Evaluation of the Item Analysis of Multiple-Choice Pediatric Exams: A College of Medicine Departmental Review

Alam Eldin M. Mustafa, MD*

ABSTRACT

Objectives: To evaluate the detailed indices of the analysis of the items of the written multiple-choice questions (MCQs) exams in the Department of Child Health, Faculty of Medicine over the last four academic years (1439-1443 AH) and to construct the outlines of a plan for improving the upcoming written MCQs exams.

Methods: This was retrospective cross-sectional study on the item analysis of the exams in the MBBS course of pediatrics-2 for both boys and girls group in the Child Health Department, Faculty of Medicine, King Khalid University, Saudi Arabia for the midterm MCQs exams in the years of 1439,1440,1441,1442 and the first semester of girl group in 1443. The total number of items studied in these 16 exams were 643 items. The data was obtained constitute the difficulty, discrimination, point biserial reliability and distractor analysis of each of the exams items. The data was tabulated and the statistical significance determined for some variables in the analysis.

Results: Total number of students enrolled in the study were 1002.The total number of items studied were 643 items. Regarding students' scores were as follows: A scored by 73 students (7.3%), B by 219 (21.8%), C by 331 (33%), D by 214 (21.4%) and F by 165 (16.5%). Difficulty index: considering a difficulty index of 80% as easy item of 30% or less as difficult item and that between 30% and 80% of moderate difficult; we obtain 3 categories of items: difficult items were 43 (6.6%)of the total items, moderate difficulty items were 343 (53.4%) and easy items 257 (40%). There was significant statistical correlation ($p \leq 0.05$) when these difficulty levels compared over the exam years.

Conclusions: Departmental exam committee needs to work comprehensively to improve the difficulty of the exams towards moderate intermediate class also the quality of the questions need extensive work on refining the distractors and revision of the correctness and the suitability of a considerable number of items.

Keywords: **Item analysis, multiple choice questions, Difficulty index, Discrimination index, Distractors, Non-functional distractor, Reliability.**

INTRODUCTION

Item analysis is the process involving evaluation of the individual student responses to the test items (questions) so as to assess and determine the quality of those questions and of the whole test. Item analysis is particularly valuable in improving items that will be needed for use again in later examinations, but also can be used to eliminate ambiguous or misleading items in a single test administration[1]. In the item analysis; statistical techniques are used for selecting and rejecting the items of the test on the basis of their difficulty value and discrimination power. Item analysis helps educators to get active feedback from students' performance and identify areas which require more emphasis, reinforcement or an alteration in teaching methodology perhaps using other learning facilities[2]. Item analysis is also important for increasing instructors' skills in test construction, and identifying specific areas of course content which need greater emphasis or clarity. Item analysis enables identifying good MCQs based on difficulty index (DIF I) or facility value (FV) or P-value, discrimination index (DI), and distractor effectiveness (DE). High quality MCQs, require well-written alternative answers[3]. Point Bi-serial correlation (PBS) is yet another important parameter which gives information about how consistent an item with the whole test is. PBS helps in indicating items which are not testing the same domain and construct as the remaining part of the test and hence aids in improving the validity and reliability of the test. Assessing the quality of items used in a test can assess test. However, the reliability coefficient and standard error of measurement help to evaluate the performance of the whole test. Reliability tells us whether a test is likely to yield the same results if administered to the same group of test-takers multiple times. The most frequently reported internal consistency reliability estimates are the K-R20 and Cronbach's alpha. The indices test the quality of items[4,5]. Multiple choice questions is preferred over other tools of assessment because of its objectivity in assessment, comparability in different settings, wide coverage of subject, and minimization of assessor's bias. Designing good MCQs is a complex, challenging, and time-consuming process. Having constructed and assessed, MCQs need to be tested for the standard or quality. Item analysis examines the student responses to individual test items (MCQs) to assess the quality of those items and test[6]. Quality educational assessment is a purposed process with student feedback assessment analysis assist in this objective[7]. Item analysis is the process of collecting, summarizing, and using information from student's responses to assess the quality of multiple-choice questions (MCQs)[8]. The best assessment method must meet five criteria which include reliability, validity, acceptability, feasibility and educational impacts on learning and practice[9].

Quantitative item analysis is a statistical technique meant to know about the test items and item concerned on the basis of three numerical indicators:

* Department of Child Health
King Khalid University, Abha, Saudi Arabia.
E-mail: alameldinmustafa641@gmail.com

Difficulty index or factor (DIF I): if test item was easy or difficult for the specific grot-JP of students. Acceptable item difficulty is how many exam takers answered the item correct. There is no a set number; If the intent is a mastery item, a difficulty level between 0.80 and 1.00 is acceptable. If the intent is a discriminating question, a range of 0.30 to 0.70 is generally acceptable items with DI below 0.3 generally need revision[10].

