Original Article

Analysis of One-Best MCQs: the Difficulty Index, Discrimination Index and Distractor Efficiency

Mozaffer Rahim Hingorjo,[1] Farhan Jaleel[2]

Department of Physiology,[1] Department of Biochemistry,[2] Fatima Jinnah Dental College, Karachi, Pakistan.

Abstract

Objective: To investigate the relationship of items having good difficulty and discrimination indices with their distractor efficiency to find how 'ideal questions' can be affected by non-functioning distractors (NF-Ds).

Methods: The cross-sectional study was conducted at Fatima Jinnah Dental College, Karachi, during Jan-Jun 2009, with 102 First Year dental students (17-20 years). Physiology paper of the first semester, given after 22 weeks of teaching general topics of physiology, was analysed. The paper consisted of 50 one-best MCQs, having 5 options each. The MCQs were analysed for difficulty index (p-value), discrimination index (DI), and distractor efficiency (DE). Items having p-value between 30-70 and DI $\geq$ 0.25 were considered as having good difficulty and discrimination indices respectively. Effective distractors were considered as the ones selected by at least 5% of the students.

Results: The mean score was 27.31 ± 5.75 (maximum 50 marks). Mean p-value and DI were 54.14 ± 17.48 and 0.356 ± 0.17, respectively. Seventy eight per cent items were of average (recommended) difficulty (mean p-value = 51.44 ± 11.11) and having DE = 81.41%. Sixty two per cent items had excellent DI (0.465 ± 0.083) with DE = 83.06%. Combining the two indices, 32 (64%) items could be called as 'ideal' (p-value = 30 to 70; DI > 0.24) and had DE = 85.15%. Overall 42% items had no Non-functioning Distractors (NF-D), while 12% had 3 NF-Ds. Excellent discrimination (DI = 0.427) was achieved with items having one NF-D, while items with 2 NF-D and no NF-D had nearly equal but lower DI (0.365 and 0.351 respectively).

Conclusion: One-best MCQs having average difficulty and high discrimination with three functioning distractors should be incorporated into future tests to improve the test score and properly discriminate among the students. Items with two NF-Ds, though easier, are better discriminators than items with no NF-D.

Keywords: One-best MCQ, Assessment methods, Difficulty index, Discrimination index, Non-functioning distractors (JPMA 62: 142; 2012).

Introduction

One-best MCQs (Multiple Choice Questions) are a form of assessment where the student selects the best possible answer from the list provided. This form of assessment has become popular in educational institutions. A large portion of curriculum is assessed in a short period of time requiring less effort on behalf of the student, although it takes a lot of effort and time spent by the examiner to make high quality one-best MCQs, as compared to descriptive questions. One-best MCQ is an efficient tool in identifying the strengths and weaknesses in students, as well as providing guidelines to teachers on their educational protocols.[1]

Properly constructed multiple choice questions assess higher-order cognitive processing of Bloom's taxonomy such as interpretation, synthesis and application of knowledge, instead of just testing recall of isolated facts.[2,3]

All this is possible if the examiner knows the correct method of formulating a question, commonly referred to as an item, consisting of a stem and several options.[4]

Item analysis is a valuable, yet relatively simple, procedure performed after the examination that provides information regarding the reliability and validity of a test item.[5]

It also tells how difficult or easy the questions were, the difficulty index, and whether the questions were able to discriminate between students who performed well on the test, from those who did not, the discrimination index.[6] Another important technique is analysis of distractors, that provides information regarding the individual distractors and the key of a test item. Using these tools, the examiner is able to modify or remove specific items from subsequent exams.[7]

Difficulty index (p-value), also called ease index, describes the percentage of students who correctly answered the item. It ranges from 0 - 100%. The higher the percentage, the easier the item. The recommended range of difficulty is from 30 - 70%. Items having p-values below 30% and above 70% are considered difficult and easy items respectively.[8] Very easy items should usually be placed either at the start of the test as 'warm-up' questions or removed altogether. The difficult items should be reviewed for possible confusing language, areas of controversy, or even an incorrect key. Inclusion of very difficult items in the test depends upon the target of the teacher, who may want to include them in order to identify top scorers.

