 DOI: 10.18231/2393-8005.2017.0007

Use of Item analysis to improve quality of Multiple Choice Questions in II MBBS

Rupal M. Patel

Associate Professor, Dept. of Microbiology, Pramukhswami Medical College, Karamsad, Aanad

***Corresponding Author:**
Email: drrplpatelmp@gmail.com

Abstract

Item analysis helps in assessing the quality of the multiple choice items and of the test as a whole.

- To analyze the quality of MCQs & to determine internal consistency reliability of test.
- To evaluate perception of teachers towards item analysis.

In a cross-sectional study, total 40 items from Microbiology First Continuous Evaluation Test of 83 students of 2nd year M.B.B.S. were analysed. Item & test analysis was done by calculating Difficulty index (DIF I), Discrimination index (DI), Distractor effectiveness (DE), Point Bi-serial correlation (PBS) and Cronbach's Alpha. Feedback of faculties was collected on a 5 point Likert scale.

Difficulty index of 12(30%) & 18(45%) items were in the ideal (50-60%) & acceptable range (30-70%) respectively, 7(17.5%) items were easy (>70%) and 3(7.5%) items were difficult (<30%). Discrimination index of 17(42.5%) items was excellent (=>0.36), 7(17.5%) items was good (0.25-0.35), 1(2.5%) item was acceptable (0.20-0.24) and 15(37.5%) items was poor (=<0.20). Out of 15 items with poor DI, 6 (40%) items were of easy and difficult level of DIF I and 3 items (20%) were poorly framed but rest of the items were appropriately framed testing higher level of cognitive domain. Out of 120 distractors analysed, 102(85%) were functional & 18(15%) were non-functional. Thirty (75%) items had acceptable PBS values (>0.15). Cronbach's Alpha improved from 0.702 to 0.794 on removal of 15 items with poor DI. Most of the faculties found item analysis useful to improve quality of MCQs.

Majority of the items had acceptable level of difficulty & discrimination index. Most of distractors were functional. Item analysis helped in revising items with poor discrimination index and thus improved the quality of items & a test as a whole.

Keywords: Item Analysis, MCQs.

Introduction

Assessment is an essential part of the learning process in education.[1] Multiple Choice Questions (MCQs) is the most commonly used tool for assessing the knowledge capabilities of medical students.[2] Designing MCQs is a complex and time consuming process in a multidisciplinary integrated curriculum.[2,3] 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.[3,4]

Item analysis is a process which examines student responses to individual test items (questions) in order to assess the quality of those items and of the test as a whole. Item analysis is especially valuable in improving items which will be used again in later tests, but it can also be used to eliminate ambiguous or misleading items in a single test administration.[5] Item analysis enables teachers to get an active feedback from students and determine areas which require emphasis, reinforcement or an alteration in teaching methodology perhaps using other learning aids.[6]

Item analysis enables identifying good MCQs based on difficulty index also denoted by facility value (FV) or *P*-value, discrimination index, and distractor effectiveness.[7] As per the literature, varying degrees of difficulty index and discrimination index have been found.[2,3,4,7,8,9] High quality MCQs, however, also needs well-written alternatives.[10] In their review of functioning and non-functioning distractors in 514 four-option MCQs assessments, Tarrant et al[10] found that only 13.8% of all items had three functioning distractors and just over 70% had only one or two functioning distractors.[10] Point Bi-serial correlation (PBS) is yet another important parameter which gives information about the 'fit' of an item with the remaining test. PBS helps us to identify items which are not testing the same domain / construct as rest of the test and thereby helps to improve the validity and reliability of the test. Assessing the quality of items used in a test can assess test as a whole. However, the reliability coefficient and standard error of measurement help to evaluate the performance of the test as a whole.[11] Reliability tells us whether a test is likely to yield the same results if administered to the same group of test-takers multiple times.[12] The most frequently reported internal consistency reliability estimates are the K-R20 and Cronbach's alpha.[11]

Medical teachers are unable to use item analysis to assess the quality of MCQs. So many defective items can be there in a test which is harmful for the students for the purpose of assessment. There is no information whether the students have learned the concept being tested and whether there is a need to re-visit the topic again. Item analysis helps the instructors to identify the effectiveness of their test items.[13] Hence the present study was taken up to analyse the quality of MCQs in the subject of Microbiology and to evaluate perception of teachers towards item analysis.

