non-functioning distractor. Source: Smart Question Bank (SQB)

significant regression [(F(1,21) = 24.81, p < 0.001)] between NFD and Poor D (Too Difficult) (Table 8).

The Correlation and Regression between NFD and Poor D (Too Easy)

Pre-clinical year, there was a significant positive correlation (r = 0.81, p < 0.001). There was a significant regression [(F (1,145) = 279.86, p < 0.001)] between NFD and Poor D (Too Easy) (Table 7), and for the clinical year, there was a significant positive correlation (r = 0.75, p < 0.001). There was a significant regression [(F (1,44) = 56.56, p < 0.001)] between NFD and Poor D (Too Easy) (Table 8).

The Correlation and Regression between NFD and Poor R (weak discriminator)

Pre-clinical year, there was no significant correlation (r = 0.24, p = 0.059) and no significant regression [(F (1,57) = 3.72, p = 0.059)] between NFD and Poor R (Table 7) and for the clinical year, had no significant correlation (r = 0.15, p = 0.308) and no significant regression [(F (1,46) = 1.06, p = 0.308)] between NFD and Poor R (Table 8).

DISCUSSION

The pre-clinical year has a higher percentage of questions with a Poor D (Too Easy), especially in the 'Tissue of the Body' and 'Metabolism' modules during Semester 1. This is because, in the pre-clinical years, many questions focus on basic knowledge from core medical science topics. This aligns with broader concerns in medical education, where MCQ tests are frequently observed to evaluate memorisation rather than deeper understanding or critical thinking (Elabd 2021), potentially leading to a higher number of "Too Easy" questions. Too easy questions result in assessments focusing primarily on understanding and memorising facts rather than requiring analysis or application. Research by Teli & Kate (2022) indicated that questions classified as too easy or difficult often exhibit poor discriminatory power, reflecting a lower level of challenge. There are increased questions with poor discrimination power (Poor R) from Year 1 to Year 2; specifically, in Year 1, semesters one and two recorded a total of 10 and 16 with Poor R questions, while in Year 2, semesters two and three recorded a total of 21 with Poor R questions (Table 2). This trend may be due to the challenge of creating high-quality questions with strong discriminating power, especially for Year 2 students. This challenge continues into the clinical years, where the quantity of Poor R questions is higher, indicating the ongoing difficulty in generating questions that effectively assess complex real-life clinical scenarios. Abdulghani et al. (2017) highlight the challenges associated with writing MCQ items that effectively capture clinical scenarios, emphasising the importance of helpful guidelines and systematic training to improve the quality of MCQ items. This view is supported by Husain et al. (2023), which underscores the significance of incorporating clinical scenarios in MCQs to maximise the impact and validity of assessments.

Relationship between NFD with Overall D and Overall R for the pre-clinical and clinical Years

The significant correlations and regressions between NFD and Overall D for both pre clinical and clinical years (Table 7 & Table 8) conclude that the number of NFDs significantly affects the D value. The results are similar to those of previous researchers, who concluded that NFD has a specific effect on an item's D, making the questions either too difficult or too easy (Puthiaparampil & Rahman 2020; Sajjad et al. 2020; Shakurnia et al. 2023).

The presence of NFD may affect the quality of OBA questions, and this is supported by Burud et al. (2019), who state that questions that are too easy will affect the quality of the assessment by failing to test higher cognitive skills, such as analysis, and simple questions will only assess lower cognitive functions such as memorisation and understanding alone; as a result, this affects the validity of the assessment method.

In both the pre-clinical and clinical years, a significant, negative correlation and regression were observed between the number of NFDs and the R for the Overall R category (Table 7 & Table 8). This concludes that increasing the number of NFDs will significantly lower the value of the R in general, which results in items having a poor discriminatory power – Poor R. These findings suggest that NFD reduces the effectiveness of item discrimination. This observation aligns with the research of Shakurnia et al. (2023) and Hingorjo and Jaleel (2012), who showed that items with a higher number of NFD generally exhibit weaker discriminating ability, causing the question to be unable to discriminate between high-achieving and low-achieving candidates.

Relationship between NFD with Good D and Good R for the Pre-Clinical and Clinical Years

A weak positive correlation was observed between NFD and the D in the "Good D" category or good questions for the pre-clinical year - Hypothesis 2 (H_2) (Table 7). This shows that the number of NFDs significantly affects the good questions' D value. However, there is no significant correlation between NFD and the "Good D" or good questions in the clinical year - H_2 (Table 9). The number of NFDs will increase the difficulty level, reinforcing findings from previous studies, indicating that with more NFDs present, the D value of the questions will be increased (Abdulghani et al. 2014; Hingorjo & Jaleel 2012; Mahjabeen et al. 2017).

