Medical education

Distractor analysis of multiple choice questions: A descriptive study of physiology examinations at the Faculty of Medicine, University of Khartoum

Afraa Musa [1], Samir Shaheen [2], Ammar Ahmed [1]

[1] Department of Physiology, [2] Department of Orthopedic Surgery Faculty of Medicine, University of Khartoum, Sudan

Abstract:

Background: Distractor analysis is an important component of item analysis. It helps in the design and construction of items with functional effective distractors which are valid for future use and development of proper question banks.

Objectives: This study aimed at investigating distractors' functionality of physiology multiple choice question (MCQ) items at the Faculty of Medicine, Khartoum University using post- examination item analysis data.

Methods: Analyzing test statistics reports produced by Remark-Software, the frequency distribution of non-functioning (NFD) or non-effective distractors (NED) on ten summative physiology MCQ exams administered to undergraduate medical students in the period from September 2015 to September 2016 was assessed. Item analysis reports provided detailed options analyses in the form of response frequency and discrimination power of each option. Each exam paper consisted of 60-80 (five-option) items and the number of the examinees ranged from 332 to 359. A total of 654 items were reviewed, including 2580 distractors. NFD/NED were defined as distractors with either response frequency<5% of the examinees or non-negative discrimination coefficient (r_{pbis}) criteria. Furthermore, the number of NFD/NED per item were computed to measure distractor efficiency (DE %) of an item as follows; 0 NED = 100 % DE, 1 NED = 75% DE, 2 NED = 50% DE, 3 NED = 25% DE, 4 NED = 0% DE.

Results: Although less than one-third (31%) of total number of distractors were non-functioning because they were chosen by less than 5% of examinees, only 4 (0.15%) distractors had a choice frequency of zero. In addition, 9.38% of total number of distractors was non-functioning because they had non-negative discrimination correlation coefficient. Defining functioning distractors using response frequency or discrimination ability criterion, the majority (61%) of our distractors were functioning effectively. The mean DE was 59.46±28.376% and the proportion of functional distractors ranged from 40.27% - 75% in the ten exams. Our study proportions of items with 100%, 75%, 50%, 25%, 0% DE were: 32.7%, 31.3%, 20.9%, 9.9% & 5.1% respectively based on response frequency criterion alone**,** compared to: 8.2%, 12.3%, 28.7%, 35.2% & 15.7% respectively when using any of the two criterion for NED definition.

Conclusion: The findings that the majority of our distractors were functioning and the high proportions of our items that have high distractor efficiency are explained by the fact that our expert staff members carefully reviewed distractors to ensure unambiguity of the correct answers and generated plausible effective distractors.

**** Corresponding author:*** *(aframusa@gmail.com)*

Introduction

Assessing students' learning is an important component of the teaching/ learning process. Assessors need to be aware of the potential biases that can influence all components of the assessment cycle from question creation to the interpretation of exam scores. Today, multiple choice questions

(MCQs) are the most commonly used tool for assessing the knowledge capabilities of medical students and they often account for a substantial portion of their course grades [1, 2]. Multiple choice questions have the advantages of providing a large number of examination items that encompass many content areas; can be administered in a relatively short period; and can be graded by computer. These factors make the administration of the examination to large numbers of trainees straight forward and standardized. However, designing good MCQs is a complex, challenging and time consuming process[3]. Item analysis uses statistics and expert judgment to evaluate tests based on the quality of individual items, item sets, and entire sets of items, as well as the relationship of each item to other items. It investigates the performance of items considered individually, either in relation to some external criterion or in relation to the remaining items on the test [4,5]. Post-exam analysis of students' performance yields many indicators that test the quality of exams by calculating difficulty index (DIFI), discrimination index (DI), distractor efficiency (DE) and their interrelationship. It gives some idea of how well the examination has performed relative to its purposes and ,thus, how future learning can be supported and directed [6-8]. On the basis of item analysis, a revision and improvement of the test can be made [9].

