International Journal of Community Medicine and Public Health
Christian DS et al. Int J Community Med Public Health. 2017 Jun;4(6):1876-1881
http://www.ijcmph.com
pISSN 2394-6032 | eISSN 2394-6040

DOI: http://dx.doi.org/10.18203/2394-6040.ijcmph20172004

Original Research Article

Evaluation of multiple choice questions using item analysis tool: a study from a medical institute of Ahmedabad, Gujarat

Donald S. Christian*, Arpit C. Prajapati, Bhavik M. Rana, Viral R. Dave

Department of Community Medicine, GCS Medical College, Ahmedabad, Gujarat, India

Received: 10 April 2017
Revised: 24 April 2017
Accepted: 25 April 2017

***Correspondence:**
Dr. Donald S. Christian,
E-mail: donald_christian2002@yahoo.com

ABSTRACT

Background: Multiple choice question (MCQ) assessments are becoming popular means to assess knowledge for many screening examinations among several fields including Medicine. The single best answer MCQs may also test higher-order thinking skills. Hence, MCQs remain useful assessment gadget. Objectives: 1) To evaluate Multiple Choice Questions for testing their quality. 2) To explore the association between difficulty index (p-value) and discrimination indices (DI) with distractor efficiency (DE). 3) To study the occurrence of functioning distractors for MCQs.
Methods: Total five MCQ test sessions were conducted among interns of a medical institute of Ahmedabad city Gujarat, between April 2016 to March 2017, as part of their compulsory rotating postings in the department. The average participation in each of the sessions was 17 interns, thus a total of 85 interns getting enrolled. For each test session, the questionnaire consisted of forty MCQs having 4 options including a single best answer. The MCQs were analyzed for difficulty index (DIF-I, p-value), discrimination index (DI), and distractor efficiency (DE).
Results: Total 85 interns attended the tests consisting of total 200 MCQ items (questions) from four major medical disciplines namely - Medicine, Surgery, Obstetrics & Gynecology and Community Medicine. Mean test scores of each test ranged from 36.0% to 45.8%.The reliability of the tests, the Kuder Richardson (KR) 20, ranged from 0.29 to 0.52. The standard error of Measurement ranged from 2.59 to 2.79.Out of total 200 MCQs, seventy nine (n=79) had Discrimination index (DI) <0.15 (poor), and 61 had DI ≥0.35 (excellent). Easy items having average DE of all tests was 20.1%.
Conclusions: Items having average difficulty and high discrimination with functioning distractors should be incorporated into tests to improve the validity of the assessment.

Keywords: Difficulty index, Discrimination index, Functioning distractors, Item Analysis, MCQs

INTRODUCTION

Multiple choice question (MCQ) assessments are becoming more and more prevalent for assessing knowledge for many professional courses including Medicine.[1] For students, the results from such assessment may determine what pathways are available to them in the future, and for teachers, the results are often subject to scrutiny by examination boards or used in benchmarking further studies.[2] Multiple-choice questions are used more and more in departmental examinations or as comprehensive examinations at the end of an academic session.[3] They may be used to determine progress or to make decisions regarding the certification of a candidate.[4] They may also be used to identify strengths and weaknesses in students as well as to provide feedback to

teachers on their educational actions. It is expected that the ability of the students to answer the questions or the items, reflect their subject knowledge. The knowledge and skills are the key competencies on which the quality of medical care would largely depend upon.[5] MCQs, whether in the format of "true/false" or "single best-answer", are expressly designed to assess knowledge. They have the advantage of sampling broad domains of knowledge efficiently and hence reliably.[6] This one characteristic of MCQs is sufficient to ensure that its edge in reliability more than compensates for some perceived failings in validity. Concerns have been voiced that most MCQs tend to measure factual recall and recognition of isolated facts. But if the MCQs are carefully made, the single best answer MCQs may also test higher-order thinking skills.[7] Therefore; MCQs remain a useful assessment instrument, despite some limitations and objections.

