Item Analysis of type "A" Multiple Choice Questions from a Multidisciplinary Units assessment in a Problem Based Curriculum

Tarik Al Shaibani, PhD* Fuad Ali, MBBS, DCH** A.Halim Deifalla, MBBCh, D Ped, MSc, MD*** Ahmed Jaradat, PhD****

ABSTRACT

Multiple choice questions (MCQs) — also known as an item — is a common tool used in assessment.

Objectives: to determine the post examination validity and reliability by item analysis of a multidisciplinary units in a problem-based learning curriculum. This includes the difficulty index (DIF-I), the discrimination index (DI) and the Distractor effectiveness (DE).

Design: Cross sectional, retrospective Study.

Setting: College of Medicine and Medical Sciences, Arabian Gulf University

Method: Item analysis of 700 items and 2800 distractors were analyzed

Result: The mean DIF-I and DI were acceptable while the mean DE was variable. The reliability was in the acceptable ranges between 0.8 and 0.9. 53.68% of the distractors were non-functional distractors (NFDs). 49.81% were easy, 44.35% was acceptable and only 5.82% were difficult. The acceptable and excellent DI was almost equal; 39.79% and 42.07% respectively, whereas the poor DI was 18.13%. 37% of the items were considered ideal with acceptable difficulty and discrimination. The DI was maximum with DIF-I in the acceptable range. DE was indirectly related to the DIF-I. However, there was no relation between DE and DI.

Conclusion: The mean DIF-I, DI, and DE were in acceptable ranges. A high percentage of items was easy, and a high percentage of distractors was NFS. These distractors need to be revised to improve the DIF-I, DI and DE parameters. The reliability of the exams was acceptable. We recommend doing item analysis after each examination to identify areas of potential weakness in each item and to improve the standard of students' assessment.

INTRODUCTION

Assessment is a core component of teaching activities. It provides the teacher with information about what students have learned, and if the learning objectives were achieved. Assessment can be either formative or summative. Formative assessments are performed during the learning activity, it advises the instructor about the students' progress toward the learning objective. In contrast, summative assessment evaluates student's learning after a course and evaluate student competencies regarding the learning objectives[1].

Multiple choice question (MCQ) — also known as an item — is a common tool used in assessment. The commonest type is “type A MCQ." It consists of a stem followed by two or more options. One option is right (key), and the other options are wrong (distractors). The distractors should be plausible; their function is to attract students who do not know the correct answer while students who know the correct answer ignore them.

MCQs allows teachers to assess many students and a wide range of content in a short time[2]. They are easy to score and analyze even by the machines. If properly constructed, they are used to assessing higher-order cognitive processing of Bloom's taxonomy such as interpretation, synthesis, and application of knowledge, instead of just testing recall of isolated facts[3]. They can discriminate between high and low achiever students[4].

However, a properly constructed MCQ is time-consuming and not easy to write, especially in a multidisciplinary curriculum. High-quality MCQ writing requires training and avoiding technical flaws such as unclear stem or implausible options[4].

The pre-examination validity of each MCQ (item) is attained by experts in the subjects who have skills in writing proper MCQs. Post examination validity and reliability are performed by item analysis. The three common item analysis measurements include difficulty index (DIF-I), discrimination index (DI) and Distractor effectiveness (DE). The data generated by these measures are used to assess the quality of each item and accordingly then either store, review or discard the items[5].

DIF-I is defined as the proportion of students who selected the correct answers. It ranges from zero to 100. Zero, means that none of the students selected the correct option while 100, means that the correct option was selected by all students. The acceptable range is from 30 - 70%, less than 30% is considered difficult while more than 70% is considered an easy item.

DI, also called point biserial correlation (PBS), is defined as how effective the question to discriminate between the low and high achiever groups of students. The low achievers are 27% of students who score the lowest marks, while the high achievers are 27% of students who

* Associate Professor
Department of Physiology,
Head of Community Service and Continuous Education Center,
Arabian Gulf University, Bahrain, E-mail: tareqas@agu.edu.bh
** Head of Assessment Unit, Department of Pediatrics
*** Professor, Dean of College of Medicine and Medical Sciences
**** Associate Professor, Department of Community and Family Medicine

score the highest marks. The DI value ranges from -1.0 to +1.0. +1.0, means that the question is answered by all the high achievers but none of the low achievers, while -1.0, means that the question is answered by all the low achievers but none of the high achievers. Usually, a value between 0.2 – 0.34 is considered as acceptable, less than 0.2 as poor, and more than 0.34 as excellent DI.