Discrimination index (DI), also called point biserial correlation (PBS), describes the ability of an item to distinguish between high and low scorers.[9] It ranges between -1.00 and +1.00. It is expected that the high-performing students select the correct answer for each item more often than the low-performing students. If this is true, the assessment is said to have a positive DI (between 0.00 and +1.00), indicating that students who received a high total score, chose the correct answer for a specific item more often than the students who had a low overall score. If, however, the low-performing students got a specific item correct more often than the high scorers, then that item has a negative DI (between -1.00 and 0.00).

The difficulty and discrimination indices are often reciprocally related. However, this may not always be true. Questions having high p-value (easier questions), discriminate poorly; conversely, questions with a low p-value (harder questions) are considered to be good discriminators.[10]

Analysis of distractor is another important part of item analysis. The distractors are important components of an item, as they show a relationship between the total test score and the distractor chosen by the student. Student's performance depends upon how the distractors are designed.[11] Distractor efficiency is one such tool that tells whether the item was well constructed or failed to perform its purpose. Any distractor that has been selected by less than 5% of the students is considered to be a non-functioning distractor (NF-D).[7] Ideally, low-scoring students, who have not mastered the subject, should choose the distractors more often, whereas, high scorers should discard them more frequently while choosing the correct option. By analysing the distractors, it becomes easier to identify their errors, so that they may be revised, replaced, or removed.[12]

Tarrent and Ware demonstrated that flawed MCQ items affected the performance of high-achieving students more than borderline students.[13] Construction of a balanced MCQ, therefore, addresses the concerns of the students of getting an acceptable average grade and that of the faculty to have an appropriate spread of the score.[14]

The objective of the study was to investigate the relationship of items having good difficulty and discriminator indices, with their distractor efficiency to find how 'ideal questions' could be affected by non-functioning distractors (NF-D).

Subjects and Methods

Conducted at Fatima Jinnah Dental College Karachi during the academic session 2009, the cross-sectional study had 102 First Year dental students aged 17-20 years. Fifty one-best MCQs were taken from the examination of physiology conducted at the end of the first semester after 22 weeks of teaching the following topics: cell, nerve and muscle, blood, respiration, and kidney. We analysed the MCQs for their level of difficulty, measured by difficulty index (p-value), power of discrimination, measured by discrimination index (DI), and distractor analysis measured by distractor efficiency (DE).

The time given for the MCQ paper was 60 minutes, and was followed by an essay type paper. The items were of one-best type, having a single stem and 5 answer options, one of them being correct and the other 4 being 'distractors'. The students were required to select the correct choice and fill in the answer sheet given separately. Each correct response was awarded 1 mark. No mark was given for blank response or incorrect answer. There was no negative marking. Thus, the maximum possible score of the overall test was 50 and the minimum 0.

High and low groups consisting of upper and lower 27 % students, respectively, were taken after arranging the scores in descending order.[15] The p-value and DI were then calculated as follows:

$$p = [(H+L) / N] \times 100,$$

$$DI = 2 \times [(H-L) / N]$$

Where, N is the total number of students in both high and low groups. H and L are the number of correct responses in the high and low groups, respectively.[8] Items with p-value between 30 - 70 and DI > 0.24 were considered as 'ideal'.

NF-Ds were those selected by less than 5% of students.[16] Distractor efficiency ranged from 0 - 100% and was determined on the basis of the number of NF-Ds in an item. Four NF-D: DE = 0%; 3 NF-D: DE = 25%; 2 NF-D: DE = 50%; 1 NF-D: DE = 75%; No NF-D: DE = 100%.

Results

A total of 102 students gave the test consisting of 50 one-best MCQs from basic physiology topics. The mean score achieved was 27.31 ± 5.75 (maximum 50 marks). Mean scores according to groups were: lower 18.25 ± 3.26; middle 27.63 ± 2.05; upper 35.85 ± 2.09. After getting the result, students were ranked in order of merit from the highest score of 40 to the lowest score of 6. The first 27% students made the high

Table-1: Difficulty index, discrimination index and distractor efficiency of the 50 one-best MCQs.

