Eldan Kapur[1,2], Almira Džuvo[1] and Maida Šahinović[1]

Assessment of functional anatomy knowledge on Faculty of sport and physical education with one-best answer multiple choice questions - an evaluation of the difficulty and discrimination indices

[1]*Faculty of Medicine, University of Sarajevo, Bosnia and Herzegovina*
[2]*Faculty of Sport and Physical Education, University of Sarajevo, Bosnia and Herzegovina*

Original scientific paper

Abstract
This paper reports the relationship between the difficulty index and the discrimination index of one-best answer type multiple-choice questions (MCQ's) in a functional anatomy paper for the students of year I of an undergraduate faculty programme of Faculty of sport and physical education. We included in this study the MCQ paper from the 1. partial examination of academic session 2009/2010. There were 104 students who sat for the 1. partial examination in session 2009/2010. This examination covered anatomy of skeletal, articular and muscular systems. There was a wide distribution of item difficulty indices in all the MCQ papers analysed. Mean difficulty index was 50,84 ± 17,30 with range 21,4 - 85,7 and mean discrimination index was 0,40 ± 0,21 with range -0,07 – 0,71. On average, 13,3 % of the MCQ items in functional anatomy paper were "very easy" (difficulty index > 70), while 13,3 were"very difficult" (difficulty index < 30). About 75 % of these very easy/difficult items had "very poor" or even negative discrimination. Our finding indicate that 73,4 % of studied MCQ's have recommended difficulty index and 76,7 % of questions have good or excellent discrimination index. 26,6 % of analyzed questions are too easy or too difficult and 13,3 % of questions have poor discrimination.

Key words: **MCQ, Item analysis, Difficulty index, Discrimination index, Functional Anatomy test**

Sažetak
U ovom radu analizirali smo odnos između indeksa težine i indeksa diskriminacije pitanja višestrukog izbora po principu jednog tačnog odgovora na ispitu iz funkcionalne anatomije u prvoj godini dodiplomskog studija na Fakultetu sporta i tjelesnog odgoja. U studiju su uključena 104 studenta koja su pristupila prvom parcijalnom ispitu u školskoj 2009/2010. godini. Parcijalnim ispitom je evaluirano znanje iz anatomije koštanog, zglobnog i mišićnog sistema. Utvrđena je široka distribucija vrijednosti indeksa težine i indeksa diskriminacije pitanja koje smo analizirali. Srednja vrijednost indeksa težine bila je 50,84 ± 17,30 sa rasponom od 21,4 - 85,7, a srednja vrijednost indeksa diskriminacije bila je 0,40 ± 0,21 sa rasponom od -0,07 – 0,71. 13,3 % pitanja višestrukog izbora na testu iz funkcionalne anatomje su bila "veoma laka" (indeks tzežine > 70), a 13,3 % "veoma teška" (indeks težine < 30). Oko 75 % ovih veoma lakih/teških pitanja imala su veoma slabu ili čak negativnu diskriminaciju. Naši nalazi su pokazali da je 73,4 % pitanja višestrukog izbora imalo preporučeni indeks težine, a 76,7 % dobar ili odličan indeks diskriminacije. 26,6 % analiziranih pitanja su prelaka ili preteška, dok je 13,3 % imalo slabu diskriminaciju.

Ključne riječi: **Pitanja višestrukog izbora, Analiza pitanja, Indeks težine, Indeks diskriminacije, Funkcionalna anatomija**

Introduction

The multiple-choice question (MCQ) is the most common type of written test item used in undergraduate, graduate, and postgraduate education (Farley, 1989). MCQ's can be used to assess a broad range of learner knowledge in a short period of time. Because a large number of MCQ's can be developed for a given content area, which provides a broad coverage of concepts that can be tested consistently, the MCQ format allows for test reliability. If MCQ's are drawn from a representative sample of content areas that constitute predetermined learning outcomes, they allow for a high degree of test validity. Critics of MCQ's argue that higher-level learning can not be tested with MCQ's. However, this criticism is more often attributed to flaws in the construction of the test items rather than to their inherent weakness. Appropriately constructed MCQ's result in objective testing that can measure knowledge, comprehension, application, and analysis (Kemp, Morrison, & Ross, 1994).

