RESEARCH ARTICLE

Assessing an assessment tool: Analysis of multiple choice questions on difficulty level and discrimination power, from an assessment in physiology

Afshan Kausar, Sayed Badar Daimi, Trupti Borulkar

Department of Physiology, JIIUS Indian Institute of Medical Science and Research, Warudi, Jalna, Maharashtra, India

Correspondence to: Trupti Borulkar, E-mail: dr.trupteeborulkar@gmail.com

Received: October 25, 2021; Accepted: November 29, 2021

ABSTRACT

Background: Multiple Choice Questions (MCQs) or “items” are frequently used in medical education to assess students’ knowledge at the end of an academic session. However, item analysis has to be done to confirm their excellence. This piece of work was aimed to analyze test items in the subject of Physiology. In view of the fact that item analysis has never been conducted in our physiology department, thus this study was planned to assess the quality of MCQs and construct a viable question bank. Aim and Objectives: To evaluate the quality of MCQs used so as to develop a pool of valid items to update the question bank. Materials and Methods: Total of one hundred 1st year MBBS students were given 40 MCQs. Analysis was done with the item difficulty and item discrimination indices. Results: Difficulty indices (p) of 14 items were easy while about two MCQ were difficult and the remaining 24 of the items were within acceptable range. The discrimination index (DI) (D) for 40 items; one item showed negative discrimination, ten were having poor discrimination, whereas 11 were having excellent discrimination. Nine items were having acceptable, and nine items were having good discrimination. The DI exhibited slight positive correlation with difficulty index (r = 0.1575 P = 0.05). However, it was not statistically significant. Conclusion: The findings show that most of the items were falling in acceptable range of difficulty and discrimination level. It concludes that Item analysis is a very useful and necessary procedure to obtain valid Items from time to time.

KEY WORDS: Item Analysis; Difficulty Index; Discrimination Index; Medical Education

INTRODUCTION

Multiplechoice questions (MCQs), evaluation has become a very important tool in today’s education system. MCQs are mainly used as an excellent comprehensive tool to provide feedback to the teachers at the end of an academic session. A good and reliable MCQ can assess higher cognitive functions such as interpretation, synthesis, and application of knowledge.[1] However, it is said that MCQs emphasize recall of factual information rather than conceptual understanding and interpretation of concepts.[2] The single best response type of MCQ is designed to assess knowledge.[3] Designing good MCQs is a complex and time-consuming process. After designing and construction of MCQs; they need to be tested for the standard or quality. They have the advantage of sampling broad domains of knowledge effectively and reliably providing some reliability for assessment. If carefully constructed, MCQs test higher-order thinking skills[4,5] Therefore, MCQs remain a useful assessment instrument; despite some limitations and objections.

“Item analysis” refers to quality assessment of MCQs. Item analysis is the process of collecting, summarizing, and using information from student’s responses to assess the quality of
test items\(^{(6)}\) and quality of test as a whole. The item analysis assists to establish the character of each item in reference to the whole test. Item analysis allows identification of good-quality MCQs based on their Difficulty Index (DIF I/P), Discrimination Index (DI).\(^{(7)}\) Difficulty Index helps to identify whether an MCQ is too easy or too difficult to attempt, DI helps differentiate in high ability and low ability students.\(^{(8)}\) The main purpose of item analysis is to improve tests by revising or eliminating ineffective items. An additional important aspect of item analysis; speaks about specifically to achievement tests.

Item analysis is especially valuable in improving items which will be used once more in future tests. The item statistics can help not only in finding out poor items needing improvement but also be helpful to remove ambiguous or misleading items. In addition, item analysis is valuable for increasing instructors’ skills in test construction and identifying specific areas of course content which need greater emphasis or clarity.\(^{(9)}\) It offers feedback to teachers to establish renovation in teaching standard. This could be beneficial for improving teaching skills in the future.

With this background, the present study had been planned with an objective, to analyze the quality of MCQs of Physiology examination by doing the item analysis by assessing difficulty index, DI, and to determine the relationship between the difficulty and discrimination indices of these MCQs. Our aim was also to create a viable question bank after revising/storing or discarding the items based on obtained results.

