IOSR Journal of Dental and Medical Sciences (IOSR-JDMS)
e-ISSN: 2279-0853, p-ISSN: 2279-0861.Volume 14, Issue 12 Ver. VI (Dec. 2015), PP 47-52
www.iosrjournals.org

Analysis of Multiple Choice Questions (MCQs): Item and Test Statistics from an assessment in a medical college of Kolkata, West Bengal

Poulomi Mukherjee[1],Saibendu Kumar Lahiri[2]

[1]3rd year postgraduate trainee, Department of Community Medicine, R. G. Kar Medical College, Kolkata, West Bengal, India

[2]Professor and Head, Department of Community Medicine, R. G. Kar Medical College, Kolkata, West Bengal,India

Abstract:

Introduction: *Multiple-choice questions (MCQs) are most widely used test format in health sciences today. The efficiency of MCQs as an efficient tool for evaluation solely rests upon theirquality which is best assessed by item and test analysis.*

Objectives:*Toassess item and test quality and to explore the relationship between difficulty index (p-value) and discrimination indices (DI) with distractor efficiency (DE).*

Materials and Methods:*The study was conducted among 40 fourth semester MBBS students in a medical college of Kolkata. Thirty MCQs administered in an internal examination in Community Medicine, were analysed for p-value, DI and DE. Reliability of the test was assessed by estimating the Kuder-Richardson 20 coefficient (KR_{20}).*

Results:*The mean score was 66.35 ± 17.29. Mean p value and DI were 61.92 ± 25.1% and 0.31 ± 0.27, respectively. DI was noted to be maximum at p value range between 40% and 60%.Combining the two indices, 14(46.67 %) items could be called 'ideal' having a p-value from 20% to 90%, as well as a DI ≥ 0.3. Overall 86.67% items had non-functional distractors (NFDs), while 80% items had functional distractors. Mean DE was 47.78 ± 32.38%. Excellent discrimination (DI = 0.404 and 0.396) was achieved with items having two and one NFD respectively while items with no NFD had lower DI (0.023). Internal consistency reliability of the test as per $KR_{20}$was 0.9.*

Conclusion: *Items having average difficulty and high discriminating power with functional distractors should be incorporated in future to improve the quality of the test.*

Keywords: *difficulty index, discrimination index, distractor efficiency, item analysis, multiple choice questions, non-functional distractor (NFD)*

I. Introduction

A multiple-choice question (MCQ) consists of a stem with a question line at its end or underneath it, followed by a number of options. One of the options is the correct or best response known as the key, [1] while the others are described as distractors. An essential characteristic of distractors is that all options shall present plausible answers and if possible none shall be incorrect.[2] Function of a distractor is to attract students who do not know the correct answer while students who know the correct answer ignore them.

Tests using MCQs are objective and easily adapted for computer delivery. Moreover, they can be used to diagnose student difficulties if the incorrect options are designed to reveal common misconceptions and they can provide a more comprehensive sampling of the subject material because of wider coverage. In addition, they are often more valid and reliable than essay tests because discrimination between performance levels is easier to determine and scoring consistency is virtually guaranteed when carried out by machine. [3]

However, some instructors believe that MCQs are “multiple-guess” items or that MCQs are only capable of testing factual information and so are ill suited for testing higher-order cognitive skills. But it is now accepted that well-constructed multiple-choice items can test many of the higher cognitive skills of Bloom’s taxonomy such as knowledge, application, analysis and synthesis; which is a necessity for the assessment of health care professionals. [4, 5]

An item is a single test element, which might be a multiple-choice question.Item analysis is a process which examines student responses to individual test items in orderto assess the quality of those items and of the test as a whole. Item analysis is especially valuable inimproving items which will be used again in later tests, but it can also be used to eliminate ambiguous ormisleading items in a single test administration. In addition, item analysis is valuable for increasinginstructors' skills in test construction and identifying specific areas of course content which need greateremphasis or clarity. [6]

DOI: 10.9790/0853-141264752

Keeping in view the widespread use of MCQs in assessment of medical students, present study was undertaken with an objective to assess item and test qualityand to explore the relationship between difficulty and discrimination indices with distractor efficiency.

