

Medical Education / Original Article

Adding Value to Multiple Choice Questions Banking : An AIIMS Patna Experience

**Tribhuwan Kumar, Yogesh Kumar*, Kamlesh Jha
and Ramji Singh**

Department of Physiology,
AIIMS, Patna

Abstract

This is an era of objectivity and Multiple Choice Questions (MCQs) are increasingly being used in evaluation of students in exams in educational institutions. Item analysis enables identifying good MCQs based on difficulty index (DIF I), discrimination index (DI), Point biserial correlation (corrected) and distracter efficiency (DE). Present study has been undertaken with an objective to evaluate MCQs (items) and develop a pool of valid items and also to revise/store or discard items based on obtained results. After 3 months of dedicated teaching 67 out of total 67 students of 1st MBBS, AIIMS Patna appeared in first terminal examination of Physiology. It comprised of 30 "single response type" MCQs. Each correct response was awarded 1 marks and each incorrect response was awarded 0, range of score being 0-30. 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. Out of 67 students, 18 were in H groups and 18 in L group; rests (31) were in middle group and not considered in the study. Total 30 MCQs and 90 distracters were analyzed and based on this data, various indices like difficulty index, discrimination index, distracter efficiency and non functional distracters were calculated. Score of 67 students ranged from 7 to 26 (out of 30). Out of 30 items, 11 had "good to excellent" level or difficulty (DIF I = 31- 60%) and 20 had "good to excellent" discrimination power (DI \geq 0.20). Nineteen questions were selected for question bank, four discarded and rest were modified and will be given for next batch and then again will be analysed.

Introduction

Every individual has to face one or other kind of examination in life. Examinations serve many functions like appraisal of student accomplishment, interrelation of abilities measured, curricular definition

etc. A teacher constricting the items of examination should have functional knowledge of principles of test construction (1). Objectivising evaluation has become very important in today's world as emphasized by guidelines of so many universities. There are several tools available which can gauge the quality of examination. One such tool is Item analysis. It is a process of analysing the performance of a multiple choice question after the examination is over. It helps us analyse how well an examination has performed relative to its purposes. Some of many purposes which item analyses accomplish are:-

***Corresponding author :**

Dr. Yogesh Kumar, Assistant Professor, Department of Physiology,
AIIMS Patna, Email :- dr.yogeshsaxena@gmail.com

(Received on April 12, 2016)

1. Whether the item is of appropriate difficulty level for the batch of students tested.
2. Can it discriminate between knowledgeable and weak (poor knowledge) students?
3. It also helps to check the effectiveness of various distracters.

This type of analysis helps in selection of ideally framed MCQ to be kept in question bank further it provides the teachers feedback of the topics which need more emphasis and reinforcement during classroom teaching.

In the present study item analysis was done to select questions for a Physiology MCQ question bank of AIIMS Patna with known level of difficulty and discriminating power, which is free from constructional errors and having functional distracters, by using principles of Item analysis on MCQ given during previous examinations.

Methodology

Data collection

The study was conducted on multiple choice questions given in the first terminal examination of first year MBBS in the Department of Physiology, AIIMS Patna for batch 2014. Only 67 students took admission in MBBS 2014 batch at AIIMS Patna. In first terminal exam thirty single response multiple choice questions were given apart from Short answer questions and long answer type questions. All multiple choice questions consisted of a stem and four choices out of which one was correct and rest three were incorrect (distracters). A correct response was awarded with 1 mark and there was no negative marking i.e. the items was dichotomously coded or binary. (2) Item analysis of all thirty MCQ was done based on responses given by 67 students who took this examination.

Item analysis:

The analysis was done manually using Microsoft

Office Excel. Each of the items were analysed for Difficulty Index, Discrimination index, Distracter efficiency and Reliability coefficient.

Difficulty index, indicated by pi , is the ratio of the number of persons who have answered the item (question) correctly to the total number of test takers. As the items are binary the difficulty index was calculated by the formula:

$$pi = \frac{Ai}{Ni}$$

Where, pi = Difficulty index of item i

Ai = Number of correct answers to item i

Ni = Number of correct answers plus number of incorrect answers to item i

The difficulty index ranges between 0 to +1 Higher the value easier the question and vice versa. Acceptable range is between 0.3–0.7.

