ORIGINALARTICLE

ITEM ANALYSIS, RELIABILITY STATISTICS AND STANDARD ERROR OF MEASUREMENT TO IMPROVE THE QUALITY AND IMPACT OF MULTIPLE CHOICE QUESTIONS IN UNDERGRADUATE MEDICAL EDUCATION IN FACULTY OF MEDICINE AT UNISZA

[1]Shahid Hassan, [2]Rahmah Mohd Amin, [3]Husbani bt. Mohd Amin Rebuan, [4]Myat Moe Thwe Aung

[1]Unit of Medical Education, [2,4]Unit of Community Medicine, [3]Unit of Radiology, Faculty of Medicine, Universiti Sultan ZainalAbidin Malaysia.

ABSTRACT

Multiple-choice question as one best answer (OBA) is considered as a more effective tool to test higher order thinking for its reliability and validity compared to objective test (multiple true and false) items. However, to determine quality of OBA questions it needs item analysis for difficulty index (PI) and discrimination index (DI) as well as distractor efficiency (DE) with functional distractor (FD) and non-functional distractor (NFD). However, any flaw in item structuring should not be allowed to affect students' performance due to the error of measurement. Standard error of measurement (SEM) to calculate a band of score can be utilized to reduce the impact of error in assessment. Present study evaluates the quality of 30 items OBA administered in professional II examination to apply the corrective measures and produce quality items for the question bank. The mean (SD) of 30 items OBA = 61.11 (7.495) and the reliability (internal consistency) as Cronbach's alpha = 0.447. Out of 30 OBA items 11(36.66%) with PI = 0.31-0.60 and 12 items (40.00%) with DI = ≥0.19 were placed in category to retain item in question bank, 6 items (20.00%) in category to revise items with DI ≤0.19 and remaining 12 items (40.00%) in category to discard items for either with a poor or with negative DI. Out of a total 120 distractors, the non-functional distractors (NFD) were 63 (52.5%) and functional distracters were 57 (47.5%). 28 items (93.33%) were found to contain 1- 4 NFD and only 2 (6.66%) items were without any NFD. Distracter efficiency (DE) result of 28 items with NDF and only 2 items without NDF showed 7 items each with 1 NFD (75% DE) and 4 NFD (0% DE), 10 items with 2 NFD (50% DE) and 4 items with 3 NFD (25% DE). Standard error of measurement (SEM) calculated for OBA has been ± 5.51 and considering the borderline cut-off point set at ≥45%, a band score within 1 SD (68%) is generated for OBA. The high frequency of difficult or easy items and moderate to poor discrimination suggest the need of items corrective measure. Increased number of NFD and low DE in this study indicates difficulty of teaching faculty in developing plausible distractors for OBA question. Standard error of measurement (SEM) should be utilized to calculate a band of score to make logical decision on pass or fail of borderline students.

Keywords: *MCQ, difficulty index, discrimination index, distraction efficiency, functional and non-functional distractors, reliability coefficient, standard error of measurement*

INTRODUCTION

Multiple-choice question as one best answer (OBA) is considered more effective tool to test higher order thinking for its reliability and validity compared to objective test (multiple true/fast) items. An OBA item consists of clinical scenario or problem followed by a clearly written lead-in or question and a list of multiple options with three to four distractors and one correct answer[1]. All distractors need to be relatively correct towards the right answer. Although less susceptible to guessing, OBA requires good number of plausible distractors to achieve reliability[2]. Item analysis allows measurement of effectiveness of individual test items[3]. Apart from psychometric evaluation of assessment it is important to perform item analysis to improve quality of items[4] by analysing difficulty index (DIF I), discrimination index (DI) and distractor efficiency (DE) based on number of non-functional distractors (NFD)[5]. To structure good OBA however, require continuing faculty development and plausible distractors. A psychometric test of OBA items is though important to perform, item analysis and its interpretation is essentially needed to provide valuable feedback to faculty who writes OBA items. The process ensures to develop question bank with quality OBA items and it is widely accepted that well-structured OBA items are time consuming and difficult to write[6].