The manner in which the test questions are prepared and put together to form an examination; and the procedure for scoring, analyzing and reporting the results; all have a bearing upon the conclusions drawn from the performance of the individuals and groups tested. Assessment of knowledge is of utmost importance in medical education along with all the other aspects.[8] It is argued that academics are generally not specialists in the research discipline of assessment, and they do not routinely analyze their assessments.[9] Statistical analysis of the multiple choice test items can ensure that items are effectively evaluating students' learning. By items one means questions, statements or scenarios that are used on an assessment instrument. A "score" is actually a reflection of what you really knew (true score) and error (things like atmosphere, nerves etc. that modify your true score). The purpose of a systematic approach to test design is to reduce error in test-taking. Sometimes, things can inflate the test score, like someone letting you see the key beforehand; or deflate the scores, like the room was too cold or the test was too early in the morning.

Item analysis provides a way of measuring the quality of questions - seeing how appropriate they were for the respondents and how well they measured their ability. Item analysis provides objective evidence of the progress of the students towards the concepts of the subject and makes the topic easier for students.[10] Many times the teaching staff find it difficult to analyse the quality of the items those are repeatedly used for grading students' performance.[11-13] Item analysis also provides a way of re-using items over and over again in different instruments with prior knowledge of how they are going to perform. The difficulty of a single response selection question in classical analysis is simply the proportion of people who answered the question incorrectly. An item contains a stem (question) and four options, of which there are one key (correct response) and three distractors (incorrect responses).[14,15]

Objectives

1. To evaluate Multiple Choice Questions for testing their quality.
2. To explore the association between difficulty index (p-value) and discrimination indices (DI) with distractor efficiency (DE)
3. To study the occurrence of functioning distractors for MCQs

METHODS

The MCQ items were first chosen by investigator faculties and were vetted for content accuracy. The vetted questions (extracted from the bank) were then chosen. The final selection of the MCQ items for an examination paper was based purely on the academic judgment and examination experience. After the final selection, the selected MCQ items were used for the study and assigned according to the subjects and topics within the subject. During the period of one year (April 2016 to March 2017) of compulsory rotating internship in the department, the posted interns were periodically subjected to MCQ test sessions at an interval of 2 to 3 months duration. This lead to a total of five such MCQ sessions conducted over the period of one year time. The total number of interns who collectively participated in sessions turned out to be 85 (for one year). Thus the average number of interns per MCQ session came out to be 17. Informed consent was obtained from each participant after explaining the purpose of the study. A single test session consisted of 40 MCQs to be competed in 40 minutes of timeline. Each MCQ questionnaire consisted of 10 MCQs each from four major subjects: Medicine, Surgery, Obstetrics & Gynecology and Community Medicine.

Each MCQ item contained a stem and four options. A true response to an item was awarded 1 mark, while an incorrect response would result in the deduction of 1 mark, and a no-attempt or blank response (indicating "I don't know") was given 0 marks. There was carrying over of negative marks from one question to another. Thus, the maximum total score for any one question was 1 mark while the minimum total score will be -1 mark.

The results of students' performance in these MCQ tests were used to determine the difficulty index and discrimination index of each MCQ item in the respective tests. In this study, the item difficulty index (DIF I) (P) refers to the percentage of the total number of correct responses to the test item. It was calculated by the formula P = R/T, where R is the number of correct responses and T is the total number of responses (i.e., correct + incorrect + blank responses).[16] Hence, the higher this index value, the lower is the difficulty, and the greater the difficulty of an item, the lower is its index. The item discrimination index (DI), however, measures the difference between the percentage of students in the upper group (PU), i.e., the top 27% scorers, who obtained

the correct response, and the percentage of those in the lower group (PL), i.e., the bottom 27% scorers, who obtained the correct response; thus $D = P_U - P_L$.[16] 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.

DIF-I define the percentage of students who answered the item correctly and ranges between 0 and 100%.[16] The criteria for DIF-I are: DIF-I >70 (Too easy); DIF-I between 30-70 (Average) and DIF-I <30 (Too difficult). DI is the ability of an item to distinguish between students of higher and lower capacities and ranges between 0 and 1. The criteria for Discrimination index (DI) s are: DI<0.15 (poor), 0.15-0.24 (marginal), DI 0.24-0.34 (good) and DI ≥0.35 (excellent). Greater the value of DI, item is more able to discriminate between students of higher and lower capacities. DI of 1 is best as it denotes to an item which perfectly discriminates between students of lower and higher abilities. There are occurrences when the value of DI can be <0 (negative DI) which simply means that the students of lower ability answer more correctly than those with higher ability. The five test sessions (a total of 200 items) had 200 correct answers and 600 distractors. The data were entered in MS Excel and were analyzed using SPSS (SPSS Inc. USA) software (Evaluation version 15). The test was analyzed for the reliability index by Kuder - Richardson Formula 20 (KR20), difficulty index (DIF I), distractor analysis (DE) and discrimination index (DI). A non-functioning distractor was defined as an option with a response frequency of <5%.[17] After collecting the basic information about the Non Functioning Distractors (NFD) and Functioning distractors, the items were categorized based on the number of NFDs; i.e. 0 NFD, 1 NFD, 2 NFDs, and 3 NFDs. The study was approved by Institutional Ethical & Scientific board of the institute.

