The levels of difficulty and discrimination indices and relationship between them in four-response type multiple choice questions of pharmacology summative tests of Year II M.B.B.S students

Bharti N Karelia, Ajita Pillai, Bhavisha N Vegada

Background: Item analysis is the process of collecting, summarising and using information from students' responses to assess the quality of test items. Difficulty index (P) and discrimination index (D) are two parameters which help to evaluate the standard of MCQ questions used in an examination, with abnormal values indicating poor quality.

Methods: In this study 200 test items of 10 MCQ tests from 2008 to 2012 were selected and analysed to obtain their difficulty and discrimination indices. The relationship between the difficulty index and discrimination index for each test item was determined by Pearson correlation analysis.

Results: Mean difficulty index scores of the individual summative tests were in the range of 47.17% to 58.08%. Twenty nine percent of total test items crossed the difficulty index of 70% indicating that those items were easy for the students. Seventy eight percent of the test items showed acceptable (> 0.2) discrimination index. Forty six percent of the test items showed excellent discrimination index. Discrimination index correlated poorly with difficulty index (r=0.11). The correlation is insignificant at 5% (p>0.10).

Conclusion: A consistent level of test difficulty and discrimination indices was not maintained from 2008 to 2012 in all the ten summative type A MCQ tests.

IeJSME 2013 7(2): 41-46

Keywords: Difficulty index Discrimination index Item analysis, Summative tests, Type A MCQ

Introduction

The educational objectives in medicine as well as in other discipline are generally allotted to three 'domains'- cognitive, psychomotor and affective. Hence, medical examination should be designed to answer whether an undergraduate has achieved the above educational objectives by answering the following three questions: (1) what does he know? (cognitive) (2) what can he do? (psychomotor) and (3) what sort of person is he? (affective). Regrettably the current medical examination system still could not completely answer these questions.[1]

Objectivising evaluation is becoming increasingly more important in the field of education, both for summative and formative purposes, as has been again and again emphasised by guidelines published by several universities. One method of achieving this purpose is the widespread use of objective written items, and the most popular form of which is the multiple choice question (MCQ).[2] Item analysis is the process of collecting, summarising and using information from students' responses to assess the quality of test items.[3] With greater usage of MCQ for this purpose, the importance of item analysis for question banking has emerged and at present, item analysis is largely used for creating a viable question bank of MCQs. In addition, many teachers use MCQ to assess class performance as a part of formative evaluation.[2]

Designing MCQs is a complex and time consuming process in a multidisciplinary integrated curriculum. MCQs are used mostly for comprehensive assessment at the end of a semester or academic session and provide feedback to the teachers on their educational action. Having constructed and assessed a test, a teacher needs to know how good the test questions are and whether the test items were able to reflect students' performance in the course related to learning. Because of their versatile character, MCQs are the most commonly used tool for assessing the knowledge capabilities of medical students.[3]

There are different types of MCQs like five-response, four-response, three-response and true/false or two-response.[4] One of the major concerns in the construction of test items for an examination is ensuring the reliability of the test items. The item statistics can help to determine those items that are good and those that need improvement or deletion from a question bank. It allows any aberrant item to be given attention and reviewed.

Department of Pharmacology, P.D.U. Govt. Medical College, Rajkot-360001Gujarat, INDIA

Address for correspondence:

Dr Bharti N Karelia, Associate Professor, Department of Pharmacology, P.D.U. Govt. Medical College, Rajkot-360001 Gujarat, INDIA
Email: nirbharti_karelia@yahoo.co.in

One of the most widely used method in investigating the reliability of test item has been Classical Test Theory (CT) item analysis. Item difficulty index is the first item characteristic in CT theory to be determined. This is a common practice as tests are often not regarded as a reliable measure of student's performance due to misfit of item difficulty with the ability of the students. In addition to item difficulty, item discrimination is an important index.[3]

The Medical Council of India (MCI), as required by the Regulation on Graduate Medical Education 1997, made it mandatory for all medical colleges to establish Medical Education Units (MEUs) or departments in order to enable faculty members to avail modern education technology for teaching. In order to boost this activity MCI has been conducting Faculty development programmes through selected regional centers since July 2009. These centres have trained manpower in Modern Education Technologies (MET).[5] Item analysis is a part of MET training. The objectives of the present research study were to analyse the quality of MCQs of Pharmacology summative tests of II M.B.B.S Students and to determine whether there is any relationship between the item difficulty and item discrimination indices of these MCQ items.

Methods

The marking format for the 1st and 2nd terminal examination for pharmacology subject at our institution consists of 80 marks theory and 50 marks practical examination. Theory examination consists of 20 multiple choice questions of 1 mark each. Two terminal examinations were held each year, A for 1st terminal and B for 2nd terminal examination.