Discrimination index (DI): How well item discriminated between high and low scorers in the test.This value is based on the top 27% scorers (HA) and bottom 27% scorers (LA) of the class on the exam. Computed by subtracting the number of successes by the low group on the item from the number of successes by the high group, and divide this difference by the size of the class. Ranges from -1.0 to +1.0; the closer to 0.0 indicates no discrimination among high- and low-performing students. Achieving closer to 1.0 discrimination index is optimal. A discrimination index of 0.3 or greater is considered highly discriminating, with no need of item revision with a discrimination index of less than 0.3 usually the item need revision or elimination[11].

Distractor analysis or efficiency (DE): If all the alternatives functioned as intended with equal distribution in choosing the wrong keys (10). The quality of MCQs items depend on these 3 indices[10].

Point biserial; generally a value of 0.2 and above is considered to have high correlation and positive association with overall performance on the assessment (i: e, acceptable ranges are 0.2 or above 0.2) lower levels are acceptable for mastery; and 0.3 or higher are best for discriminating questions A positive point biserial indicates that those scoring high on the total exam[11].

Reliability coefficient is the extent to which the test is likely to produce consistent scores and this implies the study of inter-correlations between the test items, the test size (more items more reliability) and test content. Ranges from zero (no reliability) to 1.00 (perfect reliability) but practically the range is usually from 0.5 to 0.9 the bigger this ratio the more. This Kuder-Richardson Formula 20, often abbreviated KR-20, is used to measure the internal consistency reliability of the test and is influenced by number of items in the exam[10,11].

The objectives of this study were to evaluate the detailed indices of the analysis of the items of the written multiple choice questions (MCQs) exams in the department of child health in the Faculty of Medicine over the last four academic years (1439-1443 AH) and to construct the outlines of a plan for improving the upcoming written MCQs exams in the department based on the indicators of the analysis.

MATERIALS AND METHODS

This was a retrospective cross sectional study performed by collecting the data regarding the item analysis of all midterm written MCQs exams in the MBBS course of pediatrics - 2 in the College of Medicine King Khalid University for both boys and girls group. The course of pediatrics - 2 is taught both in theoretical and practical clinical aspects. The midterm exam is one of the major assessment methods in this course which is given at level 11 in the college and evaluate the knowledge of the students regarding the common pediatrics clinical issues that are included in the curriculum. The data was obtained from the examination department and constitute the difficulty, discrimination, point biserial reliability and distractor analysis of each of the exams items for the whole years of 1439, 1440, 1441, 1442 and the first semester of girl group in 1443. Each exam consists of a range of 40 with 3 exams only with 45 items of multiple choice questions. The choices always in our exam are out of 4 best responses. There is no penalty system on wrong answers. The indices of the exams analysis were automatically calculated electronically and registered by the correcting scanner machine. The report on the exam correction will give all the necessary indices required to accomplish this study. The total numbers of items studied in these 16 exams were 643 items. All of the items were included in the study as all of the questions fulfil the criteria adopted by the medical education department at the college. The indices and variables of the item analysis of the exams evaluated in this study were the following.

The grades and scores of each group of students

The difficulty index of the items considering a difficulty index (DIF I) of 80% as easy item, of 30% or less as difficult item and that between 30% and 80% of moderate difficulty

The discrimination index (DI): discrimination of 0.4 or above was considered as a very good discriminating item; good discriminating item if the DI between 0.3 to 0.39, moderately discriminating if between 0.2 to 0.29, poorly discriminating

Point biserial index and reliability: for validity and reliability; the point biserial of the items of the exams was considered of good performance if more than 0.15.

Non-functioning distractors percentage

All of the results obtained were tabulated and the different statistical correlations were computed using SPSS program and taking the significant level when the p-value is less than 0.05. The reliability of the exams were also determined according to KR-20 method.
The ethical approval for this study was given by the King Khalid University committee of research ethics on 17.2.2022 by the approval number (ECM#2022-607). There is no disclosure of any personal student information in this study.

This article was previously posted to the Preprint from Research Square, 29 Aug 2022,
https://doi.org/10.21203/rs.3.rs-1994965/v1 PPR: PPR537322, Preprint v1.

RESULTS

In this study 16 MCQs exams were evaluated for the item analysis of the major indicators of the exams quality. Of each of these exams 4 modules containing similar questions were prepared and taken by the students as part of their evaluation in pediatrics course in level 11. These exams were constituted of 643 items or questions with average number of items of 40 per one exam. The total number of students who performed these exams were 1002 students. There were 7 groups of boys composed of 593 students (59% of the total students) and 9 groups of girls composed of 409 students (41% of the total number of the students). The average number of students per one boy group is 84, while in girl group is 45 per one group.