Aim and Objectives

1. To analyze the quality of MCQs by calculating difficulty index, discriminating index, distractor effectiveness and Point Bi-serial correlation (PBS) of an item.
2. To determine internal consistency of the test by the method Cronbach's Alpha.
3. To evaluate perception of teachers towards item analysis.

Materials and Method

The cross-sectional study was conducted in the Department of Microbiology, Pramukhswami Medical College, Karamsad between October 2015 to March 2016. Approval from Institutional Ethical committee was obtained. Total 40 items of single best response type, from General Microbiology & Immunology section, of 1st Continuous Evaluation Test (CET) examination of eighty three 2nd year MBBS students, were analysed. There was no negative marking and the time allotted was one hour. Evaluation was done out of 40 marks and 50% score was the passing mark. Individual faculty as per the topic covered by them framed MCQ items. All the faculties including co-ordinator & Head of the department have received training in framing MCQs at MCI basic course workshops. Co-ordinator and Head of the Department did pre-validation of the MCQs. To avoid possible copying from neighbouring student, the test was administered in three sets of papers with disorganized sequencing of questions. Item analysis was carried out by the co-ordinator with the help of departmental clerical staff.

Test papers were ranked in rank order from highest to lowest score. One third (n=28) of the papers with high score were selected and were referred as Higher group. One third (n=28) of the papers with low score were selected and were referred as Lower group. The middle one third (n=27) papers were not used in analysis. Each item was analyzed for DIF I, DI & DE[6] in MS Excel 2010. Point Bi-serial correlation (PBS)[12] and Cronbach's Alpha[11] were calculated in SPSS 14.

Difficulty index (P)[6,9]: It was calculated as the percentage of students who answers the item correctly using following formula:

- $P = [(H + L)/N] \times 100$
  - H = Number of correct responses in upper group.
  - L = Number of correct responses in lower group.
  - N = Total number of response in both groups.
- Difficulty index was interpreted as follows:

Value	Interpretation
30-70%	Acceptable
50-60%	Ideal
Above 70%	Very easy
Below 30%	Very difficult

Discrimination index (d)[6,9]: It was calculated as the ability of an item to differentiate between good students and not so good students with the following formula:

- $d = 2 \times [(H - L)/N]$
  - H = Number of correct responses in upper group.
  - L = Number of correct responses in lower group.
  - N = Total number of response in both groups.
- Discrimination index was interpreted as follows:

Value	Interpretation
=> 0.36	Excellent
0.25 – 0.35	Good
0.21 – 0.24	Acceptable
=< 0.20	Poor

Distractor effectiveness[6]: Any of the distractors in the item which have not attracted even 5% of the total response was considered as non-functional distractor (NFD). On the basis of number of NFDs in an item, DE ranges from 0 to 100%. If an item contains three or two or one or nil NFDs then DE would be 0, 33.3%, 66.6% and 100% respectively.

Statistical analysis: All data were expressed as mean ± SD.

Point biserial correlation (PBS)[12]:

- It is the correlation between the right/wrong scores that students receive on a given item and the total scores that the students receive when summing up their scores across the remaining items.
- PBS was calculated in SPSS 14. PBS values can range from – 1.0 to + 1.0.
- Values between 0.15 & 0.35 were considered acceptable.

Cronbach's Alpha[11]:

- It is the measure of the internal consistency reliability used to evaluate the performance of the tests as a whole.
- It can range from 0 (if no variance is consistent) to 1.00 (if all variance is consistent) with all values between 0 and 1.00 also being possible. The higher the correlation among the items, the greater the Cronbach's alpha.
- In the present study, Cronbach's alpha was calculated in SPSS 14. SPSS output computes the reliability coefficient for the test excluding one item at a time. If the reliability increases when an item is deleted, that indicates that the item is problematic and reduces the reliability instead of increasing it.
- The computation of Cronbach's Alpha when a particular item is removed from consideration is a good measure of that item's contribution to the entire test's assessment performance.

