An evaluation of a final year multiple choice questions examination at Faculty of medicine-university of Benghazi

Saleh M. Alawgali
Medical department-university of Benghazi

Abstract.

Background: Multiple choice questions (MCQs) are widely used in medical education, they have the feature of sampling broad domains of knowledge efficiently and reliably. They represent one of the most important examination tools used in undergraduate and postgraduate medical examinations. Exam analysis is an assessment tool that provides important information about validity and reliability of test. Item analysis is used to analyze the difficulty and the discrimination ability of individual items and the total test scores. The purpose of this study is to examine the quality of one of the final exam papers, that is composed of single best of five MCQs, and to examine the correlation between the difficulty and discrimination index of each item. Subject and method: The authors collected the exam results of 425 students who set for the final exam at the end of the academic year 2021-2022. The exam consisted of 70 MCQs and was aiming to assess the students` knowledge, understanding, application and analysis of the subject matter. The test was analyzed regarding content validity and reliability of the test as a whole as well as the difficulty coefficient and discrimination coefficient for each item. Results: Out of the 425 students who set for the exam about 86% has passed (scored ≥ 60% of the total marks), the test reliability was excellent (0.91) and the mean difficulty index (P) was 75.22 ± 19.4 (easy exam). Out of the 70 items six (8.57%) had a poor discrimination (D < 0.1). When comparing the exam questions with the blueprint 90% of the topics of the subject materials were either over or under represented. About 77% of the exam questions were testing either students` knowledge or understanding. The correlation between the difficulty index and discrimination index was moderately positive which was statistically significant (0.633, P = 0.000). Conclusion: although the current study showed a good exam reliability, however the exam content validity was suffering from many weakness regarding the total number of the items, and distribution of the items according to the topics and the cognitive levels, therefore the number of exam questions should be increased and a blueprint should be developed for assessment. The overall exam difficulty was low as there were too many easy questions and this could impair both exam validity and reliability. The number of average and difficult questions should be considerably increased.
تقييم امتحان أسئلة الاختيار من متعدد للسنة النهائية بكلية الطب - جامعة بنغازي

صالح محمد الأوجلي

قسم الباطنة - كلية الطب - جامعة بنغازي - ليبيا

المتى:

خلفية:

agnosis) (MCQs) على نطاق واسع في التعليم الطبي. ولديها ميزة أخذ عينات واسعة من مجالات المعرفة بكفاءة وموثوقية. وهي تمثل واحدة من أهم أدوات الفحص المستخدمة في الفحوصات الطبية لمرحلتي البكالوريوس والدراسات العليا. تحليل الاختيار هو أداة تقييم توفر معلومات مهمة حول صلاحية وموثوقية الاختيار ويستخدم تحليل كل سؤال على حدة (MCQs) لتحليل صعوبة كل سؤال والقدرة على التمييز للعناصر الفردية ومجموع درجات الاختيار. تهدف هذه الدراسة إلى فحص جودة إحدى أوراق الامتحان النهائي، المكونة من أفضل خمس اجابة (MCQs) وفحص العلاقة بين مؤشر الصعوبة والتميز لكل عنصر (سؤال). الموضوع والطريقة:

جمع الباحث نتائج امتحانات طبية 425 طالباً وطالبة تقدموا لامتحان النهائي في نهاية العام الدراسي 2022–2023. وقد تم استخدام الاختبارات (MCQs) وكان التحليل يهدف إلى تقييم معرفة الطلاب وفهمهم وتطبيقهم وتحليلهم للأسئلة. تم تحليل الاختيار من حيث صلاحية المحتوى وموثوقية الاختيار ككل وكذلك معامل الصعوبة ومعامل التمييز لكل سؤال. النتائج:

من بين 425 طالباً وطالبة تقدموا لامتحان نجح حوالي 86% (60٪ من إجمالي العلامات) وكانت موثوقية الاختبار متزامنة (0.91) ومقياس صعوبة كان 75.22 ± 19.4 (اختبار سيل). من بين 70 سؤال، كان ستة (8.5٪) يعانون من تميز ضعيف (P < 0.1). عند مقارنة أسئلة الامتحان مع المقررات الدراسية، كانت 90٪ من الاستنتاج أو تمثيلاً ناقص. وكان حوالي 77٪ من أسئلة الاختيار تمتزج مع معرفة الطلاب وفهمهم للسؤال. وكان الارتباط بين مؤشر الصعوبة ومؤشر التميز إيجابياً إلى حد ما وكان هذا دلالة إحصائية (0.633, P = 0.000) الخلاصة: على الرغم من أن الدراسة الحالية أظهرت موثوقية جيدة لامتحان، إلا أن صلاحية محتوى الاختيار كانت تعاني من العديد من نقاط الضعف فيما يتعلق بالعد الإجمالي للأسئلة، وتوزيع الأسئلة وفقاً للمقررات الدراسية والمستويات المعرفية، لذلك يجب زيادة عدد أسئلة الاختيار ووضع مخطط لتقييم. كانت صعوبة الاختيار الإجمالية منخفضة حيث كان هناك الكثير من الأسئلة السهلة وهذا يمكن أن يضعف صلاحية الاختيار وموثوقيتها. يجب زيادة عدد الأسئلة المتوسطة والصعبة بشكل كبير.
Introduction:

Multiple choice questions (MCQs) are widely used in medical education, they have the feature of sampling broad domains of knowledge efficiently and reliably. They represent one of the most important examination tools used in undergraduate and postgraduate medical examinations (Al-Rukban, 2006). Appropriately constructed MCQ-based methods are efficient, objective, capable of discrimination, and can be combined with other assessment strategies to contribute to a comprehensive student assessment strategy (Kar et al., 2015).

Like other methods of assessment, MCQs have their advantages and limitations. They have the ability to evaluate a large number of students over a broad domain of knowledge in a short period of time, and this is considered as a distinct advantage of using MCQ. (Kumar et al., 2021) Concerns have been raised that most MCQs tend to test recall and understanding of facts. However, MCQs are useful in measuring higher order thinking skills beside factual recall such as application and analysis, if they were carefully constructed (Ho et al., 1981; Sim & Rasiah, 2006). Scoring of MCQs is quick process by using various soft wares, it is also objective, and reliable. Guessing practice, cueing effect by the students (i.e., when students answer a question by identifying the right option, but couldn’t answer it in the absence of options) and no provision of feedback are considered disadvantages. (Murphy, 2007; Schuwirth et al., 1996)

Exam designers recommend that level of exam difficulty should be formulated in ascending fashion, i.e., putting easy questions at the beginning of the exam to encourage students, and hard questions that determine excellent students should be posted at the end of the exam.

Exam analysis is an assessment tool that provides important information about validity and reliability of test. Item analysis is used to analyze the easiness or difficulty of individual items and the total test scores. It uses quantitative approaches to make judgment about which questions should be adopted, revised, or discarded. Item analysis also used to provide feedback to instructors to identify course contents that require greater emphasis, and make changes and developments in the standards of teaching. Frequent usage of MCQ items during classrooms helps teachers to improve students’ performance. The purpose of teaching is not to distinguish between good and bad learners, but to make sure that all learners have learnt the lesson. So, teachers must aim at getting high difficulty indices and low discrimination values. (Shete et al., 2015). The MCQ item analysis encompass difficulty index (the percentage of students who answered the item correctly), discrimination index (discriminate between high achievers and low achievers), internal consistency reliability (Used to assess the consistency of results across items within a test). Each question is evaluated for these indices, because if a question is inaccurate, then it becomes a distractor and the assessment might fail. (Kumar et al., 2021).

Up to our knowledge there is no published research work about analysis of medical exams from Libya.
The purpose of this study is to examine the quality of one of the final exam papers, that is composed of single best of five MCQs, and to examine the correlation between the difficulty and discrimination index of each item

**Material and Methods.**

Data collection

This study was conducted by the medical education department at university of Benghazi for the final pediatric exam of the undergraduate medical program.

The authors collected the exam results of 425 students who set for the exam at the end of the academic year 2021-2022. The exam consisted of 70 MCQs and was aiming to assess the students’ knowledge, understanding, application and analysis of the subject matter.