Grades and scores of the student groups: The student scores A if he or she correctly answers 90% or more of the items, B between 80% to 90% correct, C between 70% to 80% correct, D between 60% to 70% and F if below 60%. The distribution of the students according to this score was as follows: A scored by 73 students (7.3%), B by 219 (21.8%), C by 331 (33%), D by 214 (21.4%) and F by 165 (16.5%).

In Table 1, these score values in general were of very comparable distribution between boys and girls with no significant statistical difference (p-value= 0.276). When the scores of the first semester

and second semester groups when compared there was a significant statistical difference (p=0.01)

Table 1. Scores of the students in the study group

Score	Combined students group n = 1002	Boys group n = 593	Girl group n =409
A	7.3%	8.2%	8.7%
B	21.8%	24%	26.4%
C	33%	31.2%	29.5%
D	21.4%	20.8%	20.2%
F	16.5%	15.8%	15.2%
Total %	100%	100%	100%

Difficulty evaluation of the exams: Overall when considering a difficulty index (DIF I) of 80% as easy item of 30% or less as difficult item and that between 30% and 80% of moderate difficult; we obtain 3 categories of items: difficult items were 43 (6.6%) of the total items, moderately difficult 343 (53.4%) and easy items 257 (40%) of the total items. There was significant statistical correlation ($p \leq 0.05$) when these difficulty levels compared over the exam years. In this same group of items as the recommendation of some educators and to increase the standard levels of our tests we chose to consider a difficulty index of 50 % to 60% as ideal, of 30% to 49% and 61% to 70% as acceptable, of less than 30% as very difficult item and of more than 70% as very easy item; the findings in our exam series were as follows only 11% of the items were of ideal difficulty index and 21.4% were of acceptable difficulty; so that the total of these 2 categories of reasonable difficulty constitute 32.4%. Items which were very easy constitute 61% which seems higher than the intended goal, while very difficult items were 6.6% which is not high. Difficulty indices of the studied items for boys and girl's exams and the statistical correlations are shown in tables -2 and -3 respectively (Figure 1).

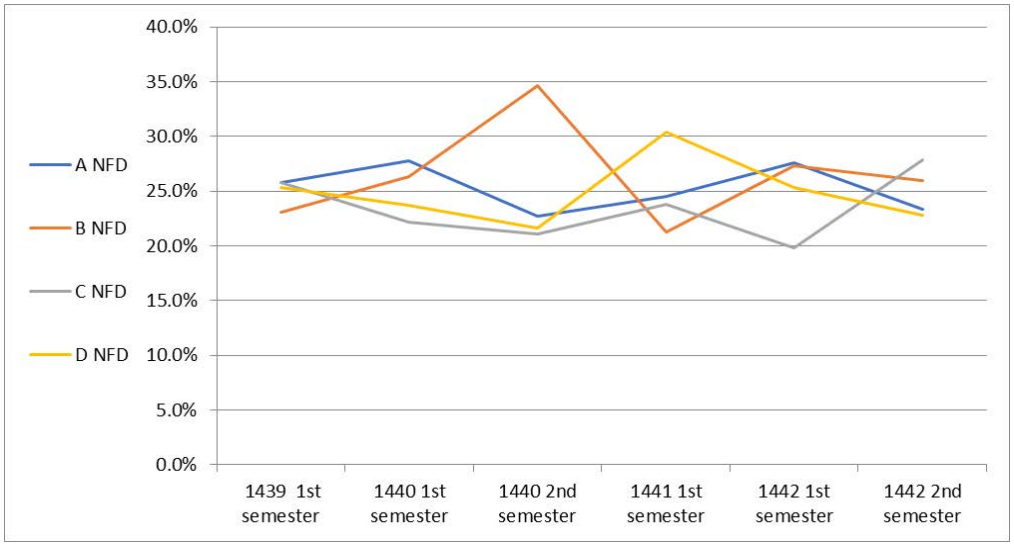


Figure 1. Different categories of the nonfunctioning distractors of the items in the exams in the boys group over the academic years (1439-1442 AH)

