Analytic assessment of multiple-choice tests

Maryam sadat kaveh tabatabaee MSc[1], Mohammad Hossein Bahreyni Toosi MD [2], Akbar Derakhshan MD [3], Mohammad Khajeh Dalloee MD [4], Hassan Gholami [5]

[1] faculty member of nursery faculty of Mashad University of Medical Science
[2] assistant professor of Mashad University of Medical Science
[3] associate professor of Mashad University of Medical Science, Director of mashad educational development center
[4] assistant professor of Mashad University of Medical Science
[5] faculty member of nursery faculty of Mashad University of Medical Science

ABSTRACT

Background: *Multiple choice tests (MCT),are widely known and applied as useful evaluation tests in the field of education especially in Medical Science. Items on a multiple-choice test consist of a stem, which is followed by a correct answer as well as three to four distracters. Items on a well-written multiple-choice test will have stems that are precise and clear, one answer that is clearly correct or best, and distracters that are plausible.*

Purpose: *The purpose of the present study is conducting item and test analysis to 24 MCTs given in first semester of 2000-2001 educational year in medical faculty of Mashad University of Medical Science.*

Methods: *Data of this descriptive study were composed of 1496 MCQs gathered from 2092 answer sheets of 24 MCTs obtained from educational department of the medical faculty. A split-half method of reliability was employed to calculate reliability coefficient for MCTs. Items Difficulty and Discrimination index also were calculated for questions. Further studies should be undertaken for developments the methods for evaluation of validity, assessment of distracters and structural principles in MCTs .*

Results: *Mean reliability coefficient of the exams was 0.72±0.13 and In more than 50% of cases, reliability coefficient was greater than 0.7. There was a significant difference between basic science exams and clinical clerkship exams in Reliability coefficient (P=0.0001). Mean standard error of measurement (SEM) was 3.51±1.11 . In 52.2% of the cases, difficulty of MCQs was inappropriate and 49.3% of questions had inadequate discriminative power to discern between poor students and good students.*

Conclusion: *Our finding indicate that only 33% of studied MCQs have desirable or acceptable item difficulty and discrimination indices both and 34.9% of those have no desirable or acceptable item difficulty neither acceptable discrimination index. Having subjects respond reliably on a measure is a great start, but there is another concept needed to get down really well named validity.*

Key Words: *MULTIPLE CHOICE QUESTION ,TEST ANALYSIS, RELIABILITY ,ITEM DIFFICULTY DISCRIMINATION INDEX*

Journal of Medical Education winter *2003;2(2):87-91*

Introduction

Evaluation of students' educational achievement is considered to be the most important evaluation in a university. It is not only effective in careful selection of students according their educational achievement cross the curricula but also seems to increase students motivation lead to more learning. University trainers can assess different kinds of learning outcome according to the results of such evaluations and also can assess their teaching quality (1). Evaluation is consisting of two steps. First step is measurement and the second step is judgement about what is measured. Thus the tools developed for measurement should be designed to lead to right judgement (2).

Writing multiple choice tests (MCT) used frequently in medical science tests for class evaluation and certification (3,4,5,6,7). In general, your assessment strategy should enable students to demonstrate the knowledge, skills or attitudes they have acquired or improved upon during the course. Clearly, in questions written in the multiple choice format, the uncertainty of the number of correct statements challenges examinees whose knowledge is not yet solid enough to make a correct response on the question (1,8).

Items on a multiple-choice test consist of a stem, which is followed by a correct answer as well as three to four distracters. Items on a well-written multiple-choice test will have stems that are precise and clear, one answer that is clearly correct

or best, and distracters that are plausible (9,10). Multiple-choice questions (MCQ), if designed carefully, can achieve satisfactory, reliability and efficacy.

It is also important to differentiate difficulty of a question from complication of the concepts on which the question is based (11).

Study of MCTs used frequently in academic settings showed that large number of these tests are not properly written and planed (12,13).