If the wrong option is selected by less than 5% of students, it is called non-functional distractors (NFD). While the functional distractors (FD) are the wrong option selected by 5% of students or more. Distractor effectiveness (DE) is determined based on the number of NFDs. For an item with a key answer and four distractors, DE ranges from zero to 100%, zero means all the four distractors are NFDs while 100% means that all the four distractors are functional or plausible. DE of 25%, 50% or 75% means the number of NFDs are three, two or one, respectively. Item reliability means that an item measures the same thing.

The reliability of the examination is measured using the Kuder Richardson Formula 20 (KR20). It ranges from zero to one, zero means not reliable while 1 means excellent reliability. A value of 0.7 or higher is usually considered as an acceptable reliability[6].

The curriculum of the College of Medicine and Medical Sciences at the Arabian Gulf University — Bahrain is Problem-Based Learning (PBL) divided into three phases. Phase I is Basic Science Phase (year 1), phase II is Medical Sciences Phase (year 2, 3 and 4) and phase III is Clinical Clerkship Phase (year 5 and 6). In phase II, the curriculum is structured around nine integrated organ/system units. 93 health problems are covered during the three years (http://www.agu.edu.bh/).

Students at the end of each phase II units are assessed by three types of examinations, MCQs, short answer examination (SAQ) and objective structured practical examination (OSPE). Pre-validation of the questions is verified by each unit's committee which consists of six academic members from the following departments: anatomy, physiology, biochemistry, pharmacology, microbiology, community medicine, and clinical departments. The questions are distributed according to a predetermined blueprint to ensure that all disciplines were included based on the learning objectives.

This is the first item analysis study conducted on a problem-based curriculum assessing multidisciplinary units. The aim is to determine if our assessment is consistently valid and reliable throughout the nine multidisciplinary units of the phase II years in the Arabian Gulf University. To analyze and compare the quality of the (type A) multiple-choice questions and to find out the relationship between the item difficulties, item discrimination Indices, and item distractor efficacies.

METHOD

Item analysis of 700 items or MCQs were analyzed for the end rotation examination of nine multidisciplinary units of phase II at AGU. The study involved the same group of students in the three academic years, 2016-2017 (unit I, II and III), 2017-2018 (unit IV, V and VI) and 2018-2019 (unit VII, VIII, and IX). The item analysis was performed using the Oracle Database, Version 10g (Oracle Corp., Redwood City, California, USA).

All examination questions were kept in the assessment unit to which only authorized individuals were allowed access via a digital security system. Each unit's examination consisted of either 75, 80 or 85 items and accordingly, the time allotted for each examination was two hours or two and a half hours. Each item consisted of a stem and five options, one correct option (key answer) and four incorrect answers (distractors) as shown in figure 1.

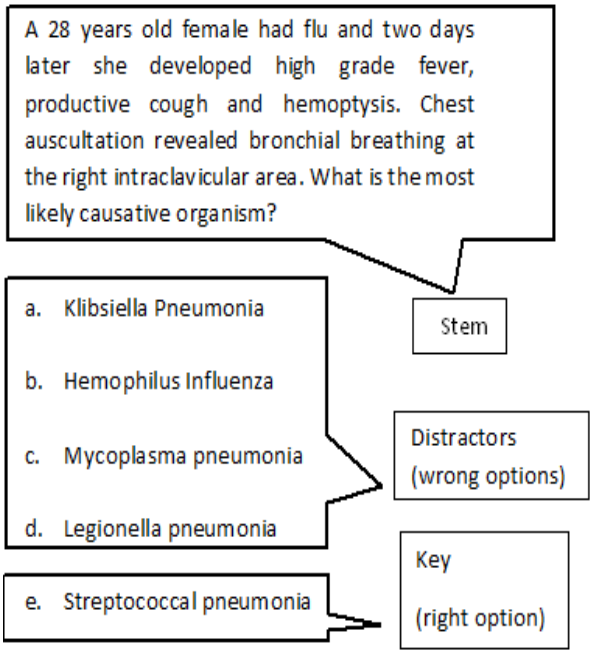

Figure 1: Component of type "A" multiple choice question

Students answered on a Scantron® optical answer sheet (Scantron Corp., Tustin, California, USA) and the results were counted by optical machine reader. There was no negative marking, and the examinations were criterion-referenced, and the passing mark was 60%. Some questions were reviewed with the students at the end of each unit.