Discrimination index is the difference between proportions of correct answers among high scorers and low scorers. First the test scores were arranged in descending order, the upper 27% were taken as upper group and lower 27% as lower group. The difference in p values for each item for the two groups was calculated. DI of <0.20 is considered poor, between 0.20-0.35 good and >0.35 is considered excellent¹.

For dichotomous items, Discrimination index can also be calculated by correlation coefficient. The formula of point biserial correlation is presented as the Pearson correlation³ between responses to a particular item and scores on the total test and defined as (3).

$$rpbi = \frac{(x_1 - x_0) \sqrt{p(1-p)}}{\sigma_x}$$

where, x_1 is the average total score for those who correctly answer an item,

x_0 is the average total score for those who incorrectly answer the item,

σ_x is the standard deviation of total scores, and P is the difficulty index for this item.

The point biserial correlation was calculated by inserting Excel function = CORREL (array1, array2), where array1 is the scores for that particular item by each candidate and array2 is the total scores of each candidate (4).

Being a correlation, Item discrimination can vary from +1 to -1 (5). Positive values indicate the students who have answered that item correctly have scored well in the exam whereas negative values indicate most students who scored less in the exam have answered that item correctly while those who did well have failed to answer. Acceptable range is 0.20 or higher.

Higher the values more are the discriminating ability of the item. Similarly, corrected point biserial correlation (4) was also calculated for comparison purpose by deleting individual item score at a time from total score.

Distracter efficiency is the proportion of students who have selected each of the response options. For the key, this proportion is equivalent to the item p-value (6). If the distracter is plausible at least some students will select the option.

Reliability:

Classic Test Theory is based on the premise that the observed score from a psychological testing is composed of an un-measurable "true score" and error i.e. $X = T + E$, where X is the observed score, T is true score and E is the measurement error.

So a measurement is said to be reliable if it reflects mostly true score, relative to the error (7). Most commonly reliability is measured by Cronbach's coefficient alpha by the formula (6)

$$\alpha = \left(\frac{k}{k-1} \right) * \left[1 - \frac{\sum (s^2_i)}{s^2_{sum}} \right]$$

where, α = Cronbach's coefficient alpha

k = no of items

s^2_i = variance for the k individual items &

s^2_{sum} = variance of sum of all items

In the Excel sheet we calculated variances by inserting function VARP (array1) (8), where array1 is the score for individual items when calculating the individual item variances and the total score when calculating variance of sum of all items.

Similarly, α was also calculated after deleting one item at a time to assess the impact of individual items on the total reliability of the test.

Results

With the aim to select questions for Physiology MCQ question bank total 30 MCQs and 90 distracters were analysed and based on this data, various indices like difficulty index, discrimination index, distracter efficiency and non-functional distracters were calculated. Score of 67 students ranged from 7 to 26 (out of 30). Out of 30 items, 11 had "good to excellent" level or difficulty (DIF I = 31- 60%) and 20 had "good to excellent" discrimination power (DI \geq 0.20). Nineteen questions were selected for question bank, four discarded and rest were modified and will be given for next batch and then again will be analysed.

Table I clearly shows that Discrimination Index was calculated by various methods and mean value by using the traditional method is 0.32 ± 0.2 , which is a good discrimination index. Point biserial coefficient of 0.29 also indicate the same. Similarly, average difficulty index of overall test is 0.54 which is excellent.

TABLE I: Difficulty level and discrimination index of whole test.

Parameter	Mean	SD
DI (discrimination index)	0.32	0.2
Point biserial correlation	0.29	0.15
Point biserial correlation (corrected)	0.19	0.15
Difficulty Index (DIF I)	0.54	0.26
Cronbach's alpha	0.6534	-

Table II indicate that 7 questions are found difficult by students indicating either concept asked is not clear to them or framing of question is not proper which is confusing them. 12 questions appear to be very easy and in four of them inefficient distractors need to be modified.