Following general guidelines were used to interpret reliability coefficients for classroom exams[6]:

Reliability	Interpretation
0.90 and above	Excellent reliability; at the level of the best standardized tests
0.80 - 0.90	Very good for a classroom test
0.70 - 0.80	Good for a classroom test; in the range of most. There are probably a few items which could be improved.
0.60 - 0.70	Somewhat low. This test needs to be supplemented by other measures (e.g., more tests) to determine grades. There are probably some items which could be improved.
0.50 - 0.60	Suggests need for revision of test, unless it is quite short (ten or fewer items). The test definitely needs to be supplemented by other measures (e.g., more tests) for grading.
0.50 or below	Questionable reliability. This test should not contribute heavily to the course grade, and it needs revision.

Post item analysis, a feedback form evaluated perception of teachers towards item analysis. In a Departmental meeting, feedback on item analysis results was provided to faculties. Items with poor DI & NFDs were reviewed and decision on retention, revision or removal of the faulty items was taken. Feedback was also given to students on their learning.

Results

Total 40 MCQs and 120 distractors were analysed. Means and standard deviations (SD) for DIF I (%), DI and DE (%) were 55.9± 15.7%, 0.29± 0.20, and 84.94± 22.58%, respectively (Table 1).

Table 1: Characteristics of the Continuous Evaluation Test (CET)

No. of items	40
No. of students	83
Mean test score % (SD)	22.5(5.3)
Range of test scores	11 - 32
Difficulty index % (P)	
mean ± SD	55.9± 15.7
Range	21.4 - 82.1
Discrimination index (d)	
mean ± SD	0.29± 0.20
Range	-0.04 - 0.68
Distractor effectiveness %	
mean ± SD	84.94± 22.58
Range	33.3 -100

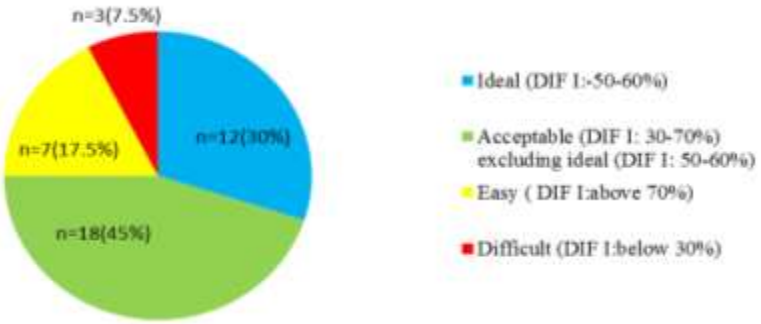

Fig. 1: Difficulty index: proportion of ideal, acceptable, easy and difficulty items (n=40)

75% of items had a difficulty index between the ranges of 30-70% where 12(30%) items had 50-60% (ideal) level of difficulty index (Fig. 1).

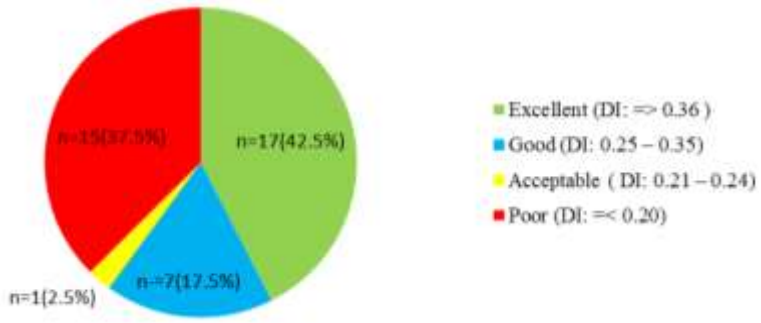

Fig. 2: Discrimination index: proportion of excellent, good, acceptable and poor items (n=40)

62.5% of the items had a discrimination index of more than >0.20 (Fig. 2). Out of 15 items with poor DI; 6 (40%) items were of easy and difficult level of DIF I and three items (20%) were poorly framed but rest of the items were appropriately framed testing higher level of cognitive domain.