Difficulty index (DIF-I) value of less than 30% is considered a difficult question, more than 70% as an easy question and between 30 - 37% as an acceptable question. Discrimination index (DI) value of less than 0.2, more than 0.34 and between 0.2-0.34 are considered as poor, excellent, and acceptable discriminating questions, respectively. The DE will be zero % if all the distractors were NFDs and DE will be 100% of all the distractors were functional. The DE will be 25, 50 or 75% if the number of NFD is three, two, or one, respectively. A value of 0.7 or more is considered as acceptable reliability.

Data was entered and analyzed using SPSS version 25.0. Categorical variables are expressed as frequencies and percentages. The quantitative variables are expressed as Means ± SD. ANOVA was used to test the statistically significant difference in a means of the DE for difficult, acceptable, and easy items. P-value < 0.05 was considered as statistically significant.

RESULTS

A total of 700 MCQs (items) and 2800 distractors for nine units in phase II in the Arabian Gulf University were analyzed. The number of weeks in each unit varies from 6 to 13 weeks. The number of students in unit-I was 178 and gradually reduced to 167 in unit-IX. The mean scores range from 62.89 ± 15.93 (out of 100) in unit III to 76.81 ± 8.38 in unit IV. The mean DIF-I of all the nine units was in the acceptable ranges, and the mean DI were in the acceptable and excellent ranges. The overall mean ± standard deviation of DIF-I and DI were 66.44 ± 19.87 and 0.31 ± 0.12, respectively. The mean ± standard deviation of the DE was variable and ranges from 43.00 ± 29.22 to 71.00 ± 26.33. The reliability of all units was in the acceptable ranges of more than 0.7. Table 1 shows the body system type, the number of weeks of each unit and whether the students were in years two, three, or four did not influence the mean DIF-I, mean DI, or mean DE.

Table 1: Systems, mean score, mean difficulty index, discrimination index, distractor efficiency and reliability for the three academic years, 2016-2017, 2017-2018 & 2018-2019 of phase II units' examinations in the Arabian Gulf University, Bahrain (N = 9)

		Score	DIF-I	DI	DE	Reliability
Unit	Organ / System	(Mean +/- SD)	(Mean +/- SD)	(Mean +/- SD)	(Mean +/- SD)	
UNIT-I	Man & his Environment	65.69 ± 15.17	65.63 ± 18.43	0.34 ± 0.12	60.31 ± 30.73	0.90
UNIT-II	Life Cycle	74.23 ± 9.78	63.82 ± 25.82	0.27 ± 0.16	50.33 ± 33.26	0.84
UNIT-III	Respiratory & Cardiovascular System	62.89 ± 15.93	62.89 ± 17.03	0.36 ± 0.12	71.00 ± 26.33	0.90
UNIT-IV	Endocrine & Reproductive System	76.81 ± 8.38	71.73 ± 18.84	0.29 ± 0.10	43.00 ± 29.22	0.85
UNIT-V	Gastroenterology & Renal System	70.72 ± 9.78	61.37 ± 19.36	0.29 ± 0.11	59.00 ± 29.82	0.85
UNIT-VI	Hematopoietic & Immune System	76.50 ± 10.26	66.38 ± 19.75	0.32 ± 0.11	54.00 ± 28.78	0.88
UNIT-VII	Musculoskeletal & Integumentary System	72.79 ± 12.29	72.35 ± 17.53	0.29 ± 0.11	46.18 ± 32.16	0.88
UNIT-VIII	Nervous System & Human Behaviors	68.07 ± 14.19	68.06 ± 18.96	0.33 ± 0.15	50.00 ± 30.62	0.90
UNIT-IX	Multi-System Integration	64.18 ± 10.79	64.26 ± 23.13	0.26 ± 0.10	50.33 ± 31.43	0.83
Overall		70.20 ± 11.84	66.44 ± 19.87	0.31 ± 0.12	53.61 ± 31.19	0.87

SD = standard deviation; DIFI = difficulty index; DI = discrimination index; DE = distractor efficiency.