Creating effective test items may be more of an art than a science and is considered as one of the greatest challenges for test developers [10]. Distractors are the incorrect alternatives or options. Each distractor is often based on a common misconception about the correct answer. Distractor analysis is achieved by reviewing the percentage/ proportion of examinees who select each of the distractors (response frequency). To be effective, distractors should look plausible to less knowledgeable students and lure them away from the keyed option, but it won't entice students who are well-informed about the topic under consideration [2]. Distractor analysis is very useful in detection of implausible distractors. If examinees consistently fail to select a given distractor, this may be the evidence that this distractor is implausible or simply too easy. Whenever the proportion of examinees who selected a distractor is greater than the proportion of examinees who selected the key, the item should be examined to determine if it has been mis-keyed or double-keyed [2, 7, 11, 12].

From a functional perspective, a distractor must meet two criteria in order to be effective [2]. First: at least some examinees must select it; if they do not, then the distractor is not luring anyone away from the keyed option, and it cannot contribute to the item's discriminatory power. Haladyna & Downing[13] have suggested that at least 5% of examinees should select each of an item's distractors, and this value is a common benchmark for distractor functionality. The second criterion relates to distractor's ability to discriminate between stronger and weaker examinees. For an MCQ item to have good discriminatory power, examinees with higher test scores must select the keyed option more frequently than those with lower scores (i.e. key option is positively correlated with total test scores). Conversely, for a distractor to be effective, the examinees with higher test scores select the distractor more frequently than those with lower scores (i.e. distractor is negatively correlated with total test scores). Then, a distractor has either a positive or a zero correlation with total test scores, it is considered as not functioning properly and detracts from an item's overall quality.

Reviewing the literature revealed that the majority of researchers [1, 14-21] have used the first criterion to measure distractors functionality. Only few studies used the two criteria [2, 22]. Distractor efficiency/ effectiveness (DE) or functionality of an item is determined by the number of functional distractors in each item and expressed as percentage values ranging from (100%) in case of items having all distractors functioning to (zero %) when having all distractors non- functioning [14, 16, 17, 20, 23]. The best is to have items with 100% DE; the worst are items with zero% DE.

It is well known that assessment drives learning. Therefore, developing a perfect tool of assessment should be the goal for any educator in an assessment position. Although when guidelines for constructing

fair and systematic tests are followed, a plethora of factors may enter into a student's perception of the test items. Yet, the creation of high-quality MCQ items is a learnable skill [2].

The aim of this study was to measure effectiveness/ efficiency of distractors of physiology MCQs' exam items in the Faculty of Medicine, University of Khartoum using post-examination analysis reports so as to guide teachers to develop appropriate question banks. Item analysis phase of test development is essential to identify potential item problems in order to improve their quality for future exams. This is achieved by doing some corrective measures (i.e. removal or modification of badly constructed items). Furthermore, it helps changing the decision of assessment by rescoring the exams after removing these items or changing the key answer of the miskeyed items. It also enables identifying misconceptions which can be corrected by counseling or by modifying the learning methods. On the other hand, it provides teachers with constructive feedback about their teaching and students' achievement [18]. Item analysis is beneficial for both the student and the teacher.

Methods:

The study was conducted at the Physiology Department, Faculty of Medicine, University of Khartoum. It retrieved ten summative MCQ exams (mid-semester and end-of-semester) of physiology that were administered in the period from September 2015 to September 2016. All tests were developed guided by a blueprint that matched each test item to the corresponding course objective and were reviewed by a team of senior expert staff prior to administration. Test content included basic undergraduate physiology systems except physiology of the central nervous system. Each exam paper consisted of 60-80 (five-option) items that comprised of a stem, one correct answer and four distractors. The total number of items was 650 out of which 5 items were excluded before marking the answer sheets because they were wrongly constructed with 2 correct options. The remaining 645 items included 3225 options (645 correct responses & 2580 distractors). The students' answer sheets were scanned using the digital optical scanner of the marking machine. The Remark Software analyzed the data using Microsoft Excel program that provided the results of performance analysis in different excel sheets for each individual exam. The Remark Software provided detailed statistical analysis of students' scores test reliability; item difficulty index; and point-biserial correlation as a measure of item discrimination. In addition, it provided distractor analysis (i.e. options analyses in the form of response frequency and discrimination power of each option) [22].