RESULTS

A total 5 test session were taken during the period of one year of and total 85 interns attended the tests consisting of total 200 best MCQ items from four major subjects.

Table 1 shows the description of tests that was taken among interns. The numbers of items in each test were 40 whereas the number of examinees in 1st test, 2nd test, 3rd test, 4th test and 5th test were 18, 15, 18, 17, and 17 respectively. Percentage of mean test scores of each test ranged from 36.0% to 45.8%.The reliability of the tests, the Kuder Richardson (KR) 20, ranged from 0.29 to 0.52. Also the standard error of Measurement ranged from 2.59 to 2.79. The mean score% achieved were 45.4 ± 8.5 (maximum 40 marks), 45.8 ± 10.1, 37.9 ± 9.2, 36.0 ± 7.7 and 44.6 ± 9.6 in 1st, 2nd, 3rd, 4th and 5th test respectively. After receiving the result, interns were classified in order of merit from the highest score to the lowest score.

As shown in Table 2, out of total 200 MCQs, 79 had Discrimination index (DI) <0.15 (poor), 47 had DI between 0.15 to 0.24 (marginal), 13 had DI between 0.24 and 0.34 (good) and 61 had DI ≥0.35 (excellent). As shown in Table 3, the distribution of difficulty and discrimination indices of the 200 items specified and their corresponding Distractor Efficiency was also worked out for all. Half of the items (50%) were of average (recommended) difficulty with a mean p-value of 49.17 ± 10.4 in first test, more than half in 2nd test and 3rd test with mean p value 49.3 ± 12.0, 46.2 ± 10.9, while in 4th & 5th tests less than half of the items with p value 47.9 ± 10.0 & 51.3 ±11.6 respectively. The relation of mean difficulty with mean distractor efficiency was also analyzed. DE was indirectly associated to the p-value with most difficult items having DE of average 85.6%, 80.0%, 80.4%, 85% and 81.3% in 1st, 2nd, 3rd, 4th and 5th test respectively while easy items having average DE of all tests was 20.1%.

Table 1: Characteristics of MCQ test sessions.

	1st Test	2nd Test	3rd Test	4th Test	5th Test	Total
No. of items	40	40	40	40	40	200
No. of examines	18	15	18	17	17	85
Percentage of mean test score ± 2SD	45.4±8.5	45.8±10.1	37.9±9.2	36.0±7.7	44.6±9.6	---
Range of test scores (%)	17.5 – 55	27.5 – 67.5	25 – 62.5	22.5 – 47.5	30 – 60	---
Kuder-Richardson 20 (KR 20)	0.39	0.52	0.42	0.29	0.51	---
Standard Error of Measurement (SEM)	2.66	2.79	2.79	2.59	2.68	---

Table 2: Classification of questions by discrimination index (DI).

Discrimination index	1st Test No. of items n=40	2nd Test No. of items n=40	3rd Test No. of items n=40	4th Test No. of items n=40	5th Test No. of items n=40	Total
≥0.35	10 (25%)	09 (22.5%)	12 (30%)	13 (32.5%)	17 (42.5%)	61
0.25-0.34	00 (0%)	13 (32.5%)	00 (0%)	00 (0%)	00 (0%)	13
0.15-0.24	17 (42.5%)	00 (0%)	14 (35%)	06 (15%)	10 (25%)	47
<0.15	13 (32.5%)	18 (45.0%)	14 (35%)	21 (52.5%)	13 (32.5%)	79
Total	40	40	40	40	40	200

Table 3: Classification of questions according to difficulty indices for all 5 tests.