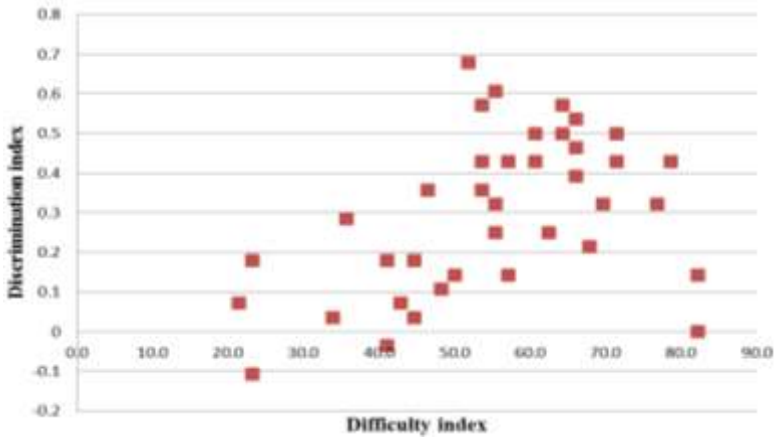

Fig. 3: Scatter plot showing relationship between difficulty & discrimination index of items (n=40)

As the DIF I increased, the DI also increased. At a DIF I between 50% and 60%, DI reached a maximum. When DIF I was more than 70%, DI decreased (Fig. 3). 42.5% of the test items with DIF I between 46.4% and 78.6% had excellent DI. 37.5% of the test items with a difficulty index ranging between 21.4 – 82.1% had poor DI.

Table 2: Distractor Performance

Number of items	40
Total distractors	120
Functional distractors	102 (85%)
Non-functional distractors(NFDs)	18 (15%)
Items with three functional (0 NFD, DE=100%) distractors	26(65%)
Items with two functional (1 NFD, DE=33%) distractors	10(25%)
Items with one functional (2 NFDs, DE=66%)distractor	4(10%)
Items with zero/none functional (3 NFDs, DE=0%) distractor	0

Out of 120 distractors analysed, 102(85%) were functional with their DE being 100%. Only 18(15%) distractors were non-functional (Table 2).

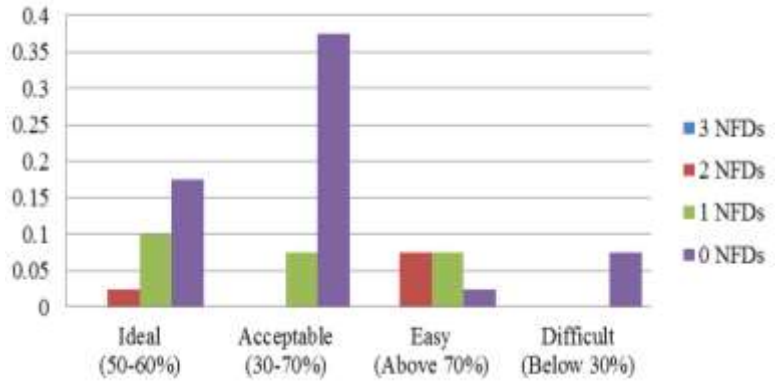

Fig. 4: Relationship between number of non-functioning distractors (NFDs) and item difficulty (n=40 items)

Presence of three functional distractors made items difficult as compared to items with one or two functional distractors, while items with only one or two functional distractors were found to be easy (Fig. 4).

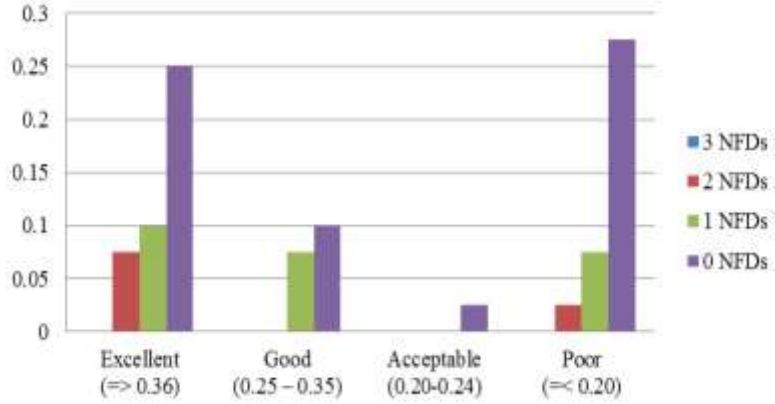

Fig. 4: Relationship between number of non-functioning distractors (NFDs) and discrimination (n=40 items)

Presence of three functioning distractors was seen in items with excellent & poor DI (Fig. 5).