When norm-referenced tests are developed for instructional purposes, to assess the effects of educational programs, or for educational research purposes, it can be very important to conduct item and test analyses. These analyses evaluate the quality of the items and of the test as a whole. Such analyses can also be employed to revise and improve both items and the test as a whole.

Analyzing of tests and items will assist the test developer in determining what is wrong with individual items. It also provides empirical data about how individual items and whole tests are performing in real test situations and determine the proportion of each item in a test.

The purpose of the present study is conducting item and test analysis to 24 MCTs given in first semester of 2000-2001 educational year in medical faculty of Mashad University of Medical Science.

Materials and Methods

In this descriptive study, 3000 coded answer sheet were sent for educational department of the medical faculty. These answer sheets were returned, which completed by students in 24 MCTs taken in first semester of 2000-2001 educational year.

After correcting the sheets in educational development center (EDC) by using optical marker reader (OMR) according answer keys provided by faculty members who designed multiple choice questions (MCQs) , all data were collected in a data base.

A split-half method of reliability was employed to calculate reliability coefficient for MCTs.Items of MCTs randomized according their number (even and unit).The Spearman-Brown formula, was used to estimate the reliability of the whole test on the basis of the correlation between scores on the two halves of the test.

Standard error of measurement (SEM) was investigated by formula as follows:

SEM=SD t (1- r tt)^0.5

SD t is standard deviation of the test and r tt is reliability coefficient of the test.Item difficulty is simply the percentage of students taking the test who answered the item correctly. The larger the percentage getting an item right, the easier the item. The higher the item difficulty, the easier the item is understood to be. To compute the item difficulty, divide the number of people answering the item correctly by the total number of people answering item. The proportion for the item is usually denoted as p and is called item difficulty.

Numeric values of Item Difficulty were classified as shown in table 1.

Discrimination index indicate power of a MCQ in discerning examinees whose knowledge is not enough to make a correct response on the question from those who have enough knowledge on that question.

The discriminating power of an item can be measured by comparing the number of people with high test scores who answered that item correctly with the number of people with low scores who answered the same item correctly. In computing the discrimination index, D, first score each student's test and rank order the test scores. Next, the 27% of the students at the top and the 27% at the bottom are separated for the analysis (11).

The discrimination index, D, is the number of people in the upper group who answered the item correctly minus the number of people in the lower group who answered the item correctly, divided by the number of people in the largest of the two groups.

The higher the discrimination index, the better the item because such a value indicates that the item discriminates in favor of the upper group, which should get more items correct.

Different amounts of discrimination index were classified as shown in table 1.

A negative discrimination index is most likely to occur with an item covers complex material written in such a way that it is possible to select the correct response without any real understanding of what is being assessed. A poor student may make a guess, select that response, and come up with the correct answer. Good students may be suspicious of a question that looks too easy, may take the harder path to solving the problem, read too much into the question, and may end up being less successful than those who guess.

Data of the study were composed of 1496 MCQs gathered from 2092 answer sheets of 24 MCTs and were analyzed using SPSS 10 software.

Results

According to the data gathered from 24 MCTs in medical faculty of mashed University of Medical Science, 16.6 % and 83.4% of the exams were

Table 1 CLASSIFICATION OF DISCRIMINATION AND DIFFICULTY INDICES

Classification of Discrimination Index		Classification of Item Difficulty	
Level of Discrimination Power	Discrimination Index range	Type of Difficulty	Item Difficulty range
Highest high	>=0.35	Desirable	0.5-0.6
High	0.25-0.34	Acceptable	0.3-0.7
Intermediate moderate	0.15-0.24	Inappropriate	<0.3 or >0.7
Low	<=0.14	-	-

conducted in the course of basic science and clinical clerkship , respectively. Thus most of the exams subjects were related to the clinical topics. The mean number of questions in each test was 66±26 and the highest and lowest numbers of questions were given medical ethics exam (31) and surgery exam (123), respectively.

