ORIGINAL ARTICLE
Volume 9 Issue 3 2017
DOI: 10.21315/eimj2017.9.3.4
ARTICLE INFO
Submitted: 26-07-2017
Accepted: 29-07-2017
Online: 30-09-2017

Use of Item Analysis to Improve the Quality of Single Best Answer Multiple Choice Question in Summative Assessment of Undergraduate Medical Students in Malaysia

Shahid Hassan[1], Rafidah Hod[2]

[1]*Centre for Education (ICE), Faculty of Medicine and Health Sciences, International Medical University, Kuala Lumpur, MALAYSIA*

[2]*Faculty of Medicine and Health Sciences, Universiti Putra Malaysia, MALAYSIA*

To cite this article: Hassan S, Hod R. 2017. Use of item analysis to improve the quality of single best answer multiple choice question in summative assessment of undergraduate medical students in Malaysia. Education in Medicine Journal. 2017;9(3):33–43. https://doi.org/10.21315/eimj2017.9.3.4

To link to this article: https://doi.org/10.21315/eimj2017.9.3.4

ABSTRACT

Background: Single best answer (SBA) as multiple-choice items are often advantageous to use for its reliability and validity. However, SBA requires good number of plausible distractors to achieve reliability. Apart from psychometric evaluation of assessment it is important to perform item analysis to improve quality of items by analysing difficulty index (DIF I), discrimination index (DI) and distractor efficiency (DE) based on number of non-functional distractors (NFD). **Objective:** To evaluate quality of SBA items administered in professional examination to apply corrective measures determined by DIF I, DI and DE using students' assessment score. **Method:** An evaluation of post summative assessment (professional examination) of SBA items as part of psychometric assessment is performed after 86 weeks of teaching in preclinical phase of MD program. Forty SBA items and 160 distractors inclusive of key were assessed using item analysis. Hundred and thirty six students' score of SBA was analysed for mean and standard deviation, DIF I, DI and DE using MS Excel 2007. Unpaired *t*-test was applied to determine DE in relation to DIF I and DI with level of significance. Item-total correlation (*r*) and internal consistency by Cronbach's alpha and parallel-form method was also computed. **Result:** Fifteen items had DIF I = 0.31–0.61 and 25 items had DIF I (≤ 0.30 or ≥ 0.61). Twenty six items had DI = 0.15 – ≥ 0.25 compared to 14 items with DI (≤ 0.15). There were 26 (65%) items with 1–3 NFD and 14 (35%) items without any NFD. Thirty nine (32.50%) distractors were with choice frequency = 0. Overall mean DE was 65.8% and NFD was 49 (40.5%). DE in relation to DIF I and DI were statistically significant with $p = 0.010$ and 0.020 respectively. Item-total correlation for most items was < 0.3. Internal consistency by Cronbach's alpha in SBA Test 1 and 2 was 0.51 and 0.41 respectively and constancy by parallel-form method was 0.57 between SBA Test 1 and 2. **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 SBA questions. This has been reflected in poor reliability established by alpha. Item analysis result emphasises the need of evaluation to provide feedback and to improve quality of SBA items in assessment.

Keywords: *Single best items, Difficulty index, Discrimination index, Distractors efficiency, Functional distractors, Non-functional distractors*

CORRESPONDING AUTHOR Shahid Hassan, Centre for Education (ICE), Faculty of Medicine and Health Sciences, International Medical University, 126 Jln Jalil Perkasa 19, Bukit Jalil, 57000 Bukit Jalil, Wilayah Persekutuan Kuala Lumpur, Malaysia | Email: shahidhassan@imu.edu.my

INTRODUCTION

Multiple choice questions (MCQs) can be designed to test simple recall such as factual knowledge to high order thinking such as appropriate analysis and evaluation. MCQs can consistently test student knowledge with high degree of reliability and validity. Single best answer (SBA) multiple-choice questions are often advantageous to use for its reliability and validity however, they are not the best test items for every circumstances. It is appropriate to use SBA when the attainment of educational objective is measured by having students' response as select answer from the list of alternatives. On validity aspect SBA covers broader sample of course content in given amount of testing time however, with reliant on experts judgement (1). SBA items ensure great variety of educational objectives adaptable to various levels of learning outcome. On reliability aspect, more objective (2) and less susceptible to guessing than multiple true and false items, SBA format of MCQ can be considered to have high degree of reproducibility. Although less susceptible to guessing SBA require good number of plausible distractors to achieve reliability.