Table 3: Distractor effectiveness with different values of DIF I & DI (N=40 items)

	Difficulty index		**Discrimination index**	
	Easy (Above 70%)	**Difficult (Below 30%)**	**Excellent (=> 0.36)**	**Poor (=< 0.20)**
No. of items	7	3	17	15
DE (%) mean ± SD	57.1± 25.2	100.0 ± 0.0	80.37 ± 26.5	88.8 ± 20.5

When viewed in relation of difficulty level of questions, DE was 100% in 3 difficult items than 57.1% in 7 easy items. However, little variation in DE was seen in items with excellent or poor DI. Mean DE was 88.8% in 15 items with poor DI compared to 80.3% in 17 items with excellent DI (Table 3).

Point biserial correlation (PBS):

- Thirty (75%) items had acceptable Point Bi-serial values (>0.15). Five items were acceptable according to PBS ranges (<0.15), but as per the discrimination index calculated manually; they were having poor DI (=<0.20).

Cronbach's Alpha:

- Cronbach's Alpha of 40 items was 0.702. Removal of 15 items with poor DI increased the value of Cronbach's Alpha from 0.702 to 0.794.

Most of the faculties found item analysis useful (Fig. 6).

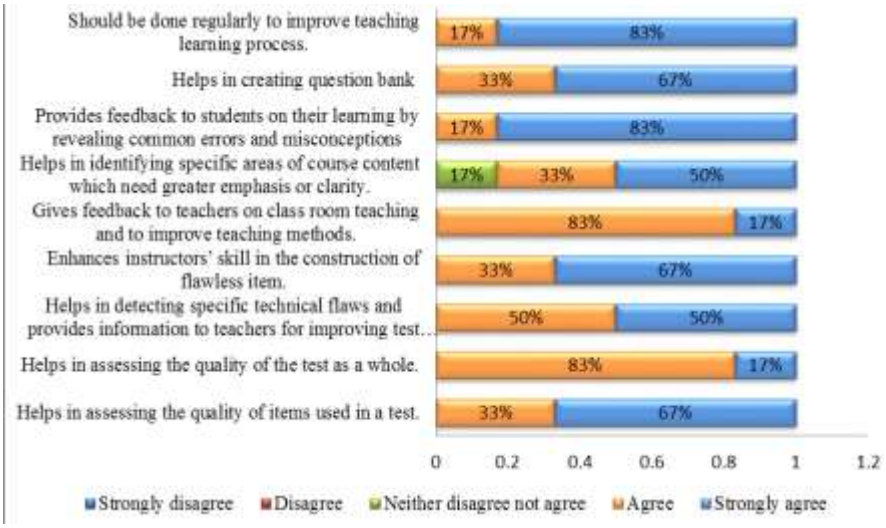

Fig. 6: Perception of faculties towards item analysis (n=10) (5 point likert scale)

Discussion

Single best 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 the analysis of item and test as a whole together referred as item and test analysis.[7]

Mean DIF I, DI & DE in the present study is comparable to the finding of Kolte et al[13] & Mehta et al[2] and actually better than the finding of the study by Gajjar et al[7] for DIF I & DI.[2,7,13]

Item difficulty is relevant for determining whether students have learned the concept being tested.[5] In the present study, a total of 30(75%) items had acceptable level of difficulty index and few items were easy & difficult. Karelia et al[4] showed 61% items in acceptable range (p 30-70%), 24 % items (p>70%) and 15 % items (p< 30%).[4] Mehta et al[2] showed that the p value of 26 (65%) items was in acceptable range (30 – 70%),10(25%) items were easy with p value >70% & 4(10%) items were difficult with p value <30%.[2] It is important to realize that easy questions need not be useless, although they are likely to be less discriminative.[8] Too easy items should be placed at the start of the test as 'warm-up' questions.[7] 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.[3] Difficult items should be reviewed for possible confusing language, areas of controversies, or even an incorrect key.[7]