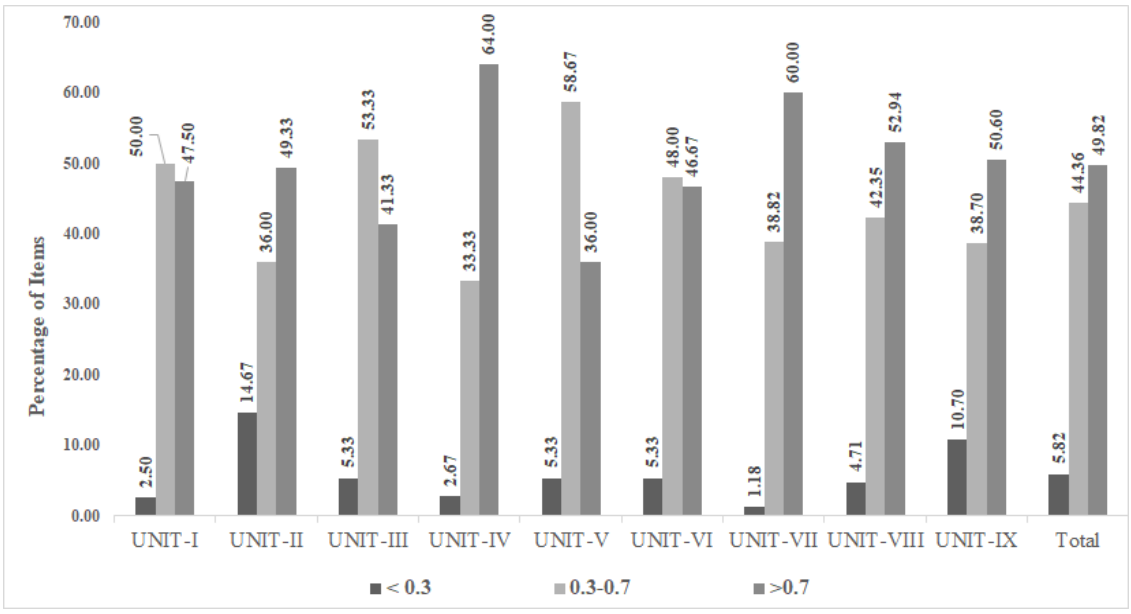

Figure 2: Distribution of the difficulty index of multiple-choice questions in phase II units' examinations at the Arabian Gulf University, Bahrain (N = 700)

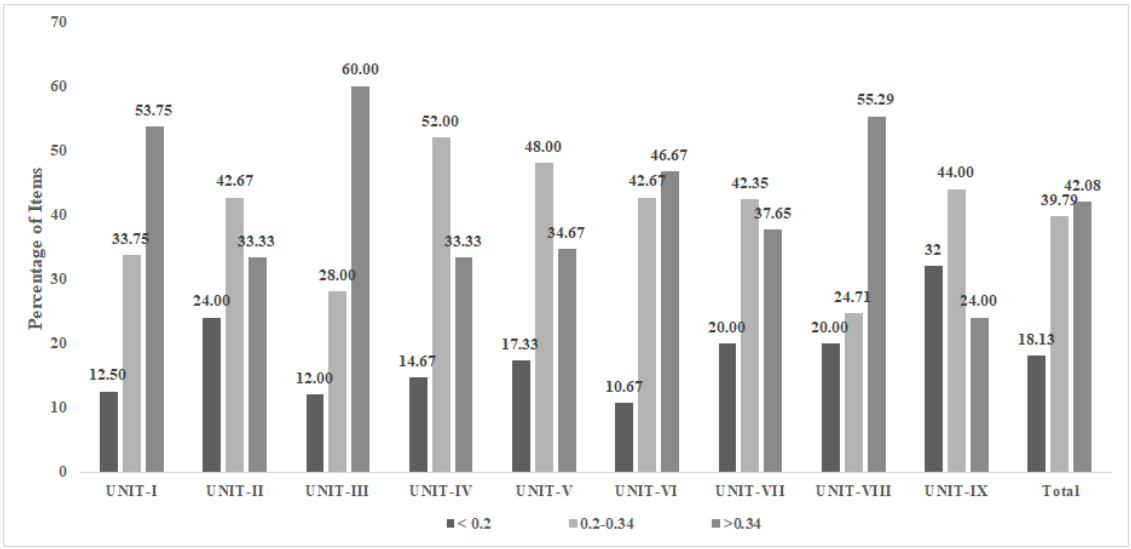

Figure 3: Distribution of the discrimination index of multiple-choice questions in phase II units' examinations at the Arabian Gulf University, Bahrain (N = 700)

Table 2: Classification of items according to difficulty and discrimination indices at the Arabian Gulf University, Bahrain (N = 700)

DIF-I	Total (%)	Poor DI (%)	Recomm-endation	Good DI (%)	Recomm-endation	Excellent DI	Recomm-endation
Difficult	40 (5.71)	25 (3.57)	Discard	11 (1.57)	Review	4 (0.57)	Review
Acceptable	310 (44.29)	51 (7.29)	Review	126 (18)	Store	133 (19)	Store
Easy	350 (50%)	51 (7.29)	Discard	140 (20)	Review	159 (22.7)	Review

DIFI = difficulty index; DI = discrimination index.