II. Materials And Methods

2.1Settings

The study was conducted in the department of Community Medicine of a medical College of Kolkata as a ‘part end’ assessment in May 2014. After developing the assessment tool pre-validation was done by establishment of content validity of the items through content validity ratio (CVR) and mean of expert’s judgment. Fourteen subject matter experts (SME) willing to act as panelists were requested to clearly indicate their judgments on the essentiality of inclusion of each items in a separate table provided to them. The judgments of the respective panelists were entered in Microsoft Excel 2010 and analyzed. Only those items, meeting the minimum CVR value (0.51) and mean value (1.5) were finally retained. [7]

The test comprising thirty MCQs was administered to 20% (i.e 38.6 ≈ 40) randomly selected MBBS students of fourth semester comprising 193 students. The time allotted was 30 minutes. The items were of one-best type, having a single stem and four answer options, one of them being correct and the other three being 'distractors'. The students were required to encircle the correct choice. Each correct response was awarded 4 marks and each incorrect response was awarded -1, range of total score being 0-120 (ignoring minus marks).

2.2 Statistical Analysis

Post validation of the paper was done by item analysis. Score of 40 students was entered in order of merit in MS Excel 2010 and simple proportions, mean, standard deviations, correlation were calculated. Items were categorized according to their difficulty index (p value), discrimination index (DI) and distractor efficiency (DE) and actions such as discard/ review /revise and store were proposed.Reliability of the test was assessedby estimating theKuder-Richardson 20 coefficient (KR_{20}).

2.3 Item statistics

2.3.1 Difficulty Index or Facility value or p value

Perhaps “item difficulty” should have been named “item easiness”; it expresses the proportion of students who answered the item correctly. The formula for the item-difficulty index is

$$p = c/n$$

where, c is the number of students who selected the correct answer and n is the total number of respondents.

The p (proportion) value statistics ranges from 0 to 1.When multiplied by 100, p value converts to a percentage, which is the percentage of students who got the item correct. The higher the p value, the easier the question.In general, items with a p value between 20 – 90% are considered as good and acceptable. Amongst these, items with p value between 40-60% are considered excellent, because discrimination index is maximum at this range. Items with p value less than 20% (too difficult) and more than 90% (too easy) are not acceptable and need modification. [3, 8-11]

2.3.2 Discrimination Index

The discrimination index (DI) is a measure of the effectiveness of an item in discriminating between high and low scorers.For this calculation, we divided the test takers into three groups according to their scores on the test as a whole: an upper group consisting of the 11(27%) who made the highest scores, a lower group consisting of the 11 (27%) who made the lowest scores and a middle group consisting of the remaining 18(46%).Discrimination index was estimated using the following formula:

$$D = P_U - P_L$$

where,P_U and P_L arethe proportions of the students in the upper and bottom group who got the item correct.

The range of values for the item discrimination index is -1.00 to 1.00.The higher the value of DI, the more effective the item is. When DIis 1.00, all test takers in the upper group and no test takers in the lower group answered the item correctly. Conversely, if none of the upper group but all of the lower group answered an item correctly; the DI value would be-1.00.In general the DI value 0.40 and greater are considered excellent items; Items with DI0.30 to 0.39 is considered reasonably good but possibly subject to improvement; those with DI 0.20 to 0.29 are considered marginal items and should be reviewed while those with DI below 0.19 are considered poor items and should be eliminated. [3, 12]

2.3.3 Distractor analysis

Non-functional distractors (NFDs) are options that are selected infrequently (<5%) by examinees and functional or effective distractor is the option selected by 5% or more students. As such, NFDs should be revised, removed or be replaced with a more plausible option.[13, 14]Distractor efficiency (DE) is determined for each item on the basis of the number of NFDs in it and ranges from 0 to 100%. If an item contains three or two or one or nil NFDs then DE will be 0, 33.3, 66.6, and 100%, respectively. [15]