Item discrimination refers to the ability of an item to differentiate among students on the basis of how well they know the material being tested.[5] In the present study, a total of 62.5% of the items had acceptable to excellent discrimination index of >0.20. Kolte et al[13] found that total 24(60%) items had excellent discriminative power, 7 (17.5%) items had good discriminative power and total 8 (20%) items had acceptable discriminative power.[13]

In the present study, fifteen (37.5%) items had poor discrimination index (=< 0.20). Earlier studies have revealed 30%[2] and 4.6%[9] of items with DI =< 0.20.[2,9] It is obvious that a question which is either too difficult (done wrongly by everyone) or too easy (attempted correctly by everyone) will have nil to poor DI.[13] In the present study; 6 (40%) items out of 15 were of easy and difficult level of DIF I and three items (20%) were poorly framed but rest of the items were appropriately framed testing higher level of cognitive domain.

There are instances when the value of DI can be less than 0 (negative DI).[7,13] In the present study, two (5%) items had negative DI values. Kolte et al[13] found that only one (2.5%) item had poor discrimination and 0% of total items had negative discriminative power.[13] Some studies have shown negative DI in 20% of items.[7] Items with negative indices should be examined to determine why a negative value was obtained.[5] Reasons for negative DI can be wrong key, ambiguous framing of question or generalized poor preparation of students.[2,7]

MCQ items with good discriminating potential tend to be moderately difficult items.[13] It has been seen that the relationship between DIF I and DI is not linear, but predicted as dome shaped.[7] When difficulty index was analysed along with discrimination index as shown in Figure 3, 42.5% of the test items with difficulty index between 46.4% and 78.6% had excellent discrimination index. Karelia et al[4] found that 46% of the test items with difficulty index between 25.93% and 80% had excellent discrimination index.[4]

Most difficult task in formatting good quality MCQs is writing appropriate options to the correct answer.[13] In the present study, out of 120 distractors analysed, 102(85%) were functional & only 12(15%) distractors were non-functional. A total of 65% of items had three functioning distractors and none of the items had zero functioning distractors. Gajjar et al[7] have shown that, in a total of 150 distracters, 133(89.6%) were functional distracters, and 17(11.4%) were NFDs which is very

similar to the findings of the present study.[7] Mehta et al[2] have shown, in their study, with fifty MCQs, having150 distracters, 53(35.33%) were found to be NFDs, 28(18.66%) were functional distracters and 69(46.01%) distracters had nil response.[2] Therefore, designing of plausible distractors and reducing the NFDs is important aspect for framing quality MCQs.[7] More NFD in an item increases DIF I (makes item easy) and reduces DE, conversely item with more functioning distractors decreases DIF I (makes item difficult) and increases DE.[7] In the present study, a total of 26 (65%), 10 (25%) and 4 (10%) items had three, two and one functional distractors and none of the items had zero functioning distractors. Mehta et al[2] in his study have shown that on the basis of number of NFDs; items with DE 66.6% were 18(54.4%), items with DE 33.3% were 9 (27.27%) and items with DE as 0 were 6(18.18%). The remaining 17 items with three functional distractors had DE as 100%.[2]

Point Bi-serial Correlation is an indicator of the item's discrimination effectiveness. The advantage of using discrimination coefficients over the discrimination index (D) is that every person taking the test is used to compute the discrimination coefficients and only 54% (27% upper + 27% lower) are used to compute the discrimination index, D.[14] Most of the items in the present study showed acceptable discrimination coefficient and items with low PBS values also shown poor DI. Cut off value for PBS & DI was 0.15 & =< 0.20 respectively. Accordingly, 10 & 15 items were found to be unacceptable by PBS & DI respectively. Five items found acceptable by PBS were actually having poor DI. A point biserial value of at least 0.15 is recommended, though it has been shown that "good" items have point biserial above 0.25.[12] A low point biserial implies that students who got the item incorrect also scored high on the test overall while students who got the item correct scored low on the test overall. Something in the wording, presentation or content of such items may explain the low point biserial correlation.[12]