The principles of writing effective MCQ's are well documented in educational measurement textbooks, the research literature, and test-item construction manuals designed for educators (Gronlund, 1998; Haladyna, Downing, & Rodriguez, 2002; Case & Swanson, 1998). Yet, a recent study from the National Board of Medical Examiners showed that violations of the most basic item-writing principles are very common in medical education tests (Jozefowicz, Koeppen, Case, Galbraith, Swanson, & Glew, 2002). Items on a multiple-choice test consist of a stem, which is folowed by a correct answer as well as four to five 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 plausabile. (Hobsley, 1999). Fowell Southgate & Bligh (1999). Multiple-choice questions, if designed carefully, can achieve satisfactory, reliability, efficacy, fidelity, and they have great educational impact (does the test instrument stimulate learning).

Study of MCT's used frequently in academic settings showed that large number of these tests are not properly wrriten and planed (Schultheis, 1998; Schuwirth & van der Vleuten, 2004). After the test is given, it is important to perform a testitem analysis to determine the effectiveness of the questions. Most machine-scored test printouts include statistics for each question regarding item difficulty, item discrimination, and frequency of response for each option. This kind of analysis gives you the information you need to improve the validity and reliability of your questions. Item difficulty is determined from the percentage of students who answered each item correctly, with the goal being to construct a test that contains only a few items that more than 90% or less than 30% of students answer correctly. Optimally, difficult items are those that about 50%–75% of the students answer correctly.

Items are considered low to moderately difficult if between 70% and 85% of the students select the correct response.

Item discrimination refers to the percentage difference in correct responses between two groups of students (generally referring to students in the top 27% and the lower 27%). The discrimination ratio for an item will fall between −1.0 and +1.0. The closer the ratio is to +1.0, the more effectively that item distinguishes students who know the material (the top group) from those who don't (the bottom group). Ideally, each item will have a ratio of at least 0,5 (Davis, 1993). An item with a discrimination of 0,6 or greater is considered a very good item, whereas a discrimination of less than 0,24 indicates a marginal or low discrimination item that needs to be revised (Vydareny, Blane, & Calhoun, 1986). An item with a negative index of discrimination indicates that the poor students answer correctly more often than do the good students, and such items should be avoided.

Methods

The MCQ items were first written by individual teachers and vetted at their respective department for content accuracy. The vetted questions (newly written or extracted from the bank) were then chosen by the departmental head before being submitted to the students on partial examination.
MCQ items taken from past Year I examinations were analysed to illustrate the principles and methods of scoring and calculation level of difficulty and power of discrimination. We included in this study the MCQ papers from the 1. partial examination of academic session 2009/2010. There were 104 students from Faculty of sport and physical education who sat for the 1. partial examination in session 2009/2010. This examination covered anatomy of skeletal, articular and muscular systems.
The MCQ paper contained 30 questions and was to be completed in 30 minutes. Each question consisted of a stem and 5 completing phrases, with one best answer. A correct response to an item was awarded 1 mark, and a no-attempt or blank response was given 0 marks. There is no deduction of marks for wrong answer. The highest possible score is 30, and lowest 0.
The mean and standard deviation of the original scores were computed by standard statistical methods. The results of students' performance in these MCQ test were then used to determine the difficulty index and discrimination index of each MCQ item in the respective test. All the 104 students were ranked in order of merit from the highest score to the lowest score (6). According to Ebel, R.L. (1965) the first 27% of the students constitute the high group (H), and the last 27% the low group (L).The difficulty index and discrimination index were then calculated according to Guilbert J-J (1957):

$$\text{DIFFICULTY INDEX} = \frac{H + L}{N} \times 100$$

where the H is the number of correct answer in the high group, L is the number of correct answer

$$\text{DISCRIMINATION INDEX} = \frac{2(H - L)}{N} \times 100$$

in low group and N is the total number of students in both groups. Hence, the higher difficulty index value, the lower is the difficulty, and the greater the difficulty of an item, the lower is its index. The higher the discrimination index, the better the item can determine the difference, i.e., discriminate, between those students with high test scores and those with low ones.

Results

Table 1 show the distribution of the original scores on 1. functional anatomy partial examination which was necessary for establishing two groups of students, generally referring to students in the top 27% and the lower 27%, according to Ebel (1967).

