Original Research Paper

Correlation between difficulty & discrimination indices of MCQs in formative exam in Physiology

Sushma S. Pande[1], Santosh R. Pande[2], Vrushali R. Parate[3], Archana P. Nikam[4], Sushil H. Agrekar[5]

Abstract

Introduction: Difficulty index (P) and discrimination index (D) are the parameters used to evaluate the standard of multiple choice questions (MCQs) in examination. Accordingly the standard of MCQs can be interpreted as excellent, acceptable or poor. This study was intended to find out the standard of MCQs in formative examination in Physiology. The study also intended to find out correlation between P and D.

Materials and Methods: There were 240 MCQ items, taken from the past 4 year batches of 100 students and were analyzed for level of difficulty and discrimination index. The relationship between them for each test item was determined by Pearson correlation analysis using SPSS 11.5.

Results: There was a wide distribution of item difficulty indices (8.57 to 95.71) and discrimination indices (-0.54 to 0.8).The mean difficulty index (P) was 52.53 ± 20.59 and mean discrimination index was 0.30± 0.18. On average, about 23% of the MCQ items were easy (P >70%), while about 15% were difficult (P <30%). The remaining 62% items were within acceptable range (30 to 70%). In all 4% of the items showed negative discrimination and 21% of the items exhibited poor discrimination. The remaining 75% of the items were in the range of acceptable to excellent discrimination. The discrimination index exhibited slight positive correlation with difficulty index (r = 0.191, P=0.003<0.01). The maximal discrimination (D=0.6-0.8) was observed with moderately easy/difficult items (P = 40% - 60%).

Conclusion: The majority (75%) of the items was acceptable as far as difficulty and discriminative indices were concerned. Moderately easy/difficult items had maximal discriminative ability. Too easy and too difficult items gave poor discrimination index. Negative discrimination was observed in only 4% of the items indicating faulty items or incorrect keys.

Keywords: *item analysis, difficulty index, discrimination index, single best response type MCQ, formative tests*

[1] *Professor & Head Department of Physiology, Coordinator, M.E.U., Dr. Panjabrao Alias Bhausaheb Deshmukh Memorial Medical College, Amravati, (MS), India*

[2] *Associate Professor, Department of Anaesthesia, Dr. Panjabrao Alias Bhausaheb Deshmukh Memorial Medical College, Amravati, (MS), India*

[3] *Assitant Professor, Department.of Physiology, Dr. Panjabrao Alias Bhausaheb Deshmukh Memorial Medical College, Amravati, (MS), India*

[4] *Demonstrator, Department of Physiology, Dr. Panjabrao Alias Bhausaheb Deshmukh Memorial Medical College, Amravati, (MS), India*

[5] *Associate Professor, Department of Community Medicine, Jawaharlal Nehru Medical College, Sawangi, Wardha, (MS), India*

Corresponding author:
Dr.Susma S.Pande
Mailing address: Dr.Sushma. S. Pande c/o Dr. S.R. Pande, Kalyannagar, Amravati, (MS), India -444606

Email: drsushmapande@gmail.com

Introduction

Formative examinations are part of the instructional process which helps to modify teaching and learning while they are happening. Timely modification can be made to improve knowledge. Knowledge of students can be assessed by MCQs dates to 1960. After 1999, in medical sciences, use of MCQs has been diversified to departmental, university and competitive examinations. In formative examinations MCQs help to understand the strength, weakness, gaps in knowledge, and provide feedback to teachers on their educational actions (Sadler, 1998; Nicol, 2006; Hubbard, 1961).

Designing good MCQs is a complex, challenging and time consuming process. Having constructed and assessed, MCQs need to be tested for the standard or quality. The "single best-response" type MCQs, are expressly designed to assess knowledge (Skakun,1979).They have advantage of sampling broad domains of knowledge effectively and reliably. This characteristic of MCQs is sufficient to ensure that, its edge in reliability more than compensates for some failings in validity. If carefully constructed, MCQs (especially single-best-response type) test higher-order thinking skills (Norman, 1995; Peitzman, 1990). Therefore, MCQs remain a useful assessment instrument, despite some limitations and objections.