Good distractors are those with relatively correct and close to key of an item. Plausible distractors are functional and is defined as the distractors selected by >5% of examinees. Non-functional distractors (NFD) are the options selected by <5% of examinees. To identify and replace NFD by functional distractors (FD) require investigation by item analysis. Item analysis determines difficulty index (p-value), discrimination index (DI) and distractor efficiency (DE)[7]. Item analysis allows measurement of effectiveness of individual question. Item analysis typically rely on classical test theory with two major statistics based on difficulty and discrimination indices based on students' score. However, it require good sample score to draw conclusion. Difficulty index or facility index are the items correctly picked up by both, upper and the lower performing group of students. It is calculated by adding the correctly answered items by upper 27% and lower 27% of students' performance[8] divided by total number of students in both the groups (see table 1).

Item difficulty can range from 0.0 or 0% to 1.0 or 100% (all the students answered the item correctly). The recommended average level of difficulty for four options OBA should range between 31%-60% (0.31-0.60)[9].Whereas discrimination index suggests the difference between the percentage of high achieving students who got the answer correct and percentage of low achieving students who got the answer correct. It is obtained by deducting the correctly responded items in upper group from the correctly responded students in lower group divided by number of students in one group[10]. Item discrimination index reflects the degree of relationship between scores on the item. It ranges from 0 to +1, depending on how students in upper group answer the item correctly. However, it may be negative (-1) when lower achievers answer the item correctly. Positive value is desirable. A discrimination index of 0.15-.0.25 is considered to be desirable.

A method to structure structured good items in OBA is to look at the number of functional vs. non-functional distractors. It has been author's experience that three options OBA are feasible to design OBA items particularly for those taking a new start to write. The theoretically calculated guessing effects with three, four and five options SBA have been 33%, 25% and 20% respectively. Higher the number of options, lesser the functional distractors are. Often the implausible distractors are the reason to produce higher number of NFDs. It is better to have less options but more functional distractors than more options and more NFD. However, to produce quality OBA items by any institution needs training, experience and effective vetting sessions by the experts. A poorly structured OBA test should not has its impact on students performance. This can be ensured by calculating the standard error of measurement (SEM) in each examination to generate a band score within 1 standard deviation (68%) in OBA items. This score can be used to adjust the borderline students either by a logical decision of straight pass for those who achieve 50% and above by addition of score calculated by SEM or alternatively considering the triangulation process based on their performance in continuous assessment. The objective of present study is to evaluats quality of 30 items OBA administered in professional II examination held in 2014 to determine the items PI, DI and DE and to apply the corrective measures to produce quality items for the question bank.

METHOD

An evaluation of post summative assessment (Professional II Examination) of OBA items was performed of MBBS program of 2015 in Faculty of Medicine at UniSZA. 30 OBA items and 120 distractors (4 options items) inclusive of key were assessed using item analysis. 30 students' score of OBA underwent descriptive statistics to determine mean and standard deviation (SD). Item analysis was peroformed using the formula for difficulty index (PI), discrimination index (DI) and distracter efficiency (DE) employing MS Excel 2007 (see table1). Reliability for internal consistency of 30 items OBA was calculated using by Cronbach's alpha. Content validity however, was taken care by the experts during the vetting session that follows the examination question blueprinting in this institution. Standard error of measurement (SEM) was also determine using the formula SD x √ (1-r) to calculate a band score within 1 SD (68%) to adjust borderline cut-off point set at ≥45% for OBA.