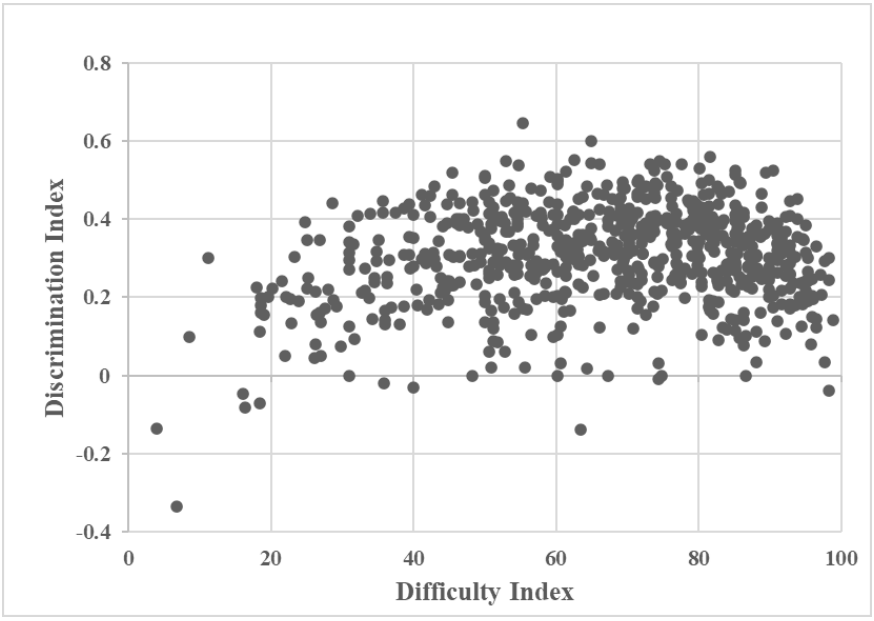

Figure 4: Scatter plot showing the relationship between difficulty index and discrimination index among multiple choice question items in the pre-clinical units' examinations at the Arabian Gulf University, Manama, Bahrain (N = 700)

Among the 2800 distractors, just over one-half (1503) (53.68%) were NFDs. The remaining 1297 (46.32) was functional distractors. Out of 700 items, the number of items with zero, 1, 2, 3, and 4 NFDs were 117 (16.71%), 175 (25.00%), 180 (25.71%), 150 (21.43%) and 78 (11.14%), respectively. The items were easier when the number of NFDs increased. However, the mean DI was not influenced by the changes in the number of NFDs. Almost half of the items (49.81%) were easy, while 44.35% was acceptable and only 5.82% was difficult. The highest percentage of easy items were in unit IV (64%) and the lowest was in unit III (41.33%) as shown in figure 2. The highest percentage of items with excellent DI were in unit III (60%) and the highest percentage of items with poor DI were in unit II (24 %) as shown in figure 3.

Analysis of DIF-I along with DI revealed that out of the 310 acceptable DIF-I items (44.29%), 126 (18%) and 133 (19%) were of acceptable and excellent DI, respectively. These items are considered ideal and should be stored. Whereas difficult and easy items with poor DI were 3.5% and 7.29%, respectively. These items should be discarded. Difficult and Easy items with acceptable or excellent DI as well as items with acceptable difficulty but poor DI should be reviewed for improvement, see table 2.

The Pearson correlation between DIF-I and DI was weak but statistically significant ($r = 0.171$, $P < 0.005$). The DI was maximum with DIF-I in the acceptable range. DI decreased as the item's DIF-I moved towards either easy or difficult ranges, see figure 4.

The distribution of difficulty, and discrimination indices and their corresponding DE was studied. DE was indirectly related to the DIF-I. Difficult items were having a high DE percentage while easy items were having low DE percentage. The DE was 81.25%, 73.55%, and 32.92% for difficult, acceptable, and easy items respectively ($P < 0.0005$). However, there was no relation between DE and DI. The DE percentage was almost the same for items with excellent, acceptable, and poor DI ($P = 0.204$), see table 3.

Table 3: Distribution of difficulty and discrimination indices and their corresponding distractor efficiency and recommendation of multiple-choice questions in phase II units' examinations at the Arabian Gulf University, Bahrain (N = 700).