TABLE II: Distribution of items in relation to Difficulty Index and actions proposed.

Cut off points	Item	Interpretation	Action
Difficulty Index			
<0.3	7	Difficult	4 discarded and 3 modified
0.31-0.40	2	Good	Selected for question bank
0.41-0.60	9	Excellent	Selected for question bank
>0.60	12	Easy	8 selected and 4 modified

Table III clearly indicate that about 20 MCQ have good to excellent ability to discriminate between good and bad students while about 10 MCQ cannot discriminate between the two.

TABLE III: Distribution of items in relation to Discrimination Index by various methods.

	Discrimination Index			Interpretation
	DI	Point-biserial correlation	Point-biserial correlation (corrected)	
<0.20	10	6	9	Poor
0.20-0.35	8	3	12	Good
>0.35	12	21	9	Excellent

Table IV shows that out of 90 total distractors around 70% are functional distractors and only 3 items have three non-functional distractors and these questions need to be reframed.

TABLE IV: Distracter analysis (N=90).

Distracter analysis	
No of items	30
Total distracter	90
Functional distracters	63(70%)
Non Functional distracters	27
Items with 1 NFD	10
Items with 2 NFD	4
Items with 3 NFD	3
Items with 0 NFD	13

NFD = Non-functional distractors.

Discussion

Evaluation of performance of medical students at the end semester and professional exams is most important aspects of any curriculum. Performance of students gives the faculty a feedback about the pace and content of the course. Therefore, the methods which we are using for assessment of students should be capable of assessing the learning objectives, core competencies and skills set as a goal by concerned faculty. Multiple choice questions (MCQs) were introduced to bring objectivity in scoring and evaluation. It increases efficiency of teachers to teach large group of students. MCQs can be recyclable and can be progressively improved and stored in the form of question banks for reuse in different combinations and settings. MCQ tests are generally used to assess lower order cognition such as the recall of discrete facts and to test student's broad knowledge of the curriculum and learning objectives. If properly framed these are equally effective to assess learning objectives of high cognitive level.

There is danger of MCQs being answered correctly by guess work if poorly designed. To assess and improve quality of MCQ tests, it is important to validate MCQ test questions for the discriminative value of items and to review and endorse MCQ use. This can be achieved by developing questions and question banks collaboratively by team of assessors.

By progressively developing MCQ bank, we can support bench marking processes and establish assessment standards having long term effects in assuring quality.

To assure that a MCQ test design contributes to the tests overall balance in terms of learning objectives across a spectrum of cognitive demand, individual MCQ should be used in a design framework. A design framework can be according to cognitive domains of Blooms taxonomy. An example of design framework used at the University of Texas is given in the table below.

<i>Cognitive domain (Blooms taxonomy)</i>	<i>Percentage of total</i>	<i>Topics</i>
Knowledge	12.5	Spread evenly
Comprehension	17.5	over all topics
Application	37.5	and subtopics
Analysis	25	
Synthesis	5	
Evaluation	2.5	

We can create our own design depending upon the Goals of the curriculum.

In our study, the point-biserial coefficient, typically known as Pearson product-moment correlation suggested by Matlock-Hetzel was selected in calculating the discrimination power. (9) Item analysis of written MCQ of first term revealed mean discrimination index of 0.29 ± 0.15 which is good discrimination index. Further analysis of individual items revealed 6 items had poor discrimination index, 3 were good while 21 items were excellent discriminators. Those questions which have poor DI means topic is either not understood by whole of the class or every student has understood the topic equally well. These items should be highlighted for reviewing purpose. A similar study was conducted by El-Uri and Malas to analyse undergraduate examination in obstetrics and gynaecology, who reported that 38% of the test items had the discrimination coefficient less than 0.2 with 23 questions obtained negative discrimination. (10) Items with negative coefficient values should be removed from the question bank. Si Mui & Rasiah and Matlock-Hetzel (9) proposed that student in the low achievement group often make a guess in answering the easy question and it is by chance that they come up with the correct answer while students in the higher achievement group embark upon the easy question too vigilantly and end up choosing the wrong answer. El-Uria, Malas and Ovwigbo also suggested that items with negative discrimination coefficient should be eliminated from the test.