In contrast, a weak negative correlation was found between NFD and the R in the Good R category for the pre-clinical year - H_2 is supported (Table 7), indicating that a higher number of NFD will reduce the ability of items to discriminate between different levels of student performance effectively. This effect was not observed in the clinical year – H_2 is not supported (Table 8), where no significant correlation was found, indicating that NFD does not affect discriminative power for the Good R category. However, NFDs do have a significant relationship with Overall R for both pre-clinical and clinical years (Table 7 & Table 8), with a higher total number of Overall R with NFDs (Table 9). The larger dataset, especially for the Overall R with NFDs, may influence the analysis result of these categories.

There is a higher percentage of NFDs that has 0 or 1 per item seen in the pre-clinical years, indicating that the smaller number of NFDs in an item will be effective in discriminating between student performance, as there were higher percentages of Overall R seen (0 NFD - 137, 34.9%) (1 NFD - 149, 37.9%) (Table 9). This supports the notion that reducing the number of NFDs can improve the discriminatory ability of an item, which results in a higher quality of OBA. This is consistent with research by Abdul

TABLE 9: The number of NFDs with different categories of discrimination index (R) for pre-clinical and clinical years

Number of NFD	Discrimination Index (R)					
	Overall R		Good R		Poor R	
	Pre-clinical n (%)	Clinical n (%)	Pre-clinical n (%)	Clinical n (%)	Pre-clinical n (%)	Clinical n (%)
0	137 (34.9)	25 (23.6)	129 (38.6)	19 (32.8)	8 (13.6)	6 (12.5)
1	149 (37.9)	37 (34.9)	134 (40.1)	22 (37.9)	15 (25.4)	15 (31.3)
2	77 (19.6)	30 (28.3)	64 (19.2)	16 (27.9)	13 (22)	14 (29.2)
3	30 (7.6)	14 (13.2)	7 (2.1)	1 (1.7)	23 (39)	13 (27.1)
Total	393	106	334 (85)	58 (54.7)	59 (15)	48 (45.3)

D: difficulty index; R: discrimination index; NFD: non-functioning distractor. Source: Smart Question Bank (SQB)

Ghani et al. (2014), Puthiaparampil & Rahman (2020) and Shakurnia et al. (2023) that supports fewer distractions to maintain or improve item quality without compromising rating integrity. Overall, these findings highlight the influence of NFD on the validity and effectiveness of OBAs, emphasising the need for careful distractor analysis to optimise assessment tools in medical education.

Relationship between NFD and Poor D (Too Difficult) for the Pre-Clinical and Clinical Years

In the pre-clinical year, there was no significant correlation between the number of NFDs and the D in the Poor D (Too Difficult) – Hypothesis 3 (H_3) is not supported: (Table 7). This is because there are fewer Poor D items with NFDs (Total - 43, 10.9%): - (1 NFD - 11, 25.6%) and no Poor D that has either 2 NFDs (0, 0%) or 3 NFDs (0, 0%) (Table 10). Although there is no statistically significant relationship, the results conclude that an item with fewer NFDs will cause the item to become 'Too Difficult' even if they are smaller in number. These findings are similar to other research Rezigalla et al. (2024) and Shakurnia et al. (2023), where there is a significant positive correlation between the number of NFDs and the D of items, suggesting that items with fewer NFDs tend to be difficult, especially items with efficient distractors (fewer NFDs) correlate with lower D values (<0.35), indicating that effective distractors will improve the difficulty of the questions and of course eliminate any potential test-wiseness among test takers.

A strong negative correlation between NFD and Poor D (Too Difficult) - H_3 is supported: (Table 8), was observed for the clinical year, indicating that fewer NFDs corresponded to a lower D (<0.35), making items more challenging or 'Too Difficult'. This increased complexity is due to fewer NFDs in the items, leading to more plausible distractors that effectively distract the test-taker and hinder the ingenuity of the test, thereby making the item to be difficult.