RESULT

The descriptive statistics of 30 items OBA included, mean (SD) = 61.11 (7.495) and the reliability (internal consistency) as Cronbach's alpha = 0.447. Item analysis suggest (see table 2), 1 (3.33%) item as excellent with PI (0.41-0.60) and DI (≥.30), 7 (23.33%) items as good to excellent and 11(36.66%) items as good with PI (0.30 – 0.60) and DI (≥0.15). However, another 11(36.66%) items inclusive of 3 (10.00%) items with negative DI were having PI = ≤0.30 or ≥0.61 and DI = ≤0.15 (see table 1).Out of 30 OBA items 12 (40.00%) with DI = ≥0.19 were placed in category to be retained, 6 (20.00%) in category to be revised and another 12 items (40%) in category to be discarded with DI = ≤0.19 (see table 2 and 3).

Out of a total 120 distractors, the non-functional distractors (NFD) were 63 (52.5%) and functional distracters were 57 (47.5%). 28 items (93.33%) were found to contain 1- 4 NFD and only 2 (6.66%) items were without any NFD (see table 3). Items with 1-4 distractors varied in number and the maximum numbers of items 10 (33.33%) were the one with 2 distractors and 7 (23.33) each were the items with 1 or 4 distracters. Out of remaining 6 items 4 contained 3 (10.00%) distractors and 2 items were without any NFD (see table 4).Distractors with choice frequency = 0 were 7 (23.33%) (see table 3).

Distracter efficiency (DE) result of 28 items with NDF and only 2 items without NDF showed 7 items each with 1 NFD (75% DE) and 4 NFD (0% DE), 10 items with 2 NFD (50% DE) and 4 items with 3 NFD (25% DE) in this study (see table 3). Total Item with NFD across PI were 28(62 NFD) and across DI were 27 (64 NFD). These items with their NFD were varyingly distributed among the different range of PI and DI (seetable 5). SEM calculated for OBA has been ± 5.51 and considering the borderline cut-off point set at

≥45%, a band score within 1 SD (68%) is generated for OBA (see table 6).

Table 1: Item analysis formula to calculate difficulty index (PI) and the discrimination index (DI), functional distractor (FD), non-functional distractor (NFD), distractors efficiency (DE)

Item Analysis	Formula to calculate difficulty and discrimination indices and DE
Difficulty Index (PI)	PI = No. of students in upper group + lower group with correct answer / Total number of students in both groups
Discrimination Index (DI)	DI = No. of students in upper group - lower group with correct answers / Total number of students in one group.
Functional (FD)/ non functional (NFD) Distractor	Functional distractor (FN) and NFD are the distractors determined by >5% and <5% of examinees selecting a distractor out of option list respectively.
Distractor Efficiency (DE)	Determined by number of NFD in an item and it ranges from 0 to 100% having 4, 3, 2, 1 or nil NDF.4 NDF = 0 DE, 3 NDF = 25% DE, 2 NFD = 50% DE, 1NDF = 75% DE and 0 NDF = 100% DE.

DISCUSSION

OBA is an effective instrument to measure the students' analytic reasoning skills and in-depth performance in outcome based education practiced in an integrated curriculum[11]. However, qualities of OBA items depend upon faculty development to write good items that discriminate students with higher and poor abilities[12]. Items with more NFD are implausible and of little valueand that DE is determined by number of NFDs present in an item and it ranges from 0%-100%[13]. Selection or rejection of items for question bank is best guided by DE.

Overall there was low mean score and larger SD may partly be attributed to structuring of quality questions by the faculty in OBA test in assessment. Reliability coefficient as internal consistency was analyzed using Cronbach's alpha for OBA was = 0.449, which is extremely low for alpha statistics and needs to improve in subsequent assessment to produce quality OBA in future. However, a low alpha may also be attributed to small sample size of merely 3o students taking the examination in the 1st batch of MBBS program.