2.3.4 Test Reliability

Internal consistency reliability of the test was measured by Kuder-Richardson 20 (KR_{20}) coefficient.The formula for KR_{20} for a test with K test items numbered i=1 to K is

$$r = \frac{K}{K-1}\left[1 - \frac{\sum_{i=1}^{K} p_i q_i}{\sigma_X^2}\right]$$

where, p_i is the proportion of correct responses to test item i, q_i is the proportion of incorrect responses to test item i (so that $p_i + q_i = 1$), and the variance for the denominator is

$$\sigma_X^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n}.$$

where, n is the total sample size, X_i is the score of individual students and $\overline{X}$ is the mean total score. The value of KR_{20} can range from 0 to 1, with numbers closer to 1 reflecting greater internal consistency indicating that the items are all measuring the same thing or general construct. The widely-accepted cut-off value of KR is greater than or equal to0.7. [2, 16]

III. Results

Total 30 MCQs and 90 distractors were analyzed.Score of 40 students ranged from 30 to 91(out of 120).The mean score achieved was 66.35 ± 17.29. Mean scoresaccording to groups were: lower 43.73±7.80;middle 68.06 ± 7.32; upper 86.18± 3.66.Means and standard deviations (SD) for p value (%), DI and DE (%) were 61.92 ± 25.1%, 0.31 ± 0.27, and 47.78 ± 32.38%, respectively[Table 1].

Table 1: Analysis of MCQ paper based on various item indices (N=30)

Parameter	Mean	Standard deviation (SD)
Difficulty index (p value)	61.92	25.10
Discrimination index (DI)	0.31	0.27
Distractor efficiency (DE)	47.78	32.38

Table 2 shows the distribution of difficulty and discriminationindices of the items and their corresponding DE. Majority of theitems (83.33%) were of acceptable level of difficultywith p value within the range of 20% to 90% while seven items among them had excellent p value (40-60%).Two items (6.67%) were identified to be too difficult (p value <19%) and three items (10%) were too easy (p value >90%).Similarly,majority of items (46.67%) had good to excellent discrimination indices (DI≥0.3), with 16.67% and 36.67% items having marginal and poor DI respectively.

A combination of the two indicesrevealed that 14(46.67%) items could be called 'ideal'having a p value from 20 to 90%, as well as a DI ≥ 0.3.However, if only the items with excellent p value (40-60%) and excellent DI (≥0.4) are considered,there were five (16.67%) items as excellent. Amongst these 14 MCQs, seven items had DE 66.67 %, six items had DE 33.33 % and one item had DE 0%.Mean DE was 47.62 which is close to items having two NFDs .Two items had negative DIs, with item 8 being a very easy one (p = 93%), and item 11, relatively difficult (p = 28%).

Fig. 1 shows the relationship between difficulty index and discrimination index. First, as p increases, the DI also increases, but at a p value between 40% and 60%, DI reaches a maximum. When p is more than 60%, DI decreases. Over the range 40% - 60%, the DI is more than 0.5.