Q.No	P	DI	DE (%)	Q No.	P	DI	DE (%)
1	89	0.14	25	26	64	0.43	100
2	66	-0.11	25	27	39	0.43	75
3	54	0.35	100	28	71	0.35	50
4	52	0.60	50	29	86	0.28	25
5	11	0.14	100	30	70	0.46	50
6	71	0.50	75	31	66	0.46	100
7	86	0.28	25	32	66	0.46	100
8	43	0.57	75	33	73	0.39	75
9	45	0.53	50	34	55	0.32	75
10	64	0.35	75	35	23	0.25	100
11	57	0.50	100	36	50	0.42	100
12	46	0.50	75	37	48	0.17	75
13	54	0.57	75	38	57	0.35	100
14	50	0.42	75	39	48	0.53	75
15	41	0.46	100	40	63	0.53	75
16	54	0.14	50	41	54	0.21	100
17	57	0.57	100	42	34	0.35	100
18	84	0.11	50	43	21	-0.14	100
19	63	0.68	75	44	30	0.32	100
20	54	0.21	75	45	38	0.17	100
21	86	0.21	25	46	43	0.50	100
22	30	0.25	100	47	55	0.46	100
23	52	0.18	25	48	61	0.28	75
24	70	0.46	75	49	43	0.42	100
25	38	0.46	100	50	32	0.35	75

P = Difficulty Index, DI = Discrimination Index, DE = Distractor Efficiency.

Table-2: Classification of questions according to difficulty and discrimination indices.

p-value	Interpretation	Mean p-value	DE (%)	No. of items (n = 50)
> 70	too easy	80.75	43.75	8 (16%)
30-70	Average	51.43	81.41	39 (78%)
< 30	too difficult	18.33	100	3 (6%)
DI	**Interpretation**	**Mean DI**	**DE (%)**	**No. of items (n = 50)**
≥ 0.35	Excellent	0.469	83.06	31 (62%)
0.25 - 0.34	Good	0.285	71.42	7 (14%)
0.15 - 0.24	Marginal	0.196	66.66	6 (12%)
< 0.15	Poor	0.047	58.33	6 (12%)

p-value	DI and DE				No. of items (n = 50)
	DI ≤ 0.24	DE (%)	DI ? 0.25	DE (%)	
<30	2 (4%)	100.0	1 (2%)	100.0	3 (6%)
30 - 70	7 (14%)	64.28	32 (64%)	85.15	39 (78%)
>70	3 (6%)	33.33	5 (10%)	50.00	8 (16%)
Total	12 (24%)	62.50	38 (76%)	77.24	50 (100%)

p-value = Difficulty index; DI = Discrimination index; DE = Distractor Efficiency;
*Marginal question - revise; + Poor question - most likely discard.

Table-3: Distractor analysis.

Item	P, DI, DE	Group	A	B	C	D	E
1	P = 89; DI = 0.14;	Lower	0 (0%)	3 (11%)	1 (3%)	23 (82%)	1 (3%)
	DE = 25%	Upper	0 (0%)	0 (0%)	1 (3%)	27 (96%)	0 (0%)
2	P = 66; DI=- 0.10;	Lower	2 (7%)	1 (3%)	0 (0%)	20 (71%)	5 (18%)
	DE = 25%	Upper	0 (0%)	0 (0%)	0 (0%)	17 (60%)	1 1
(39%)							
13	P = 54; DI=0.57;	Lower	8 (28%)	4 (14%)	6 (21%)	2 (7%)	8 (28%)
	DE = 75%	Upper	0 (0%)	2 (7%)	23 (82%)	1 (3%)	2 (7%)
43	P = 21; DI=-0.14;	Lower	10 (35%)	4 (14%)	1 (3%)	5 (18%)	8 (28%)
	DE = 100%	Upper	4 (14%)	1 (3%)	3 (11%)	16 (57%)	4 (14%)

P = Difficulty index; DI = Discrimination index; DE = Distractor efficiency
Correct alternative for each item is represented by bold letters.

group and the last 27%, the low group.

The p-value, DI and DE were analysed for each MCQ (Table-1). Two items had negative DIs, with item 2 being a relatively easy one (p = 66), and item 43, very difficult (p = 21).

The distribution of difficulty and discrimination indices of the 50 MCQs given and their corresponding DE was also worked out (Table-2). Majority of the items (78%) were of average (recommended) difficulty with a mean p-value of 51.44 ± 11.11. Similarly, majority of items (62%) had excellent DI (0.465 ± 0.083), with items having marginal and poor DI being low in number. A combination of the two indices revealed that 32 (64%) items could be called 'ideal' having a p-value from 30 to 70, as well as a DI > 0.24.