The reliability of a test refers to the extent to which the test is likely to produce consistent scores. Reliability coefficient computed by Cronbach's alpha theoretically range in value from zero (no reliability) to 1.00 (perfect reliability).[5] High reliability means that students who answered a given question correctly were more likely to answer other questions correctly. If a parallel test were developed by using similar items, the relative scores of students would show little change. Low reliability means that the questions tended to be unrelated to each other in terms of who answered them correctly. The resulting test scores reflect peculiarities of the items or the testing situation more than students' knowledge of the subject matter.[5] In the present study, Cronbach's Alpha of 40 items was 0.702. Removal of problematic items (misfitting items, poorly written items, multi-dimensional items) will increase the overall test reliability.[12] If the reliability increases when an item is deleted, that indicates that the item is problematic and reduces test reliability instead of increasing it.[12] In the present study, when 15 items with poor DI were excluded from the calculation, Cronbach's Alpha increased from 0.702 to 0.794 and thus overall test reliability improved significantly.

Items having poor DI & NFDs were discussed with faculty members in the department and required modifications were done to improve the questions. Most of the faculty members found item analysis useful (Fig. 6).

Conclusions

Majority of the items had acceptable level of difficulty & discrimination index. Most of distractors were functional. Item analysis provided valuable data for question improvement and helped in revising items with poor discrimination index and thus improved the quality of items & a test as a whole. Item analysis therefore should be incorporated into the process of test development and review.

Limitations

In the present study, item analysis of only one test was done. More convincingly; a study may be carried out involving teachers who framed MCQs of subject, interacting with them the data on one test, improving their items in the next tests, comparing the data of next test with previous, taking the feedback of teachers on their experience with item analysis in improving items and so on can be planned. Such study will use Item analysis as truly intervention and evaluate the impact/effect on expected key area from previous test to next test on same topics but on different batch of students.

Acknowledgement

I offer my unending thanks to faculty members of Department of Microbiology, PSMC, Karamsad & NHLMMC, Ahmedabad for their constant support & guidance for this project. I am grateful to the statistical department of HMPCMCE, Karamsad for their help in the analysis of the data.

References

1. Baig M, Ali SK, Ali S, Huda N. Evaluation of Multiple Choice and Short Essay Question items in Basic Medical Sciences. Pak J Med Sci. 2014;30(1):3–6.
2. Mehta G, Mokhasi V. Item Analysis of Multiple Choice Questions- An Assessment of the Assessment Tool. IJHSR. 2014;4(7):197-202.
3. Mitra N K, Nagaraja H S, Ponnudurai G, Judson J P. 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. Karelia BN, Pillai A, Vegada BN. The levels of difficulty and discrimination indices and relationship between them in four responses type multiple choice questions of

pharmacology summative tests of year II MBBS students. IeJSME 2013;6:41-46.
5. Scorepak®: Item analysis. Available from: www.washington.edu./oea/score1/htm . [Last Accessed on 13 September 2015].
6. Ananthakrishnan N. Item Analysis validation and banking of MCQs. In: Medical Education Principles and Practices. 2nd ed. Ananthkrishnan N, Sethuraman KR, kumar S. Alumini Association of National Teacher Training Centre, JIPMER, Pondichery, India. p,131-137.
7. 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 journal of Community Medicine 2014;39:17-20.
8. Ho TF, Yip WC, Tay JS. 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.
9. Chauhan P, Ratrhod S, Chauhan B, Chauhan G, Adhvaryu A, Chauhan A. Study of Difficulty Level and Discriminating Index of Stem Type Multiple Choice Questions of Anatomy in Rajkot, BIOMIRROR 2013; Volume 4:1-4.
10. Tarrant M, Ware J, Mohammed AM. An assessment of functioning and nonfunctioning distractors in multiple choice questions: a descriptive analysis. BMC Medical Education 2009;9:1-8.
11. Singh T, Anshu. Item analysis and Question Banking In: Principles of Assessment in Medical Education. 1st ed. New Delhi: Jaypee Brothers Medical Publishers (P) Ltd: 2012.p,116-127.
12. Verma S. 1999 Preliminary Item Statistics Using Point Biserial Correlation And P-Values. Education Data Systems Inc., Morgan Hills, USA.
13. Kolte V. Item Analysis of Multiple Choice Questions in Physiology Examinations. Indian J of Basic & Applied Medical Research 2015;4(4):320–326.
14. Matlock-Hetzel S. Basic Concepts in Item and Test Analysis. US: Texas A & M University; 1997.