Item analysis is a process which examines students' responses to individual test items in order to assess the quality of those items and quality of the test as a whole. It is of great help in improving the quality of items which may be used again in subsequent tests. It also nurtures a thought in the mind of the instructor to improve the skill in the construction of test items, and also helps identify course content which needs greater emphasis or clarity. Nonetheless, it also provides feedback to teachers to instill changes in the standard of teaching. The item statistics can help find out poor items which need improvement or deletion. It allows any aberrant items to be given attention and reconstructed. Although some basic form of item analysis of the MCQ tests might have been carried out routinely there has been no evidence that the data generated have been used to help develop or select subsequent MCQ items (Si-Mui Sim, 2006; Zubairi, 2006). How "good" were our MCQs? Were they really able to discriminate the student's performance in the examinations? We tried to answer these questions in this study. We also tried to find out the relationship which existed between the difficulty index (P) and discrimination index (D) of these MCQs.

Materials and Methods

1. Construction and Selection of MCQ Items

The MCQ items were constructed by all teachers and vetted at department for content accuracy every year from 2006. The vetted questions were selected by the Departmental Head and formatted for an examination paper.

2. Data Collection

In this study, 240 test MCQs taken from the past 4 year for I MBBS Physiology First term and Preliminary Examinations (Paper I & II) were analyzed. Each examination was carried out at the end of the term. A hundred students appeared for each the examination. Each term the examination covered different topics, grouped generally according to the systems. However, some repetition of the topics did occur. Each MCQ consisted of a stem and four responses and the students were asked to select one best answer from these four choices.

3. Scoring of MCQs

The MCQ paper contained 20 questions drawn from different systems. It formed a part of 2 ½ - hour written paper to be completed in the first 20 minutes. A correct response to an item was awarded ½ mark and the wrong one zero, no negative marks allotted.

4. Item Analysis

The results of students' performance in these MCQ tests were then used to determine the level of difficulty (P-scores in percent) and power of discrimination (D-scores) using Microsoft Office Excel.

Steps for item analysis were:

1. Scoring whole test for all students
2. Rank students in order of merit based on their test scores
3. Top third were taken as high achievers (h) & bottom third (*l*) were as low achievers
4. Table was prepared for each item to get the value of h and the calculations were made using the following formulae from the book of Medical Education. (Ananthkrishnan, 2000; Tejinder, 2009).

a. Difficulty Index (P) = *h + l / n X 100*

b. Discrimination Index (D) = *h - l / n X 2*

where,

h = Number of students answering correctly in high achievers group

l = Number of students answering correctly in low achievers group

n = Total number of students in two groups including non-responders

5. Interpretation

Difficulty Index (P) if

P< 30 %	Difficult
P 30% to 70 %	Acceptable
P >70%	Easy

Discrimination Index (D) if

D = Negative	Defective item / Wrong key
D < 0- 0.19	Poor discrimination
D between 0.2-0.29	Acceptable discrimination
D between 0.3-0.39	Good discrimination
D > 0.4	Excellent discrimination

Hence, the higher the difficulty index value, the lower is the difficulty, and the lower the difficulty index value, the greater is the difficulty of an item. For discrimination index, higher the index, better the item can discriminate between those students with high test scores and those with low ones (Ananthkrishnan, 2000; Tejinder, 2009; Mitra, 2009).

6. Statistical Analysis

The data are reported as % and mean ± SD of n items. The relationship between the item difficulty index and discrimination index values for all items was determined using Pearson correlation analysis and using SPSS 11.5. *P* value of <0.05 was considered to indicate statistical significance.

Results

As seen in Table 1, Mean Difficulty index (P) was 52.53 $\pm$ 20.59 while mean discrimination index (D) was 0.30 $\pm$ 0.18. There was a wide distribution among difficulty indices (Range 8.57 to 95.71) and discrimination indices (Range -0.54 to 0.8) in all 240 MCQ items analyzed.

Table 1: Descriptive Analysis for Difficulty Index (P) and Discrimination Index (D) for M CQ Paper Analyzed for Formative Examination (n =240 test items)

Item Analysis Parameter	Mean ± SD	Range
Difficulty Index P %	52.53 $\pm$ 20.59	8.57 to 95.71
Discrimination index D	0.30 $\pm$ 0.18	- 0.54 to 0.8

Figure 1: Proportion of "easy" (P >70%), "acceptable" (P between 30-70%) and " Difficult" (P<30%) indices for N = 240 Items.