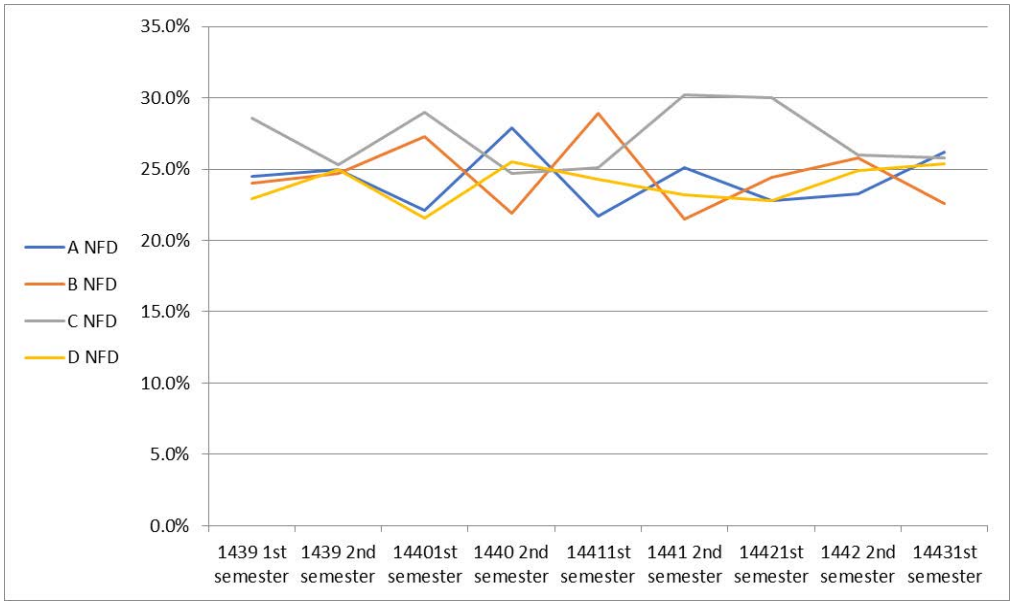


Figure 2. Different categories of the nonfunctioning distractors of the items in the exams in the girls group over the academic years (1439-1443 AH)

Table 2. The different important item analysis of the exams in boys groups years (1439-1442 H) and their statistical significance ratio

	1439		1440		1441		11442		P value
	1st semester	2nd semester	1st semester	2nd semester	1st semester	2nd semester	1st semester	2nd semester	
Difficulty index									
Ideal	7.5%	2.2%	7.5%	13.4%	10.6%		9.1%	16.9%	P < 0.05
Acceptable	18.3%	2.2%	27.5%	23.3%	23.3%		18.2%	34.4%	P < 0.05
Very difficult	20%	11.1%	10.8%	0%	6.7%		1.1%	0.6%	P < 0.05
Very easy	54.2%	84.4%	54.2%	63.3%	59.4%		71.6%	48.1%	P < 0.05
Discrimination index									
Very good test	52.5%	51.1%	57.5%	64.4%	49.4%		55.7%	58.1%	N. S
Good test	2.5%	15.6%	8.4%	4.5%	6.7%		1.7%	12.5%	P < 0.05
Moderately discriminating	15%	8.9%	19.1%	21.1%	22.2%		27.9%	19.4%	N. S
Not discriminating (marginal item)	17.5%	17.8%	1.7%	0%	5%		0%	0%	P < 0.05
Poor quotation	12.5%	6.7%	13.3%	10%	16.7%		14.7%	10%	N. S
Distractor index									
A Non-Functioning Distractors (NFD)	25.8%		27.8%	22.7%	24.5%		27.6%	23.3%	N. S
B Non-Functioning Distractors (NFD)	23.1%		26.3%	34.6%	21.3%		27.3%	26%	N. S
C Non-Functioning Distractors (NFD)	25.8%		22.2%	21.1%	23.8%		19.8%	27.9%	N. S
D Non-Functioning Distractors (NFD)	25.3%		23.7%	21.6%	30.4%		25.3%	22.8%	N. S

Table 3. The different important item analysis of the exams in girls groups years (1439-1442 H) and their statistical significance ratio