**Construction and Selection of MCQs.**

The MCQ items were constructed by individual teachers at their departments for content accuracy. Then the items were selected by members of the exam committee based on their examination experience and academic judgments.

The questions were designed in ‘best of five’ format (a, b, c, d, e), where students choose the best answer from five possible alternatives. There was no predetermined exam blueprint. Therefore, in order to assess the balance of the questions between the different topics the authors revised the course specification document and created a course blueprint based on the number of lectures contact hours of each topic.

**MCQs scoring.**

All the answer sheets were marked by an Optical Mark Reader (OMR) By Remark® From Gravic, Inc.. The OMR output was processed by computer and scores were allocated according to the student’s responses as following:

- **a)** One mark is awarded (+1) for each correct answer,
- **b)** No mark is awarded or subtracted for an incorrect answer (zero).
- **c)** No mark is awarded or subtracted if a question was left unanswered (zero).
- **d)** No mark is awarded or subtracted if more than one answer was recorded or the answer is not sufficiently clear (zero).
- **e)** No negative marks allotted
Item analysis.

The software automatically generates exam statistics including test reliability, median, mean, standard deviation, difficulty coefficient and discrimination coefficient for each item.

1. Difficulty coefficient (P) was calculated according to the following formula: 
   \[ P = \frac{T}{N} \]
   
   Where:
   - T: Number of students who answered the question correctly.
   - N: The total number of students who answered the question.

   Difficulty index (P):
   - a.  \( P = < 50\% \) → Difficult
   - b.  \( P = 50-75\% \) → Average
   - c.  \( P = >75\% \) → Easy

2. Discrimination index was calculated by the following formula:
   \[ D = \frac{(X-Y)}{0.25N} \]
   
   Where:
   - X = Number of students who answered the question correctly from the upper class.
   - Y = Number of students who answered the question correctly from the lower class.
   - N = Sample size.

   Discrimination index (D):
   - a.  Good discrimination: >0.3
   - b.  Fair discrimination: 0.1-0.3
   - c.  Poor discrimination: < 0.1 or Negative (Defective item).

   Hence, the higher the difficulty index value; the lower is the difficulty of an item and the lower the difficulty index value; the greater is the difficulty of an item. For discrimination; the higher the discrimination index, better the discriminative ability of an item i.e., it can discriminate among students with high test scores and those with low ones.

   Exam reliability was assessed by calculating the coefficient of internal consistency.
Statistical analysis:
The relationship between the item difficulty index and discrimination index values for all items was determined using Pearson correlation analysis. P value of < 0.05 was considered to indicate statistical significance. The data were reported mean ± standard deviation of all questions.

Results.
Out of the 425 students who set for the exam about 86% has passed (scored ≥ 60% of the total marks), (table-1), the test reliability according to the coefficient of internal consistency (Cronbach alpha) was 0.91 and the mean difficulty index (P) was 75.22 ± 19.4 while the mean discrimination index (D) was 0.38 ± 0.18, (table-2). Out of a total 70 items, 48 of the items (68.5%) were easy (P ≤ 0.75). Furthermore, there was a wide range of difficulty level among MCQs, as they range from 0.10 (very difficult items) to 0.95 (very easy items) (figure-1). Out of the 70 items six (8.57%) had a poor discrimination (D < 0.1) five of them showed a negative discrimination index (figure-2). When comparing the exam questions with the blueprint 9 out of the 10 (90%) topics of the subject materials were either over or under represented in the exam (table-3). About 77% of the exam questions were testing either students’ knowledge or understanding (table-4). The correlation between the difficulty index correlated and the discrimination index was moderately positive and which is statistically significant (0.633, P = 0.000) (figure-3).