**MATERIALS AND METHODS**

The present cross-sectional study was conducted in the Department of Physiology, JIIU’S IIMSR Medical College Warudi Jalna Maharashtra. The study was approved by the Institutional Ethics Committee. At the end of one term, the first MBBS students were given an examination on Physiology. In total one hundred 1\(^{\text{st}}\) year MBBS students were given 40 MCQs. All MCQs were single best response type with four distracters. The MCQs were constructed by all faculties of the department. Experts had done Pre-validation of all the items. Half mark allotted each for correct answer. No negative marking was there for the wrong answer. All students’ scores were arranged in order of merit. The upper one-third students were considered as high achievers and lower third as low achievers. High achiever group \((n = 33)\)-Starting from the highest rank, 1/3\(^{\text{rd}}\) of the papers with high scores were selected. Low achiever group \((n = 33)\)-lower 1/3\(^{\text{rd}}\) of the papers with low scores were selected. Middle third \((n = 34)\) were set aside.

Each item was analyzed for:

1. Difficulty index or \(P\) value

“Item difficulty” perhaps supposed to be described as “item easiness”; it states the proportion of students answering the item correctly. The formula for the item-difficulty index (denoted as “\(P\)”) is

\[
p = \frac{H+L/N}{100}
\]

\(H\) = Number of students answering the item correctly in the high achievers group

\(L\) = Number of students answering the item correctly in the low achievers group

\(N\) = Total number of students in the two groups (including non-responders).

In general, items with a difficulty index <30\% are considered as difficult. If an item has a “\(P\)-value” between 30 and 70\% it is considered as acceptable. Items with a difficulty index greater than 70\% are considered as easy.

2. DI denoted as “D” measures the ability of an item to discriminate between students. To calculate this index the numbers of responders to the “key” was taken into account. Here the difference between the two groups was found out. For any item; the higher the difference between high achievers and low achievers, the greater will be its discrimination power.

DI or \(D\) value was calculated using the formula

\[
D = \frac{H-L}{2/N}
\]

\(H, L, N\) represent the same values as mentioned above.

The DI ranges from −1 to +1. An index value of +1 means the item has maximum discriminative power. An item having DI “0” cannot discriminate between two \((H\) and \(L)\) groups. An item having Negative DI ranging from −1 to 0 has poor discriminative power.

\begin{align*}
\text{DI (D):} & \\
D= & \text{ Negative } \rightarrow \text{Defective item/wrong key} \\
D= & 0–0.19 \rightarrow \text{Poor discrimination} \\
D= & 0.2–0.29 \rightarrow \text{Acceptable discrimination} \\
D= & 0.3–0.39 \rightarrow \text{Good discrimination} \\
D= & 0.4 \rightarrow \text{Excellent discrimination}. \\
\end{align*}

**Statistical Analysis**

The data were reported as % and mean ± standard deviation of all items. The relationship between the item difficulty index and DI values for all items was determined using Pearson correlation analysis. \(P < 0.05\) was considered to indicate statistical significance.

**RESULTS**

Mean difficulty index \((P)\) was 58.47 ± 13.7, and mean DI was 0.27 ± 0.15 [Table 1]. The difficulty index, out of 40 MCQs; about 24 Items (60\%) were in acceptable range \((P = 30–70\%)\) while 14 (35\%) were too easy \((P > 70\%)\), 2 (5\%) were too difficult \((P < 30\%)\) [Figure 1]. Hence 16 items that means
40% needs revision and could be added to MCQ bank after modification [Table 2].

While in terms of DI 9 MCQs (22.5%) were in acceptable range, (D = 0.2-0.29), 9 (22.5%) were in good range (D = 0.3–0.39), 8 (20%) were in excellent range (D ≥ 0.4) and 10 (25%) were in poor range (D < 0.19). One item showed negative discrimination [Figure 2]. Hence, a total of 29 MCQ (9+9+11) were in acceptable range in terms of DI [Table 3].

There were 24 and 29 MCQs in acceptable range in terms of difficulty and DI respectively [Tables 2 and 3]. Out of 24 Items, 18 were common in acceptable range of both the indices. While there were 11 MCQs with DI in acceptable range but too easy difficulty index. There were 6 items having poor DI but difficulty index in acceptable range.

