Did faculty training on item construction and peer review improve the quality of MCQs? An observational study in a medical school

Bahaeldin Hassan  
King Khalid University

Abdulaziz A. Alamri  
King Khalid University

Hassan A. Alzahrani  
King Khalid University

Mohammad Tauheed Ahmad  
King Khalid University

Ayed A. Shati  
King Khalid University

Ayyub A. Patel  
King Khalid University

Muhammad Khan  
King Khalid University

Omer A Elfaki (✉ oa.elfaki@amc.edu.sa)  
Al Rayan College of Medicine

Research Article

Keywords: MCQs, item analysis, peer review, training program, Medical education

Posted Date: March 27th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2680534/v1

License: ☑️ This work is licensed under a Creative Commons Attribution 4.0 International License. 
Read Full License
Abstract

Background: multiple choice questions (MCQs) test is the most popular assessment tool that used in medical schools, faculty training on construction of MCQs and peer review are the main factors that contributing to have a good quality MCQs test. This study aimed to evaluate the effect of these two factors on the quality of MCQs test.

Methods: an extensive training program was conducted for the staff members who participated in teaching and assessment of undergraduate medical students of clinical biochemistry and psychiatric courses during the academic year 2020-2021. Post-test item analysis was used to compare pre-training with post-training item analysis parameters. The Chi-square test was applied, the level of significance was 5%.

Results: The difficulty index showed great improvement in the number of items which lie in the acceptable range of difficulty from 16% and 25% before the training program, to 61% and 42% after the training program for psychiatry and clinical biochemistry courses respectively. This is significantly different (P<0.05). In psychiatry course the number of questions with negative discrimination power decreased from 36% before training program to 10% after it (P value >0.05). No significant improvement was noticed in biochemistry course regarding this aspect.

The percentage of test items where all distracters were functioning was increased from 11% to 41% in psychiatry course (P value=0.0002), and from 12% to 46%, in biochemistry course (P value=0.0001). This is a significant improvement. Reliability after the training program and peer review processes also increased in both courses. However, this is not statistically significant.

Conclusions: A dramatic improvement was observed in the quality of MCQs after peer review and training program. We recommend adopting training program on construction of MCQs to be mandatory for all medical school staff members in Saudi Arabia.

Introduction

Multiple choice questions (MCQs) are a very common and well-accepted method of evaluating the professional competencies of medical science education (1). Construction of high-quality MCQs items is a very challenging task for the faculty members who have never undergone precise and dedicated training (2).

Psychometric parameters like discrimination index (DI) and difficulty index (P-value), and validity and reliability of the examination, will be affected by ignorance of item construction guidelines (3, 4).

Evaluating the quality of assessment within a course requires review at multiple levels, starting from institutional level and passing through policies, guidelines, and peer review as they affect the educational outcomes (5).
For each MCQs test, reliability refers to the generalization and reproducibility of the test, many research reports refer to the Kuder-Richardson coefficient (KR-20). Reliability coefficient varies from zero to 1, with more reliability when it is closer to 1. Classroom tests are classified to have acceptable reliability when they score 0.7 or more (6).

The difficulty index (p-value) is the percentage of students who answered the item correctly. Its value lie in the range 0–100%. The higher the percentage, the easier the item. The 30–70% p-value is considered acceptable. Items with p-values below and above, are considered difficult and easy items respectively (7).

The discrimination index (DI), also known as the point biserial correlation, identifies discrimination between students with different levels of achievement. An ideal item should have a DI of > 0.2(8).

The function of distractors reflects the ability of the teacher to design a good MCQ item with suitable options that can distract the less knowledgeable examinee from the key answer. Any distractor that has been selected by less than 5% of the students is considered to be a non-functioning distractor (NF-D) (9).

In our department of medical education, college of medicine, King Khalid University (KKU), Abha, Saudi Arabia, the assessment unit was established since the academic year 2013. Faculty development workshops to train the faculty members to write high-quality MCQs items were conducted at a rate of two workshops every semester. Item writing guidelines were established, and a pretest checklist was adopted for every test in the college.