Table 1. *The distribution of the original scores*

Original scores	Number of students	% of students
6 - 9	16	15,5
10 - 13	12	11,5
14 - 17	42	40,5
18 - 21	24	23
22 - 25	10	9,5
Total	104	100

Mean difficulty index was 50,84 ± 17,30 with range 21,4 – 85,7 (Table 2 and 5). In 73,4 % of the MCQ's difficulty index was acceptable (table 3), with 16,7 % of ideal questions (difficulty index 50 – 60) (Table 4, Figure 1). On the other hand 13,3 % of questions were too difficult (difficulty index < 30, mean 24,1 ± 5,4), while 13,3 were to easy (difficulty index >70, mean 78,6 ± 6,5) (Table 3).

Table 2. *The difficulty index and discrimination index of the 30 MCQ's*

Question number	Difficulty index	Discrimination index
1.	39,3	0,64
2.	82,1	0,36
3.	71,5	0,28
4.	39,3	0,36
5.	60,7	0,64
6.	60,7	0,5
7.	64,3	0,43
8.	21,4	0,29
9.	64,3	0,57
10.	57,1	0,71
11.	53,6	-0,07
12.	46,4	0,64
13.	75	0,36
14.	28,6	0,29
15.	35,7	0,43
16.	39,2	0,64
17.	85,7	0,14
18.	46,4	0,36
19.	17,9	-0,07
20.	35,7	0,29
21.	60,7	0,5
22.	64,3	0,43
23.	67,9	0,36
24.	28,6	0
25.	42,9	0,29
26.	50	0,43
27.	50	0,57
28.	42,9	0,57
29.	50	0,71
30.	42,9	0,43

Table 3. *Distribution of the difficulty index of the 30 MCQ's*

Difficulty index	Interpretation*	Question number	Number of Questions	% of Questions
>70	Too easy	2,3,13,17	4	13,3
30 - 70	Average, recommended	1,4,5,6,7,9,10,11, 12,15,16,18,20, 21,22,23,25,26, 27,28,29,30	22	73,4
<30	Too difficult	8,14,19,24	4	13,3
		TOTAL	30	100

Table 4. *Distribution of ideal guestions of the 30 MCQ's*

Difficulty index	Interpretation*	Question number	Number of Questions	% of Questions
50 - 60	Ideal questions	10,11,26,27,29	5	16,7
Difficulty index	Interpretation*	Question number	Number of Questions	% of Questions
50 - 60	Ideal questions	10,11,26,27,29	5	16,7

Figure 1. *Amplitude of difficulty index*

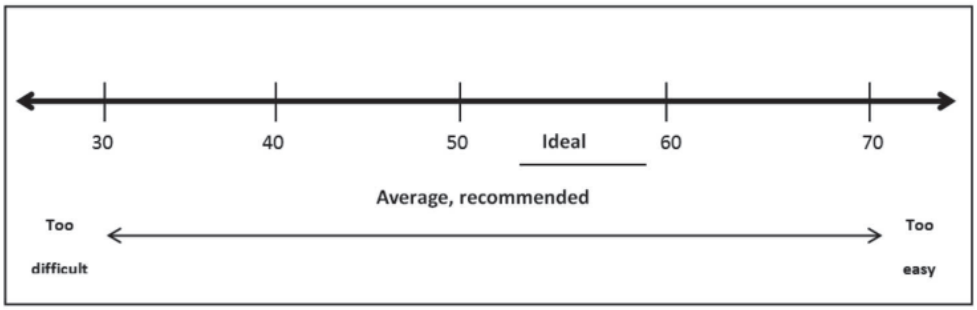

Table 5. *Mean difficulty index MCQ paper analysed for 1. partial examinations (n = 30 test items)*

Academic session	Partial examination	Number of students	Difficulty index	
			Mean ± SD	Range
2008/2009	2	52	50,84 ± 17,30	21,4 - 85,7

Mean discrimination index was 0,40 ± 0,21, with range – 0,07 – 0,71 (Table 7). Analysis of MCQ's shows that we had 86,7 % of questions with good or excellent discrimination index (Table 6, Figure 2). 13,3 % of questions had low discrimination index (<0,15) (Table 6). Note that two questions had a negative discrimination index (question 11, -0,07; question 19, -0,07) (Table 2).