		1st Test			2nd Test			3rd Test			4th Test			5th Test		
P value	**Interpretation**	**Mean P value**	**DE (%)**	**No. of items n=40(%)**	**Mean P value**	**DE (%)**	**No. of items n=40 (%)**	**Mean P value**	**DE (%)**	**No. of items n=40 (%)**	**Mean P value**	**DE (%)**	**No. of items n=40 (%)**	**Mean P value**	**DE (%)**	**No. of items n=40 (%)**
>70	Too easy	84.03 ±8.1	15.97 ± 8.1	08 (20)	81.11 ±6.6	18.89 ±6.6	06 (15)	79.6 ±3.2	20.4 ± 3.2	03 (7.5)	78.4 ±6.1	21.6 ± 6.1	06 (15)	76.5 ± 9.6	23.5 ± 9.6	10 (25)
30-70	Average	49.17 ±10.4	50.83 ± 10.4	20 (50)	49.3 ±12.0	50.7 ±12.0	23 (57.5)	46.2 ±10.9	53.8 ± 10.9	22 (55.0)	47.9 ±10.0	52.1 ± 10.0	14 (35)	51.3 ± 11.6	48.7 ±11.6	14 (35)
<30	Too Difficult	14.35 ±8.7	85.65 ± 8.7	12 (30)	20.00 ± 7.3	80.0 ±7.3	11 (27.5)	19.6 ±8.4	80.4 ± 8.4	15 (37.5)	15.0 ±8.6	85.0 ± 8.6	20 (50)	18.8 ±8.4	81.3 ±8.4	16 (40)

Table 4: Classification of test MCQs according to discrimination indices (DI).

		1st Test		2nd Test		3rd Test		4th Test		5th Test	
Discrimination index (DI)	**Interpretation**	**Mean P value**	**DE (%)**	**Mean P value**	**DE (%)**	**Mean P value**	**DE (%)**	**Mean P value**	**DE (%)**	**Mean P value**	**DE (%)**
≥0.35	Excellent	45.6 ± 22.4	54.4 ± 22.4	45.9 ± 10.8	54.1 ± 10.8	36.6 ± 12.6	63.4 ± 12.6	48.9 ± 18.7	51.1 ± 18.7	49.8 ± 22.9	50.1 ± 22.9
0.25-0.34	Good	---	---	53.9 ± 26.6	46.2 ± 26.6	---	---	---	---	---	---
0.15-0.24	Marginal	54.6 ± 24.9	45.4 ± 24.9	---	---	33.7 ± 19.8	66.3 ± 19.8	55.9 ± 30.9	44.1 ± 30.9	48.2 ± 29.1	51.8 ± 29.1
<0.15	Poor	34.2 ± 28.3	65.8 ± 28.3	40.4 ± 22	59.6 ± 22	45.6 ± 23.7	54.4 ± 23.7	22.4 ± 18.9	77.6 ± 18.9	34.8 ± 24.5	65.2 ± 24.5

P value= Discrimination index (DI), DE= Distractor Efficiency.

As per Table 4, majority of items (42.5%) had marginal or poor DI. The DE was directly related to the discrimination index. Items with good and excellent discrimination had DE of 71.42% and 83.06% respectively. The relation of mean discrimination indices with mean distractor efficiency was also analyzed. Incorrect key, confusing framing of questions or generalized poor preparation of students may the causes for negative discrimination index. Items with negative DI decrease the validity of the test and should be discarded from the collection of MCQs.

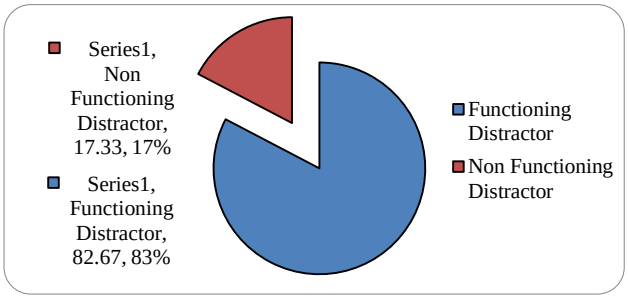

Figure 1: Distractor performance (n=600).