We observed the poor quality of MCQs tests that come from psychiatry and clinical biochemistry courses during the academic year 2020–2021. This alarmed us to conduct a training program for the staff members who are participate in these courses. The program included an extensive workshop on how to construct MCQs items and peer review enhancement.

**Methodology**

**Study design**

An interventional, quasi experimental study design was adopted.

**Data Collection**

The participants were the academic staff members who participated in teaching and assessment of psychiatry and clinical biochemistry courses during the academic year 2020–2021. The participants attended an extensive workshop on how to construct high quality MCQs items. The workshop was delivered by experts in medical education. Peer review processes including blueprint were explained to the participants and models of practices were applied. Two MCQs tests were conducted prior to the intervention of faculty training and peer review practice were compared with another two tests which were conducted after training. All items were single best answer questions, each item contains four options...
with no penalties for incorrect answers. Item analysis parameters were used to assess the quality of the MCQs tests. Permission was taken from the Medical Education and other concerned departments to access these data for research purposes.

**Data analysis**

Acceptable reliability was set at 0.7 or more. The recommended range of difficulty was considered at a range of 30–70%. Items having p-values below 30% and above 70% were considered difficult and easy items respectively and desirable DI was > 0.2. Any distractor that has been selected by less than 5% of the students is a non-functioning distractor (NF-D). Frequencies and percentages were calculated, chi-square test was used to measure the p-values between variables. P < 0.05 was considered as significant value.

**Results & Discussion**

Regarding difficulty index, the psychiatry course test showed great improvement in the number of questions which lie in the acceptable range of difficulty. The acceptable percentage of difficult questions increased from 16% before the training program to 61% after the training program. This difference was statistically significant (P value = 0.00001).

The number of easy questions decreased in the post-training test from 82–31%. This difference positively affected the quality of the MCQs test. The difference was statistically significant (P value = 0.00001) (Table 1, Fig. 1).
<table>
<thead>
<tr>
<th></th>
<th>Before</th>
<th>After</th>
<th>p-values</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Difficulty Index</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Difficult Question (&gt;70)</td>
<td>2%</td>
<td>8%</td>
<td>0.133</td>
</tr>
<tr>
<td>Easy Questions (&lt;30)</td>
<td>82%</td>
<td>31%</td>
<td>0.00001</td>
</tr>
<tr>
<td>Acceptable Questions (30–70)</td>
<td>16%</td>
<td>61%</td>
<td>0.00001</td>
</tr>
<tr>
<td><strong>Discrimination Index</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Negative values</td>
<td>36%</td>
<td>10%</td>
<td>0.0008</td>
</tr>
<tr>
<td>Zero values</td>
<td>6%</td>
<td>28%</td>
<td>0.0014</td>
</tr>
<tr>
<td>Equal or above 0.2</td>
<td>48%</td>
<td>55%</td>
<td>0.4449</td>
</tr>
<tr>
<td><strong>Distractor Index</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 NFD</td>
<td>36%</td>
<td>18%</td>
<td>0.0270</td>
</tr>
<tr>
<td>2 NFD</td>
<td>36%</td>
<td>3%</td>
<td>0.00001</td>
</tr>
<tr>
<td>1 NFD</td>
<td>17%</td>
<td>38%</td>
<td>0.0103</td>
</tr>
<tr>
<td>0 NFD</td>
<td>11%</td>
<td>41%</td>
<td>0.0002</td>
</tr>
<tr>
<td><strong>KR-20</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.73</td>
<td>0.79</td>
<td>0.44</td>
</tr>
</tbody>
</table>