Table 6. *Distribution of the discrimination index of the 30 MCQ's*

Discrimination index	Interpretation*	Question number	Number of Questions	% of Questions
≥0,35	Excellent discrimination	1,2,4,5,6,7,9,10,12 13,15,16,18,21,22, 23,26,27,28,29,30	21	70
0,25 – 0,34	Good discrimination	3,8,14,20,25	5	16,7
0,15 – 0,24	Marginal discrimination		0	0
<0,15	Poor discrimination	11,17,19,24	4	13,3
		TOTAL	30	100

Figure 2. *Amplitude of discrimination index*

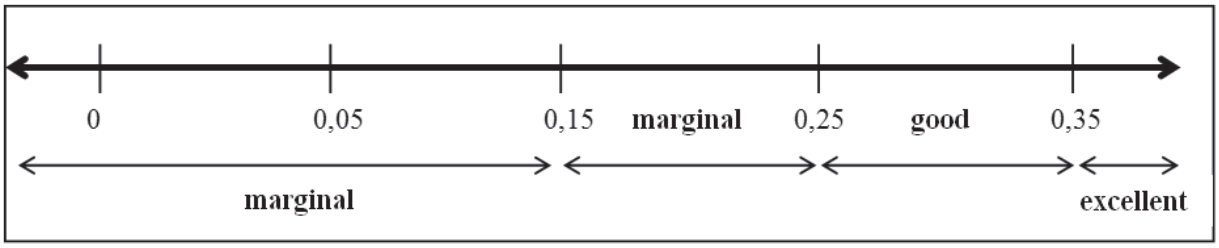

Table 7. *Mean discrimination index MCQ paper analysed for 1. partial examinations (n = 30 test items)*

Academic session	Partial examination	Number of students	Discrimination index	
			Mean ± SD	Range
2008/2009	2	52	0,40 ± 0,21	-0,07 – 0,71

Table 8 shows that difficult and easy questions account gor great majority (75 %) of the MCQ's of low discrimination index (<0,15). Note that of 26 MCQ's with good or excellent discrimination value, 84,6 % were acceptable questions (difficulty index 30 – 70), while 14,4 % were difficult and easy questions (Table 9).

Table 8. *Relation between difficulty index and discrimination index (MCQ's of low discrimination value; discrimination index ≤ 0,24)*

Difficulty index	Number of Questions	% of Questions
≥70	1	25
30-70	1	25
<30	2	50
Total	4	100

Table 9. *Relation between difficulty index and discrimination index (MCQ's of good or excellent discrimination value; discrimination index ≥0,25).*

Difficulty index	Number of Questions	% of Questions
≥70	2	7,7
30-70	22	84,6
<30	2	7,7
Total	26	100

Discussion

As with other health professional training, the effective measurement of knowledge is an important component of both, education and practice (Case & Swanson, 1998).). Furthermore, the methods used to analyse the evidence resulting from the tasks (i.e., interpretation) need to be aligned with the aspects of achievement that are to be assessed (i.e., cognition) and the tasks used to collect evidence about students' achievement (i.e., observation). Therefore, it is important for us to evaluate our MCQ items to see how effective they are in assessing the knowledge of our students in the functional anatomy, and in predicting their total test scores. Many methods have been developed to calculate the discriminatory power of individual items; e.g., difficulty index, discrimination index, biserial correlation coefficient, point biserial correlation coefficient, and phi coefficient (Kelley, 1939). The basic purpose of the methods is to give a numerical value to the relationship between scores for the total MCQ test and the score for a single item. This numerical value is the difficulty index and index of the discriminatory effectiveness of the item. Although there are various similar ways of calculating the discrimination index, we used the simplified technique of selecting the upper and lower 27%, which have been demonstrated by Kelley (1939) to be the most efficient fraction. Difficulty and discrimination indices are important in that poor discriminatory items are a valuable signpost towards ambiguous wording, grey areas of opinion and perhaps, even wrong keys. However, we must recognise that there may be other factors that need to be taken into account when using difficulty and discrimination indices to categorise MCQ's as "good" or "bad"(Guilbert, 1957). In this paper, item difficulty and discrimination indices were calculated for each one of MCQ's because it is well known fact that quality of a MCT is dependent of its items quality. An ideal MCQ should have item difficulty in range from 50 to 60, and discrimination index more than 0,35 (Kemp, Morrison, & Ros, 1994; Hobsley, 1999; Schulthei, 1998). In this survey, 16,7 % of MCQ's difficulty ranged between 50 and 60 (ideal questions), but 73,4 % of MCQ's had appropriate difficulty index. Our finding indicate that 73,3 % of questions

recommended or acceptable item difficulty and discrimination indices both, and 26,7 % of those had no acceptable item difficulty neither acceptable discrimination index. Test items with very poor discrimination indices should be reviewed by the respective disciplines. It serves as an effective feedback to the departments concerning their educational activities. When a test item appears to be very difficult (i.e., difficulty index is very small), it may be that the topic tested is inappropriate at this stage of students' training, or that it is not taught well or not taught at all in this particular academic session. It is interesting to note that despite the lack of written guidelines or the use of item analysis to help the lecturer in constructing the MCQ test items, a consistent level of test difficulty (and hence, standard) appears to be maintained from term to term and from year to year.