Index	n (%)	Mean	DE	*P* value	Recommendation
DIF-I					
Difficult	40 (5.82)	21.86	81.25		Review
Acceptable	310 (44.35)	53.33	73.55	< 0.0005*	Review
Easy	350 (49.81)	83.04	32.92		Discard
DI					
Poor	127 (18.13)	0.11	55.12		Discard
Acceptable	277 (39.79)	0.27	51.08	0.204**	Review
Excellent	296 (42.07)	0.42	55.49		Store

DE = distractor efficiency; DIF-I = difficulty index; DI = discrimination index

** There is a statistically significant difference in means of the DE for difficult, acceptable and easy items (P-value < 0.0005). Post hoc analysis reveals that they is a pairwise significant differences in the means of DE (Easy vs. Acceptable, P-Value=0.049; Easy vs. Difficult, P-Value=0.049; Difficult vs. Acceptable, P-Value < 0.0005).*

*** There is no statistically deference in the means of DE for poor, acceptable and excellent DI items.*

DISCUSSION

Item analysis is an important post validation tool which provides information regarding the validity and reliability of the examination. It improves the quality of MCQs and so the learning process and students' performance. Appropriate assessment is important to improve the cognitive clinical abilities of students, and ultimately their competencies in patients' care[7].

MCQs should be designed according to the learning objectives and consistent with the higher-order skills of Bloom's taxonomy. They should assess students' ability to create, evaluate, analyze and apply their knowledge rather than just understanding and recalling certain facts by the students[3].

The number of students was 178 in the year two and ended with 167 in the year four, only 11 students dropped out during the three years of the phase II, giving a retention rate of 93.82%. This high rate is probably due to strict admission criteria and the filtration process during phase I (year 1). We used the same group of students during the three years to limit the selection bias by analyzing the items of the nine units taken by the same group of students.

The mean scores range from 62.89 ± 15.93 to 76.81 ± 8.38 (out of 100). The passing mark was 60%. High mean score indicates either that the teaching was effective, students were highly motivated, or the exam was easy. Vice versa, low mean score means either the exam was too difficult, teaching was not effective, students were not motivated, or objectives were unrealistic[8]. There was no relation between the mean score of each unit and the duration of the unit, specific unit body system, or whether the students were in year two, three or four.

The DIF-I determines whether the students understand the learned objective being tested. The DI measures how high achiever students are doing versus low achiever students on a particular MCQ. DE is a useful tool in evaluating the effectiveness of each distractor (wrong option).

The mean difficulty and discrimination indices of all units were all in the acceptable ranges although each unit consists of different body systems and has multi-disciplinary subjects. The overall mean ± standard deviation of DIF-I and DI were 66.44 ± 19.87 and 0.31 ± 0.12, respectively. The mean ± standard deviation DE was variable among all the units, being highest in unit III (71.00 ± 26.32) where the DI was also the highest (0.36 ± 0.12).

For item analysis in a single subject, Gajjar et al. reported a mean DIF-I, DI, and DE of 39.4 ± 21.4, 0.14 ± 0.19 and 88.6 ± 18.6%, respectively in community medicine[9]. Mukherjee and Laheri et al. reported means DIF-I, DI and DE of 61.92 ± 25.1%, 0.31 ± 0.27, and 47.78 ± 32.38%, respectively in 30 items in Community Medicine[10]. Zia- ul-Islam and Usmani reported means DIF-I and DI of 51.63 and 0.28, respectively in anatomy[11]. Keralia et al, reported mean DIF-I and DI between 47.17-58.08% and 0.29-0.38 in 200 test items of 10 summative papers in pharmacology[12].

Studies in multidisciplinary examinations reported mean DIF-I and DI of 0.59 and 0.28 from 3,138 items taken from 66 multidisciplinary examinations in the Faculty of Medicine in Tunis[8]. Another study conducted on 12 summative tests for Foundation 1 multidisciplinary course at international Medical University - Malaysia, reported a mean DIF-I ranging from 64% to 89% and a mean DI of 0.2 or higher[13].
An ideal item is the one that has an acceptable DIF-I with acceptable or an excellent DI. Our study revealed that 310 out of 700 items (44.28%) were of acceptable DIF-I, out of which 51 and 126 items were of acceptable and excellent DI respectively (37%). Similar studies reported ideal items' percentage of 30%9, 27.02%8, 64%14, and 46.67%10. These items need to be stored in the bank.

Items with low or high DIF-I should be revised if their DIs were acceptable or excellent. In our study, 11 items (1.57%) were difficult items (low DIF-I), and 140 items (20%) were easy items (high DIF-I). Both were of acceptable DI. These items need to be reviewed to improve their item analysis before using them in the next exams. Whereas 25 (3.57%) were difficult items and 51 items (7.29%) were easy items. Both were of poor DI. These items should be discarded from the question bank.