TABLE 10: The number of NFDs with different categories of difficulty index (D) for pre-clinical and clinical years

Number of NFD	Difficulty Index (D)							
	Overall D		Good D		Poor D (Too Easy)		Poor D (Too Difficult)	
	Pre-clinical n (%)	Clinical n (%)	Pre-clinical n (%)	Clinical n (%)	Pre-clinical n (%)	Clinical n (%)	Pre-clinical n (%)	Clinical n (%)
0	137 (34.9)	25 (23.6)	94 (46.3)	14 (37.8)	11 (7.5)	1 (2.2)	32 (74.4)	10 (43.5)
1	149 (37.9)	37 (34.9)	91 (44.8)	19 (51.4)	47 (32)	7 (15.2)	11 (25.6)	11 (47.8)
2	77 (19.6)	30 (28.3)	18 (8.9)	4 (10.8)	59 (40.1)	24 (52.2)	0 (0)	2 (8.7)
3	30 (7.6)	14 (13.2)	0 (0)	0 (0)	30 (20.4)	14 (30.4)	0 (0)	0 (0)
Total	393 (100)	106 (100)	203 (51.7)	37 (34.9)	147 (37.4)	46 (43.4)	43 (10.9)	23 (21.7)

D: difficulty index; R: discrimination index; NFD: non-functioning distractor. Source: Smart Question Bank (SQB)

Relationship between NFD with Poor D (Too Easy) for the Pre-Clinical and Clinical Years

In the pre-clinical year, a notably high number of "Too Easy" questions pose a risk of transforming OBA questions into One Choice Answer (OCA) types, primarily testing memorisation and recall instead of application and analysis skills. This shift undermines the original intent of OBA methods, which aim to assess higher cognitive functions such as analysis by Belay et al. (2022) suggest that a high presence of NFDs in questions significantly lowers their difficulty, reducing their ability to differentiate between high and low performers. To enhance the validity of MCQ-based assessments, it is crucial to minimise NFDs, ensuring questions accurately and authentically assess candidates' knowledge and skills.

In the clinical year, the significant relationship between NFD and D in the Poor D (Too Difficult) (H_3) and Poor D (Too Easy) (H_4) - (Table 8), this shows that the correlation of NFD leading to item becoming either too difficult or too easy for clinical year indicating challenges in drafting questions that require clinical reasoning and higher-level thinking.

Hift (2014) underscores the necessity of well-constructed questions to boost the validity of assessments and better prepare students for clinical practice. Further studies align with these findings, showing that excessive NFD makes items either too difficult or too easy, negatively impacting the quality of OBA questions (Puthiaparampil & Rahman 2020; Sajjad et al. 2020; Shakurnia et al. 2023). Burud et al. (2019) notes that overly simplistic questions fail to test higher cognitive skills, such as analysis, and focus only on lower cognitive functions like memorisation and understanding, thereby compromising the assessment's validity.

Alternatives to OBA, such as very short answer questions (VSAQ) and context-rich short answer questions (CR-SAQs), are suggested to evaluate students' knowledge and critical thinking better. VSAQs, as described by Puthiaparampil and Rahman (2020), require students to provide brief, direct answers that assess their knowledge, critical thinking and expression. Similarly, CR-SAQs, according to Bahner et al. (2012), involve short answers within a contextual framework, allowing a deeper exploration of students' understanding, problem-solving and critical thinking skills.

Relationship between NFD and Poor R for the Pre-Clinical and Clinical Years

In the pre-clinical and clinical years, no significant relationship exists between the number of NFDs and the R in the 'Poor R' category. However, there was a significant negative correlation and regression between the number of NFDs and the 'Overall R' and 'Good R' categories for pre-clinical years (Table 7). The negative correlation shows that fewer NFDs lead to a lower R value, meaning the items could not differentiate between high- and low-achieving test-takers. This aligns with a study by Higorjo et al. (2012), which states that NFDs in MCQs can impact the item's discriminatory capabilities.

Furthermore, the negative relationship indicates that more NFDs reduce discrimination effectiveness. This finding aligns with Mahjabeen et al. (2017) and Abdulghani et al. (2014), who stated that items with more NFD are generally easier and have lower discriminative power. Additionally, research by Deepak et al. (2015) highlights that fewer distractors correlate with lower reliability and that MCQs with only three functional options barely meet acceptable psychometric standards.

