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

Copyright: © the author(s), publisher and licensee Medip Academy. This is an open-access article distributed under the terms of the Creative Commons Attribution Non-Commercial License, which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

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. 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. Multiple-choice questions are used more and more in departmental examinations or as comprehensive examinations at the end of an academic session. They may be used to determine progress or to make decisions regarding the certification of a candidate. 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.\textsuperscript{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.\textsuperscript{6} This one characteristic of MCQs is sufficient to ensure that its edge in reliability more than compensate 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.\textsuperscript{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.\textsuperscript{8} It is argued that academics are generally not specialists in the research discipline of assessment, and they do not routinely analyze their assessments.\textsuperscript{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.\textsuperscript{10} Many times the teaching staff find it difficult to analyse the quality of the items those are repeatedly used for grading students’ performance.\textsuperscript{11-13} Item analysis also provides a way of reusing 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).\textsuperscript{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).\textsuperscript{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 $D = P_U - P_L$. 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 that was categorized based on 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 1<sup>st</sup> test, 2<sup>nd</sup> test, 3<sup>rd</sup> test, 4<sup>th</sup> test and 5<sup>th</sup> 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 1<sup>st</sup>, 2<sup>nd</sup>, 3<sup>rd</sup>, 4<sup>th</sup> and 5<sup>th</sup> 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 2<sup>nd</sup> test and 3<sup>rd</sup> test with mean p value 49.3 ± 12.0, 46.2 ± 10.9, while in 4<sup>th</sup> & 5<sup>th</sup> 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%, 80.4%, 85% and 81.3% in 1<sup>st</sup>, 2<sup>nd</sup>, 3<sup>rd</sup>, 4<sup>th</sup> and 5<sup>th</sup> test respectively while easy items having average DE of all tests was 20.1%.

**Table 1: Characteristics of MCQ test sessions.**

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

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

<table>
<thead>
<tr>
<th>Discrimination index</th>
<th>1&lt;sup&gt;st&lt;/sup&gt; Test</th>
<th>2&lt;sup&gt;nd&lt;/sup&gt; Test</th>
<th>3&lt;sup&gt;rd&lt;/sup&gt; Test</th>
<th>4&lt;sup&gt;th&lt;/sup&gt; Test</th>
<th>5&lt;sup&gt;th&lt;/sup&gt; Test</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of items</td>
<td>n=40</td>
<td>n=40</td>
<td>n=40</td>
<td>n=40</td>
<td>n=40</td>
<td></td>
</tr>
<tr>
<td>≥0.35</td>
<td>10 (25%)</td>
<td>09 (22.5%)</td>
<td>12 (30%)</td>
<td>13 (32.5%)</td>
<td>17 (42.5%)</td>
<td>61</td>
</tr>
<tr>
<td>0.25-0.34</td>
<td>00 (0%)</td>
<td>13 (32.5%)</td>
<td>00 (0%)</td>
<td>00 (0%)</td>
<td>00 (0%)</td>
<td>13</td>
</tr>
<tr>
<td>0.15-0.24</td>
<td>17 (42.5%)</td>
<td>00 (0%)</td>
<td>14 (35%)</td>
<td>06 (15%)</td>
<td>10 (25%)</td>
<td>47</td>
</tr>
<tr>
<td>&lt;0.15</td>
<td>13 (32.5%)</td>
<td>18 (45.0%)</td>
<td>14 (35%)</td>
<td>21 (52.5%)</td>
<td>13 (32.5%)</td>
<td>79</td>
</tr>
<tr>
<td>Total</td>
<td>40</td>
<td>40</td>
<td>40</td>
<td>40</td>
<td>40</td>
<td>200</td>
</tr>
</tbody>
</table>
### Table 3: Classification of questions according to difficulty indices for all 5 tests.

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

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

<table>
<thead>
<tr>
<th>Discrimination index (DI)</th>
<th>Interpretation</th>
<th>Mean P value</th>
<th>DE (%)</th>
<th>Mean P value</th>
<th>DE (%)</th>
<th>Mean P value</th>
<th>DE (%)</th>
<th>Mean P value</th>
<th>DE (%)</th>
<th>Mean P value</th>
<th>DE (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>≥0.35</td>
<td>Excellent</td>
<td>45.6±22.4</td>
<td>54.4±22.4</td>
<td>45.9±10.8</td>
<td>54.1±10.8</td>
<td>36.6±12.6</td>
<td>63.4±12.6</td>
<td>48.9±18.7</td>
<td>51.1±18.7</td>
<td>49.8±22.9</td>
<td>50.1±22.9</td>
</tr>
<tr>
<td>0.25-0.34</td>
<td>Good</td>
<td>---</td>
<td>---</td>
<td>53.9±26.6</td>
<td>46.2±26.6</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>0.15-0.24</td>
<td>Marginal</td>
<td>54.6±24.9</td>
<td>45.4±24.9</td>
<td>---</td>
<td>---</td>
<td>33.7±19.8</td>
<td>66.3±19.8</td>
<td>55.9±30.9</td>
<td>44.1±30.9</td>
<td>48.2±29.1</td>
<td>51.8±29.1</td>
</tr>
<tr>
<td>&lt;0.15</td>
<td>Poor</td>
<td>34.2±28.3</td>
<td>65.8±28.3</td>
<td>40.4±22</td>
<td>59.6±22</td>
<td>45.6±23.7</td>
<td>54.4±23.7</td>
<td>22.4±18.9</td>
<td>77.6±18.9</td>
<td>34.8±24.5</td>
<td>65.2±24.5</td>
</tr>
</tbody>
</table>

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.

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%. 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. 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 distractors.

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. Present study shows discrimination index correlate poorly with difficulty index. The relationship between difficulty index and discrimination index is dome shaped rather than linear. First, as difficulty index increases, the discrimination index 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.

**DISCUSSION**

The present item analysis uses single best response questions as it is seen as an efficient way of evaluation in academics. 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.

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%. 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. 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 distractors.

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. Present study shows discrimination index correlate poorly with difficulty index. The relationship between difficulty index and discrimination index is dome shaped rather than linear. First, as difficulty index increases, the discrimination index 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.

**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

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