The mean PI is not in desirable range of 0.31-0.60 or comparable with other study[14]. PI not comparable with other study is in-fact due to non-functional distractors, making it feasible for both, upper and lower performers to respond. Overall DI was below the excellent power of discrimination except in one item (see table 2) and is attributed to unexpected number of NFD in many items. The number of NFD was found high in present study and it had its impact on DIF I, DI and DE. Overall 12 (40.00%) items were considered unacceptable for professional examination. 3 (10.00%) items with negative DI in OBA paper were not considered for revision after carefully reviewed to ensure that a wrong key was not the case. Most of the OBA items were comparatively easy, shown by the presence of >NFD. Such assessment has no motivation for low performers. DI was also slightly below the desirable power of discrimination again due to >NFD, some with 3-4 NFD Based on PI and DI items were categorized into poor, good and excellent to decide to retain, revise or discard item and develop a pool of valid items for future use.

The total number of non-functional distractors (NFD) were 63 (52.5%) out of a total number of 120 distractors. 28 (93.33%) were found to contain 1- 4 NFD and only 2 (6.66%) items were without any NFD. Items with 1-4 distractors varied in number and the maximum numbers of items 10 (33.33%) were the one with 2 distractors and only 2 items were without any NFD.Distractors with choice frequency = 0 were 7 (23.33%) both in upper and lower achievers (see table 3). In a corrective measure all these non-functional distractors need to be replaced with more plausible distractors while reviewing the items by experts. A misconception about the distractors exist is that more the distractors better will be the OBA item. High numbers of distractors do not determine the quality of OBA items. Research has provided the evidence that none of the five options had four functioning distractors[15] and it is not an easy for those structuring OBA to develop 4 equally plausible distractors. On the contrary it has been established that items with 2 plausible distractors are better than items with three or four implausible distractors[16, 17]. The argument for choosing the number of distractors for single best answer MCQ has often been in favour of having more options to minimise guessing effect. This however, has been researched and found that three options are optimal for MCQs in most setting[18]. Research has also established that the psychometric properties such as reliability and validity of a test are not affected if the number of options is reduced to three distractors[19].

Table 2: Item analysis of OBA Items in professional 2, 2014 examination to evaluate the items in terms of difficulty and discriminating indices and its outcome

OBA Item No.	Difficulty Index (PI)	Discrimination Index (DI)	Distractor Efficiency (DE)	Recommendation for disposal of items		
				Retain	Revise	Discard
1	0.37	0.17	75%		√	
2	0.23	0.32	50%	√		
3	0.87	0.02	25%			√
4	0.67	0.26	75%	√		
5	0.27	0.07	75%			√
6	0.13	0.22	100%	√		
7	0.63	0.32	75%	√		
8	0.40	0.55	75%	√		
9	0.60	-0.02	50%			√
10	0.83	0.38	50%	√		
11	0.70	0.26	50%	√		
12	0.93	0.15	0%		√	
13	0.47	0.24	50%	√		
14	0.73	0.36	50%	√		
15	0.73	0.03	50%			√
16	0.60	0.16	50%		√	
17	0.93	0.15	25%		√	
18	0.33	0.17	75%		√	
19	0.97	0.08	0%			√
20	0.97	0.00	0%			√
21	0.37	0.27	50%	√		
22	0.93	0.15	25%		√	
23	0.07	-0.08	100%			√
24	0.10	0.02	50%			√
25	0.50	0.24	75%	√		
26	0.97	0.08	0%			√
27	0.97	0.00	0%			√
28	1.00	0.00	0%			√
29	0.57	-0.19	0%			√
30	0.50	0.29	25%	√		
	Number of items retained, revised or discarded.			12	6	12

Table 3: Distribution of items in relation to difficulty index and discrimination index and action recommended retaining, revising or discarding the items.