The total number of distractors were 200 (4 per item) out of which 47 (23.5%) were NF-Ds. Twenty nine (58%) items had NF-Ds, while 21 (42%) items had effective distracters. Items with 3 NF-Ds had a high p-value (77.5%) and poor DI (0.160); items with 2, 1 and no NF-D were of average (recommended) difficulty having p-values 62.66, 54.94 and 44.38, and with excellent DI: 0.365, 0.427 and 0.351 respectively. This also highlights better discrimination potential (DI = 0.427) of items with one NF-D, as compared to items with all distracters that were functioning (DI = 0.351).

The relation of mean difficulty and discrimination indices with mean distractor efficiency was also analysed (Table-2). DE was indirectly related to the p-value with most difficult items having DE of 100% and easy items having DE 43.75%. Items with average difficulty had DE of 81.41%. The DE was directly related to the discrimination index. Items with good and excellent discrimination had DE of 71.42% and 83.06% respectively.

Distractor analysis gives an opportunity to study the responses made by the students on each alternative of the item. The analysis of 4 questions selected on the basis of p-value and DI gave a varied result (Table-3). The first MCQ was a very easy question with poor discrimination, as both upper and lower groups

selected, nearly equally, the correct answer. The DE was 25% as 95% students selected the correct response, making the rest of the alternatives useless. The second MCQ was selected for its negative DI with more of the lower group choosing the correct response. The DE was 25%. Alternatives B and C served no purpose at all, as hardly any student selected them, making it easier to choose from the reduced number of choices. Alternative E was considered the right answer by many in the upper group and needs to be revised to properly discriminate from the correct choice. MCQ No 13 had a p-value of 54, DI of 0.57, and DE of 75%, showing that it was moderately difficult and being able to differentiate students into different strata. From the upper group, 82% students selected the correct response C, while students of the lower group were distributed among all the choices. NCQ No 43 was very difficult and gave a negative DI, as most of the students in the upper category chose the alternative D, instead of the right answer E. This question had a DE of 100% as the distractors were not clear to many students, making it a very difficult item.

Discussion

The assessment tool is one of the strategies which should be designed according to the objective. If the objective is not clear, we would be unable to strategise the assessment tool and the test will be a failure. One-best MCQs, if properly written and well constructed, are one of the strategies of the assessment tool that quickly assess any level of cognition according to Bloom's taxonomy.[2] 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. A distractor used by less than 5% of students is not an effective distractor and should be either replaced or corrected as it affects the overall quality of the question. An NF-D makes the question easier to answer, thereby affecting the assessment of the student. Items having NF-D can be carefully reviewed and, with some alterations, given as the initial item on the test, as a 'warm-up' question. However, they would not be able to differentiate among students, if that is the purpose. Assessment of MCQs by these indices highlights the importance of assessment tools for the benefit of both the student and the teacher.[17]

The DE of difficult items in our study was 100% which was expected, as difficult items would require a lot of guesswork on the part of the student, thereby using all the distractors.

The numbers of NF-Ds also affect the discriminative power of an item. It is seen that reducing the number of distractors from four to three decreases the difficulty index, while increasing the DI and the reliability.[18] We observed that items having one NF-D had excellent discriminating ability (DI = 0.427) as compared to items with all four functioning distracters (DI = 0.351). This compares well with other studies favouring better discrimination by three distracters as compared to four.[7] This can be because writing items with four distractors is a difficult task and while writing the fourth distractor we are mostly trying to fill the gap, allowing it to become the weakest distractor. It was also observed that items having good difficulty index (p-value = 30 to 70) and good/excellent discrimination index (DI > 0.24), considered to be 'ideal questions,' had DE of 85.15%, which is close to items having one NF-D.