	1439		1440		1441		1442		1443	P value
	1st semester	2nd semester	1st semester	2nd semester	1st semester	2nd semester	1st semester	2nd semester	1st semester	
Difficulty index										
Ideal	10%	15.5%	16.7%	8.9%	16.7%	8.5%	12.8%	5.6%	13.8%	N. S
Acceptable	28.3%	24.5%	24.2%	26.7%	21.1%	16.5%	27.2%	15.6%	11.3%	N. S
Very difficult	8.3%	4.4%	8.3%	4.4%	6.7%	5.1%	6.1%	6.3%	6.3%	N. S
Very easy	53.4%	55.6%	50.8%	60.0%	55.5%	69.9%	53.9%	72.5%	68.7%	N. S
Discrimination index										
Very good test	50.0%	60.0%	50.0%	51.7%	46.1%	33%	62.2%	41.2%	45.6%	N. S
Good test	6.7%	0%	15.0%	7.7%	17.8%	35.8%	16.7%	0%	27.5%	P < 0.05
Moderately discriminating	7.5%	0%	6.7%	0%	0%	0%	0%	0%	0%	P < 0.05
Not discriminating (marginal item)	0%	0%	0%	0%	0%	0%	0%	0%	0%	N. S
Poor quotation	35.8%	40%	28.3%	40.6%	36.1%	31.2%	21.1%	58.8%	26.9%	N. S
Distractor index										
A Non-Functioning Distractors (NFD)	24.5%	25.0%	22.1%	27.9%	21.7%	25.1%	22.8%	23.3%	26.2%	N. S
B Non-Functioning Distractors (NFD)	24.0%	24.7%	27.3%	21.9%	28.9%	21.5%	24.4%	25.8%	22.6%	N. S
C Non-Functioning Distractors (NFD)	28.6%	25.3%	29.0%	24.7%	25.1%	30.2%	30.0%	26.0%	25.8%	N. S
D Non-Functioning Distractors (NFD)	22.9%	25.0%	21.6%	25.5%	24.3%	23.2%	22.8%	24.9%	25.4%	N. S

N.S: not significant statistically

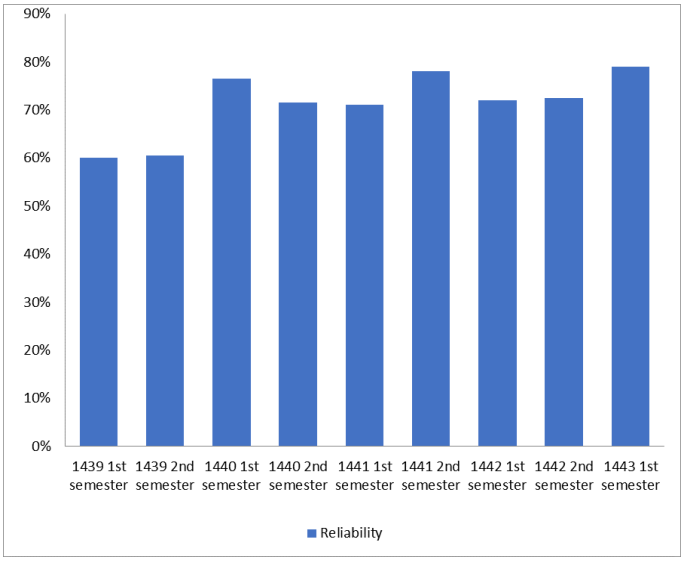
Discrimination index (DI): in this study an item with discrimination of 0.4 or above was considered as a very good discriminating item ; good discriminating item if the DI between 0.3 to 0.39, moderately discriminating if between 0.2 to 0.29, poorly discriminating if between 0.1 to0.19 and not discriminating if below 0.1.The average discrimination of the items in the study according to this scale was as follows ; very good DI in 331 items (51.5%), good DI in 73 items (11.3%), moderate DI in 61 items (9.4%), poor DI in 17 items (2.7%) and unacceptable DI in 161 items (25.1%). Therefore, acceptable discrimination of the items in this study was (72.2%) with (27.8%) of the items being of poor or no discriminating function. Discrimination indices of the studied items for boys and girls' exams and the statistical correlations are shown in Tables 2 and 3 respectively.

Point biserial index and reliability: as an indicator of exam discrimination and internal consistency and also for validity and reliability; the point biserial of the items of the exams was of good performance (more than 0.15) in average of 444 items (69% of the items) and not performing well in 199 items (31%) indicating necessary revision of the correctness of the keys in this group. Reliability detection according to point biserial was compared statistically between the girl and boys groups and the difference was significant (p=0.002) (Table 4). The average KR-20 calculated for the whole exams was about (71%). KR-20 for the years (1439-1441AH). The reliability according to this index is shown in Figure 3 and the statistical association of reliability compared to the different exam years and also to the gender of the students was not significant (Figure 2).

Non-functioning distractors percentage: Percentages of zero non-functioning distractors (ANFD), 1NFD (BNFD), 2NFD (CNFD) and 3 NFD (DNFD) of the items of the exam series are shown in Table 2 for male groups and Table 3 for female groups. These different 4 categories of distractor indices range from 20% to 35% in males and from 20% to 30% in female groups and the levels in the exams as compared over the last 4 years was not statistically significant (Tables 2 and 3)

DISCUSSION

Multiple choice questions (MCQ) form useful assessment tool in measuring factual recall and if carefully constructed can test higher order of thinking skills which is very important for a medical graduate. The method of assessment should be regularly evaluated. Developing an appropriate assessment strategy is a key part in curriculum



Association between years and Reliability (P value =0.691)
Association between sex and Reliability (P value =0.336)
Figure 3. Reliabilty estimation (KR-20) of the items of most of the exams studied (1439-1441 AH)

development. It is important to evaluate MCQ items to see how effective they are in assessing the knowledge of students[5]. This the most widely used test format in health sciences today. The efficiency of MCQs as an efficient tool for evaluation depends mainly on their quality which is best assessed by item and test analysis[12].