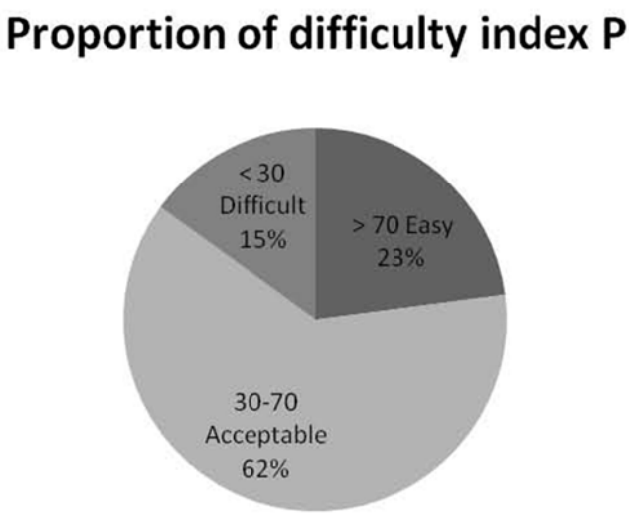
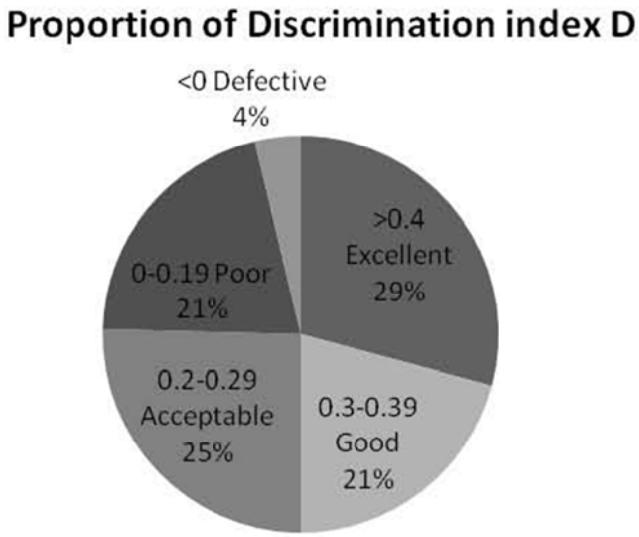


Figure 2: Proportion of excellent, good, acceptable, poor and defective Discrimination indices for N=240 items



discrimination indices for N= 240 item.

Figure 3: Correlation between Difficulty Index (P) and Discrimination Index (D)