Good SBA is generally more difficult to write than other types of MCQ items. Continuing faculty development to construct quality SBA with plausible distractors also requires skills and experience. An approach to write quality SBA begins with teaching faculty vigilantly involved to document case scenarios as problem, regularly attend vetting of exam questions and participate in expeditious evaluation of test result after each examination. An evaluation and feedback on SBA items is as important as writing a SBA question. Apart from psychometrics of SBA items it is important to perform item analysis and to interpret result to provide valuable feedback to faculty who writes SBA questions. The process though tedious, ensures to develop question bank with quality SBA items. It is widely accepted that well-constructed MCQ items are time consuming and difficult to write (3).

A SBA item consists of clinical scenario or problem followed by a clearly written lead-in or question and a list of multiple options with 3–4 distractors and one correct answer (4, 5). All distractors need to be relatively correct and close to key of an item. Producing plausible distractors, which are functional and best, defined as the distractors selected by > 5% of examinees is considered important of quality SBA items. Non-functional distractors (NFD) are the options selected by < 5% of examinees. To identify NFD and to replace those with functional distractors (FD) require investigation by item analysis and subsequent action to retain, revise or discard the items. Item analysis determines difficulty index (DIF I) (*p*-value), discrimination index (DI) and distractor efficiency (DE) (6). Item analysis allows measurement of effectiveness of individual test items. Item analysis typically utilises classical test theory with two major statistics based on difficulty and discrimination to determine item quality based on students' score. However, the conclusions drawn depend very much on the sample score used to collect information.

DIF I, actually determines as easy or facility index with which items are 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 (7) divided by total number of students in both the groups. DIF I expressed in percentage or decimals. Item difficulty can range from 0.0 or 0% (none of the students answered the item correctly) to 1.0 or 100% (all the students answered the item correctly). The recommended average level of difficulty for four options SBA should range between 31%–60% (0.31–0.60) (8). DI on the other hand reflects 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 (9). Item DI is the point-biserial correlation that reflects the degree of relationship between scores on the item. It ranges from 0 to +1, if more 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. An item with a difficulty of 0 or 1 will always have DI of 0 and DI is maximised when DIF I is close to 0.50. A DI of 0.15–0.25 is considered to be desirable.

A low DIF I (very difficult item) and a low DI (poorly discriminating item) ultimately lead to apply the corrective measures to readjust and administer the item again. An attempt to restructure the items may show improvement of difficulty and discrimination indices and regardless to say that the exercise is a good learning experience for faculty development and their learning curve. In order to keep these items in question bank, it definitely requires to be restructured by sending it back to experts to achieve desired difficulty and discrimination indices. In multiple items instrument a varying range of low to high difficulty and discrimination indices are acceptable provided it is not on the extreme ends. A very easy difficulty index with poor discrimination is often the case when items test the procedural skills such as in Objective Structured Practical Examination (OSPE) or Objective Structured Clinical Examination (OSCE) questions that students have mastered the practical or clinical skills contained in those tasks. Therefore to retain, revise or reject the items require a number of aspects to be considered for logically sound decision. If properly constructed SBA are able to test high level of cognitive reasoning and can accurately discriminate between high and low achieving students (10).

A good method to decide on a well-structured vs. poorly structured items in SBA is to look at the number of functional vs. non-functional distractors. It may not be practical to produce SBA items without zero frequency choice distractors. However, items with functional distractors should at least be selected by > 5% of examinees. It has been author's experience that three options SBA are feasible for faculty members to design SBA items particularly for those taking a new start to write SBA question. However, the argument that guessing effect is more common with three than with four or five options may compel institution to adopt four or even five options SBA. The theoretically calculated guessing effects with three, four and five options SBA have been 33%, 25% and 20% respectively.