In contrast, in the clinical year, there is no significant correlation ($r = 0.24$, $p = 0.059$) and regression ($F = 3.24$, $p = 0.059$) between Poor R ($R<0.15$) and NFD (Table 8). However, there is only a significant relationship between NFDs in the 'Overall R' category (Table 8). This is due to the Overall R would exhibit a higher total number of questions (48 questions, 45.28%) (Table 9) that includes both Good R and Poor R with NFDs (1 NFD - 37%, 34.9%), (2 NFD - 30%, 28.3%), and (3 NFD - 14%, 13.2%) (Table 9). The significance of this relationship is again possibly due to the larger dataset (Overall R - Total: 106) (Table 9) available compared to the smaller

dataset in the other categories separately, such as in the Good R - Total: 58 and Poor R - Total: 48 (Table 9). To address this problem of poor discrimination in clinical year OBA questions, it is recommended that the faculty enhance the pre-existing item writing workshop, particularly for new trainees. This enhancement should include targeted workshops emphasising the construction of questions that effectively assess clinical reasoning. Furthermore, these workshops should address common flaws compromising item quality and discriminatory power. We suggests that the faculty should address the Poor R item according to our recommended post-item analysis checklist flaws includes test-wiseness, irrelevant difficulty, followed by a process to conduct a major overview of items stem, lead-in and distractor. These recommendations are experience-based rather than evidence-based and should be interpreted with limitation in mind. Nevertheless, the importance of addressing flaws in distractor is supported by Ali & Ruit (2015), who highlight, the quality of distractors is paramount, as ineffective distractors contribute to low discrimination by failing to challenge students effectively. Their study on item flaws emphasises that poorly functioning distractors can lead to high-performing students selecting correct answers even in the absence of sound knowledge due to a lack of credible alternatives, thereby reducing the overall discriminatory power of the assessment. Therefore, alongside enhanced training, the faculty may ensure that question developers adhere to clear item writing guidelines or templates tailored for different clinical question types. Finally, to further safeguard question quality, it is advisable to mandate that all questions undergo two levels of vetting: departmental and faculty.

Research Strength

There are several key strengths in this study's methodology. Firstly, the purposive selection of the year of study for the OBA questions that underwent vetting at departmental and faculty levels enhances the reliability and relevance of the data. Serial regression was employed to examine the relationships between variables. The exclusion of multivariate regression techniques, which could address multiple independent variables, was a deliberate decision to maintain a focused analysis on the predefined hypothesis, with the number of NFDs as the sole independent variable in this study.

Limitations

This study is limited by its focus on a single institution. Therefore, further study is needed by expanding the scope of the dataset to include additional OBA questions from other academic years and other medical programs from different institutions, thus enhancing representativeness and allowing for broader generalisability of the findings. Employing a random sampling technique could also mitigate potential biases introduced by the purposive sampling method used in this study. The study recognises the significance of reliability scores in ensuring the consistency and dependability of assessment results; this study's primary focus is on a detailed item analysis of the 499 OBA questions. Conducting a comprehensive reliability analysis for each question would necessitate a separate, extensive investigation, falling outside the defined scope of this current research. A recommendation for future studies might consider integrating multivariate regression analyses to explore the impact of multiple independent variables simultaneously. Addressing the variability in the number of OBA questions across different modules or professional exams could be achieved by standardisation.

Recommendation and Suggestions

To address the prevalence of "Too Easy" items ($D > 0.75$) in the pre-clinical years, institutions should investigate the potential influence of "test-wiseness" among students. For the "Too Difficult" items ($D < 0.35$) observed in the clinical years, a critical review for sources of irrelevant difficulty is warranted. A systematic revision of the item's stem, lead-in, and distractors is

advised to improve the higher number of poor discriminator items (R < 0.15), in the clinical years. The ultimate aim is to have items with a good R, a good D, and zero NFDs where the item will be retained in the Question Bank. It is important to acknowledge, however, that while this study effectively showed the associations between the D, R and NFDs through correlation and regression analyses, the development of concrete, actionable recommendations for post-item analysis interventions was not within the scope of this study. Though informed by the invaluable, experience-based insights of expert medical educationists, the suggestions presented were not empirically tested as part of this research. Therefore, future studies are essential to rigorously validate the effectiveness of these proposed strategies in improving OBA question quality and ultimately strengthening medical student assessments.

CONCLUSION

The influence of NFDs on the D causes items to be perceived as 'Too Easy' in pre-clinical years and 'Too Difficult' in clinical years for their respective set of OBA questions. Additionally, NFDs will impact the value of the R by impairing the ability of items to differentiate between high- and low-achieving students, particularly in the clinical years. Effective item analysis is crucial to identifying factors that compromise the validity of OBA assessments. Enhancing the quality of questions ensures that assessments are both challenging and fair, thereby accurately measuring student competence and improving the overall validity of future assessments. Future study recommendations include expanding the research to incorporate previous academic years and other institutions within the medical program, to determine if the trends observed at UKM are consistent in other contexts. This would help validate the findings and enhance their applicability.