Positive correlation was noticed \( r = 0.1575 \) \( P = 0.05 \) between difficulty index and DI in all 40 MCQs [Figure 3]. First, as \( p \) increases, the DI also increases, but at \( P \) value between 40% and 80%, DI reaches a maximum. Similarly, 18 common MCQs of acceptable range shows positive correlation between DI and P, \( r = 0.30308 \) [Figure 4]. However, a negative correlation was found between DI and P of 14 too easy MCQs \( r = −0.38541 \) [Figure 5].

### Table 1: Difficulty index and DI of MCQs

<table>
<thead>
<tr>
<th>Item analysis parameters</th>
<th>Mean±SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Difficulty index (P)</td>
<td>58.47±13.7</td>
</tr>
<tr>
<td>Discrimination index</td>
<td>0.27±0.15</td>
</tr>
</tbody>
</table>

SD: Standard deviation, DI: Discrimination index, MCQs: Multiple choice questions

### Table 2: Categories of items according to difficulty index

<table>
<thead>
<tr>
<th>Difficult index “P”</th>
<th>Total number of Items/40 (%)</th>
<th>Interpretation</th>
<th>Action taken</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;30%</td>
<td>2 (5)</td>
<td>Too difficult</td>
<td>Revise</td>
</tr>
<tr>
<td>30–70%</td>
<td>24 (60)</td>
<td>Acceptable</td>
<td>In MCQ bank</td>
</tr>
<tr>
<td>&gt;70%</td>
<td>14 (35)</td>
<td>Too easy</td>
<td>Revise</td>
</tr>
</tbody>
</table>

MCQ: Multiple choice question

### Table 3: Categories of items according to DI

<table>
<thead>
<tr>
<th>DI</th>
<th>Total number of items/40 (%)</th>
<th>Interpretation</th>
<th>Action taken</th>
</tr>
</thead>
<tbody>
<tr>
<td>Negative</td>
<td>1 (2.5)</td>
<td>Defective item/wrong key</td>
<td>Revise</td>
</tr>
<tr>
<td>0–0.19</td>
<td>10 (25)</td>
<td>Poor discrimination</td>
<td>Revise</td>
</tr>
<tr>
<td>0.2–0.29</td>
<td>9 (22.5)</td>
<td>Acceptable discrimination</td>
<td>In MCQ bank</td>
</tr>
<tr>
<td>0.3–0.39</td>
<td>9 (22.5)</td>
<td>Good discrimination</td>
<td>In MCQ bank</td>
</tr>
<tr>
<td>≥0.4</td>
<td>11 (27.5)</td>
<td>Excellent discrimination</td>
<td>In MCQ bank</td>
</tr>
</tbody>
</table>

DI: Discrimination index, MCQ: Multiple choice question

**DISCUSSION**

The present study was undertaken to analyze the MCQs with the help of item analysis and select the items which are good enough to assess the knowledge of subject physiology and to discriminate, the high achievers from the low achievers. If it is so, then a valid pool of MCQs can be created and incorporated into future question bank with reliability.
In our study, 24 of the items had acceptable difficulty indices ($P = 30–70\%$), 14 were easy ($P > 70\%$), whereas two of the items with $P < 30\%$ were difficult. Our findings correspond with the previous studies by Kolte done in physiology having a mean of Difficulty Index as $57.92 \pm 19.58$. The 26 items were in acceptable range, ten items were easy and four items were difficult.$^{[10]}$ In a similar type of study by Shete et al., mean difficulty index was $0.54 \pm 0.26$ and 18 of the items had acceptable, ten were easy while 12 of the items were difficult.$^{[11]}$

It was noticed that many of easy items were from “Must know” portion of syllabus. Those topics were taught and discussed in greater detail during theory lecture. Test paper must include easy questions also. These questions definitely help in boosting up confidence of students of all types. Similarly placing difficult questions in test paper, help to select toppers. On reviewing, it was also observed that difficult items were from “Desirable to know” or “Nice to know” portion of syllabus. Doubts related to these questions were cleared to students by helping them in proper understanding of the topic in depth.