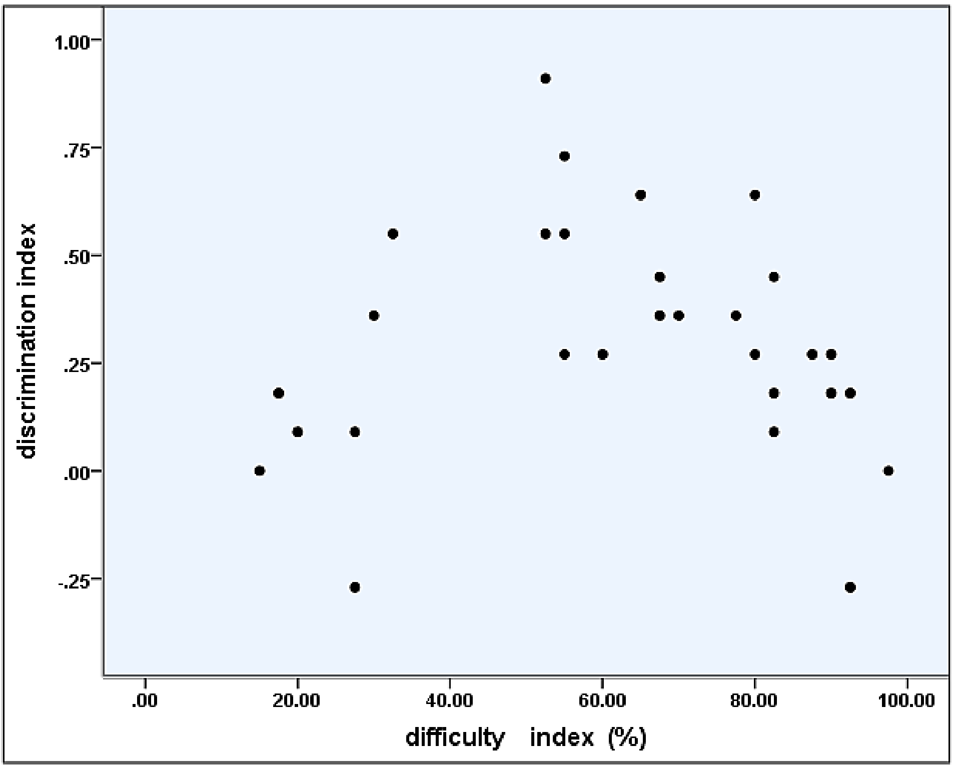

Figure 1: Scatter plot showing relationship between difficulty index and discrimination index of items.

The relation of mean difficulty anddiscrimination indices with mean distractor efficiency was also analyzed [Table-2]. DE was directly relatedto the p-value with most difficult items having DE of83.34% and most easy items having DE 22.22%. Items withgood difficulty had DE of 48%.Howeverno relation with DE with increasing DI of items was found. Items withexcellent and good discrimination had DE of 43.33% and 58.34% respectively whereas items with marginal and poor DI had 53.33% and 45.45% DE respectively.

Table 2: Distribution of items according to difficulty and discrimination indices and actions proposed (N = 30)

Difficulty index (p value)	Interpretation	No. of items (%)	Distractor Efficiency (DE)	Action
20-90	Good	25(83.33)	48.00	Store /review distractors
<20	Too difficult	2(6.67)	83.34	Review for confusing language and revise
>90	Too easy	3(10.0)	22.22	Discard
Discrimination index (DI)				
≥0.40	Excellent	10 (33.33)	43.33	Store/review
0.30-0.39	Good	4 (13.33)	58.34	Store
0.20-0.29	Marginal	5 (16.67)	53.33	Review and revise
≤0.19	Poor	11 (36.67)	45.45	Discard

Just over one-half (52.2%) of all distractors were NFDs. Twenty six (86.67%) items had NFDs (eleven items had one NFDs, nine items had two NFDs and six items had three NFDs), while 24 (80%) items had functional distractors. The mean functional distractor per item was 1.43. Items with three NFDs were easier with a high p value (87.08%) and poor DI (0.195); items with two and one NFD were of good difficulty level having pvalues 71.11% and 51.36%; and with excellent DI: 0.404 and 0.396 respectively. Items with no NFD on the other hand had higher difficulty level having p values 32.50% and poor DI (0.023). This suggests better discrimination potential of items with two and one NFD, as compared to items with all NFDs or no NFDs [Table 3].

Table 3: Non-functioning distractors (NFDs) and Distractor Efficiency (DE) of individual items (N = 30)

Parameter	Items with 0 NFDs	Items with 1 NFD	Items with 2 NFDs	Items with 3 NFDs
Number (%)	4(13.33)	11(36.67)	9(30.0)	6 (20.0)
DE (%)	100	66.67	33.33	0
Mean p value (%)	32.50	51.36	71.11	87.08
Mean DI	0.023	0.396	0.404	0.195

The reliability of the test as measured by the above mentioned formula of KR_{20}coefficient was 0.9.

IV. Discussion

Single correct response type MCQ is an efficient tool for evaluation; however, this efficiency solely rests up on the quality of MCQ which is best assessed by item and test analysis.[17] The difficulty and discrimination indices are among the tools to check whether the MCQs are well constructed or not. Another tool used for further analysis is the distractor efficiency which analyses the quality of distractors and is closely associated with difficulty and discrimination indices.