Data collection

MCQ items were taken from the 10 summative test papers from the years 2008-2012. A total of 200 test items were selected for the item analysis. Each MCQ consisted of a stem and four choices and the students were to select one best answer from these four choices. A correct response to an item was awarded 1 mark, while an incorrect response would result in negative 0.25 marks and a no- attempt or blank response was given no marks.

Item analysis

The results of the examinee's performance in the summative tests were used to analyse the difficulty index and discrimination index of each MCQ item. First scoring of the whole test for all students was done, then students were ranked based on their total score. The bottom third were taken as low achievers and upper third as high achievers. The difficulty index is calculated as percentage of the total number of correct responses to the test item.[3] It is calculated using the formula P= (H+L/N)*100, where P is the item difficulty index, H is the number of students answering the item correctly in the high achieving group, L is the number of students answering the item correctly in the low achieving group and N is the total number of students in two groups. An item was considered difficult when the difficulty index value was less than 30% and considered easy when the index was more than 70% and the value between 30-70% was acceptable (between 50-60% are ideal).[2] The item discrimination index measure the differences between the percentages of students in the upper group with that of the lower group who obtained the correct response.[3] The discrimination index was calculated using the formula d= (H-L/N)*2. Items with a discrimination index between 0.25-0.35 were considered good, those with indices more than 0.35 were excellent, between 0.20-0.24 were acceptable and below 0.20 were poor.[2]

Statistical analysis

All data were expressed as mean ± SD. The relationship between the item difficulty index and discrimination index for each test item was determined by Pearson correlation analysis and the coefficient of determinates is given by r.[6] A P-value of < 0.05 was considered to be statistically significant.

Results

From Table 1, 24±8.43% (Mean ± SD) of the 20 MCQ items in each paper had a difficulty index of >70% ("Very easy" items), 15±7.07% items had a difficulty index of <30% ("Very difficult" items), while 61±8.43% items had a difficulty index between 30 - 70% ("Acceptable" items). An average 20±4.08% of the 20 MCQ items in each paper had a difficulty index between 50 - 60% ("Ideal" items).

Table 1: Proportion of "Ideal" (P between 50 - 60%), "Acceptable" (P between 30 - 70%), "Very Easy" (P >70%) and "Very Difficult" (P <30%) items for each MCQ paper analysed (n=20 test items)

Academic Year	Internal Examination A = 1st internal B = 2nd internal	"Ideal" items % (no.)	"Acceptable" items % (no.)	"Very easy" items % (no.)	"Very difficult" items % (no.)
2008	A	25 (5)	55 (11)	20 (4)	25 (5)
	B	20 (4)	60 (12)	30 (6)	10 (2)
2009	A	20 (4)	70 (14)	20 (4)	10 (2)
	B	25 (5)	60 (12)	30 (6)	10 (2)
2010	A	20 (4)	65 (13)	30 (6)	05 (1)
	B	20 (4)	55 (11)	25 (5)	20 (4)
2011	A	20 (4)	55 (11)	25 (5)	20 (4)
	B	10 (2)	55 (11)	20 (4)	25 (5)
2012	A	20 (4)	55 (11)	35 (7)	10 (2)
	B	20 (4)	80 (16)	05 (1)	15 (3)
	Mean ± S.D. (%)	20 ± 4.08	61 ± 8.43	24 ± 8.43	15 ± 7.07

On average, 46±9.37% (Mean ± SD) of the 20 MCQ items in each paper had a discrimination index of >0.35 ("Excellent" items), 22±11.11% items had a discrimination index between 0.25 to 0.35 ("Good" items), 10±5.27% items had a discrimination index between 0.20 to 0.24 ("acceptable" items), while 22±8.23% items had a discrimination index of <0.20 ("Poor" items), as shown in Table 2.

Table 2: Proportion of "Excellent" (D > 0.35), "Good" (D between 0.25 - 0.35), "Acceptable" (D between 0.20 - 0.24) and "Poor" (D < 0.20) items for each MCQ paper analysed (n=200 test items)

Academic Year	Internal Examination A = 1st internal B = 2nd internal	"Excellent" items % (no.)	"Good" Items % (no.)	"Acceptable" items % (no.)	"Poor" items % (no.)
2008	A	40 (8)	20 (4)	15 (3)	25 (5)
	B	40 (8)	15 (3)	20 (4)	25 (5)
2009	A	60 (12)	05 (1)	05 (1)	30 (6)
	B	50 (10)	40 (8)	05 (1)	05 (1)
2010	A	60 (12)	15 (3)	10 (2)	15 (3)
	B	35 (7)	30 (6)	05 (1)	30 (6)
2011	A	50 (10)	25 (5)	10 (2)	15 (3)
	B	35 (7)	25 (5)	10 (2)	30 (6)
2012	A	40 (8)	35 (7)	05 (1)	20 (4)
	B	50 (10)	10 (2)	15 (3)	25 (5)
	Mean ± S.D (%)	46 ± 9.37	22 ± 11.11	10 ± 5.27	22 ± 8.23