Item analysis is especially valuable in improving items which will be used again in later tests, but it can also be used to eliminate ambiguous or misleading items in a single test administration. In addition, item analysis is valuable for increasing instructors' skills in test construction, and identifying specific areas of course content which need greater emphasis or clarity. Separate item analyses can be requested for each raw score[13]. Such statistics must always be interpreted in the context of the type of test given and the individuals being tested[14].

According to the results of this study many aspects of the MCQs exams in the department of pediatrics require further deep study and post-test re-evaluation as for consideration in the same test or the future use of

Table 4. Point biserial indicators of exam reliability in the item series with comparison between boys and girls group

Sex	Point of Biserial	Interpretation	1439 Sem1	1439 Sem2	1440 Sem1	1440 Sem2	1441 Sem1	1441 Sem2	1442 Sem1	1442 Sem2	1443 Sem1
Boys	(≥ 0.15)	Performing well	66.7%		61.8	88.9	75.5	MCQs Exam Not done	90.3	92.5	Not included
	(< 0.15)	Should be examined for a possible incorrect key	33.3%		38.2	10.1	24.5	MCQs Exam Not done	9.7	7.5	Not included
	Total		100%	100%	100%	100%	100%	100%	100%	100%	100%
Girls	(≥ 0.15)	Performing well	51%	55.6	66.7	66.7	63.9	69.3	81.7	42.5	70.6
	(< 0.15)	Should be examined for a possible incorrect key	49%	44.4	33.3	33.3	26.1	30.7	18.3	57.5	29.4
	Total		100%	100%	100%	100%	100%	100%	100%	100%	100%

p = 0.002

the items. Optimizing the different important item analysis indices is an essential aim of medical exams constructors. A study by Poulomi et al[13] concluded that items with average difficulty, high discrimination and functional distractors are the best to be incorporated in the exams[12]. Avery similar conclusion was reached to by the study performed by Durgesh et al[15]. Determination of the item analysis indices can -to certain extent- differ according to the cognitive levels of Bloom taxonomy; Serpil et al[16] in their study found that difficulty and discrimination were more associated with the remembering and understanding levels than with the applying level[16]. In any case, all indices should be considered together before making decisions or revisions[17].

Comparable and similar results to this study were also reported by other authors when they evaluated the MCQs tests in their courses. Al Shaibani et al[18] concluded that the mean DIF- I, DI and DE were in the acceptable ranges. A high percentage of items were 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[18]. Vrunda et al[19] study results were as such: difficulty index of analysed MCQs ranged from 6.25% (lowest) to 90.6% (Highest) & discriminative index ranged from 0 (lowest) to 0.63 (Highest). Total 65% items were in acceptable range of difficulty level ('p' value 30 – 70%) and 10 % items were very difficult which later discussed with students. Discrimination index of 60% items was excellent (d value>0.35). No item had Negative discriminative power. About 47.5% items had100% Distracter Efficiency (DE) whereas 7.5% items had 0% DE[18]. These two studies have figures of indices of item analysis which may resemble or differ from the figures in this study which indicate the need of individualization of the analysis of every institutional exam and looking for underlying factors of the analysis result. Many other similar studies in medical schools in the Gulph or other regional parts obtain useful results and indicators for improvement of their MCQs exams; for example the studies done by Rao et al, Prashant et al and Kheyami et al (8, 20,21). Gajjar et al (5) in their detailed study with objectives close to this study in assessing the analysis indices stated that out of 50 items, 24 had "good to excellent" DIF I (31 - 60%) and 15 had "good to excellent" DI (> 0.25). Mean DE was 88.6% considered as ideal/ acceptable and non-functional distractors (NFD) were only 11.4%. Mean DI was 0.14. Poor DI (< 0.15) with negative DI in 10 items indicates poor preparedness of students and some issues with framing of at least some of the MCQs. Increased proportion of NFDs (incorrect alternatives selected by < 5% students) in an item decrease DE and makes it easier. There were 15 items with 17 NFDs, while rest items did not have any NFD with mean DE of 100% compared with this present study there was no negative DI which is a positive finding in our setting. Actually they proposed the cause for negative DI in their sample that it can be wrong key, ambiguous framing of question or generalized poor preparation of students as was the case in their study where the overall score of their students was very poor (0-33/100) and none of them secured passing marks. Items with negative DI are not only useless, but actually serve to decrease the validity of the test (5).