In the present study, the mean p value was 61.92 ± 25.1 % which was close to the excellent level of difficulty (p=40 to 60 %);[3] Our findings was corroborative with that of Mehta G and Mokhasi V who reported mean p value of 63.06.[18]However few studies have reported lower mean p values. [15, 19]

The mean DI found in this study was 0.31 ± 0.27 which is considered reasonably good.[12]However a substantial proportion of items (36.67%) had poor DI. An earlier study reported mean of DI of 0.33± 0.18. Items with DI > 0.35 were 52%, DI between 0.2 and 0.34 were 18% and DI <0.2 were 30%.[19]Two items (6.67%)had negative DI which means students of lower ability answered more correctly than those with higher ability. Some studies have shown negative DI in 20% and 4% items. [15, 17] Reasons for negative DI can be wrong key, ambiguous framing of questions or generalized poor preparation of students. Items with negative DI decrease the validity of the test and should be removed from the collection of questions.

Difficulty index and discrimination index are often reciprocally related except for extreme situations where the difficulty index is either too high or too low. It has been seen that the relationship between them is not linear, but predicted as dome shaped. [19, 20]The findings of this study corroborated the same with maximum DI of items between p value of 40-60%.

Analyzing the distractors is done to determine their relative usefulness in each item. If students consistently fail to select certain multiple choice alternatives it may be that the alternatives are probably totally implausible and therefore of little use as decoys in multiple choice items. [14]Therefore, designing of plausible distractors and reducing the NFDs is important aspect for framing quality MCQs. [21] Mean DE in the present study was 47.78 ± 32.38%; much lower than the DE of 88.6 and 63.97% reported in similar type of studies. It was so because a high proportion of the distractors (52.2%) were NFDs. The number of MCQ items having NFDs was also greater (86.67%) in this study with more number of items having two and three NFDs with DE 33.3% and 0% respectively. [15, 18]

More NFD in an item increases p value (makes item easy) and reduces DE, conversely item with more functioning distractors decreases p value (makes item difficult) and increases DE. This was also reflected in the findings of the present study. Items with three NFDs were easier with a high mean p value (87.08%) whereas items with two, one and no NFD showed mean p values of 71.11%, 51.36% and 32.5% respectively. On the other hand, the mean DE was 83.34% in very difficult items and as low as 22.22% in very easy items. Items with average difficulty had DE of 48%.

The numbers of NFDs also affect the discriminative power of an item. [17]Our observation that items having one and two NFDs had excellent discriminating ability (DI = 0.396 and 0.404 respectively) as compared to items with nil NFD (DI = 0. 0.023) was similar to a previous study.[18] However mean DE showed little variation in items with high (≥0.40)or low DI (≤0.19).

We found 14items (46.67 %) to be 'ideal' having a good p value (20 to 90%), as well as good to excellent DI (≥ 0.3). Hingorjo MR found 32 items (64%) as ideal having a p-value from 30 to 70, as well as a DI > 0.24. [17] Other researchers have reported 30% and 24% items to be ideal in their studies. [15, 18]

Reliability of our test was 0.9 which suggests that this was a highly reliable test with excellent internal consistency .Though much data are not available regarding the reliability of the tests from various studies done on item and test analysis, one rule of thumb states that values greater than or equal to 0.70 are acceptable. [2, 16]

V. Conclusion

Items analyzed in the study had optimum difficulty level but poor distractor efficiency. Though overall discriminating power was good but a substantial proportion of items had poor DI. Maximum DI was seen at p value range between 40% and 60%.Items with two NFDs, though easier, were better discriminators than items with no NFD. However findings of the present study have to be interpreted cautiously in the light of certain limitations; the number of items in this test was less and other semester students were not included. Future studies with larger number of items having average difficulty and high discrimination with functioning distractors administered to a bigger sample will add to the findings of this study.

References

[1]. Cizek GJ, O'Day DM. Further investigations of nonfunctioning options in multiple-choice test items. EducPsycholMeas1994; 54(4):861-72.