Data analysis:

Data analysis was carried out using the software Statistical Package for the Social sciences (IBM SPSS statistics) version 23. Item difficulty is the proportion of examinees answering the question correctly while point-biserial correlation coefficient measures the association between the test item and the total test score. For this study, we used the two criteria to evaluate distractor functionality. We identified the number of NED per item (based on distractor response frequency<5% criteria alone) and number of NED/ item (based on distractor discrimination power criteria alone) and further identified the number of NED per item (based on the two criteria together). Then distractor effectiveness or efficiency (DE) is calculated as follows based on response frequency criteria: (0 NED=100% DE, 1 NED=75% DE, 2 NED=50% DE, 3 NED=25% DE, 4 NED=0% DE).

Ethical clearance:

The research study was approved by the Research Ethics Committee of the Faculty of Medicine, University of Khartoum. As the software provided anonymous data [deals with exam numbers rather than examinees' names], there was no need for full board review and informed consent.

Results:

The ten analyzed tests consisted of 645 items (2580 distractors). The number of items on the tests ranged from (60–80), while the number of examinees

ranged from (332–359). Mean percent scores varied from (48.62-65.67%). Our tests were characterized by very high reliability that ranged from (0.86- 0.95) as measured by Kuder-Richardson Formula 20.

Proportions of functional/non-functional distractors according to definition criteria

Regarding the ten exams, there were total of 2580 distractors. Table 1 shows less than one-third (31%) of total number of distractors were non-functioning because they were chosen by less than 5% of examinees; only 4 of them (0.15%) were very implausible (i.e. had a choice frequency of 0). On the other hand, 9% distractors were non-functioning because they had non-negative discrimination correlation coefficient (i.e. equal to or greater than zero). Considering distractor as non-effective only if it fulfills any of the two criterions increased the proportion of NED to 39%. In other words, when distractor efficiency/functionality was calculated based on the response frequency criterion alone, 69% of total distractors were effective (functioning) distractors and 91% when using the discrimination criterion alone. In case of using any of the criteria for distractor functionality definition, the proportion of functioning distractors was reduced to 61%.

Distractor efficiency of items according to definition criteria

Our exams' items (n=645) were characterized by high efficiency as their mean distractor efficiency (DE %) was very high (69.53, 90.78 & 59.46) using the different methods of definition of functional distractor, shown in table 2.

Distractor analysis of individual exams

Distractor analysis of individual exams comprised of frequency distribution of NED/NFD using different definitions. It also showed the proportion of properly functional distractors which varied across exams, ranging from as low as 40.27% (exam 4) to as high as 75% (exam 9). In addition, it revealed proportions of items having different number of functional distractors (distractor efficiency) in each exam and that the modal number of functional distractors was three for almost all of the exams as shown in table 3.

Proportions of items' distractor efficiency according to definition criteria

Regarding total number of items, analysis of the number of NED per item revealed that items with 100% DE had the highest frequency followed sequentially by items with 75%, 50%, 25% DE and the lowest frequency was for items with zero functioning distractors based on response frequency criterion alone, as shown in figure 1 & on discrimination criterion alone as shown in figure 2. Using any of the two criteria (i.e. response frequency or discrimination) to define number of NED in an item, figure 3 revealed that the highest frequency was for items with 75% DE (i.e. the modal number of functional distractors was three); 8.2% of the total items had none of their distractors functioning properly (0% DE); while only 15.7% of the items, had fully functioning distractors (100% DE).