Further analysis of the data indicated that there was a wide spectrum of level of difficulty and discriminating power among the MCQ items in all the papers. The difficulty index of these papers ranged from as low as 5 – 26% ("extremely difficult" items) to as high as 84.48 – 97.92% ("extremely easy" items). The discrimination index ranged from as low as -0.25 – 0.15 ("Poor" items) to as high as 0.52 – 0.69 ("Excellent" items) as shown in Table 3. The mean difficulty index of all the tests were found in the range between 47.17 – 58.08% and the mean discrimination index ranged between 0.29 – 0.38 as shown in Table 3.

When difficulty index was analysed along with discrimination index, 22% of the test items with poor discrimination index had a difficulty index ranging between 5 – 97.92%. Forty six percent of the test items with excellent discrimination index had a difficulty index ranging between 25.93 – 80.00%. Pearson correlation between difficulty and discrimination indices showed that discrimination index correlate poorly with difficulty index (r=0.11).The correlation is insignificant at 5% (p>0.10).

Table-3: Mean Difficulty index (P) and Discrimination index (D) for each MCQ paper analyzed (n=200 test items)

Academic Year	Internal Examination A = 1st internal B = 2nd internal	No. of students	Difficulty index (P) (%)		Discrimination index (D)	
			Mean ± S.D	Range	Mean ± S.D	Range
2008	A	52	50.87 ± 23.40	9.62 to 88.46	0.31 ± 0.17	-0.08 to 0.69
	B	54	56.57 ± 20.12	24.67 to 90.74	0.32 ± 0.18	0.07 to 0.63
2009	A	48	52.08 ± 20.75	12.50 to 97.92	0.32 ± 0.21	-0.25 to 0.63
	B	54	56.39 ± 18.94	20.37 to 85.19	0.35 ± 0.10	0.15 to 0.52
2010	A	60	58.08 ± 19.33	11.67 to 90.00	0.38 ± 0.17	0.10 to 0.67
	B	58	52.67 ± 21.07	15.52 to 84.48	0.29 ± 0.15	0.07 to 0.52
2011	A	50	57.20 ± 19.52	26.00 to 96.00	0.35 ± 0.18	0.04 to 0.68
	B	48	50.10 ± 24.14	8.33 to 89.58	0.31 ± 0.17	0.04 to 0.63
2012	A	60	58.00 ± 24.00	5.00 to 88.33	0.30 ± 0.14	-0.10 to 0.57
	B	60	47.17 ± 19.77	5.00 to 88.33	0.32 ± 0.18	0.03 to 0.60

Discussion

As with other health professional training, the effective measurement of knowledge is an important component of both medical education and practice. 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 medical students, and in predicting their total test scores.[7]

Many methods have been developed to calculate the discriminatory power of individual items; e.g. discrimination index, biserial correlation coefficient, point biserial correlation coefficient, and phi coefficient. 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 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 to be the most efficient fraction. However, the main limitation of the use of this method in estimating discrimination power is that it cannot be used for small sample size.[7]

Out of 12 summative tests conducted from 2008 to 2012, the mean difficulty index score of the individual tests ranged from 47.17 - 58.08% which indicated that all items had an acceptable level. None of the tests had the mean difficulty index value more than 70% and lower than 30% which meant that in these tests, there were neither very easy nor very difficult items. This observation was opposite to a study of item analysis of a type A MCQ of pre clinical semester 1 multidisciplinary summative tests reported by Mitra *et. al.* (2009), who found that two tests had mean difficulty index value more than 80% and had easy MCQs where most of the students got full score in the tests.[3] Our results showed that 24% of the total test items had difficulty index score crossing 70%. Li *et. al.* (1999) who performed item analysis of basic medical science items of registered nurse licensure examination in Taiwan, found item difficulty in the range of 10 - 93% with a mean of 48%.[3] Our results showed that item difficulty index ranged between 5 - 97.92%.

Any discrimination index of 0.2 or higher is acceptable and the test item would be able to differentiate between the weak and good students.[2] In the present study, it was shown that 78% of the MCQ from ten tests had a discrimination index of more than 0.2. Thus it showed that most of the MCQ used in all these summative tests were good or satisfactory questions which did not need any modification or editing as these questions were able to differentiate good and weak students. Three (3) out of 10 tests showed mean discrimination index equal to or more than 0.35, indicating that these MCQ items were excellent test items for differentiating between poor and good performers.