Parameter (Range)	Interpretation	Items (N=30)	Action
Difficulty Index			
≤0.30	Difficult	5 (16.66%)	Revise to retain or discard
0.31-0.40	Good	4 (13.33%)	Retain in question pool
0.41-0.60	Excellent	6 (20.0%)	Retain in question pool
≥0.61	Easy	15 (50.0%)	Revise to retain or discard
Discrimination Index			
<0.19	Poor	18 (60.66%)	Revise to retain or discard
0.19-0.29	Marginal	7 (23.33%)	Revise to improve
0.30-0.39	Good	4 (13.33%)	Retain in question pool
≥0.40	Excellent	1 (3.33%)	Retain in question pool

Table 4: Distractors established as FD, NFD and DE in a 4 options response of one correct answer and 3 distractors in OBA items.

Item/Distractor	No. of Items	No. of NFD	Total Items with NFD (DE)
Total items	30	Items with 1 NFD	7 / (75%)
Total distractors	120	Items with 2 NFD	10 / (50%)
Total Functional Distractors (FDs)	57 (47.5%)	Items with 3 NFD	4 / (25%)
Total Nonfunctional Distractors (NFDs)	63 (52.5%)	Items with 4 NFD	7 / (0%)
Items with no NFD	2 (6.66%)	Items with 0 NFD	2 / (100%)
Distracter with choice frequency = 0	7 (23.33%)	Total items with NFD	28

Table 5: Items with non-functional distractors (NFDs) and their relationship with PI and DI in 30 items OBA.

Difficulty Index (PI)	Items (NFDs) Across PI	Discrimination Index (DI)	Items (NFDs) Across DI
≤0.30	3 (5), 2 (0)	<0.15	10 (31), 1 (0)
0.31-.040	4 (5)	0.15-0.29	12 (25), 2 (0)
0.41-0.60	6 (9)	≥0.30	5 (8)
≥.61	15 (43)	-	-
Total Item (NFD) across PI	28 (62)	Total Items (NDF) across DI	27 (64)

Determined by number of NFD in an item and it ranges from 0 to 100% having 4, 3, 2, 1 or nil NDF.4 NDF = 0 DE, 3 NDF = 25% DE, 2 NFD = 50% DE, 1NDF = 75% DE and 0 NDF = 00% DE. DE is reported as 0%, 75%, 50%, 25% and 100% depending on items containing 4, 3, 2, 1 and 0 NFD respectively. DE was also lower than reported for OBA items in literature and it is determined by number of NFDs present in an item. Items with –ive or 0 DE were recommended to be discarded and such were 9 (30.00%) out of 30 items. It has been noted that easier the item more are the NFD. Out of a total 63 NFD 39 NFD were reported with PI ≥0.61 (14 items) leading to low DE.Similarly lower the DI more are the NFD and out of 63 NFD 31 NFD were with DI <0.15 (10 items) again leading to low DE.DE was widely varied among the items performing in various range of PI and DI (see table 4).. Difficult the item more is the DE or less are the NFD. Higher the power of discrimination more is DE or less are the NFD.

Table 6: Standard error of measurement calculated as an evidence to consider adjustments of borderline score for a logical decision of straight pass or triangulation.

Formula to calculate standard error of measurement (SEM)	SD x √ (1-r)
Defining borderline	45% to 49.5%
Calculating SEM	7.495 x √ (1- 0.449) 7.495 x √ (0.551) 7.495 x 0.742 = 5.561
Borderline +/- SEM (1 SD)	45% +/- SEM = 45 +/- 5.56 1 SD (68%) = 50.56 to 40.44
Standard setting for logical decision	Readjusted SEM score provides evidence for borderline students to consider straight pass or alternatively a triangulation process to relook into student’s performance in continuous or formative assessment to decide on pass or fail.

Standard error of measurement (SEMs) calculated have been ± 5.51 for OBA. Considering the borderline cut-off point set at ≥45%, a band score within 1 SD (68%) is generated for OBA. It is observed that students within one ± SD when allowed SEM to combine with cut-off point (i.e., 45+- 5.551) produced a band of true score (50.56 to 40.44), which is well beyond passing marks. This justifies allowances for borderline candidate, either another assessment (borderline viva/written test) or alternatively considering a triangulation process that justifies deciding on pass or fail.