Studies on different options and guessing effect have shown no difference between 3–5 options (11). Faculty should focus more on evaluation and its interpretation of SBA items in terms of DIF I, DI, DE and NFD rather than falling in controversies of choices of options and its guessing effects that might compromise SBA reliability of items. A practical observation has been that the three options SBA is feasible and less time consuming to produce items with more functional distractors. It has been observed that higher the number of options, lesser the functional distractors are. Often the implausible distractors are the reason to produce higher number of NFDs. Adopting to restricted options list can be offset by increasing the number of SBA items to increase the reliability of this assessment format. It is better to have less options but more functional distractors than more options and more NFDs.

METHODS

Present item analysis of SBA is part of regularly held post examination evaluation of Medical Degree (MD) program assessed by two professional (summative) examinations held at the end of preclinical and clinical phase in a medical school in Malaysia. The current evaluation comprised of 40 items multiple choice SBA administered to 136 students in their preclinical phase of an integrated curriculum of basic and clinical sciences content. SBA test items were developed

based on a blueprint of examination questions with predefined weighting across the modules and learning outcome guided by Bloom's taxonomy. SBA as an instrument has been recently included in written tests of summative examination after having practiced this instrument in faculty development workshops to write quality SBA items. Regular evaluation and feedback to faculty has been a feature of SBA items added to multiple true and false items in equal numbers in 1st professional examination. SBA test has been administered as two different tests (Test 1 and Test 2) in succession covering nine modules taught in two years. The item analysis follows a psychometric evaluation to monitor reliability and validity of instruments.

All SBA items have a stem or problem followed by a clearly written lead-in and four options with single correct answer (key) and three distractors. A vetting of SBA items in the presence of experts, module and phase coordinators, medical educationist and senior faculty members is meticulously done to select plausible distractors with varying degree of correctness. Item analysis is performed immediately after the examination and presented before the examination body inclusive of external examiners on board.

Score of 136 students' was entered in MS Excel 2007 and reorganized in descending order. The students' score of upper 27% (37 students) of high achievers and lower 27% (37 students) of poor achievers were included in analysis. A total of 40 SBA items with 120 options of three distractors and one key were analysed to calculate DIF I, DI, DE and NFD. DIF I and DI were calculated using the following formula.

$$\text{DIF I} = \frac{\text{No. of students in upper group + lower group with correct answer}}{\text{Total number of students in both groups}}$$

$$\text{DI} = \frac{\text{No. of students in upper group + lower group with correct answer}}{\text{Total number of students in one groups}}$$

DIF I is the ease or facility with which a student can answer the item and it is recorded in decimal (0 – +1). Higher the value easier is the item and lower the value difficult is the item. DI is the ability of an item that differentiates between students of higher achievers and low achiever groups. DI is also recorded in decimal (–1 +1) and higher the value, more is the discrimination power of an item.

NFD is the alternative from an option list other than the correct answer and it is determined by < 5% of examinees selecting a distractor and it may not exactly be of 0 frequency choice. Compared to this a functional distractor in an item is the alternative selected by ≥ 5% of examinees. DE for any item is calculated by numbers of NFD contained in it and it ranges from 0%–100%. DE is expressed as 0%, 33.3%, 66.6% and 100% depending on number of NFD as 3, 2, 1 and 0 respectively (9). Items with undesirable DIF I and DI and DE are sent for corrective measures and NFDs are replaced with better plausible distractors by experts involved in rewriting the items.

Pearson correlation coefficient (*r*) was also performed to establish the item-total correlation coefficient of individual items with total score in each test. It ranges between –1.00 to +1.00 and an item having a score < .30 is considered to be unreliable in context of what item purports to test. Internal consistency as reliability of the overall test was determined by Cronbach's alpha. Cronbach's alpha of > 0.6 to 0.7 is considered acceptable, ≥ 0.5 to 0.6 is poor and < 0.5 is unacceptable. Constancy as short-term stability was computed by parallel-form method for 2-tests of SBA measuring the same construct and administered in succession. Key to parallel-form method of reliability is to develop alternate test that is equivalent in content, item format and response process. A total of 40 SBA were administered in succession on

same day as Test 1 and Test 2 with 20 items each. There was no negative mark for wrong answer and 1 mark for each correct answer.

RESULT

Mean (standard deviation) for DIF I, DI and DE were 0.64 (±0.25), 0.24 (±0.18) and 65.8 (±32.4) respectively (see Table 1). Out of 40 SBA 15 items had good to excellent level of difficulty (DIF I = 0.31–0.61) and remaining 25 items were either difficult (DIF I ≤0.30) or easy (DIF I ≥0.61) as established in this item analysis (see Table 2). There are 26 (65%) items had good to excellent power of discrimination (DI = 0.15–≥0.25) compared to 14 items with poor (≤0.15) DI (see Table 2). The 14 items viewed together with excellent DIF I (0.31–0.60) and DI (≥0.25) and 14 (35%) were perfectly well structured and recommended to be included in question bank. No item was found with negative DI in this evaluation.

Distractors with choice frequency = 0 were 24 (40%) in upper achievers and 15 (25%) in lower achievers and in all there were 39 (32.5%) distractors with choice frequency = 0 in both upper and lower achievers (see Table 3). Functional distractors (FD) were 71 (59.1%) in overall 40 items SBA (see Table 4). The total number of NFD was 49 (40.8%) and out of this 26 (65%) items were with 1–3 NFD and 14 (35%) items without any NFD. Items with 1–3 distractors varied in number and the maximum numbers of items 14 (35%) were the one with 2 distractors (see Table 4).

DE was widely varied between 0%, 33.3%, 66.6% and 100% with 3, 2, 1 and 0 NFD respectively. There are 27 items with NFD were found to have mean DIFI = 0.73 and mean DI = 0.19 (see Table 5). Remaining 14 items were without NFD had a mean DIF I = 0.41 and DI = 0.16. Distractors efficiency viewed in relation to difficulty level of items showed mean DE high (91.65) in 4 difficult items than mean DE (49.16) in 21 easy items (see Table 6). Similar was the case with discrimination index, which showed a high mean DE (72.70%) with good to excellent DI in 22 items compared to mean DE (47.45%) in 14 items with poor DI. These differences in DE in both cases were statistically significant (see Table 6).

Item-total correlation (*r*) as internal consistency of individual items with total score ranged from –0.001 to 0.100 in SBA Test 1 and 0.164 to 4.00 in SBA Test 2. Items with r = > 0.30 were 0 items in SBA Test 1 and 6 items in SBA Test 2. Consistency as a whole determined by Cronbach's alpha showed 0.512 for 20 items SBA Test 1 and 0.417 for another 20 items Test 2. Reliability in terms of parallel-form estimate computed for 2-test of SBA administered in succession was 0.572.

Table 1: Descriptive statistics of 4 options response of one correct answer and 3 distractors in 40 SBA items in 1st Professional Examination 2014

Parameter	Test	Mean	SD	Test 1 + Test 2 Mean (SD)
Difficulty Index (DIF I)	Test 1	0.6290	0.2418	0.6400 (0.2560)
	Test 2	0.6510	0.2753	
Discrimination Index (DI)	Test 1	0.2590	0.1878	0.2498 (0.1879)
	Test 2	0.2405	0.1924	
Distractor Efficiency (DE %)	Test 1	63.30	35.70	65.80 (32.46)
	Test 2	68.30	29.57	

Table 2: Distribution of items in relation to difficulty index and discrimination index and action recommended in SBA items in 1st Professional Examination 2014

Parameter (Range)	Interpretation	Items (N = 40)	Action
Difficulty Index			
≤30	Difficult	4 (10%)	Revise/Discard
31–40	Good	5 (12.5%)	Store in Question Bank
41–60	Excellent	10 (25%)	Store in Question Bank
≥61	Easy	21 (52.5%)	Revise/Discard
Discrimination Index			
<0.15	Poor	14 (35%)	Revise/Discard
0.15–0.24	Good	4 (10%)	Store in Question Bank
≥0.25	Excellent	22 (55%)	Store in Question Bank

Table 3: Number of distracters employed in a 4 options response of one correct answer and 3 distracters in SBA questions vs. number of distracters used by upper and lower performers

Upper and lower performers	SBA Test 1		SBA Test 2	
	Number of distracters	Distracters with choice frequency = 0	Number of distracters	Distracters with choice frequency = 0
Upper 27% (N = 37)	60	24 (40%)	60	22 (36.66%)
Lower 27% (N = 37)	60	15 (25%)	60	13 (21.66%)
Total	120	39 (32.50%)	120	35 (29.16%)

Table 4: Distractors analysis in a 4 options response of one correct answer and 3 distractors in SBA questions in 1st Professional Examination 2014

Distractors Analysis	SBA Test 1	SBA Test 2	SBA1 + SBA2	
Total number of items	20	20	40	
Total number of distractors	60	60	120	
Functional Distractors (FDs) Distractors selected by ≥ 5% students	34 (56.6%)	37 (61.6%)	71 (59.1%)	
Non-functional Distractors (NFDs) Distractors selected by < 5% Students	26 (43.3%)	23 (38.3%)	49 (40.8%)	
Items with 1 NFD (DE = 33.3%)	3 (3 NFDs)	4 (4 NFDs)	7	Total
Items with 2 NFD (DE = 66.6%)	7 (14 NFD)s	8 (16 NFDs)	15	26 (65%)
Items with 3 NFD (DE = 0.00%)	3 (9 NFDs)	1 (3 NFDs)	4	–
Items with 0 NFD (DE = 100%)	7 (0 NFD)	7 (0 NFD)	14 (35%)	
Overall mean DE	63.3	68.3	65.8	

Table 5: Items with non-functional distractors (NFDs) and their relationship with DIFI and DI in SBA items in 1st Professional Examination 2014

Difficulty Index (DIF I)	Items with NFDs	Discrimination Index (DI)	Items with NFDs
≤0.30	1	<0.15	13
0.31–0.40	3	0.15–0.24	1
0.41–0.60	4	≥0.25	13
≥0.61	19	–	–
Total DIF I	27 (0.73)	Total (Mean DI)	27 (0.19)

Table 6: Distractor efficiency (DE) of items (N = 40) with different values of DIF I and DIS I in SBA in 1st Professional Examination 2014

Parameter	Difficulty Index (DIF I)		Discrimination Index (DI)	
	Difficult (≤ 30)	Easy (≥ 0.61)	Poor (< 0.15)	Excellent (≥ 0.25)
Number of Items	4 (10%)	21 (52.5%)	14 (35%)	22 (55%)
DE Mean ± SD	91.65 ± 16.70	49.16 ± 29.07	47.57 ± 33.85	72.70 ± 28.44
Unpaired *t*-test	$t = 2.80$; df = 23; $p = 0.010$		$t = -2.39$; df = 34; $p = 0.020$	

DISCUSSION

Professional (summative) examination is an ongoing process in medical education and needs right selection of instrument to assess students' performance with minimum error of measurement. SBA as multiple choice questions is an effective instrument to measure the students' analytic reasoning skills and in-depth performance in outcome based education practiced in an integrated curriculum. MCQ format allows teachers to efficiently assess larger number of candidates and to test a wide range of content (12). However, quality of SBA items depend upon faculty development to write good items that test higher order thinking across the content to discriminate students with higher abilities from the students with poor abilities. To ensure that quality of SBA used in assessment were of sufficiently good quality, evaluation of each item for its DIF I, DI and DE is considered important and needs to be consistently done and interpreted for teaching faculty to improve their learning curve in writing SBA questions. Technical training to write SBA alone is not sufficient to produce good SBA items and it requires continuing experience and evidence-based evaluation to write quality SBA. Faculty must be informed of SBA items performed in assessment with regards to level of difficulty and power of discrimination between high abilities and poor abilities students. The process helps them to accept or discard the items or apply the corrective measure for subsequent inclusion of items in SBA question bank with acceptable DIF I, DI and DE for summative assessment.