Mean number of examinees studied in each exam was 87.2±30.12. Obstetrics and gynecology exam had the lowest number of attendants (31) and toxicology and legal medicine exam had highest number (161).

Mean reliability coefficient of the exams was 0.72±0.13.lung diseases exam reliability coefficient was the lowest (0.4) and histology exam reliability coefficient was the highest (0.93) reliability coefficient.

In more than 50% of cases, reliability coefficient was greater than 0.7 (table2).

Table 2 THE DISTRIBUTION OF 1496 MCQS IN MEDICAL FACULTY OF MASHAD UNIVERSITY OF MEDICAL SCIENCE ACCORDING TO THEIR RELIABILITY COEFFICIENTS

Test						Reliability Coefficient
Total		Clinical Clerkship		Basic Science		
%	No.	%	No.	%	No.	
8.33	2	8.33	2	-	-	0.40-0.49
4.77	1	4.77	1	-	-	0.50-0.59
20.83	5	20.83	5	-	-	0.60-0.69
33.33	8	33.33	8	-	-	0.70-0.79
29.17	7	16.76	4	12.5	3	0.80-0.89
4.17	1	-	-	4.17	1	≥0.90
100	24	83.33	20	16.67	4	Total

Reliability coefficient for basic science exams was 0.88±0.03 while for clinical clerkship exams, it was equal to 0.68±12.there was a significant difference between basic science exams and clinical clerkship exams in Reliability coefficient (P=0.0001).

Mean standard error of measurement (SEM) was 3.51±1.11. SEM for toxicology and legal medicine was the lowest (2.15) and SEM for the surgery exam was the highest standard error of measurement (5.78). in 41.5 % of exams, SEM was between 3 and 3.99 (table3). There was a significant difference between basic science exams and clinical clerkship exams in SEM (P=0.0001).

TABLE 3 THE DISTRIBUTION OF 24 MCTS IN MEDICAL FACULTY OF MASHAD UNIVERSITY OF MEDICAL SCIENCE ACCORDING TO THEIR STANDARD ERRORS OF MEASUREMENT(SEM)

Test		Standard Error of Measurement(SEM)
Percent	Number	
34.4	9	2-2.99
41.5	10	3-3.99
7.8	2	4-4.99
16.3	3	>=5
100	24	Total

In 52.2% of the cases, difficulty of MCQs was inappropriate. On the other hand,.6% of questions were extremely difficult (item difficulty>0.7) while 44.6% were extremely easy (item difficulty <0.3).only difficulty of 16.5 % of questions ranged between 0.5 and 0.6(desirable item difficulty) (table 4).

TABLE 4 THE DISTRIBUTION OF 1496 MCQS IN MEDICAL FACULTY OF MASHAD UNIVERSITY OF MEDICAL SCIENCE ACCORDING TO THEIR DIFFICULTY INDICES

MCQ		Item Difficulty
Percent	Number	
34.4	114	<=0.3
15.9	238	.03-0.49
16.5	247	0.5-0.6
15.4	230	0.61-0.7
44.6	667	≥0.7
100	1496	Total

Physiology exam had highest number of questions with item difficulty lesser than 0.3 questions

(17.5%) while obstetrics and gynecology exam had lowest number of extremely difficult questions (2.5%).
Toxicology and legal medicine exam had highest number of questions with item difficulty greater than 0.7 (65%) while pharmacology exam had lowest number of extremely easy questions (15%).
49.3% of questions had inadequate discriminative power to discern between poor students and good students.
Among 24 studied MCTs, histology exam had highest number of questions with highest high and high level of discrimination index(70%) while orthopedics exam had lowest number of questions with highest and high level of discrimination index(33%).
Orthopedics exam had highest number of questions with Negative discrimination Index (21%). Discrimination Index for 20% of questions of semiology exam was equal to 0.