Sim abd Rasiah (2006)[7] found that the maximum discrimination occurred with difficulty index between 40 – 74%. In the present study, 46% of the test items with difficulty index between 25.93% and 80% had excellent discrimination index. Subjective judgment of item difficulty by item writer and the vetting committee may allow faulty items to be selected in the item bank. Items with poor discrimination index and too low or too high difficulty index should be reviewed by the respective content experts. This serves as an effective feedback to the respective departments in a medical school about the quality control of various tests. When the difficulty index is very small, indicating difficult question, it may be that the test item is not taught well or is difficult for the students to grasp. It also may indicate that the topic tested is inappropriate at that level for the student.[3] The wide scatter of item discrimination values for questions with a similar level of difficulty may reflect that some extent of guessing practices still occurred despite penalty marking.

The quality of test items may be further improved based on action taken in reviewing the distractors by the item writer based on the calculated discrimination and difficulty index values. Some common causes for the poor discrimination are ambiguous wording, grey areas of opinion, wrong keys and areas of controversy. Items showing poor discrimination should be referred back to the content experts for revision to improve the standard of these test items. It is important to evaluate the test items to see how effective they are in assessing the knowledge of the students based on the difficulty and discrimination indices of the test items.

Administration of an objective test and use of item analysis at the end of the period of instruction, sometimes even as small as a single lecture, has great advantages for the teacher. It enables him to get an active feedback from the students and determine areas which require emphasis, reinforcement or an alteration in teaching methodology perhaps using other learning aids. Although every aspect of an instructional exercise cannot be reduced to MCQ, use of items frequently during classroom teaching especially in areas of problematic learning considerably helps the teacher in improving his and his students' performance. In the ranking situation, usually items which have a good positive discrimination and moderate difficulty are chosen. In fact, teachers must aim at getting high facility values and low discrimination indices, as the aim of classroom teaching is not to distinguish between good

and bad students but to ensure that all students have learnt the lesson correctly.[2]

Developing the perfect test is an unattainable goal for anyone in an evaluative position. Even when guidelines for constructing fair and systematic tests are followed, a plethora of factors may enter into a student's perception of the test items. Looking at an item's difficulty and discrimination will assist the test developer in determining what is wrong with individual items. Item and test analysis provide empirical data about how individual items and whole tests are performing in real test situations.[8]

Conclusion

Most of the test items have acceptable levels of difficulty index and excellent discrimination index. The test items that demonstrated excellent discrimination tend to be in the moderately difficult range and those that demonstrated poor discrimination tend to be have wide variety of difficulty index. Discrimination index correlate poorly with difficulty index. The results of this study should initiate a change in the way MCQ test items are selected for any examination and there should be a proper assessment strategy as part of the curriculum development. Much more of these kinds of analysis should be carried out after each examination to identify the areas of potential weakness in the one best answer type of MCQ tests to improve the standard of assessment.

REFERENCES

1. Ho TF, Yip WCL, Tay JSH. The use of multiple choice questions in medical examination: An evaluation of scoring and analysis of results. Singapore Med J 1981; 22(6): 361-7.
2. Ananthakrishnan N. Item analysis-validation and banking of MCQs. In: Ananthakrishnan N, Sethuraman KR, Kumar S, editors. Medical Education principles and practice. 2nd ed. JIPMER, Pondicherry: 131-7.
3. Mitra NK, Nagaraja HS, Ponndurai G et al. The levels of difficulty and discrimination indices in type A multiple choice questions of pre-clinical semester 1 multidisciplinary summative tests. IeJSME 2009; 3(1): 2-7.
4. Understanding item analysis reports. Available from: http://www.washington.edu/oea/service/scanning-scoring/item_analysis.html Accessed: December 20, 2012.
5. Faculty Development programmes. Available from: http://www.mciindia.org/information desk/for colleges/faculty development program.aspx. Accessed: December 20, 2012.
6. Mahajan BK. Methods in biostatistics.7th ed. New Delhi: Jaypee Brothers Medical publishers; 2010.
7. Sim SM, Rasiah RI. Relationship between item difficulty and discrimination indices in true/false-type multiple choice question of a para clinical multidisciplinary paper. Ann Acad med Singapore 2006; 35: 67-71.
8. Item Analysis Assumptions (Difficulty & Discrimination Indexes). Available from: http://com.ksau-hs.edu.sa/eng/images/DME_Fact_Sheets/fs_24.doc. Accessed: 20 December 2012.

Figure-I: Scatter plot showing relationship between difficulty index and discrimination index of items. Also showed is the Pearson Correlation value. Correlation was tested between individual item's difficulty index and discrimination index score.