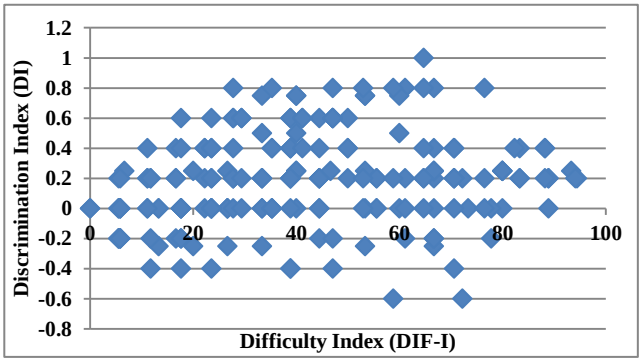

Figure: 2 Relationship between difficulty index (DIF-I) and discrimination index (DI).

Table 5: Functioning distractors per item

Tests	Item having One Distractor	Item having Two Distractor	Item having Three Distractor
1st Test	03 (7.5%)	15 (37.5%)	22 (55%)
2nd Test	05 (12.5%)	13 (32.5%)	22 (55%)
3rd Test	01 (2.5%)	13 (32.5%)	26 (65%)
4th Test	02 (5%)	12 (30%)	26 (65%)
5th Test	09 (22.5%)	15 (37.5%)	16 (40%)

Overall, 200 items and 600 distractors were assessed. 104 distractors (17.3%) had a choice frequency of <5%. A considerable percentage of distractors were so implausible (17.3%) and therefore they were not selected by anybody as shown in Figure 1. As Table 5 shows 496 (82.7%) of all distractors were classified as functioning. The proportion of items with three functioning distractors ranged from 40% on 5th test to 65% on both 3rd and 4th Test. Overall 56% of items had three functioning distractors. Figure 2 shows that discrimination index correlate poorly with difficulty index.

DISCUSSION

The present item analysis uses single best response questions as it is seen as an efficient way of evaluation in academics.[17] It is important to align cognitive domain with tasks to be achieved. Item analysis provides one such method of analyzing observation and interpretation of the knowledge gained by the students.[18]

In present study the ranges of difficulty index for 1st test was 5.5% to 94.4%, for 2nd test 6.7% to 93.3%, for 3rd test 5.6% to 83.3%, for 4th test 0.0% to 88.2% and for 5th test 5.9% to 94.1% respectively. In several studies, the ranges of difficulty index were found to be 41-60%.[19] For discrimination index, the limits of upper 27% and lower 27% was used by Kelley et al and is used widely for calculating the same.[20] In present study, overall, 200 items and 600 distractors were assessed. 104 distractors (17.3%) had a choice frequency of <5% while in similar study conducted by of 100 students with 100 items with 300 distractors were studied with total 24% items with NFD and remaining 76% items were with functional distracters.[21]

Mean DE in present study for 1st test was 54.30 ±26.04, for 2nd test was 54.0 ±21.8, for 3rd test was 61.25 ± 19.5, for 4th test was 64.0 ±24.8 and for 5th test was 55.4 ± 25.1, lower than DE of 81.4%reported elsewhere in a similar type of study.[17] Present study shows discrimination index correlate poorly with difficulty index. The relationship between difficulty index and discrimination index is dome shaped rather than linear.[16] First, as difficulty index increases, the discrimination index also increases, but at a p value between 40% and 60%, DI reaches a maximum. When p is more than 60%, DI falls. Over the range 40% - 60%, the DI is more than 0.5.[22]

CONCLUSION

Out of total 200 MCQs, 79 had discrimination index (DI) <0.15 (poor), 47 had DI between 0.15 to 0.24 (marginal), 13 had DI between 0.24 to 0.34 (good) and 61 had DI ≥0.35 (excellent). Majority of items (42.5%) had Marginal or poor DI. Overall, 200 items and 600 distractors were assessed. 104 distractors (17.3%) had a choice frequency of <5%. Substantial proportions of distractors were so improbable (17.3%) and were not selected by anybody. Though the number of items in each test was less and the number of interns in each test was also limited, the study surely provided valid information regarding the quality of the questions used.

Recommendations

The MCQ Item Analysis should be practiced more often in academics, as they provide the insight to the quality of

questions being asked. Finding from this research emphasized the significance of item analysis that included difficulty and discrimination indices and distractor analysis, which are often overlooked for many such examinations. Items having average difficulty and high discrimination with functioning distractors should be incorporated into tests to improve validity of the tests as well to improve effectiveness of the questions.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the interns for their contributions to the test sessions.