The DI (D) serves as an effective feedback to teachers about the quality of each item. Items with poor discrimination should be reviewed. According to Crocker and Alagina$^{[12]}$ D > 0.2 is acceptable and able to discriminate between good and weak students.

In study done by Kolte,$^{[10]}$ the mean of DI was $0.33 \pm 0.15$. Total 24 (60\%) items had excellent, 7 (17.5\%) items had good, total of 8 (20\%) items had acceptable whereas 1 (2.5\%) item showed poor discrimination, 0\% of total items had negative discriminative power. The present study showed that 11 of the items had excellent discrimination, 18 of items showed good and acceptable discrimination, ten items were having poor discrimination. In all, only one of the items had negative discrimination. Similar results are found in the study by Ho et al.$^{[13]}$ and Shete et al.$^{[11]}$ showing that five of the items had excellent discrimination, 20 of items showed good and acceptable discrimination, 13 items were having poor discrimination, two of the items had negative discrimination.

Study by Gajjar et al.$^{[7]}$ have shown negative DI in 20\% and by Hingorjo et al. 4\%$^{[8]}$ items. Probable explanation for negative DI could be wrong key, due to poor understanding of difficult topics and ambiguous framing of questions or personal variation in forming the MCQs and may also be a clue to variation in student’s intelligence level. Items with negative DI decrease the validity of the test and should be removed from the question bank. Hence, these items should be revised, reframed, and reconsidered for future question bank.

Our study showed a slight positive correlation between difficulty and DI. Discrimination was noted to be maximum in the upper range of acceptable difficulty (40\% and 80\%); slightly lower in the very easy items and maximal lowering occurs at the level of the very difficult items, thus the relation is not linear. Too easy and too difficult items showed poor discrimination, whereas maximal discrimination found with moderately easy/difficult items. So too difficult and too easy items needed further modification. Similar observation was reported by Badkur et al.$^{[14]}$ Our findings corresponded with the Pande et al.$^{[15]}$ study findings which showed strong curvilinear relation. Similar results were depicted in other studies such as Shete et al.$^{[11]}$ ($r = 0.11, P > 0.10$) by Karelia et al.$^{[16]}$ and Badkur et al.$^{[14]}$.

Our study had also found that 18 common MCQs of acceptable range showed positive correlation between discrimination and difficulty index. Badkur et al.$^{[14]}$ had similar findings for MCQs in acceptable range of both indices. So we can say that our 18 MCQs are reliable enough to assess the knowledge of physiology.

The findings of this article have significance for teachers and test developer. They should be very careful while selecting items. The size of an acceptable item will depend upon the purposes for which the test has been designed, the length of the test, and the range of difficulty indices. The poor items
are improved or removed for inclusion in the final test. The principle function of an item used in any educational research is to infer student’s capacities. It also offers information on the bases of which the correct decisions can be made in proper framing and administering. Other subjects can repeat similar work to develop a good item bank for student community. Indeed item analysis is an important phase in the development of a test or instrument.

CONCLUSION

Nowadays test Items are used more frequently as an assessment tool while assessing student’s learning abilities and skills. For a valid and reliable assessment, quality Items are required. In the present study, the majority of items fulfilled the criteria of acceptable difficulty and good discrimination. Maximum DI was seen at p value range between 40% and 80%. This means the MCQs selected were of good quality.

Items with negative and poor discrimination will be reviewed, reconstructed, and added to the departmental MCQ bank. It concludes that Item analysis is a very useful and necessary procedure to obtain valid MCQs Bank for future use.

ACKNOWLEDGMENT

I would like to thank all the teachers in Physiology Department for contributing MCQs and their valuable support.

REFERENCES


How to cite this article: Kausar A, Daimi SB, Borulkar T. Assessing an assessment tool: Analysis of multiple choice questions on difficulty level and discrimination power, from an assessment in physiology. Natl J Physiol Pharm 2022;12 (Online First). DOI: 10.5455/njppp.2022.12.10387220
2129112021

Source of Support: Nil, Conflicts of Interest: None declared.