In our study about 30% MCQs have excellent difficulty index 9% were good while 22% questions were difficult. Items having difficulty index < 0.3 are considered very difficult and these should not be generally used in exams. These questions provide the teachers with the insight about topics which need

further explanation and emphasis in future classes. Content of such classes should be rechecked and modified accordingly. Questions having DIF I of > 0.60 are considered easy, such items are also having poor discriminating power. Instructional Assessments Resources (IAR) insinuates the usage of easy question as warm up questions in assessing student mastery. Ideally we should avoid these items during the exam but a few questions can be kept as first few questions of the paper as these help in boosting the confidence of the students.

We have also analysed a total of 90 discriminators and found 27 of them as non functional meaning less than 5% of students opted them as correct answer. Questions having 3 non functional discriminators were discarded outright while items having 1 or 2 of them were reframed and will be given in future exams and again analysed before rejecting or including them in our question bank.

To construct reliable and efficient MCQ exam, teachers need to be trained in teachers training technology courses with emphasis given on ideal MCQ construction techniques. This kind of item analysis should be regularly done after every internal and professional examination. Questions included in the bank should be given at an interval every 2-3 years and reanalysed to check the changing trend in relevance of the topic.

Conclusion:

MCQs are one of the very powerful tools used to screen and judge students quickly in various fields. To be specific and accurate, the MCQs need to get examined and re-examined in scientific ways. It is not uncommon to encounter wrongly framed MCQs even by the faculty members/subject experts involved in teaching and framing MCQ's for years. Present study focused upon the item analysis of 30 MCQs given in the internal examinations of physiology at AIIMS Patna. Out of 30 questions 19 questions with difficulty index of ≥ 0.31 and discrimination index of ≥ 0.20 were selected for physiology question bank. Four questions were discarded and seven (with one non-functional distractor) were reframed and will be

used in future examinations. As with time relevance and importance of topics keep on changing, one need to undertake the item analysis for the MCQs as a continuous process to weed out the irrelevant topics/items and incorporate newer ones. As the purpose of the examination is not to discriminate between

'good' and 'bad' students but to test the knowledge acquired by them, the study recommends that the experts may be given exposure to the technical aspects of MCQ framing and item analysis through workshops so as to get MCQs with high facility value (difficulty index) and low discrimination power.

References

1. N Ananthkrishnan, Item Analysis. In: N. Ananthkrishnan, K.R Sethuraman, Santosh Kumar. Medical Education: Principles and practice: Second Edition, 2000; 131–137.
2. V Krishnan. The Early Child Development Instrument (EDI): An item analysis using Classical Test Theory (CTT) on Alberta's data. www.cup.ualberta.ca/wp-content/uploads/2013/04/ItemAnalysisCTT_CUPWebsite_10April13.pdf. assessed on 20.07.2015
3. Lin Ding and Robert Beichner, Approaches to data analysis of multiple-choice questions. *PHYS. REV. ST PHYS. EDUC. RES.* 5, 020103: 2009, 020103-1 - 020103-17.
4. Seema Verma, Preliminary Item Statistics Using Point-Biserial Correlation and P-Values. www.eddata.com/resources/publications/EDS_Point_Biserial.pdf assessed on 20-07-2015.
5. Mhairi McAlpine. A Summary of Methods of Item Analysis. www.caacentre.ac.uk/dldocs/BP2final.pdf assessed on 21.07.2015.
6. Professional Testing Inc. Conduct the Item Analysis. www.proftesting.com/test_topics/pdfs/steps_9.pdf assessed on 21.07.15
7. Statsoft. *Reliability and Item Analysis*. www.statsoft.com/Textbook/Reliability-and-Item-Analysis assessed on 21.07.2015.
8. Real statistics using excel. www.real-statistics.com/reliability/cronbachs-alpha assessed on 21.07.2015.
9. Sim, S.M & Rasiah R.I. 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.
10. El-Uri FI, Malas N. Analysis of Use of A Single Best Answer Format in An Undergraduate Medical Examination. *Qatar Medical Journal* 2013; 1.