Table 1 shows exam statistics

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of students</td>
<td>425</td>
</tr>
<tr>
<td>Number of exam items</td>
<td>70 representing 82 hours of lectures.</td>
</tr>
<tr>
<td>Pass score</td>
<td>42 (60%)</td>
</tr>
<tr>
<td>Success rate</td>
<td>85.65% (364/425)</td>
</tr>
<tr>
<td>Excellent</td>
<td>(36%) 131/364</td>
</tr>
<tr>
<td>Very good</td>
<td>(37%) 135/364</td>
</tr>
<tr>
<td>Good</td>
<td>(19%) 69/364</td>
</tr>
<tr>
<td>Pass</td>
<td>(8%) 29/364</td>
</tr>
<tr>
<td>Median score</td>
<td>56 (80%)</td>
</tr>
<tr>
<td>Mean score</td>
<td>52.65±10.10.23 (75.21%)</td>
</tr>
</tbody>
</table>
Table 2: Difficulty index and discrimination index

<table>
<thead>
<tr>
<th></th>
<th>Difficult Index</th>
<th>Discrimination Index</th>
</tr>
</thead>
<tbody>
<tr>
<td>N Valid</td>
<td>70</td>
<td>70</td>
</tr>
<tr>
<td>Missing</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Mean</td>
<td>75.2173</td>
<td>.3853</td>
</tr>
<tr>
<td>Std. Deviation</td>
<td>19.47715</td>
<td>.18100</td>
</tr>
<tr>
<td>Range</td>
<td>84.24</td>
<td>.79</td>
</tr>
<tr>
<td>Minimum</td>
<td>0.10</td>
<td>-.14</td>
</tr>
<tr>
<td>Maximum</td>
<td>0.95</td>
<td>.65</td>
</tr>
<tr>
<td>Test reliability</td>
<td></td>
<td>0.91</td>
</tr>
</tbody>
</table>

Figure 1 item difficulty

Figure 1: proportion of difficulty index

Figure 2: proportion of discrimination index
Table-3: Post exam blueprint

<table>
<thead>
<tr>
<th>Subject</th>
<th>Lectures hours</th>
<th>Number of exam questions</th>
<th>Assessment</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Neonate</td>
<td>8 (9.75%)</td>
<td>9 (12.8%)</td>
<td>Over</td>
</tr>
<tr>
<td>2 Respiratory</td>
<td>6 (7.31%)</td>
<td>8 (11.42%)</td>
<td>Over</td>
</tr>
<tr>
<td>3 Cardiovascular</td>
<td>7 (8.53%)</td>
<td>4 (5.71%)</td>
<td>Under</td>
</tr>
<tr>
<td>4 GIT</td>
<td>7 (8.53%)</td>
<td>5 (7.14%)</td>
<td>Under</td>
</tr>
<tr>
<td>5 CNS</td>
<td>10 (12.19%)</td>
<td>11 (15.71%)</td>
<td>Over</td>
</tr>
<tr>
<td>6 Renal</td>
<td>6 (7.31%)</td>
<td>6 (8.75%)</td>
<td>Over</td>
</tr>
<tr>
<td>7 Infectious</td>
<td>6 (7.31%)</td>
<td>6 (8.75%)</td>
<td>Over</td>
</tr>
<tr>
<td>8 Hematology</td>
<td>7 (8.53%)</td>
<td>6 (8.75%)</td>
<td>Balanced</td>
</tr>
<tr>
<td>9 Endocrine</td>
<td>8 (9.75%)</td>
<td>6 (8.75%)</td>
<td>Under</td>
</tr>
<tr>
<td>10 Others</td>
<td>17 (20.73%)</td>
<td>9 (12.85%)</td>
<td>Under</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>82</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table-4: cognitive levels covered by the exam questions

<table>
<thead>
<tr>
<th>Knowledge</th>
<th>Understanding</th>
<th>Application</th>
<th>Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>48 (68.57%)</td>
<td>6 (8.57%)</td>
<td>8 (11.42%)</td>
<td>8 (11.42%)</td>
</tr>
</tbody>
</table>

Figure 3: Correlation of difficulty and discrimination index

The scattered diagram Figure 3 represents the relationship between the difficulty index and the discrimination index of 70 MCQs. The difficulty index correlated positively $r = 0.633$ with the discrimination index ($P = 0.000$) which is statistically significant.
Discussion

Item analysis is a simple procedure that analyze and interpret students’ response to single test items, the quality of the item, and a test as a whole. Item analysis should be carried out routinely to measure the quality of questions and to reconstruct proper MCQs for subsequent exams (Tavakol & Dennick, 2011).