Funding: No funding sources
Conflict of interest: None declared
Ethical approval: The study was approved by the Institutional Ethics Committee

REFERENCES

1. Hubbard JP, Clemans WV. Multiple-choice Examinations in Medicine: A Guide for Examiner and Examinee; Lea & Fabiger, Libraries Australia; 1961: 186.
2. Crisp GT, Palmer EJ. Engaging Academics with a Simplified Analysis of their Multiple - Choice Questions (MCQ) Assessment Results. J University Teach Learn Practice. 2007;4(2):88-106.
3. Clemans WV. Multiple-choice Examinations in Medicine: A Guide for Examiner and Examinee. London: Lea & Fabiger, WHO Library. 1965;18:61.
4. De Champlain AF, Melnick D, Scoles P, Subhiyah R, HoltzmanK, Swanson D, et al. Assessing medical students' clinical sciences knowledge in France: a collaboration between the NBME and a consortium of French medical schools. Acad Med. 2003;78:509-17.
5. 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 of Ahmedabad, Gujarat. Indian J Community Med. 2014;39 (1):17-20.
6. Norman G. Evaluation methods: A resource handbook. In: Shannon S, Norman, G, editors. Chapter 4.1. Multiple choice questions, The Program for Educational Development, McMaster University. Hamilton, Canada: McMaster University; 1995: 47-54.
7. Peitzman SJ, Nieman LZ, Gracely EJ. Comparison of "fact-recall" with "higher-order" questions in multiple-choice examinations as predictors of clinical performance of medical students. Acad Med. 1990;65:59-60.
8. Ross MM, McDonald B, McGuinness J. The palliative care quiz for nursing (PCQN): the development of an instrument to measure nurses' knowledge of palliative care. J Adv Nurs. 1996;23:126-37.
9. Knight, P. The local practices of assessment. Assessment & Evaluation in Higher Education. 2006;31(4):435-52.
10. Biswas SS, Jain V, Agrawal V, Bindra M, Small group learning: effect on item analysis and accuracy of self-assessment of medical students, Educ Health (Abingdon). 2015;28(1):16-21.
11. Fowell SL, Southgate LJ, Bligh JG. Evaluating assessment: the missing link? Medic Education. 1999;33:276-81.
12. Kehoe J. Basic item analysis for multiple-choice tests. Practical Assessment, Research & Evaluation. 1995;4(10):20-4.
13. Maunder P. In support of multiple choice questions: some evidence from Curriculum 2000, Paper presented at the Annual Conference of the British Educational Research Association, University of Exeter, England, Education Line; 2002: 12-14.
14. Matlock-Hetzel S. Basic concept in item and test analysis, Presented at annual meeting of the Southwest Educational Research Association, Austin; 1997: 1-27.
15. Eaves S, Erford B. The Gale group: The purpose of item analysis, item difficulty, discrimination index, characteristic curve. 2007 edition, Online Source: Available at: www.education.com/reference/article/itemanalysis/. Accessed on 21 April 2017.
16. Sim S, 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.
17. Hingorjo MR, Jaleel F. Analysis of one-best MCQs: The difficulty index, discrimination index and distracter efficiency. J Pak Med Assoc. 2012;62:142-7.
18. Pellegrino J, Chudowsky N, Glaser R, (Book) Knowing What Students Know: The Science and Design of Educational Assessment. Washington, DC: National Academic Press; 2001: 103-108.
19. Guilbert JJ. Educational Hand-Book for health professionals, WHO offset Publication 35, 1 ed. Geneva: World Health Organization; 1981: 461-465.
20. Kelley TL. The selection of upper and lower groups for the validation of test items. J Educ Psychol. 1939;30:17-24.
21. Patil VC, Patil HV, Item Analysis of Medicine Multiple Choice Questions (MCQs) for Under Graduate (3rd year MBBS) Students. RJPBCS. 2015;6(3):1242
22. Mukherjee P, Lahiri SK, Analysis of Multiple Choice Questions (MCQs): Item and Test Statistics from an assessment in a medical college of Kolkata, West Bengal, IOSR. J Dental Med Sci. 2015;14(12):47-52.

Cite this article as: Christian DS, Prajapati AC, Rana BM, Dave VR. Evaluation of multiple choice questions using item analysis tool: a study from a medical institute of Ahmedabad, Gujarat. Int J Community Med Public Health 2017;4:1876-81.