[2]. Saudi Commission for Health Specialties. Item writing manual for multiple-choice questions. [Internet] [cited 2015 June 12]. Available from:http://www.scfhs.org.sa/education/HighEduExams/Guidlines/mcq/Documents/MCQ%20Manual.pdf

[3]. Hotiu A.The relationship between item difficulty and discrimination indices in multiple-choice tests in a Physical science course [MSc thesis]. Boca Raton, Florida: Florida Atlantic University; 2006 [cited 2015 June 14]. Available from:http://www.physics.fau.edu/research/education/A.Hotiu_thesis.pdf

[4]. Designing and managing MCQs: Appendix C: MCQs and Blooms taxonomy. [Internet] [cited 2015 June 15]. Available from:http://www.u.arizona.edu/~jag/POL602/Designing-Managing-MCQs.pdf

[5]. Case SM, Swanson DB. Constructing written test questions for the basic and clinical sciences [internet].Philadelphia: National Board of Medical Examiners; 1998 [cited 2015 July 16]. Available from:http://www.uclouvain.be/cps/ucl/doc/adef/documents/EVA_Res_Ext_Questions_QCM.pdf

[6]. Understanding item analysis reports. [Internet] [cited 2015 July 4]. Available from: http://www.washington.edu/oea/services/scanning_scoring/scoring/item_analysis.html

[7]. Ehlers L. A validated model of the South African labour relations system: Research Methodology [Internet] [cited 2015 July 5]. Available from: http://upetd.up.ac.za/thesis/available/etd-10122007 082729/unrestricted/02chapter2.pdf

[8]. Item difficulty item discrimination. [Internet] [cited 2015 July 5]. Available from:http://www.omet.pitt.edu/docs/OMET%20Test%20and%20Item%20Analysis.pdf

[9]. Varma S. Preliminary item statistics using point-biserial correlation and p-values.[Internet] [cited 2015 July 6]. Available from:http://www.eddata.com/resources/publications/EDS_Point_Biserial.pdf

[10]. BhoopatirajC,ChellamaniK.Analysis of test items on difficulty level and discrimination index in the test for research in education.IRJC 2013; 2 (2):189-93.

[11]. Item analysis. [Internet] [cited 2015 July 6]. Available from:http://ctl.utexas.edu/sites/default/files/iar-assesslearning-exams-item_analysis.pdf

[12]. Ebel, RL, Frisbie, DA. Essentials of educational measurement. Englewood Cliffs, NJ: Prentice-Hall; 1986.

[13]. Tarrant M, Ware J, Mohammed AM. An assessment of functioning and non-functioning distracters in multiple-choice questions: a descriptive analysis. BMC Med Educ 2009; 9: 40.

[14]. Haladyna TM, Downing SM: Validity of a taxonomy of multiple choice item-writing rules. ApplMeasEduc 1989; 2(1):51-78.

[15]. Gajjar S, Sharma R, Kumar P, Rana M. Item and test analysis to identify quality multiple choice questions (MCQS) from an assessment of medical students Ahmedabad, Gujarat. Indian J Community Med 2014; 39:17-20.

[16]. Kuder–Richardson Formula 20. [Internet] [cited 2015 July 6]. Available from: https://en.wikipedia.org/wiki/Kuder%E2%80%93Richardson_Formula_20

[17]. Hingorjo MR, Laleel F. Analysis of One -Best MCQs: the Difficulty Index, Discrimination Index and Distractor Efficiency. J Pak Med Assoc 2012; 62:142-47.

[18]. Mehta G, Mokhasi V. Item analysis of multiple choice questions- an assessment of the assessment tool. Int J Health Sci Res. 2014; 4(7):197-202.

[19]. Karelia BN, Pillai A, Vegada BN. 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 MBBS students. IeJSME 2013; 7(2):41-6.

[20]. Pande SS, Pande SR, Parate VR, Nikam AP, Agrekar SH. Correlation between difficulty and discrimination indices of MCQs in formative exam in Physiology. South East Asian Journal of Medical Education 2013; 7(1):45-50.

[21]. Rodriguez MC. Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. Educational Measurement: Issues and Practice 2005; 3-13.