Content validity:

Eighty two hours of lectures are potentially capable of producing at least 410 questions (5 questions per 1 hour lecture). Seventy MCQs are too few to cover all the subject matter which impairs the exam content validity. Nine out of the ten topics in the course were either over or under represented which again impairs the content validity. The absence of course blueprint for exam, impairs the examiner ability in judging the suitable number of questions for each subject. About 77% of the questions were assessing lower order thinking skills (memorization and understanding) and mostly (about 69%) were assessing the students’ ability for memorization. The effective measurement of knowledge and understanding is an essential component of medical education, however, the percentage of questions that assess these cognitive levels should not constitute this high percentage of the exam. A well-constructed MCQs can assess higher order thinking skills such as application and analysis. This method of assessment should be regularly evaluated to know how effective they are in assessing various levels of cognitive domains. (Shete et al., 2015)

Reliability

Reliability refers to the extent to which repeated measurements of a relatively stable phenomenon fall closely to each other. The internal consistency reliability of the test was 0.91 which is excellent. Reliability coefficients theoretically range in value from zero (no reliability) to 1.00 (perfect reliability). A reliability coefficient of 0.7 or higher is considered “acceptable”. High reliability means that the students who answered a given question correctly were more likely to answer other questions correctly.

Difficulty index:

The item difficulty index ranges from 0 to 1; the higher the value, the easier the question. It is recommended that the level of difficulty of each questions should range between 0.50-0.75. In the current study, there were too many easy questions (about 69%), leading to a mean exam difficulty (Mean Percent Score) of 75.22 which is considered an easy exam. These too many easy questions in the exam were on the expense of the average questions and difficult questions and this could impairs both the exam validity and reliability. The exam designers recommend putting some easy questions at the beginning of the exam to encourage students, but some hard questions that determine strong students are posted at the end of the exam. Very easy and very difficult questions need to be reevaluated and reconstructed. The process of item analysis should be used as atool to generate valid MCQ bank. The use of item analysis at the end of the academic period, has great advantages for the instructors. It makes
teachers got constructive feedback from the students and identify areas which require emphasis, and alteration in teaching methods. Usually, items which have moderate difficulty and good discrimination are chosen. (Sim & Rasiah, 2006)

**Discrimination index:**

The discrimination index (D) is used as effective feedback to instructors about the quality of individual item (Pande et al., 2013). Although the overall discrimination ability of the exam was fair-to-good. However 8.57% of the items (6 questions) had a poor discrimination index (D < 0.1), which is a high percentage that affects the exam quality. Values for point-biserial range from -1.00 to 1.00. (Varma, 2006). Questions with negative discrimination index means that more students from the lowest 25% group have answered the question than students in the highest 25% groups, which might indicates either inappropriate key or cheating has occurred, so that they have to be reviewed and examined for possible incorrect key and to be rewritten before being reused again. This could have been due to ambiguity inwordings of the questions, poor understanding of difficult topics, or inappropriate key. Items with poor discrimination ability should be reviewed for reconstruction or deletion.

**Correlation between difficulty and discrimination**

In this study, there was wide scatter of discrimination and difficulty indices which indicate that students might have practiced guessing because no negative marks allotted to wrong answers and this have caused this variation. Same issue was reported from Malaysia (Sim & Rasiah, 2006) and Mitra et al. (N K, 2009). The correlation between difficulty and discrimination indices showed that discrimination index correlates moderately with difficulty index (r= 0.633). The correlation is significant at 0.000 level (2-tailed). Positive correlation signifies that with increasing difficulty index, there is increase in discrimination index. In other words, as the items get easy (above 0.75), the level of discrimination index increases consistently (Figure 3). This finding is against the findings of Mitra et al. (N K, 2009) who showed that discrimination index correlate poorly with difficulty index. This Negative correlation signifies that with increasing difficulty index values, there is decrease in discrimination index.

**Conclusion.**

The current study showed a good exam reliability, however the exam content validity was suffering from many weakness regarding the total number of the items, and distribution of the items according to the topics and the cognitive levels, therefore the number of exam questions should be increased and a blueprint should be developed for assessment. The overall exam difficulty was low as there were too many easy questions and this could impair both exam validity and reliability. The number of average and difficult questions should be considerably increased.
References:


