Original Article

Item and Test Analysis to Identify Quality Multiple Choice Questions (MCQs) from an Assessment of Medical Students of Ahmedabad, Gujarat

Sanju Gajjar, Rashmi Sharma, Pradeep Kumar, Manish Rana

Department of Community Medicine, Gujarat Medical Education and Research Society Medical College, Sola, Ahemdabad, Gujarat, India

ABSTRACT

Background: Multiple choice questions (MCQs) are frequently used to assess students in different educational streams for their objectivity and wide reach of coverage in less time. However, the MCQs to be used must be of quality which depends upon its difficulty index (DIF I), discrimination index (DI) and distracter efficiency (DE). **Objective:** To evaluate MCQs or items and develop a pool of valid items by assessing with DIF I, DI and DE and also to revise/ store or discard items based on obtained results. **Settings:** Study was conducted in a medical school of Ahmedabad. **Materials and Methods:** An internal examination in Community Medicine was conducted after 40 hours teaching during 1st MBBS which was attended by 148 out of 150 students. Total 50 MCQs or items and 150 distractors were analyzed. **Statistical Analysis:** Data was entered and analyzed in MS Excel 2007 and simple proportions, mean, standard deviations, coefficient of variation were calculated and unpaired t test was applied. **Results:** 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%. **Conclusion:** Study emphasizes the selection of quality MCQs which truly assess the knowledge and are able to differentiate the students of different abilities in correct manner.

Keywords: Difficulty index, discrimination index, distractor efficiency, multiple choice question or item, nonfunctional distractor (NFD), teaching evaluation

Introduction

Economic and social prosperity of a nation depends upon the human resources development (HRD) of health related manpower and medical schools play

Access this article online

Quick Response Code:

Website:
www.ijcm.org.in

DOI:
10.4103/0970-0218.126347

crucial role in this process. Assessment of medical undergraduate student gives insight about their learning and competencies. Multiple choice questions (MCQs) or "items" are being increasingly used in such assessments. "Item analysis" examines student responses to individual test items (MCQs) to assess the quality of those items and test as a whole[1,2] to improve/ revise items and the test.[3] A good item can assess cognitive, affective, as well as psychomotor domain and is preferred over other methods for its (1) objectivity in assessment, (2) comparability in different settings, (3) wide coverage of subject, and (4) minimization of assessor's bias. MCQ

Address for correspondence:
Dr. Pradeep Kumar, A 1/7, Swagat City, Adalaj, Gandhinagar-382 421, Gujarat, India. E-mail: drpkumar_55@yahoo.com

based evaluation apart from assessing knowledge also evaluates understanding and analyzing power of students. Item analysis enables identifying good MCQs based on difficulty index (DIF I) also denoted by FV (facility value) or *P*-value, discrimination index (DI), and distractor efficiency (DE).[1,3-7] Keeping in view the widespread use of MCQs in assessment of students, present study has been undertaken with an objective to evaluate MCQs or items and develop a pool of valid items by assessing with DIF I, DI, and DE and also to revise/store or discard items based on obtained results.[1,2,8]

Materials and Methods

148 out of total 150 students of 1st MBBS after 40 hours of didactic teaching of Community Medicine appeared in an internal examination in April 2012. It comprised of 50 "single response type" MCQs. The examination time was 60 minutes and marks allotted were 100. All MCQs had single stem with four options/responses including, one being correct answer and other three incorrect alternatives (distractor). Each correct response was awarded 2 marks and each incorrect response was awarded -1, range of score being 0-100 (ignoring minus marks). To avoid possible copying from neighboring student, they were administered one of three paper sets which were prepared with disorganized sequencing of questions.

Data analysis

Data obtained was entered in MS Excel 2007 and analyzed. Score of 148 students was entered in descending order and whole group was divided in three groups. One group consisting of higher marks was considered as higher ability (H) and other group consisting of lower marks was considered as lower ability (L) group.[7,9,10] Out of 148 students, 49 were in H group and 49 in L group; rests (50) were in middle group and not considered in the study. Total 50 MCQs and 150 distractor were analyzed and based on this data, various indices like DIF I, DI, DE, and nonfunctional distractor (NFD) were calculated with following formulas:[2,7,9]

1. DIF I or p value = $[(H + L)/N] \times 100$ and
2. DI = $2 \times [(H - L)/N]$

Here N = total number of students in both high and low groups and H and L are the number of correct responses in high and low groups, respectively.

DIF I describe the percentage of students who answered the item correctly and ranges between 0 and 100%.[1,3,4,6] DIF I is a misnomer as bigger is the value of DIF I, easier is the item and vice versa; hence, it is also called by some authors as ease index.[9] DI is the ability of an item to differentiate between students of higher and lower

abilities and ranges between 0 and 1. Higher the value of DI, item is more able to discriminate between students of higher and lower abilities. DI of 1 is ideal as it refers to an item which perfectly discriminates between students of lower and higher abilities.[2] There are instances when the value of DI can be <0 (negative DI) which simply means that the students of lower ability answer more correctly than those with higher ability. Such situations though undesirable, happen due to complex nature of item making it possible for students of lower ability to select correct response without any real understanding. Here a student of lower ability by guess select correct response; while a good student suspicious of an easy question, takes harder path to solve and end up to be less successful.[3]

An item contains a stem and four options including one correct (key) and three incorrect (distractor) alternatives.[3,4] Here, NFD in an item is option (s) (other than key) selected by <5% of students; alternatively functional or effective distractors are those selected by 5% or more participants.[2,6,9] DE is determined for each item on the basis of the number of NFDs in it and ranges from 0 to 100%. If an item contains three or two or one or nil NFDs then DE will be 0, 33.3, 66.6, and 100%, respectively.[6,9] Items were categorized as poor, good, or excellent and actions such as discard/revise and store were proposed based on the values of DIF I and DI as suggested [Table 1].

Results

Total 50 MCQs and 150 distractors were analyzed. Score of 148 students ranged from 0 to 33 (out of 100); 20% students scored zero or less than zero, but their negative marks were ignored. Means and standard deviations (SD) for DIF I (%), DI and DE (%) were 39.4 ± 21.4%, 0.14 ± 0.19, and 88.6 ± 18.6%, respectively [Table 1]. Out of 50 items, 24 had "good to excellent" level of difficulty (DIF I = 31-60%) and 24 had "good to excellent" discrimination power (DI ≥ 0.15) [Table 2]. When these two are considered together, there were 15 items as ideal which could be included in question bank. However, if only the items with "good to excellent DIF I and excellent DI (≥0.25) are considered, there were 10 items as ideal. Besides this, 10 (20%) items had negative DI. Out of 150 distractors, 17 (11.4%) were NFDs present in 15 items (13 had 1 and two had 2) with DE varying between 33 and 66%. Remaining 35 items had no NFDs with their DE being 100%. No item had DE of 0% [Table 3]. Fifteen items with NFDs had mean values

Table 1: Assessment of 50 items based on various indices amongst 148 students

Parameter	Mean	Standard deviation (SD)
Difficulty index (DIF I) (%)	39.4	21.4
Discrimination index (DI)	0.14	0.19
Distractor efficiency (DE) (%)	88.6	18.6

of DIF I and DI as 53.5% and 0.18, respectively [Table 4]. Remaining 35 items without NFDs had lower mean values for DIF I and DI as 33.3% and 0.13, respectively. When viewed in relation of difficulty level of questions, DE was high (91.7%) in 16 difficult items than 79.3% in 10 easy items. However, DE showed little variation amongst items with changing DI. Mean DE was 88.9% in 15 items with excellent DI compared to 88.4% in 26 items with poor DI; difference in DE in both cases was statistically not significant [Table 5].

Discussion

Quality medical care depends upon the development of knowledgeable, skilled, and competent medical personnel. Any assessment whether formative or summative has intense effect on learning and is an important variable in directing the learners in a meticulous way.[5] Single correct response type MCQ is an efficient tool for evaluation;[9] however, this efficiency solely rests up on the quality of MCQ which is best assessed by the analysis of item and test as a whole together referred as item and test analysis.

Each item (MCQ) while being used in assessment must be evaluated based on DIF I, DI, and DE because if an item is flawed then this itself becomes distracting and the assessment can be false. Item difficulty rather than defining difficulty due to some intrinsic characteristic of the item is defined in terms of frequency with which those taking the test choose the correct response.[3] Mean DIF I in this study was 39.4 ± 21.4% well within the acceptable range (31-60%) identified in present study. Others have proposed this range as 41–60%.[11] Too difficult items (DIF I ≤ 30%) will lead to deflated scores, while the easy items (DIF I > 60%) will result into to inflated scores and a decline in motivation.[4] Too easy items should be placed either at the start of the test as 'warm-up' questions or removed altogether, similarly difficult items should be reviewed for possible confusing language, areas of controversies, or even an incorrect key.[9] DI of an item indicates its ability to differentiate between students of higher and lower abilities. It is obvious that a question which is either too difficult (done wrongly by everyone) or too easy (attempted correctly by everyone) will have nil to poor DI. Value of DI normally ranges between 0 and 1. Mean DI in present study was 0.14 ± 0.19 less than the acceptable cut off point of 0.15.[2] It was so because 10/50 items had DI less than zero (negative DI). Another study[9] had the mean DI of 0.36 ± 0.17 and had only 2/50 items with negative DI. As mentioned earlier, in such cases, students of lower ability answer questions correctly than those with higher ability. Reasons for negative DI can be wrong key, ambiguous framing of question[7] or generalized poor preparation of students as was the case in present study where the overall score

Table 2: Distribution of items in relation to DIF I and DI and actions proposed

Cut off points	Items (*N* = 50)	Interpretation	Action
Difficulty index (DIF I)			
≤30	16	Difficult	Revise/discard
31-40	12	Good	Store
41-60	12	Excellent	Store
≥61	10	Easy	Revise/discard
Discrimination index (DI)			
<0.15	26	Poor	Discard/revise
0.15-0.24	9	Good	Store
≥0.25	15	Excellent	Store

Table 3: Distractor analysis (*N* = 150)

Distractor analysis	
Number of items	50
Total distractors	150
Functional distractors	133 (89.6%)
Nonfunctional distractors (NFDs)	17 (11.4%)
Items with 1 or 2 NFDs (DE between 33 and 66%)	15 (30%)
Items with 1 NFD (DE=33.3%)	13
Items with 2 NFD (DE=66.6%)	2
Items with 0 NFD (DE=100%)	35
Overall mean DE (mean±SD)	88.6±18.6%

Table 4: Items (*N* = 15) with nonfunctional distractors (NFDs) and their relationship with DIF I and DI

DIF I (%)	Items with NFDs	DI	Items with NFDs
≤30	2	<0.15	6
31-40	2	0.15-0.24	4
41-60	4	≥0.25	5
≥61	7		
Mean DIF I (%)	53.5	Mean DI	0.18

DIF I: Difficulty index

Table 5: Distractors efficiency (DE) of items (*N* = 50) with different values of DIF I and DI

Grading	Difficulty index (DIF I)		Discrimination index (DI)	
	Difficult (≤30%)	Easy (≥61%)	Poor (<0.15)	Excellent (≥0.25)
No of items	16	10	26	15
DE (%) mean±SD	91.7±14.7	79.3±21.1	88.4±18.7	88.9±15.7
Unpaired *t*-test	*t*=1.77; df=24; *P*=0.089		*t*=0.08; df=39; *P*=0.93	

SD: Standard deviation, DF: degrees of freedom

of 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.[3] Based on the cut off points for "good to excellent" for DIF I and DI, there were 15 items as ideal compared to 32 (out of 50) in another study.[9]

Analyzing the distractors (incorrect alternatives) is done to determine their relative usefulness in each

item. Items need to be modified if students consistently fail to select certain distractors. Such alternatives are probably implausible and therefore of little use as decoys.[3] Therefore, designing of plausible distractors and reducing the NFDs is important aspect for framing quality MCQs.[12] More NFD in an item increases DIF I (makes item easy) and reduces DE, conversely item with more functioning distractors decreases DIF I (makes item difficult) and increases DE. Higher the DE more difficult the question and vice versa, which ultimately relies on presence/absence of NFDs in an item. Mean DE in present study was 88.6 ± 18.6% higher than DE of 81.4% reported elsewhere in a similar type of study.[9] It showed little variation in items with high or low DI [Table 5].

DIF I and DI are often reciprocally related except for extreme situations where the DIF I is either too high or too low. The relationship between the two is not linear, but is dome-shaped.[13] Questions with high DIF I (easier questions), discriminate poorly; conversely, questions with a low DIF I (difficult questions) are good discriminators unless they are so difficult that not even good students can attempt them correctly.[7,10]

Conclusion and Recommendations

Item analysis is a simple yet valuable procedure performed after the examination providing information regarding the reliability and validity of an item/test by calculating DIF I, DI, and DE and their interrelationship. An ideal item (MCQ) will be the one which has average difficulty (DIF I between 31 and 60%), high discrimination (DI ≥ 0.25) and maximum DE (100%) with three functional distractors. Items analyzed in the study were neither too easy nor too difficult (mean DIF I = 39.4%) which is acceptable but the overall DI was 0.14. Therefore, items were acceptably difficult but were not good at differentiating higher and lower ability students. DI was poor due to the 10 items with negative DI; situation was further complicated by poor preparedness of students. Items with negative DI and NFDs must be identified and removed from future assessment. While preparing ideal items, level of preparedness of students must be kept in mind and more efforts be made to replace NFDs with ideal/plausible distractors.

References

1. Scorepak®: Item analysis. Available from: www.washington. edu./oea/score1/htm [Last citied on 2013 Apr 13].

2. 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. p. 70-77.

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

4. Eaves S, Erford B. The Gale group: The purpose of item analysis, item difficulty, discrimination index, characteristic curve. Available from: www.education.com/reference/article/item-analysis/ [Last cited on 2013 Apr 15].

5. Sarin YK, Khurana M, Natu MV, Thomas AG, Singh T. Item analysis of published MCQs. Indian Pediatr 1998;35:1103-5.

6. Tarrant M, Ware J, Mohammed AM. An assessment of functioning and non-functioning distractors in multiple-choice questions: A descriptive analysis. BMC Med Educ 2009;9:1-8.

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

8. Manual of Basic Workshop in medical education technologies. Ahmadabad: Regional Training Centre Smt. NHL Medical College; 2012.

9. 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.

10. Mohammad S. Assessing student performance using test item analysis and its relevance to the State Exit Final Exams of MAT0024 Classes-An Action Research Project from Mathematics dept, Miami dade college, USA. Available from: www.mdc. edu/hialeah/Polygon2008/articles/shakil3.pdf [Last cited on 2013 Apr 13].

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

12. Rodriguez MC. Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. Educ Measure 2005:3-13.

13. Sim SM, Rasiah RI. Relationship between item difficulty and discrimination indices in true/false-type multiple choice questions of a para-clinical multidisciplinary paper. Ann Acad Med Singapore 2006;35:67-71.