DIF I of item as it shows percentage of students both in the upper and the lower group answering the items correctly has been found with a mean DIF I of 0.64±0.25 (see Table 1) in present study. This is not within the desirable range of DIF I = 0.31–0.60 or comparable with other study (13). This indicates that SBA items were comparatively easy to answer, which may be due to NFD, making it feasible for both, upper and lower performers. Such situations have no motivation for students with low abilities (5). Items should be within the acceptable range and not too difficult (≤30) nor too easy ≥0.61, since difficult items are not good even for students with high abilities and all such items should be rewritten with corrective measure to remove the flaws before deciding to accept or

discard the items for keeping in item bank. DIF I of 16 (40%) items was in the range of good to excellent (see Table 2). Items established as difficult (≤30) were only 4 (10%). Difficult items should be reviewed for its language and grammar, ambiguity and controversial statements. Highly difficult items should also be checked for correct answer as it might have been fed with wrong key.

DI is a good parameter to differentiate between good and poor performers and 0.15–0.25 is considered good discriminating items. DI <0.15 is considered poor for items discrimination power whereas items ≥0.25 is regarded excellent. In this study DI was 0.23±0.24 (see Table 1) slightly below the excellent power of discrimination and it is attributed to unexpected number of NFD in many items that some of these items have been shown to have 3 NFD. However, there was no negative DI. Reasons for negative DI are ambiguity, wrong key or poor preparation of students (6). Items were also categorised into poor, good and excellent items (see Table 2) depending upon their DIF I and DI for decision being made to revise, discard or store (14).

Items with NFD (<5% examinees selected the distractor) are important to establish DE. Plausible distractors are important for quality SBA. It has been observed that number of NFD increases with increasing number of distractors (15). Plausible distractors are even more difficult for SBA items developed from basic sciences compared to SBA from clinical science by teaching faculty. DE is indirectly proportional to NFD and items with more functional distractors increase the DE. The number of NFD was found high in present study and it had its impact on DIF I and DE. DIF I is increased (easy items) and DE is decreased with increasing number of NFDs in items. Items with more NFD are implausible and of little value (16) and were revised or removed from the items pool. Number of 0 frequency distractors attempted in upper and lower achievers varied however, comparable in Test 1 and Test 2 (see Table 3). Distractors with choice frequency = 0 were 24 (40%) in upper achievers and 15 (25%) in lower achievers. However, over all 39 (32.50%) distractors with choice frequency = 0 in both upper and lower achievers (see Table 3) was expected, considering the minimal training of teaching faculty in writing SBA items. In the given circumstances, low proportion of items with three functioning distractors (zero NFD) of 7 items in each Test 1 and Test 2 (see Table 4) was reasonably well. The reason for items with increasing number of NFD may be due to lack of experience of teaching faculty writing the SBA items after attending 1–2 hands-on workshop only. Incidentally other researchers have similar finding and even professionally developed items have been reported to rarely have more than 2 functional items (17).

A research study suggests that none of the five options had four functioning distractors (18). It is not easy for faculty members to develop 2 or 3 equally plausible distractors. However it has been established that items with 2 plausible distractors are better than items with three or four implausible distractors (19, 20). 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 (21). Over all functional distractors (distractor selected by ≥ 5% examinees) were 71 (59.1%) compared to NFD 49 (40.8%) in present study (see Table 4) is not unexpected for a teaching faculty beginning to write SBA items. Number of items with NFD in relation to varying level of DIF I and DI were not the same. More items with NFD were observed with high DIF I (easy item) whereas equal number of items with NFD were seen with both, low and high discrimination powers (see Table 5) in present study. Research has established that the psychometric properties of the test remain similar and there is no reduction in the reliability and validity of a test when

number of options is reduced to three distractors (22).

DE is determined by number of NFDs present in an item and it ranges from 0%–100%. Selection or rejection of items for question bank is best guided by DE. Items with 0% DE should be discarded whereas those with varying percentages should be revised by replacing the distractors with better choices to be reused in future examinations. Mean DE in present study = 65.80±32.46 (see Table 1) is lower than DE reported for SBA items in literature (23). However, distractors efficiency viewed in relation to difficult and easy items showed mean DE high (91.65±16.70) in 4 difficult items than mean DE (49.16±29.07) in 21 easy items (see Table 6). This obviously is due to the difference of increasing number of NFDs between these two difficulty indices. Similar was the case with discrimination index, which showed a high mean DE with excellent DI in 22 (72.70±28.44) items compared to mean DE = 47.45±33.85 in 14 items with poor DI. These differences in DE in both cases were statistically significant with p = 0.010 and 0.020 respectively (see Table 6).