NFDs are wrong options that were selected by less than 5% of students. NFDs deny the chance to test the student's ability to learn. They need to be modified into more plausible distractors. More than half of all 2800 distractors were NFDs (53.68%). Previous studies reported 11.4% NFD out of 150 distractors[9] and 23.5% out of 200 distractors[14]. Writing plausible distractors and decreasing the number of NFDs will improve the quality of MCQs. The distractors need to be modified if students constantly avoid choosing them.

Out of the 700 items, 150 (21.43%) and 78 (11.14%) were having 3NFD (25%DE) and 4 NFD (0% DE). Items with more NFDs were easier. However, the mean DI was not influenced by the changes in the number of NFDs. A similar study showed that Items with 3 NFDs had a high DIF-I (77.5%) and poor DI (0.160); whereas items with 2, 1 and zero NFD had acceptable DIF-I of 62.66, 54.94 and 44.38, and excellent DI of 0.365, 0.427 and 0.351, respectively[14]. Constructing good quality MCQs have a better assessment of student performance. For high achiever students, NFDs add little to the performance of a test item; in contrast, increasing the number of distractors decreases the likelihood of students accidentally choosing the correct answer by guesswork.

DE was indirectly related to the DIF-I. Difficult items were having higher DE than the acceptable and easy items. The DE was 81.25%, 73.55% and 32.92% for difficult, acceptable, and easy items, respectively ($P<0.0005$). However, there was no relation between DE and DI; the DE was almost the same for items with excellent, acceptable, and poor DI. Similar findings were reported by Gajjar et al where the DE was higher (91.7%) in difficult items than easy items (79.3%). However, DE showed little variation among items with different DI. Mean DE was 88.9% in items with excellent DI compared to 88.4% in items with poor DI; difference in DE in both cases was statistically not significant[9].

On the contrary, Mukherjee and Laheri et al, reported a significant relation between DE with DIF-I and DI. The mean DE was 83.34% in very difficult items, 48% in average difficulty items and as low as 22.22% in very easy items[10]. The same study reported excellent DI of 0.396 and 0.404 for items having only one and two NFDs, respectively, as compared to DI of 0.023 for items with zero NFD. A similar finding was reported by Higorio with difficult, average difficulty and easy items having DE of 100%, 81.41%, and 43.75%, respectively. Items with excellent, acceptable, and poor discrimination had DE of 83.06%, 71.42%, and 58.33%, respectively[14].

About half the items were easy (49.81%), the highest percentage was in unit IV (64%) and the lowest was in unit V (36%). This is considered a high percentage and can lead to inflate scores. Easy items should be revised for the possibility of a large number of NFDs. It is advisable to place some of these easy items at the start of the test “warm-up” to boost the confidence of the students, and they should cover the core topics that students must know. Only 5.82% of the items were difficult. Unit II has 14.67%, and unit VII has only 1.18% of difficult items. These items should be placed at the end of the test and can be used to

identify the top scorers of students[6]. At the same time, these difficult items should be reviewed for possible ambiguity, controversy topics, or even incorrect key answers8. Kaur et al. reported 76% of 50 items in pharmacology examination were of acceptable difficultly, 22% easy and only 2% difficult items[15].

Almost an equal percentage of the acceptable DI (39.79%) and excellent DI (42.07%) were found among all the units. 127 out of 700 items (18.13%) were of poor DI; ten of them were of negative DI. Kaur et al., reported 62% excellent, 24% acceptable and 14% poor DI[15]. Poor DI means that low achievers' group answered the item — most likely by guessing — but the high achiever group failed to select and right option. This can be due to either wrong answer key, very difficult question or the stem of the item was ambiguous. When negative marking is not applied, students will usually guess the answer even if they do not know the right option. In a multidisciplinary exam, student performance in certain subjects such as anatomy does not reflect their performance in another subject such as physiology. Negative DI item is useless and decreases the validity of the test; they should be revised or even discarded[9].

The Pearson correlation between mean DIF-I and DI was of dome-shaped. Easy and difficult items discriminate poorly; the DI was maximum at an acceptable difficulty range. DI is low in very difficult items because these items cannot be answered even by high achievers' students. Similar findings reported by several other reports[8,9,10,13,14.]

Reliability is the degree to which an assessment tool produces stable and consistent results. In psychometrics, the Kuder–Richardson Formula 20 (KR-20), first published in 1937, is a measure of internal consistency reliability for measures with dichotomous choices[16]. The scores for KR-20 range from zero to one, where zero, means the test is not reliable and one means it is very reliable. In general, a score of above 0.7 is usually considered acceptable[6]. The mean reliability of all units was more than 0.7, indicating good examination reliability, some units' reliability was 0.9 indicating excellent internal consistency.