Author contributions: Conceptualisation, methodology, data analysis: MNO, MNAB, MAK; Manuscript-original draft: MNO, MNAB; Data interpretation analysis: MNO, MNAB, MAK, MNY; Manuscript-review and editing: MNO. All authors have approved the final manuscript.

Conflict of interest: The authors declare no conflicts of interest.

Funding: The authors did not receive funding for this study.

Acknowledgement: The authors would like to thank the Faculty of Medicine UKM for granting the permission to conduct this study and to express gratitude to all who were involved and supported this study.

Ethical statement: Ethical approval was obtain from UKM Research Ethical Commitee (JEP-2024-322). This study involved retrospective analysis of anonymised, system-generated item statistics generated from the Smart Question Bank, with no access to individual student scores, grades, or identifiable information.

REFERENCES

Abdulghani, H.M., Ahmad, F., Ponnamperuma, G.G., Khalil, M.S., Aldrees, A. 2014. The relationship between non-functioning distractors and item difficulty of multiple-choice questions: A descriptive analysis. *J Health Spec* **2**(4): 148.

Ali, S.H., Ruit, K.G. 2015. The Impact of item flaws, testing at low cognitive level, and low distractor functioning on multiple-choice question quality. *Perspect Med Educ* 4(5): 244-51. https://doi.org/10.1007/s40037-015-0212-x

Balaha, M.H., El-Ibiary, M.T., El-Dorf, A.A., El-Shewaikh, S.L., Balaha, H.M. 2022. Construction and writing flaws of the multiple-choice questions in the published test banks of Obstetrics and Gynecology: Adoption, caution, or mitigation? *Avicenna J Med* **12**(03): 138-47. https://doi.org/10.1055/s-0042-1755332

Bahner, D.P., Hughes, D., Royall, N.A. 2012. I-AIM: A novel model for teaching and performing focused sonography. *J Ultrasound Med* **31**(2), 295-300. https://doi.org/10.7863/jum.2012.31.2.295

Belay, L.M., Yigzaw, T., Abebe, F. 2022. Quality of multiple-choice questions in medical internship qualification examination determined by item

response theory at Debre Tabor University, Ethiopia. *BMC Med Educ* **22**(1): 635. https://doi.org/10.1186/s12909-022-03687-y

Brown, G.T.L., Abdulnabi, H.H.A. 2017. Evaluating the quality of higher education instructor-constructed multiple-choice tests: Impact on student grades. *Front Educ (Lausanne)* **2**: 24. https://doi.org/10.3389/feduc.2017.00024

Burud, I., Nagandla, K., Agarwal, P. 2019. Impact of distractors in item analysis of multiple-choice questions. *Int J Res Med Sci* **7**(4): 1136-9. https://doi.org/10.18203/2320-6012.ijrms20191313

Considine, J., Botti, M., Thomas, S. 2005. Design, format, validity and reliability of multiple choice questions for use in nursing research and education. *Collegian* **12**(1): 19-24.

Deepak, K.K., Al-Umran, K.U., Al-Sheikh, M.H., Dkoli, B.V., Al-Rubaish, A. 2015. Psychometrics of multiple-choice questions with non-functioning distracters: Implications to medical education. *Indian J Physiol Pharmacol* **59**(4): 428-35.

Elabd, K. 2021. Enhancing effectiveness of residents' virtual medical education during COVID-19 pandemic. *MedEdPublish* 10: 1-11.

Ebel, R.L., Frisbie, D.A., 1972. Essentials of educational measurement (1st ed.). Upper Saddle River, NJ: Prentice Hall.

Haladyna, T.M., Downing, S.M., Rodriguez, M.C. 2002. A review of multiple-choice item-writing guidelines for classroom assessment. *Appl Meas Educ* **15**(3): 309-33.

Haladyna, T.M., Downing, S.M. 1989. Validity of a taxonomy of multiple-choice item-writing rules. *Appl Meas Educ* **2**(1): 51-78.

Hassan, S., Ibrahim, M.S., Hassan, N.G. 2018. The structural framework, implementation strategies and students' perception of team- based learning in undergraduate medical education of a medical school in Malaysia. *EIMJ* **10**(1): 53-66.