Discussion

According to the results of this study, more than 50% of MCTs given in second semester of 2000-2001 educational year were greater than 0.7. Having subjects respond reliably on a measure is a great start, but there is another concept needed to get down really well. That's validity. There are many kinds of validity, but they all refer to whether or not what you are manipulating, or what you are measuring, truly reflect the concept you think it does. Therefore, although acceptable level of reliability is necessary for confirming the validity of a MCT but it should not be looked as an enough evidence for validity. Evaluation of MCTs' Validity should be also based on professional judgments given by experts.
Factors such as environmental and psychological factors may affect outcomes of a MCT which students are taken. such factors should also be included in assessments for more clear and valid results. in this survey standard deviations were calculated . Calculation of standard deviations helps us to determine upper and lower limits of students' scores. For more clear assessments, students' score should be considered as a score ranged between upper and lover limits of given Significance interval. In this case, students' scores are calculated as follows:
Score=score±SD(1,2)
Low amounts of SD indicate low level of standard error of measurement and high level of reliability.
In addition, item difficulty and discrimination index were calculated for each one of MCQs because it is well known fact that quality of a MCT is dependent on its items' quality.
An ideal MCQ should have item difficulty equal to 0.5 and discrimination index equal to +1(1).
In this survey, only 16.5 % of MCQs' difficulty ranged between 0.5 and 0.6 and 52.2 % of MCQs' had inappropriate item difficulty. Another fact that should be mentioned here, is the existence of some clues in a MCQ that may guide the respondents with inadequate knowledge on the question to choose the correct statement.
Brazo(1984) showed that 44% of 1220 questions written by faculty members, have some kind of such clues and over 70% of respondents were able to answer these questions correctly.
Our finding indicate that only 33% of studied MCQs have desirable or acceptable item difficulty and discrimination indices both and 34.9% of those have no desirable or acceptable item difficulty neither acceptable discrimination index.
Analysis of tools that are frequently used to facilitate the evaluation process helps us to improve their quality and lead us to modify items and test that have not been properly designed. Modified and corrected test can provide us more clear information about the students' educational achievement and quality of teaching in an academic setting.
However, Further studies should be undertaken for developments the methods for evaluation of validity, assessment of distracters and structural principles in MCTs .

References

1.Seyf AA. Measurement and assessment tools in education. Daavaran publ, Tehran, Iran, 1999[Farsi].
2.Gilbert JJ. Instruction for training of the primary health care providers.1st ed. Tehran: academic publication center,1985.[Farsi translation by Arfaa F.]
3.Amini Nik S, Azoudi P, Djahanpour F. Study of methods implied by faculty members of Bousher Unversity of Medical Science for assessing students' academic achievement.Bulletin of 4th congress of medical education 2000:13-14[Farsi].
4.Arab M, Assessment of multiple choice questions in Hamedaan Medical University during second semester of 1999-2000 educational year. Bulletin of 4th congress of medical education 2000:51[Farsi].

5.Yuosefi Mashoof R. Study of current methods used for evaluating knowledge and practice of the medical students. Bulletin of 4th congress of medical education 2000:74-75 [Farsi].
6.Mavis BE, Cole BL, Hoppe RB, A survey of students assessment in U.S medical schools: The balance of Breath versus fidelity. Teaching and Learning in Medicine 2001;13(2):74-79.
7.Tabatabaee MSK,Toosi MHB, Modaber MD, Ebrahimzadeh S, Toosi VB, toosi KB.Study of students, attitude on methods implied for assessing their educational achievements during clinic clerkship [Dissertation].Mashad Unversity of Medical Science 2001.
8.Matlock HS. Basic concepts in item and test analysis.1997[Online]. Available form URL:http://ericae.net/ft/tamu/Espy.htm No:ED406441.
9.Denti JA, Harden RM. A practical guide for medical teachers.Churchill Livingstone, London, England ,2001.
10.Shoer LA. Measurement and assessment in education. Besat Publ, Tehran, Iran, 1998.[Farsi translation by Ganji H].