Reliability can be influenced by the length of the exam, the time provided for the exam, difficulty of the exam, the student condition on the day of the exam, the environment of the exam and finally the standards used to score the exam[17].

CONCLUSION

Item analysis is an important tool to discriminate the high, average, and low achiever students. It enhances the teaching process and improves the quality of MCQs bank. It is concluded from this study that the mean DIF- I, DI and DE were in the acceptable ranges. A high percentage of items was easy, and a high percentage of distractors were NFDs. These distractors need to be revised to improve the DIF-I, DI and DE parameters. The reliability of the exams was acceptable. We recommend doing these analyses after each examination to identify the areas of potential weakness in each item to improve the standard of students' assessment.

Authorship Contribution: All authors share equal effort contribution towards (1) substantial contributions to conception and design, acquisition, analysis and interpretation of data; (2) drafting the article and revising it critically for important intellectual content; and (3) final approval of the manuscript version to be published. Yes.

Potential Conflict of Interest: None.

Competing Interest: None.

Sponsorship: None.

Acceptance Date: 12 March 2021

Informed Consent: The work that we have submitted to the Bahrain Medical Bulletin is our own original work and has not previously been published by any other party in any other format. We have the authority to enter into agreement with the editor and publishers of Bahrain Medical Bulletin to the exclusive right to publish our work first.

REFERENCES

1. Black P, Wiliam D. Assessment and Classroom Learning. Assessment in Education 1985; 5(1):7-74.
2. McCoubrie P. Improving the fairness of multiple-choice questions: a literature review Med Teach 2004; 26(8):709-12.
3. Case SM, Swanson DB. Constructing written test questions for the basic and clinical sciences. National Board of Medical Examiners 3rd ed. 2010.
4. Tarrant M, Ware J. Impact of item writing flaws in multiple-choice questions on student achievement in high stakes nursing assessments. Med Educ 2008; 42:198-206.
5. Tarrant M, Ware J, Mohammed AM. An assessment of functioning and non-functioning distractors in multiple-choice questions: A descriptive analysis. BMC Med Educ 2009; 9:1-8.
6. El-Uri FI, Malas N. Analysis of use of a single best answer format in an undergraduate medical examination. Qatar Med J 2013; 1:3-6.
7. Considine J, Botti M, Thomas S. Design, format, validity and reliability of multiple-choice questions for use in nursing research and education. Collegian 2005; 12(1):19-24.
8. Hermi A, Achour W. Item analysis of examinations in the Faculty of Medicine of Tunis. Tunis Med 2016; 94(4):247-52.
9. Gajjar S, Sharma R, Kumar P, et al. Item and test analysis to identify quality multiple-choice questions (MCQs) from an assessment of medical students of Ahmedabad, Gujarat. Indian J Community Med 2014; 39(1):17-20.
10. Mukherjee P, Lahiri SK. Analysis of multiple-choice questions (MCQs): Item and test statistics from an assessment in a medical college of Kolkata, West Bengal. IOSR J Dent Med Sci 2015; 14(12):47-52.
11. Zia-ul-Islam, Usmani A. Psychometric analysis of Anatomy MCQs in Modular examination. Pak J Med Sci. 2017; 33(5):1138-43.
12. Karelia BN, Pillai A, Vegada BN. The levels of difficulty, and discrimination indices and the relationship between them in four-response type multiple choice questions of pharmacology summative tests of year II M.B.B.S students. IJSME 2013; 7(2):41-6.
13. Mitra NK, Nagaraja HS, Ponnudurai G, et al. The levels of difficulty, and discrimination indices in type A multiple-choice questions of pre-clinical semester 1 multidisciplinary summative tests. Int EJ Sci Med Educ 2009; 3(1):2-7.
14. Hingorjo MR, Jaleel F. Analysis of one-best MCQs: The difficulty index, discrimination index, and distractor efficiency. J Pak Med Assoc 2012; 62(2):142-7.
15. Kaur M, Singla S, Mahajan R. Item analysis of in-use multiple choice questions in pharmacology. Int J Appl Basic Med Res 2016; 6(3):170-3.
16. Kuder GF, Richardson MW. The theory of the estimation of test reliability. Psychometrika 1937; 2(3):151-60.
17. Meshkani Z, Hossein Abadie F. Multivariate analysis of factors influencing reliability of teacher made tests. J Med Educ 2005; 6(2):149-52.