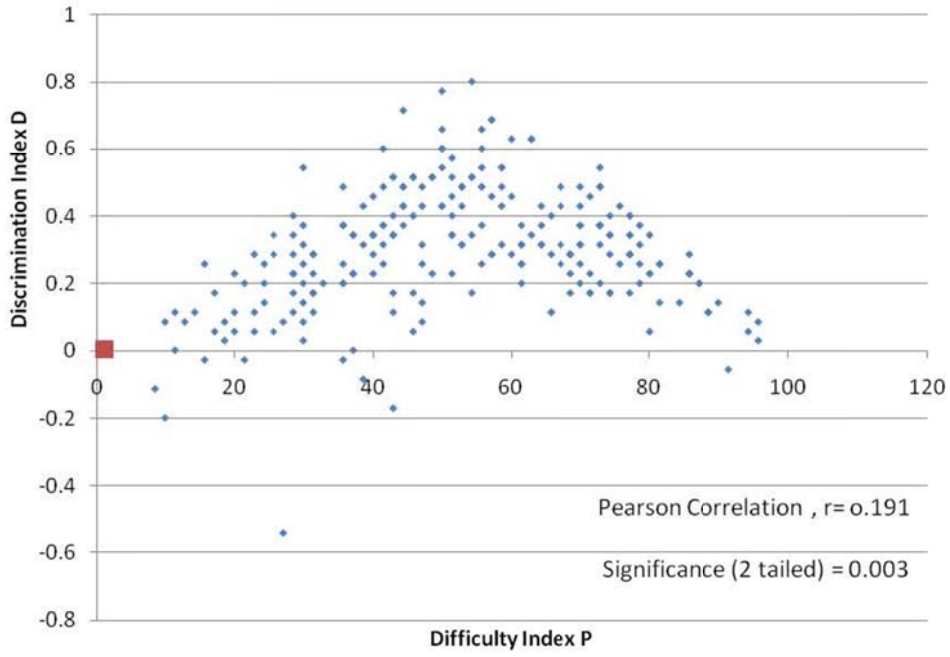


Figure 1 shows out of a total of 240 items, difficulty indices of 23% MCQ items were easy (P>70%), while about 15% were difficult (P <30%) and the remaining 62% of the items were within an acceptable range (30% to 70%).

As seen in Figure 2, the discrimination indices for 240 items, showed 4% negative discrimination. 21% of the items were having poor discrimination power (0-0.19), while 29% of the items exhibited excellent discrimination (>0.4). The remaining 46% were acceptable and good, out of which 25% of the items had an acceptable range (0.2 to 0.29) and 21% of the items showed good discrimination (0.3-0.39).

The scattered diagram (Figure 3) represents the relationship between the difficulty index (P) and discrimination index (D) of 240 MCQ items. It is not linear, but more or less pyramidal in shape which is as predicted a dome shape. The discrimination index correlated positively with the difficulty index (r = 0.191, P=0.003<0.01). The maximal discrimination (D =0.6-0.8) had moderate easy/difficult items (P=40%-60%). It is seen that in 4% of the items had negative discrimination indices value ranging from 0 to-0.06 with the corresponding difficulty index between 0% to 30%. This may be due to faulty items or a wrong key.

Discussion

The effective measurement of knowledge acquired is an important component of medical education. MCQs form are useful assessment tools in measuring factual recall and if carefully constructed can test higher order of thinking skills which is very important for a medical graduate (Norman, 1995; Peitzman, 1990). The method of assessment should be regularly evaluated. Fowell and coworkers have stressed the importance of this step of assessment which is often omitted (Fowell, 1999). Developing an appropriate assessment strategy is a key part in curriculum development. It is important to evaluate MCQ items to see how effective they are in assessing the knowledge of students. Items that discriminate poorly should be reviewed for possible corrections and reconstruction or deletion. Some basic forms of item analysis may be carried out routinely and the data generated should be used regularly to test the quality of the questions or for the development of multiple choice questions for the subsequent tests. The present study was conducted with the same objective.

In this study the wide scatter of difficulty and discrimination indices was observed indicating some guessing practices, probably because no negative marks were allotted to wrong answers. Same observations were reported by Si-Mui Sim et al., (2006) in their study on True/False questions and MCQs in paraclinical multidisciplinary examination and Mitra et al., (2009) in a study conducted on 120 Type A MCQs of preclinical semester 1 multidisciplinary summative tests.

In present study 62% of the items had acceptable difficulty indices (P= 30-70%), 23% were easy (P > 70%) while 15% of the items with P <30 % were difficult. This could have been due to poor understanding of difficult topics, ambiguity in wordings of the questions or even inappropriate key or personal variation in forming the MCQs and may also be due to personal variations in students' intelligence level.

The discrimination index (D) serves as an effective feedback to teachers about quality of each item. Items with poor discrimination should be reviewed. According to Brown (1983) and Algina (1986) D>0.2 is acceptable and able to discriminate between good and weak students. The present study shows that 29% of the items had D > 0.4, which is excellent discrimination, 46% of items showed good and acceptable discrimination. In all 21% of the items had poor discrimination and 4% of the items showed negative discrimination.

A similar type of study which reported by Ho et al. (1981) showed the difficulty and discrimination indices in 50 MCQs on Physiology of CNS (40 true/ false question and 10 multiple completion). They also reported that too easy or too difficult items poorly discriminated. However, the sample size in their study was quite small: n=50 as compared to n =240 in this present study. Moreover, they have not worked out the correlation between the two indices. In our study we have tried to find the correlation between difficulty and discrimination indices. It was seen that the relationship between difficulty and discrimination indices was not linear but more or less pyramidal in shape which is widely accepted. Positive correlation (r=0.191, P=0.003<0.01) was noted in difficulty and discrimination indices. Maximal discrimination (D=0.6-0.8) occurred with moderately easy/difficult items (P=40% - 60%).Very easy and very difficult items showed poor discrimination.