In present study item-total correlation ranged from –0.001 to 1.00 in SBA Test 1 and 0.164 to 4.00 in SBA Test 2. Items with r = > 0.30 were 0 items in SBA Test 1 and only 6 items in SBA Test 2. This suggests that most of the items were unable to show what it purports to test as an individual item. Many items in SBA Test 1 showed negative correlation to total score and should be discarded. This was also reflected by internal consistency of the test as a whole determined by Cronbach's alpha, which showed 0.51 for 20 items SBA Test 1 and 0.41 for another 20 items in SBA Test 2. Small number of items (20 in each test) may also be the reason of low alpha. However, Cronbach's alpha of ≥0.5 to 0.6 is considered poor and <0.5 is unacceptable and by this criteria internal consistency of items have been established to be poor in SBA Test 1 and unacceptable in Test 2. Reliability in term of parallel-form estimate was also computed for 2-tests of SBA measuring the same construct administered in succession. Cronbach's alpha by parallel-form method was also low (alpha = 0.57). Key to parallel-form method of reliability determined as short-term stability was developed as alternate test, which is equivalent in content, item format and response process. An overall low score of SBA may be attributed to, a new instrument for faculty and their inability to structure items with plausible distracters, indicated by low DI and DE. However, it has been made mandatory for all faculty members to attend hands-on workshop before they are invited to write SBA items.

A high DIF I (easy items) and low DI (poor discrimination) is expected to gradually improve to produce good SBA items as the teaching faculty gain experience in writing SBA items. However, practicing standard setting method set at passing score of 50% may not be the logical decision and under the circumstances it will be a wise decision to set a passing score relative to difficulty of test using "Angoff method", which is comparatively easy to implement (24, 25). SBA with suspected low DIF I, DI and DE in professional examination should employ standard error of measurement (SEM) to calculate a band of score to protect borderline students. Any flaw in item structure should not be allowed to affect students' performance with false result. SEM can easily be calculated with standard deviation and reliability coefficient alpha statistics already computed. An estimated SEM band score can be used to decide on borderline students for triangulation of their performance in professional examination with performance in other test score, for instance semester examination, continuous assessment, PBL evaluation, logbooks and case write up in deciding whether or not they should pass the examination.

CONCLUSION

The policy of item analysis as part of greater evaluation and psychometrics of professional examination has been a valuable step to identify SBA items for its difficulty and discrimination indices and DE. Interpretation of evaluation must aim to provide feedback to teaching faculty. Writing a quality SBA item primarily needs training and experience complimented with just-in-time evaluation and feedback to faculty. The high frequency of difficult or easy items and moderate to high frequency of poorly discriminating items in present study suggest continuing corrective measure to improve the quality of SBA items and storing them in question bank.

Increased number of non-functional distractors subsequently affecting the DE in this study has been due to difficulty of teaching faculty to produce plausible distractors for single best answer questions with four distractors. However, it is recommended that until such time that the faculty is experienced to produce quality SBA, standard error of measurement (SEM) should be utilised to calculate a band of score to handle borderline students with care. SEM estimate can be used for triangulation of borderline students' performance in professional examination with their performance in other test score such as semester examination or continuous assessment in deciding whether or not they should pass the examination. Item analysis result emphasises the need of regular evaluation of instruments to provide feedback to teaching faculty in order to improve the quality of SBA items in summative assessment.