Table 1: Proportions of functional/ non functional distractors using the different criteria for definition of functional distractor

Criteria of definition	Number of NED/NFD	% of NFD/ NED out of total distractors	% of functional distractors out of total distractors
Based on response frequency	794	31	69
Based on discrimination power	242	9	91
Based on either response frequency or discrimination power	1000	39	61

* Total number of distractors (2580)

Table 2: Distractor efficiency of the total number of items using the different criteria for definition of functional distractor

	DE % based on response frequency	DE% based on discrimination power	DE% based on either response frequency or discrimination power criteria
No of items	645	645	645
Mean	69.53	90.78	59.46
SD	28.89	15.83	28.376
Minimum	0	0	0
Maximum	100	100	100

Table 3: Distractor analysis of individual exams

Exam	Items No	Total No of distractors	No (%) of NFD with			No (%) of Functional Distractors based on either response frequency or discrimination criteria	% Functional of Distractors/ item based on either response frequency or discrimination criteria				
			Frequency <5%	Pbi ≥0	≥1 of the criterion		0	1	2	3	4
End S2 Nov 2015	70	280	90 (32.14%)	25 (8.92%)	112 (40%)	168 (60%)	7.1	18.6	24.3	31.4	18.6

End S3 Nov 2015	79	316	88 (27.848%)	38 (12.02%)	122 (38.6%)	194 (61.4%)	6.3	13.9	21.5	44.3	13.9
End S3 April 2016	59	236	57 (24.15%)	31 (13.13%)	84 (35.59%)	152 (64.41%)	3.4	8.6	41.4	22.4	24.1
End S4 April 2016	80	320	158 (49.37%)	44 (13.75%)	190 (59.37%)	130 (40.27%)	21.3	18.8	37.5	17.5	5.0
Mid S2 Sep 2015	60	240	63 (26.25%)	20 (8.33%)	83 (34.58%)	157 (65.42%)	1.7	8.3	33.3	40.0	16.7
Mid S3 Sep 2015	60	240	68 (28.33%)	18 (7.5%)	84 (35%)	156 (65%)	3.3	10.0	28.3	40.0	18.3
Mid S2 Sep 2016	60	240	66 (25.7%)	21 (8.75%)	101 (42.08%)	139 (57.92%)	25.0	10.0	30.0	35.0	0.0
Mid S4 Sep 2016	60	240	58 (24.166%)	17 (7.08%)	73 (30.41%)	167 (69.59%)	1.7	6.7	26.7	43.3	21.7
Mid S3 Feb 2016	58	232	47 (20.25%)	17 (7.32%)	58 (25%)	174 (75%)	1.7	1.7	17.2	50.0	29.3
Mid S4 Feb 2016	59	236	99 (41.949%)	10 (4.23%)	106 (44.92%)	130 (55.08%)					
Total	**645**	**2580**									

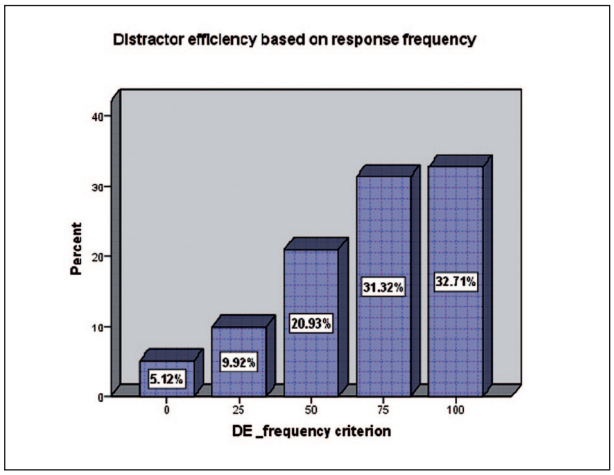

Figure 1: Proportions of items' distractor efficiency based on response frequency criterion. (Items=645)