Analysis of tools that are frequently used to facilitate the evaluation process help us to improve their quality and lead us to modify items and test that have not been properly designed (Fowell, Southgate, & Bligh, 1999). Modified and corrected questions can provide more clear information about the students educational achievement and quality of teaching and curriculum in an academic setting (Kavel, 2003).

Conclusion

There are two important questions that always come to the mind of an examiner when he/she sets a MCQ. Firstly, is the MCQ too difficult, too easy or just about right? The second question is closely related to the first, and that is whether the MCQ could differentiate "good" students from "poor" students. Obviously questions which are too difficult or too easy have poor discrimination value. It is difficult and very often impossible to know the answers of these two questions before the test is administered to a group of students. Hence it is important for teachers to find out by calculating the difficulty and discrimination indices of all the MCQ's after marking the test paper. By analysing difficulty and discrimination indices of a particular MCQ, we could evaluate the response of the students to that particular question so that we could ascertain not only whether that MCQ is too difficult, too easy or just about right, but also it could differentiate "good" students from "poor" students. Finally, critical evaluation of MCQ's results would also enable the teachers to identify areas of deficiency that need remediation or further learning, determine final grades or make promotion decisions and identify areas where course/curriculum is weak.

References

Farley, JK. (1989). The multiple choice test: writing the questions. Nurse Educ, 14:10–12, 39.

Kemp, JE., Morrison, GR., & Ross SM. (1994). Developing evaluation instruments. In: Designing effective instruction. New York, NY: MacMillan College Publishing, 180–213.

Gronlund, NE. (1998). Assessment of student achievement. Boston, Mass: Allyn & Bacon.

Haladyna, TM., Downing, SM., & Rodriguez, MC. (2002). A review of multiple-choice item-writing guidelines. Appl Meas Educ, 15:309–333.

Case SM, Swanson DB. (1998). Constructing written test questions for the basic and clinical sciences. Philadelphia, Pa: National Board of Medical Examiners.

Jozefowicz, RF., Koeppen, BM., Case, S., Galbraith, R., Swanson, D., & Glew, H. (2002). The quality of in-house medical school examinations. Acad Med, 77: 156–161.

Hobsley, M. (1999). Counting apples with oranges: a limitation of the discrimination index. Med Educ, 33:192-6.

Fowell, SL., Southgate, LJ., & Bligh, JG. (1999). Evaluating assessment: the missing link? Med Edu, 33:276-81.

SchultheisNM. (1998). Writing cognitive educational objectives and multiple-choice test questions. Am J Health Syst Pharm, 55:2397–2401.

Schuwirth , LW., & van der Vleuten CP. (2004). Different written assessment methods: what can be said about their strengths and weaknesses? Med Educ, 38:974–979.

Davis, BG. (1993). Multiple-choice and matching tests. In: Tools for teaching. San Francisco, Calif: Jossey-Bass, 262–271.

Vydareny, KH, Blane, CE, & Calhoun, JG. (1986). Guidelines for writing multiple-choice questions in radiology courses. Invest Radiol, 21:871–876.

Ebel, RL. (1965). Measuring education achievement. New Jersey, USA:Prentice Hall.

Guilbert, J-J. (1957) Educational handbook for healt personel. WHO Offset Publication No.35 Geneva, World Health Organisation.

Hubbard, JP, & Clemans, WV. (1961). Multiple-choice Examinations in Medicine: A Guide for Examiner and Examinee. London: Lea & Fabiger.

Kelley, TL. (1939). The selection of upper and lower groups for the validation of test items. J Educ Psychol, 30:17-24.

Kavel. T., MS et al. (2003). Analytic assessment of multiple choice tests. J Med Educ, (2):87-91

Submitted: February 25, 2010.
Accepted: May 19, 2010.

Correspodence to:
Ass. Prof. Eldan Kapur, PhD
Faculty of Medicine, University of Sarajevo
Čekaluša 90,
71 000 Sarajevo, Bosnia and Herzegovina
Phone: +387 33 203 670
E-mail: eldan_kapur@hotmail.com