Grades and scores of the student groups in the present study: the curve of grade distribution is acceptable as most of the students lie in the central part of the curve giving somewhat a normal distribution; however there is tendency of F results to double the A scores which is a source of some annoyance. The difference between the student scores in the first compared to the second terms may be attributable to the different choice level of students when they are initially accepted to the college as students with higher levels at entrance were those of the first term.

Difficulty evaluation of the exams: to calculate the Index of Difficulty we need three parameters. They are a) number of higher achievers (HA), b) number of lower achievers (LA) and c) total number of respondents (N). At the end of the Item Analysis report, test items are listed according their degrees of difficulty (easy, medium, and hard) and discrimination (good, fair, and poor). These distributions provide a quick overview of the test, and can be used to identify items which are not performing well and which can perhaps be improved or discarded with this high percentage of very easy items in our midterm exams of 61% measures are needed to improve the difficulty levels of the items (although the difficulty is lower than average in some of the groups). Part of the problem may be attributed to the repetition of certain questions in a way or another and part may be due to possession of the students of large question banks with consumption of many questions used by the department. This necessitates renewal of the questions bank in the departmental exam committee and more effort of the faculty members to re –edit or replace the questions. Ideal difficulty levels for 4 response multiple-choice items in terms of discrimination potential is approximately 74%. This is highly recommended to improve the levels of the exams (13). There are instances when the value of DI can be 60% will result into to inflated scores and a decline in motivation. Too easy items should be placed either at the start of the test or better to be removed altogether, similarly difficult items should be reviewed for possible confusing language, areas of controversies, or even an incorrect key (5). A study by Deepak et al (22) concluded that even if a single distracter is non-functioning or poor, it would seriously affect the psychometric properties and reliability of the test paper in a 3-option format. The present data clearly provides evidence that the items with three non-functioning distracters have serious psychometric inadequacy (22). Ideally, if the distractors are properly designed, it should lead to LA group selecting the incorrect options more often than the HA group (23). In their review of functioning and non-functioning distractors in 514 four option MCQs assessments, Tarrant et al (3) found that only 13.8% of all items had three functioning distractors and just over 70% had only one or two functioning distractors (3).

A reliability coefficient between 70% and 80% is good for classroom tests Good for a classroom test; in the range of most. There are probably a few items which could be improved. While that of level 60% to70% is 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. Between 50% and60% there is a need to revise the whole test (1). Factors affecting the reliability of the test are multiple and include the length of the test with improved reliability with more number of questions, proportion of correct and incorrect responses, item difficulty as very easy and very difficult items do not discriminate properly and also number and individual factors regarding the examinee (17). Thus through item analysis, the instructors can improve their skill in constructing valid MCQs in the future. In addition it also directs the curriculum administrators to identify specific areas of the course content that needs revision or further clarification as evidenced by poor mastery of the subject (23). In a study by Saxena et al; after the analysis of their MCQs test and considering the DIFI, DI and DE together, 23.33% items were validated for MCQ Bank, 56.67% items would be re-validated after revision and modification, and 20% items were discarded (24).

Limitations of this study were minimal in form of getting the needed permission for the data obtainance which required some process. Also, further study of the final written exams is needed to complete the picture of item analysis evaluation in the department.

CONCLUSIONS AND RECOMMENDATIONS

Based on the findings of this study that the percentage of easy items approaching that of those with average difficulty, non-discriminating items in one fourth of the tests, average reliability just over 70% and average of 3 NFD of about 24%; it is of vital importance to the exam committee in the department to get benefits of these item analysis indicators to improve the quality and educational and assessment yield of the written MCQs exams. The followings are suggested practical steps for the committee towards that goal.

Faculty member extensive training on proper construction of the stems, keys and distractors of the MCQs in a detailed professional manner to improve the pre-test preparation of the exam and improve the sense of item adjusting regarding the most important indices at the expected desired level.

Committee exam meetings to revise every item in context of the ideal MCQs construction guidelines with efforts of all of the members to have their inputs in putting the best possible items.

The post-test meeting for discussion of the exam item analysis is a neglected important part of the job. With consensus of the committee some poor items can be omitted form the correction and some of the good items can be selected to enrich the departmental questions bank for reuse in the future exams; with rejection of items with improper indices as not accepted for developing the departmental bank items.