Same observation was reported by Si-Mui Sim et al., (2006) in their study, Mitra et al., (2009) showed that the discrimination index correlated poorly with the difficulty index (r= -0.325). The negative correlation signified that

with increasing difficulty index values, there was a decrease in the discrimination index indicating that low performance students were more likely to get the correct answer. As the items got easier (above 75%), the level of discrimination index decreased consistently (Mitra, 2009).

Very difficult and very easy items need to be properly reconstructed and revalidated. We hope item analysis will serve as helpful tool to generate MCQs question banks at departmental and university levels which will provide items with acceptable difficulty and discrimination indices.

Conclusion:

In our study the majority of items fulfilled the criteria of acceptable difficulty and good discrimination, which means the MCQs selected were of good quality. Moderately easy/difficult items had maximal discriminative ability. Very easy and very difficult items displayed poor discrimination. Even negative discrimination was observed in very difficult items.

Items with negative and poor discrimination will be reviewed, reconstructed and added to the departmental MCQs bank. Items from this bank will be revalidated for the next 4 years The results will be compared with this study to see the effects of reconstruction of items and will be subsequently represented. This type of study should also be conducted at the university level for summative examinations of all subjects.

Acknowledgements

Authors are thankful to Dr. D.S. Jane, the Dean of the Institution for permitting and encouraging the study.

References

Sadler, D.R. (1998) Formative assessment revisiting the territory. *Assessment in Education: Principles, Policy & Practice,* 1465-329X, 5, 1, pp. 77-84.

Nicol, D.J. & Macfarlane-Dick, D. (2006) Formative assessment and self-regulated learning: a model and seven principles of good feedback practice: *Studies in Higher Education,* 31, 2, pp. 199-218.

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

Skakun, E.N., Nanson, E.M., Kling, S. & Taylor, W.C. A preliminary investigation of three types of multiple choice questions. *Medical Education*, 1979, 13, pp. 91-96.

Norman, G. (1995) Evaluation methods: A resource handbook. In: Shannon, S., Norman, G., editors. Chapter 4.1 Multiple choice question: *The Program for Educational Development*, McMaster University. Hamilton, Canada: McMaster University, pp. 47-54.

Peitzman, S.J., Nieman, L.Z & Gracely, E.J. (1990) Comparison of "fact-recall" with "higher-order" questions in multiple-choice examinations as predictors of clinical performance of medical students. *Acad Med,* 65, pp. 59-60.

Si-Mui Sim, Rasiah R.I. (2006) Relationship between item difficulty and discrimination indices in true/false type multiple choice questions of a para-clinical multidisciplinary paper. *Ann Acad Med* Singapore, 35, pp. 67-71.

Zubairi, A.M & Kassim. N.L.A. (2006) Classical and Rasch analysis of dichotomously scored reading comprehension test items. *Malaysian J of ELT Res*, 2, pp. 1-20.

Ananthkrishnan, N. (2000) The item analysis: Medical Education Principles and Practice, 2nd Ed, pp. 131-137.

Tejinder, S., Piyush, G & Daljit, S. (2009) Principles of Medical Education, 3rd Ed, pp. 70-77.

Mitra, N.K., Nagaraja, H.S., Ponnudurai, G. & Judson, J. P. (2009) The levels of difficulty and discrimination indices in type A multiple choice questions of Pre-clinical Semester 1 multidisciplinary summative tests. *IeJSME*, 3, 1, pp. 2-7.

Fowell, S.L., Southgate, L.J & Bligh, J.G. (1999) Evaluating assessment: the missing link? *Medical Education*, 33, pp. 276-81.

Brown, F.G. (1983) Principles of educational and psychological testing. 3rd Ed, New York: Holt, Rinehart and Winston.

Crocker, L. Algina, J. (1986) *Introduction to classical and modern test theory*. New York: Holt, Rinehart and Winston.

Ho, T.F., Yip, W.C.L. & Tay, J.S.H. (1981) The use of multiple choice questions in medical examination: An evaluation of scoring and analysis of results. *Singapore Medical Journal,* 22, 6, December, pp. 361-367.