In clinical biochemistry course, only the number of questions which had an accepted range of difficulty (30–70), increased significantly from 25–42% (P value = 0.0200) (Table 2, Fig. 2).
<table>
<thead>
<tr>
<th></th>
<th>Before</th>
<th>After</th>
<th>p-values</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Difficulty Index</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Difficult Question (&gt; 70)</td>
<td>5%</td>
<td>10%</td>
<td>0.22</td>
</tr>
<tr>
<td>Easy Questions (&lt; 30)</td>
<td>54%</td>
<td>48%</td>
<td>0.43</td>
</tr>
<tr>
<td>Acceptable Questions (30–70)</td>
<td>25%</td>
<td>42%</td>
<td>0.0200</td>
</tr>
<tr>
<td><strong>Discrimination Index</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Negative values</td>
<td>5%</td>
<td>4%</td>
<td>0.75</td>
</tr>
<tr>
<td>Zero values</td>
<td>56%</td>
<td>47%</td>
<td>0.24</td>
</tr>
<tr>
<td>Equal or above 0.2</td>
<td>39%</td>
<td>48%</td>
<td>0.23</td>
</tr>
<tr>
<td><strong>Distractor Index</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 NFD</td>
<td>32%</td>
<td>14%</td>
<td>0.005</td>
</tr>
<tr>
<td>2 NFD</td>
<td>30%</td>
<td>28%</td>
<td>0.77</td>
</tr>
<tr>
<td>1 NFD</td>
<td>26%</td>
<td>12%</td>
<td>0.02</td>
</tr>
<tr>
<td>0 NFD</td>
<td>12%</td>
<td>46%</td>
<td>0.00001</td>
</tr>
<tr>
<td><strong>KR-20</strong></td>
<td>0.71</td>
<td>0.82</td>
<td>0.09</td>
</tr>
</tbody>
</table>

Point biserial reflects the ability of the test item to discriminate between high and low scoring students and is desirable to be more than 0.2. In the psychiatry course, the number of questions with negative discrimination power decreased from 36% before the training program to 10% after training (P value > 0.05) (Table 1). The quality of the psychiatry test was improved as it includes less negatively discriminating questions. No significant improvement in the discrimination index was noticed in the biochemistry course.

Both courses showed improvement in the function of distracters. The percentage of test items where all distracters were functioning was increased from 11–41% in psychiatry course (P value = 0.0002) Table 1, Fig. 1) and from 12–46% in biochemistry course (P value = 0.0001) (Table 2, Fig. 2).

The two courses included in the study, showed increased reliability after the training program and peer review process but this improvement is not statically significant (Table 1&2).
Overall, the results showed improvement in the quality of MCQs test after faculty training on test item construction and peer review. The improvement was evident by the increased number of difficult questions within the acceptable range of difficulty, discrimination power and function of distractors. Similar findings were observed by a published study in our region (2). Moreover, the importance of faculty training was reported by many authors (10, 11, 12).

The length of the training program was reported to be insignificant in a previous study (13). However, other studies reported non-effectiveness of single short training session on correction of flaws in writing MCQs (14).

Our study adds to studies which approved the importance of peer review in the improvement of the quality of MCQs (15, 16, 17).

Conclusions

The quality of MCQs tests assessed by psychometric properties showed marked improvement after peer review and faculty training programs. We recommend adoption of the training programs on the construction of MCQs and peer review to be a mandatory license for all medical school staff members in our school and similar schools in Saudi Arabia.

Abbreviations

MCQs: multiple choice questions
DI: discrimination index
P-value: difficulty index
KR-20: Kuder-Richardson coefficient
NF-D: non-functioning distractor
KKU: King Khalid University

Declarations

Ethics approval and consent to participate

All methods were carried out in accordance with relevant guidelines and regulations. This study was approved by the Ethics and Internal Review Board Committee-King Khalid University (WCM# 202/9). Informed consent was obtained from all subjects.

Consent for publication
Availability of data and materials

The data that support the findings of this study are available from the corresponding author upon reasonable request and with permission of the medical education department in KKU.

Competing interests

The authors declare that they have no competing interests.

Funding

No funding of this research from any source.

Authors' contributions

B. H and O. E. work on conception and design. A. A. and H. Z. collect data and prepare manuscript. M. T., A. P., A. S. and M. K. analyzed and interpreted the data. All authors contribute in writing and reviewing of manuscript.

Acknowledgements

Not applicable.

References


**Figures**
Figure 1
Satisfactory ranges of item analysis parameters pre-Vs-post-training for psychiatric course

Figure 2
Satisfactory ranges of item analysis parameters pre-Vs-post-training for Biochemistry course