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 Conflicts of Interest: None

Competing Interest: None

Acceptance Date: 22-04-2024

REFERENCES

1. Scorepak®: Understanding Item analysis. Available from: www.washington.edu./oea/score1/htm. [Last Accessed on 13 September 2015]. https://www.washington.edu/assessment/scanning-scoring/scoring/reports/item-analysis/
2. Ananthakrishnan N. Item Analysis validation and banking of MCQs. In: Medical Education Principles and Practices. 2nd ed. Ananthkrishnan N, Sethuraman KR, kumar S. Alumini Association of National Teacher Training Centre, JIPMER, Pondichery, India. p,131-7.
3. Tarrant M, Ware J, Mohammed AM. An assessment of functioning and nonfunctioning distractors in multiple choice questions: a descriptive analysis. BMC Med Educ 2009; 9:1-8.
4. Patel RM. Use of item analysis to improve quality of multiple choice questions in II MBBS. J Educ Technol Health Sci 2017; 4(1): 22-9.
5. 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 Journal of Community Medicine 2014; 39:17-20.
6. Singh T, Gupta P, Singh D. Test and Item analysis. In: Principles of Medical Education. 3rd ed. New Delhi: Jaypee Brothers Medical Publishers (P) Ltd. 2009; 70-7.
7. Hervas G. Features and qualities of educational assessment. Center for Teaching and Learning International Christian University. 2020; 24(2):8-13.
8. Rao C, Kishan Prasad H L, Sajitha K, et al. Item analysis of multiple choice questions: Assessing an assessment tool in medical students. International Journal of Educational and Psychological Researches. 2016; 2: 201-4.
9. Van der Vleuten C. Validity of final examinations in undergraduate medical training. BMJ; 2000; 321(7270):1217-9.
10. Zurawski RM. Making the most of exams: procedures for item analysis. The National Teaching & Learning Forum. 1998; 7(6): 1-12.
11. Ermie E. Psychometrics 101: Know what your assessment data is telling you. Slides presented at the Exam Soft Assessment Conference 2017, Denver, Colorado. 2017.
12. Lord, FM. The Relationship of the Reliability of Multiple-Choice Test to the Distribution of Item Difficulties, Psychometrika, 1952; 18, 181-94.
13. Poulomi M, Saibendu KL. Analysis of Multiple Choice Questions (MCQs): Item and Test Statistics from an assessment in a medical college of Kolkata, West Bengal IOSR Journal of Dental and Medical Sciences (IOSR-JDMS) 2015; 14(12): 47-52.
14. Pande SS, Pande SR, Parate VR, et al. Correlation between difficulty and discrimination indices of MCQ's informative exam in physiology. South East Asian J Med Educ 2013; 7:45-50.
15. Durgesh PS, Rakesh S. Item and distracter analysis of multiple choice questions (MCQs) from a preliminary examination of undergraduate medical students ; International Journal of Research in Medical Sciences. Int J Res Med Sci. 2017; 12: 5351-5.
16. Serpil K, Nejdet K, Murat DŞ. Analysis of the Difficulty and Discrimination Indices of Multiple-Choice Questions According to Cognitive Levels in an Open and Distance Learning Context; Turkish online Journal of Educational Technology. 2016; 15(4): 16-24.
17. Suskie, L. Making Multiple Choice Tests More Effective. Schreyer Institute for Teaching Excellence. The Pennsylvania State University. 2017.
18. Al Shaibani T, Fuad A, Deifalla AH, et al. Item Analysis of type "A" Multiple Choice Questions from a multidisciplinary Units assessment in a Problem Based Curriculum Bahrain Medical Bulletin, 2021; 43(2); 471-6.
19. Vrunda K. Item analysis of Multiple-Choice Questions in Physiology examination; Indian Journal of Basic and Applied Medical Research, 2015; 4(4): 320-6.
20. Prashant NE, Souza FD. Item Analysis (Postvalidation) of MCQs in Anatomy for the First MBBS Course. Indian Journal of Anatomy, 2016; 5(1): 29-32.
21. Kheyami D, Jaradat A, Al-Shibani T, et al. Analysis of Multiple Choice Questions at the Department of Paediatrics, Arabian Gulf University, Manama, Bahrain, Sultan Qaboos Univ Med J. 2018; 18(1): e68–e74**.**
22. Deepak KK, Al-Umran KU, Al-Sheikh MH, et al. Psychometrics of Multiple Choice Questions with Non-Functioning Distracters: Implications to Medical Education. Indian J Physiol Pharmacol 2015; 59(4); 428-35. https://ijpp.com/ijpp_archives_new.php/
23. Senthil Velou M, Ahila E. Refine the multiple-choice questions tool with item analysis. IAIM, 2020; 7(8): 80-5.
24. Saxena S, Srivastava PC, Mallick AK, et al. Item analysis: An unmatched tool for validating MCQs in Medical Education. Indian Journal of Basic and Applied Medical Research, 2016; 5(4): 263-9.