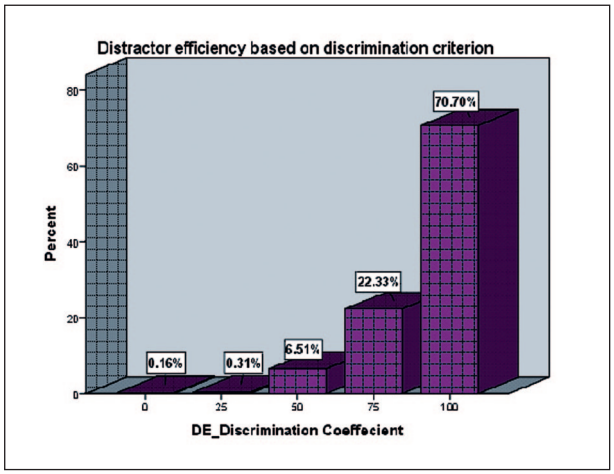

Figure 2: Proportions of items' distractor efficiency based on discrimination ability criterion. (Items=645)

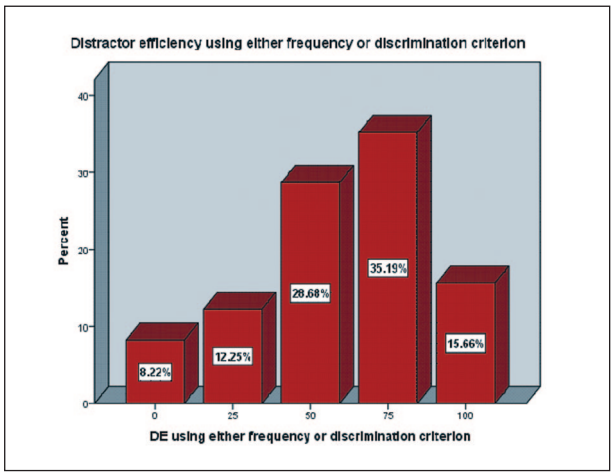

Figure 3: Proportions of items' distractor efficiency using any of the response frequency or the discrimination criteria for definition of functional distractor. (Items=645)

Discussion:

Distractor analysis was done to estimate their relative usefulness in each item by studying the students responses to each alternative or option. Items' evaluation is achieved by looking into their individual distractors; then followed by removal or modification and replacement of implausible options (NFD/ NED) with more plausible options. When modifying items for future use, content experts should follow the guidelines for writing effective distractors by Haladyna [24] & McDonald[25].

Using the definition of non-functional (i.e. the option with response frequency <5%), less than one third (31% out of 2580) of our distractors [***table 1***] were identified as non-functional which was consistent with the large Chinese study by Tarrant et al. who identified (35.1%) out of 1542 as non-functional distractors [22]. Few studies that included small number of distractors reported lower proportions of NEDs. They showed variable results: (6% out of 300 distractors) [19]. (11.4% out of 150 distractors) [15]. (14.06% out of 300 distractors) [16]. (17.8% out of 90 distractors) [20]. (18% out of 150 distractors [18] and (23.5% out of 200 distractors[14]. On the other hand, some studies revealed much higher proportions of NED among their distractors: 41% (out of 300 distractors) [21]. 76.66% (out of 60 distractors) [17] . and 81.3% (out of 150 distractors) NEDs [1].

Although nearly one third of our distractors were chosen by <5% of the examinees, only 3 of them (0.12%) were very implausible (i.e. not attempted by any of the examinees) which was extremely lower than the proportions of distractors having zero response frequency reported by Tarrant et al. (10.2%) [22], Haliker et al. (31.66%) [17] and (46.06%) by Mehta & Mokhasi [1]. This might yet reflect a degree of attractiveness of these distractors to some of the low performers despite of their bad construction.

Our study revealed 61% out of 2580 were functional distractors [table 1] when using any of the two criteria of definition of a properly functioning

distractor. Though using the same definition for a functional distractor, our proportion of functioning distractor was higher than the two large Chinese and Canadian studies which reported (52.2% out of 1542 distractors) [22] and (55% out of 3819 distractors) [2] respectively.

The distribution of our five-option items was ideal [figure 1& figure 2], showing the highest proportion having fully functioning distractors and the lowest proportion having none of their distractors functioning compared to the other studies [1, 14-18, 20, 21] that analyzed smaller numbers of four-option items and revealed variable results of items distractor efficiencies. Our study revealed [figure 1] that nearly one third of items (32.7%) were fully functioning/100% DE (using the response frequency definition of NED) which was higher than proportions reported by other studies; 26% [21], 13.8% [22], 11.33% [1] . Yet again, our findings were extremely better than other studies that showed none of their items (zero%) had 100% DE [13, 17]. Analyzying only one exam, few studies reported higher proportions of fully functioning items; (62% out of 50 items) [18] and (42% out of 50 items) [14]. Though it is difficult to construct items (especially five-option type) with four functioning distractors, our high proportion of items with fully functioning distractors compared to other studies, was not surprising given the fact that these exams were constructed by qualified staff members who carefully reviewed the items so as to ensure generation of plausible distractors. In addition, another greater portion of our distractors (31.3%) had 3 functioning distractors (75% DE/ 1 NED), which confirms the fact that some questions have naturally different number of feasible plausible distractors depending on item contents. Usually the number of options that items should contain is determined by the institutional guidelines. It is often difficult for teachers to develop three or more equally plausible distractors; therefore, additional distractors are often added as "fillers" to conform to the guidelines [2, 22]. Fortunately, we have very minimal proportion 5.12 % of 645 items having (0% DE) [figure 1], compared to other studies which reported higher proportions; (12.3%) [22], (18.18%) [1] and up to (50%) [17].

Based on the definition of NED using any of the two criteria [figure 3] , the modal number of functional distractors was three as the proportions of our items having (0,25,50,75,100% DE) were as follows: 8.2%, 12.3%, 28.7%, 35.2% and 15.7% respectively. That was comparable to the 12.3%, 34.8%, 39.1%, and 13.8% proportions of items containing 0, 1, 2, and 3 functioning distractors respectively, obtained from the large Chinese study which analyzed 514 four-option items and used the same definition of NED. Their study emphasized that only 13.8% of all items had three functioning (fully functioning) distractors and just over 70% had only one or two functioning distractors [22]. Also Haladyna & Downing found that approximately two-thirds of all four-option items had only one or two functioning distractors and none of the five-option items had four functioning distractors. Therefore, they concluded that "three options may be a natural limit for multiple-choice item writers in most circumstances" [13].

Limitations of study:

Limitation of this study may be due to the fact that MCQs evaluation was conducted in only one department. In order to reach a conclusion of generalizability about the quality of MCQs' assessment tool, we need to recruit the rest of the Faculty's departments in such evaluation.

Conclusions & Recommendations:

The high quality of our distractors may be attributed to the expertise of the departmental examination committee leading to proper construction of exams' items. This meticulous scrutiny and careful review of items' distractors ensured unambiguity of the correct answers and generated plausible effective distractors.

Results from this study raise the awareness of the importance of item analysis after exam administration as an essential objective step to ensure quality of multiple choice questions. To improve items'

distractor efficiency, exam constructors should perform some distractors' modification following guidelines for writing effective distractors.

Further analysis is necessary to assess the relationship between the number of functioning distractors per item and item psychometric characteristics (i.e. difficulty and discrimination indices) to provide accurate guidelines to exam constructors. In addition, we can utilize the results of distractor analysis to determine the optimal recommended number of MCQ options required for effective students' assessment.

Acknowledgements:

The researchers acknowledge the support provided by the academic and administrative staff of the Physiology Department and the Educational Development Center of the Faculty of Medicine, Khartoum University for their great help and support in data acquisition. Special thanks to the Secretary of Dean's Office, Awad Gabir who performed the task of scoring the Examinees' answer sheets using the Remark Software.

Conflict of Interest:

The authors declare that they have no conflict of interest.