REFERENCES

1. Polit DF, Hunglern BP. Nursing research: principles and methods. Philadelphia: Lippincort, Williams and Wikkins; 1999.
2. Haladyna TM. Developing and validating multiple choice test items. New Jersey: Lawrence Eribaum Associate; 1999.
3. Farley JK. The multiple-choice test: writing the questions. Nurse Educ. 1989;14(6):10–12. https://doi.org/10.1097/00006223-198911000-00003.
4. Cizek GJ, O'Day DM. Further investigations of non-functioning options in multiple-choice test items. Educ Psychol Meas. 1994;54(4):861–72. https://doi.org/10.1177/0013164494054004002.
5. Eaves S, Erford B. The Gale group. The purpose of item analysis, item difficulty, discrimination index, characteristic curve. [cited 2013 Apr 15]. Available from: www.education.com/reference/article/itemanalysis.
6. Hassan S, Mohd Amin R, Mohd Amin Rebuan H, Aung MMT. 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. Malaysian Journal of Public Health Medicine. 2016;16(3):7–15.
7. Linn RL, Miller MD. Measurement and assessment in teaching. Upper Saddle River, New Jersey: Pearson Education Inc.; 2005. p. 348–65.
8. 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. https://doi.org/10.1186/1472-6920-9-40.
9. Ebel RL, Frisbie DA. Essentials of educational measurements. 5th ed. Englewood Clipps, NJ: Prentice Hall; 1991.
10. Drowning SM. Assessment of knowledge with written test forms. In: Norman GR, Van der Vleuten C, Newble DI, editors. International handbook of research in medical education Volume II. Dorcrecht: Kluwer Academic Publishers; 2002. p. 647–72.

11. Haladyna TM, Downing SM. Validity of taxonomy of multiple choice item-writing rules. Appl Meas Educ. 1989;2(1):51–78. https://doi.org/10.1207/s15324818ame0201_4.

12. McCoubrie P. Improving the fairness of multiple-choice questions: a literature review. Med Teach. 2004;26(8):709–12. https://doi.org/10.1080/01421590400013495.

13. Guilbert JJ. Educational hand book for health professionals. WHO offset Publication 35. Geneva: World Health Organization; 1981.

14. Gajjar S, Sharma R, Kumar P, Rana M. Item and test analysis to identify multiple choice questions (MCQ) from an assessment of medical students of Ahmedabad, Gujrat. Indian J of Community Med. 2014;39(1):17–20. https://doi.org/10.4103/0970-0218.126347.

15. Sidick JT, Barrett GV, Doverspike D. Three-alternative multiple choice tests: an attractive option. Peers Psychol. 1994;47(4):829–35. https://doi.org/10.1111/j.1744-6570.1994.tb01579.x.

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

17. 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. https://doi.org/10.1186/1472-6920-9-40.

18. Haladyna TM, Downing SM. How many options is enough for a multiple-choice test item? Educ Psychol Meas. 1993;54(4):999–1010. https://doi.org/10.1177/0013164493053004013.

19. Schuwirth LWT, Vlueten CPM van der. Different written assessment methods: what can be said about their strengths and weaknesses? Med Educ. 2004;38(9):974–79. https://doi.org/10.1111/j.1365-2929.2004.01916.x.

20. Crehan KD, Haladyna TM, Brewer BW. Use of an inclusive option and the optimal number of options for multiple-choice items. Educ Psychol Meas. 1993;53(1):241–247. https://doi.org/10.1177/0013164493053001027.

21. Rodriguez MC. Three options are optimal for multiple-choice items: a meta- analysis of 80 years in research. Educ Meas Issues Pract. 2005;24(2):3–13. https://doi.org/10.1111/j.1745-3992.2005.00006.x.

22. Trevisan MS, Sax G, Micheal WB. The effects of the number of options per item and student ability on test validity and reliability. Educ Psychol Meas. 1991;51(4):829–37. https://doi.org/10.1177/001316449105100404.

23. Hingorjo MR, Jaleel F. Analysis of one-best MCQs: the difficulty index, discrimination index and distracter efficiency. J Pak Med Assoc. 2012;62:142–7.

24. Norcini JJ. Setting standards on educational tests. Med Educ. 2003;37(5):464–9. https://doi.org/10.1046/j.1365-2923.2003.01495.x.

25. Case SM, Swanson DB. Constructing written test questions for the basic and clinical sciences. 3rd ed. Philadelphia, PA: National Board of Medical Examiners; 2001.