Tarrent et al found three-option items performing equally well as four-option items and have suggested to write three-option items as they require less time to be developed.[19] Similar observations were made by literature review conducted by Vyas and Supe.[20] We observed that items with two NF-Ds, comparable to three-option items, had difficulty index within acceptable range (p-value = 62.66) and were better discriminating (DI = 0.365) than items with no NF-D. Owen and Froman, however, did not see any difference in the difficulty and discrimination indices while reducing the number of distractors from five to three.[21]

Results from this study highlighted the importance of item analysis that included difficulty and discrimination indices and distractor analysis. Items having average difficulty and high discrimination with functioning distractors should be incorporated into future tests to improve the test development and review. This would also improve the overall test score and properly discriminate among the students.[22]

Conclusion

The study concluded that items with 3 distractors perform best in discriminating among the students. Items with two NF-Ds, though easier, are better discriminators than items with no NF-D.

References

1. Tan LT, McAleer JJ; Final FRCR Examination Board. The introduction of single best answer questions as a test of knowledge in the final examination for fellowship of the Royal College of Radiologists in Clinical Oncology. Clin Oncol (R Coll Radiol) 2008; 20: 571-6.
2. Carneson J, Delpierre G, Masters K. Designing and managing MCQs:

Appendix C: MCQs and Bloom's taxonomy. (Online) 2011 (Cited 2011 Feb 2). Available from URL: http://web.uct.ac.za/projects/cbe/mcqman/mcqappc.html.

3. Case SM, Swanson DB. Constructing written test questions for the basic and clinical sciences, National Board of Medical Examiners 3rd ed. (Online) 2010 (Cited 2011 Feb 5). Available from URL: http://www.nbme.org/publications/item-writing-manual.html
4. Collins J. Education techniques for lifelong learning: writing multiple-choice questions for continuing medical education activities and self-assessment modules. Radiographics 2006; 26: 543-51.
5. Considine J, Botti M, Thomas S. Design, format, validity and reliability of multiple choice questions for use in nursing research and education. Collegian 2005; 12: 19-24.
6. Sim SM, Rasiah RI. Relationship between item difficulty and discrimination indices in true/false - type multiple choice questions of a para-clinical multidisciplinary paper. Ann Acad Med Singapore 2006; 35: 67-71.
7. 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.
8. Miller MD, Linn RL, Gronlund NE, editors. Measurement and assessment in teaching. 10th ed. Upper Saddle River, NJ: Prentice Hall; 2009.
9. Fowell SL, Southgate LJ, Bligh JG. Evaluating assessment: the missing link? Med Educ 1999; 33: 276-81.
10. Carroll RG. Evaluation of vignette-type examination items for testing medical physiology. Am J Physiol 1993; 264: S11-5.
11. Dufresne RJ, Leonard WJ, Gerace WJ. Making sense of student's answers to multiple-choice questions. Phys Teach 2002; 40: 174-80.
12. Gronlund NE, Linn RL. Measurement and evaluation in teaching. 6th ed. New York: Macmillan publishing Co; 1990.
13. Tarrant M, Ware J. Impact of item-writing flaws in multiple-choice questions on student achievement in high-stakes nursing assessments. Med Educ 2008; 42: 198-206.
14. Carroll RG. Evaluation of vignette-type examination items for testing medical physiology. Am J Physiol 1993; 264: S11-5.
15. Kelley TL. The selection of upper and lower groups for the validation of test items. J Educ Psychol 1939; 30: 17-24.
16. Sarin YK, Khurana M, Natu MV, Thomas AG, Singh T. Item analysis of published MCQs. Indian Pediatrics 1998; 35: 1103-5.
17. Pellegrino J, Chudowsky N, Glaser R, editors. Knowing What Students Know: The Science and Design of Educational Assessment. Washington, DC: National Academic Press, 2001.
18. Rodriguez MC. Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. Educational Measurement: Issues and Practice 2005; 24: 3-13.
19. Tarrant M, Ware J. A comparison of the psychometric properties of three- and four-option multiple-choice questions in nursing assessments. Nurse Educ Today 2010; 30: 539-43.
20. Vyas R, Supe A. Multiple choice questions: a literature review on the optimal number of options. Natl Med J India 2008; 21: 130-3.
21. Owen SV, Froman RD. What's wrong with three-option multiple choice items? Educ Psychol Meas 1987; 47: 513-22.
22. Wallach PM, Crespo LM, Holtzman KZ, Galbraith RM, Swanson DB. Use of a committee review process to improve the quality of course examinations. Adv Health Sci Educ 2006; 11: 61-8.