CONCLUSION

The high frequency of difficult or easy items and moderate to poor discrimination suggest the need of items corrective measure. Increased number of NFD and low DE in this study indicates difficulty of teaching faculty in developing plausible distractors for OBA question. Item analysis has been a valuable step to identify OBA items for its PI, DI and DE. Writing a quality OBA item primarily needs training and experience, complimented with just-in-time evaluation and its interpretation to teaching faculty to produce expert assessors. The high frequencies of difficult or easy and below expectation discriminating items in present study suggest continuing corrective measure to improve the quality of OBA items. Increased number of non-functional distractors has been due to difficulty of teaching faculty to produce plausible distractors. Standard error of measurement (SEM) should be utilized to calculate a band of score to handle borderline students with care. Any flaw in item structuring should not be allowed to affect students' performance due the error of measurement.

REFERENCES

1 Farley JK: The multiple-choice test: writing the questions. Nurse Educ 1989, 14(6): 10-12.

2 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 (40): 2-8.

3 McCoubrie P: Improving the fairness of multiple-choice questions: a literature review. Med Teach 2004, 26(8): 709-712.

4 Haladyna TM, Downing SM: Validity of taxonomy of MCQ - writing rules. ApplMeasEduc 1989, 2(1): 51-78.

5 Haladyna TM: Developing and validating multiple choice test items. Lawrence Eribaum Associate, New Jersey; 1999.

6 Farley JK: The multiple-choice test: writing the questions. Nurse Educ 1989, 14(6): 10-12.

7 Scranton Guides – Item Analysis, adapted from Michigan State University website and Barbara gross devils tools for teaching. [Last cited on 2013 Apr 13]. Available from: www.freepdfb.com/pdf/item-analysis-scranton

8 Linn R. L., Miller M. D. Measurement and assessment in teaching. Pearson Education, Inc., Upper Saddle River, New Jersey, 2005, 9th edition; 348-365

9 Tarrant M, Ware J, Mohammed AM. An assessment of functioning and non functioning distractors in multiple choice assessment: A descriptive analysis. BMC Med Educ. 2009; 9:1-8. (PMC free article) (PubMed).

10 Ebel RL, Frisbie DA. Essentials of educational measurements. 5th edition. Englewood Clipps, N.J: Prentice Hall; 1991.

11 Polit DF, Hunglern BP: Nursing research: Principles and methods. Lippincort, Williams and Wikkins, Philadelphia; 1999.

12 Linn R. L., Miller M. D. Measurement and assessment in teaching. Pearson Education, Inc., Upper Saddle River, New Jersey, 2005, 9th edition; 348-365.

13 Matlock-Hetzel S. Presented at annual meeting of the Southwest Educational Research Association, Austin, January; 1997. [Last cited on 2013 Apr 13]. Basic concept in item and test analysis. Available from: www.ericae.net/ft/tamu/espy.htm

14 Guilbert JJ. 1st ed. Geneva: World Health Organization; 1981. Educational Hand Book for health professionals. WHO offset Publication 35.

15 Haladyna TM, Downing SM: How many options is enough for a multiple-choice test item? EducPsycholMeas 1993, 54(4): 999-1010

16 Schuwirth LWT, Vlueten CPM van der: Different written assessment methods: what can be said about their strengths and weaknesses? Med Educ 2004, 38(9): 974-9-79.

17 Crehan KD, Haladyna TM, Brewer BW: Use of an inclusive option and the optimal number of options for multiple-

choice items. EducPsycholMeas 1993, 53(1): 241-247.

18 Rodriguez MC: Three options are optimal for multiple-choice items: A meta-analysis of 80 years in research. EducMeas Issues Pract 2005, 24(2): 3-13.

19 Trevisan MS, Sax G, Micheal WB: The effects of the number of options per item and student ability on test validity and reliability. EducPsycholMeas 1991, 51(4): 829-837