References:

1. Mehta G, Mokhasi V. Item analysis of Multiple Choice Questions-An assessment of the assessment tool. *Int J Health Sci Res*. 2014;4:197-202.
2. DiBattista D, Kurzawa L. Examination of the Quality of Multiple-Choice Items on Classroom Tests. *Canadian Journal for the Scholarship of Teaching and Learning*. 2011;2:4.
3. Epstein RM. Assessment in medical education. *New England Journal of Medicine*. 2007;356:387-96.
4. Thompson B, Levitov JE. Using Microcomputers to Score and Evaluate Items. *Collegiate Microcomputer*. 1985;3:163-68.*Co*
5. McCowan RJ, McCowan SC. Item Analysis for Criterion-Referenced Tests. Online Submission. 1999.
6. Eaves S, Erford B. Item Analysis. Education.com, Inc.; [updated Dec 23, 2009; cited 2017 5th December]; Available from: http://www.education.com/.
7. Matlock-Hetzel S. Basic Concepts in Item and Test Analysis. 1997.
8. McCoubrie P. Improving the fairness of multiple-choice questions: a literature review. *Medical teacher*. 2004;26:709-12. Epub 2005/03/15.
9. Tavakol M, Dennick R. Post-examination analysis of objective tests. 2011;33:447-58. *Medical teacher*
10. Downing SM. Twelve steps for effective test development. Handbook of test development. 2006:3-25.
11. Tucker S. Using remark statistics for test reliability and item analysis. 2014.
12. McAlpine M. A summary of methods of item analysis. *CAA Centre, Luton.* 2002:b21.
13. Haladyna TM, Downing SM. How many options is enough for a multiple-choice test item? *Educational and Psychological Measurement*. 1993;53:999-1010.
14. Hingorjo MR, Jaleel F. Analysis of one-best MCQs: the difficulty index, discrimination index and distractor efficiency. *JPMA-Journal of the Pakistan Medical Association*. 2012;62:142.
15. Gajjar S, Sharma R, Kumar P, Rana M. 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* : *official publication of Indian Association of Preventive & Social Medicine*. 2014;39:17-20. Epub 2014/04/04.

16. Patil VC, Patil HV. Item analysis of medicine multiple choice questions (MCQs) for under graduate (3 rd year MBBS) students. *Res J Pharm Biol Chem Sc*i. 2015;6:1242-51.

17. Halikar SS, Godbole V, Chaudhari S. Item Analysis to Assess Quality of MCQs. *Indian Journal of Applied Research*. 2016;6.

18. Kaur M, Singla S, Mahajan R. Item analysis of in use multiple choice questions in pharmacology. *International Journal of Applied and Basic Medical Research* . 2016;6:170.

19. Mehta M, Banode S, Adwal S. Analysis of multiple choice questions (MCQ): important part of assessment of medical students. *International Journal of Medical Research and Review*. 2016;4.

20. Patil R, Palve SB, Vell K, Boratne AV. Evaluation of multiple choice questions by item analysis in a medical college at Pondicherry, India. *International Journal of Community Medicine and Public Health*..2016;3:1612-6.

21. Elfaki OA, Bahamdan KA, Al-Humayed S. Evaluating the quality of multiple-choice questions used for final exams at the Department of Internal Medicine, College of Medicine, King Khalid University. *Sudan Medical Monitor*. 2015;10:123.

22. Tarrant M, Ware J, Mohammed AM. An assessment of functioning and non-functioning distractors in multiple-choice questions: a descriptive analysis. *BMC medical education.* 2009;9:1.

23. Christian DS, Prajapati AC, Rana BM, Dave VR. Evaluation of multiple choice questions using item analysis tool: a study from a medical institute of Ahmedabad, Gujarat. *International Journal of Community Medicine and Public Health*. 2017;4:1876-81

24. Haladyna TM. Developing and validating multiple-choice test items: *Routledge*; 2012.

25. McDonald ME. The nurse educators guide to assessing learning outcomes: Jones & Bartlett Learning; 2017.