Hift, R.J. 2014. Should essays and other "open-ended"-type questions retain a place in written summative assessment in clinical medicine? *BMC Med Educ* **14**: 1-18. https://doi.org/10.1186/1472-6920-14-40

Hingorjo, M.R., Jaleel, F. 2012. Analysis of one-best MCQs: The difficulty index, discrimination index and distractor efficiency. *J Pak Med Assoc* **62**(2): 142.

Husain, S.F., Wang, N., McIntyre, R.S., Tran, B.X., Nguyen, T.P., Vu, L.G., Vu, G.T., Ho, R.C., Ho, C.S. 2023. Functional near-infrared spectroscopy of medical students answering various item types. *Front Psychol* **14**: 1178753. https://doi.org/10.3389/fpsyg.2023.1178753

Kheyami, D., Jaradat, A., Al-Shibani, T., Ali, F.A. 2018. Item analysis of multiple choice questions at the department of paediatrics, Arabian Gulf University, Manama, Bahrain. *Sultan Qaboos Univ Med J* **18**(1): 68.

Kim, H.Y. 2013. Statistical notes for clinical researchers: Assessing Normal Distribution using Skewness and Kurtosis. *Restor Dent Endod* **38**(1): 52-4.

Linn, R.L., 2000. Measurement and Assessment in Teaching. 10th Edition, Pearson Education Ltd., Upper Saddle River.

Mahjabeen, W., Alam, S., Hassan, U., Zafar, T., Butt, R., Konain, S., Rizvi, M. 2017. Difficulty index, discrimination index and distractor efficiency in multiple choice questions. *Annals of Pakistan Institute of Medical Sciences-Shaheed Zulfiqar Ali Bhutto Medical University* **13**(4): 310-5.

Pham, H., Besanko, J., Devitt, P. 2018. We are examining the impact of specific types of item-writing flaws on student performance and the psychometric properties of the multiple-choice question. *MedEdPublish* **7**: 225.

Puthiaparampil, T., Rahman, M.M. 2020. Very short answer questions: A viable alternative to multiple choice questions. *BMC Med Educ* **20**(1): 141. https://doi.org/10.1186/s12909-020-02051-8

Rezigalla., A.A., Eleragi. AMESA., Elhussein, A.B., Alfaifi. J., AlGhamdi. M.A., Al, Ameer. A.Y., Yahia. A.I.O., Mohammed. O.A., Adam, M.I.E. 2024. Item analysis: The impact of distractor efficiency on the difficulty index and discrimination power of multiple-choice items. *BMC Med Educ* **24**(1): 445. https://doi.org/10.1186/s12909-024-03445-y

Sajjad, M., Iltaf, S., Khan, R.A. 2020. Nonfunctional distractor analysis: An indicator for the quality of Multiple-choice questions. *Pak J Med Sci* **36**(5): 982.

Sam, A.H., Wilson, R., Lupton, M., Melville, C., Halse, O., Harris, J., Meeran, K. 2019. Clinical prioritisation questions: A novel assessment tool to encourage tolerance of uncertainty? *Med Teach* **42**(4): 416-21.

Schober, P., Boer, C., Schwarte, L.A. 2018. Correlation coefficients: Appropriate use and interpretation. *Anesth Analg* **126**(5): 1763-8. https://doi.org/10.1213/ANE.0000000000002864

Shakurnia, A., Ghafourian, M., Khodadadi, A., Ghadiri, A., Amari, A., Shariffat, M. 2023. Evaluating functional and non-functional distractors and their relationship with difficulty and discrimination indices in four-option multiple-choice questions. *EIMJ* **14**(4): 55-62.

Tarrant, M., Ware, J., Mohammed, A.M. 2009. An assessment of functioning and non-functioning distracters in multiple-choice questions: A descriptive analysis. *BMC Med Educ* **9**: 40. https://doi.org/10.1186/1472-6920-9-40

Teli, C., Kate, N. 2022. Item analysis of multiple-choice questions in anatomy for first year MBBS. *Natl J Physiol Pharm Pharmacol* **12**(10):

1529-32.

Wahab, I.A., Shamsuddin, N., Alwi, S., Wahab, M.S.A., Ali, M., Long, C.M., Ahmad Hisham, S., Jamil, N. 2022. Reliability of online simulation-based assessment to measure cognitive performance and its acceptance among pharmacy students. *EIMJ* **14**(4): 43-53.