EDITORIAL

What does Item Analysis Tell Us? Factors affecting the reliability of Multiple Choice Questions (MCQs)

(Item Analysis)

Mohamed Elhassan Abdalla, MB.BS, MHPE

Corresponding Author
Mohamed Elhassan Abdalla
Medical Education Unit
Faculty of Medicine
Jazan University
Saudi Arabia
e.mail: hason75@yahoo.com

Abstract

Introduction: Item analysis is a very important part of the examination cycle, as it can be used to judge the difficulty and discrimination of test items as well as their reliability.
Objective: The objective of this study is to determine the relation between reliability and other components of the item analysis of Multiple Choice Question (MCQs).
Methodology: The item analysis of 38 modules conducted at the Faculty of Medicine-Jazan University in the academic year 2010-2011 were further analyzed descriptively and correlation were made between the reliability co-efficient (KR-20) and other elements of the item analysis such as the difficulty and discrimination indices
Results: All the items have five options, the percentage of nonfunctioning options is between 27.6% and 70% across items. The facility index is < 35% is found in 18.3% of the items and facility index of > 90% is a maximum of 75%. All exams have items with a discrimination index <19% .

There is a positive correlation among the KR-20, number of items and the percentage of items with discrimination index >60% and negative correlation with percentage of the nonfunctioning options, percentage of items with facility index <35%, percentage of items with facility index >90% and percentage of items with discrimination index <19%.

Further analysis of the modules in each phase (Introductory, organ system and clinical) and qualitative analysis of the items are needed.

Introduction

The Faculty of Medicine-Jazan University in Saudi Arabia is adopting an Integrated-Organ system based curriculum. Established in 2004, three batches have now graduated. The programme in the faculty has 64 modules in six academic years. The school, through its Student Assessment Committee (SAC), is adopting a quality assurance process for its examination. This process is started by issuing guidelines for item construction and through construction of the exam by an independent committee and revision of the exam

by a spate committee before offering it to students. The final step in this process is item analysis and feedback.

Item analysis is a process that uses quantitative methods to judge the quality of items in an objective test - most commonly used with the multiple-choice questions. It is an important part of the of the examination cycle for several reasons.[1] The process can estimate the ease of difficulty of questions.[2, 3] Also, item analysis can relate the individual test score with the global test score. [1]

Reliability, which is concerned with the reproducibility and stability of the test, is one of the most important measures. It also assesses how individual items are functioning to measure the construct intended to measure by the test.[1] There are many ways to measure this reliability. In the item analysis process, the reliability the Kuder-Richardson (20) co-efficient is the most often used. [1, 2, 3, 4]

Reliability of test scores depends on internal and external factors. The internal factors refer to the item sampling, item construction, and the scoring system. The external factors are related to any environmental, personal or administrative issues that can affect the test.[1]

Much research work has been done on the effect of item construction on the exam reliability. The present study addresses the question of whether the test reliability and the other components of the item analysis are related.

Objective

The objective of this study is to determine the relation between reliability and other components of the item analysis of Multiple Choice Question (MCQs).

Methodology

All the items were obtained from MCQs from the final exams conducted in the academic years 2010-2011. There were 51 examinations from 38 modules (13 modules are offered for males and females separately), 10 modules are introductory, 9 are organ system modules and 19 are clinical and community medicine modules.

The item analyses were analyzed, using SPSS 17.0. Descriptive analysis was conducted for the following variables: number of items in each exam, number and percentage of non-functioning options (Non-functioning options are considered as the distracters that are used by >5% of the students or those have positive discrimination index), reliability of the exam as represented by KR20, percentage of items with facility index less that 35% and those with facility index more that 90%, percentage of items with discrimination index less that 19%, and those with facility index more than 60%. Correlations were conducted among the variables and the reliability coefficient.

Results

EDITORIAL

The modules analyzed are 38 (61%) of the total modules in the programme. The minimum number of items per exam is 18 while the maximum is 80; the mode is 40 items per exam. Within the 51 exams there are 2448 items; each of those items has five options. None of the items has four or three options, and the total number of options in the whole set of items we dealt with was 12240 options.

The number of nonfunctioning options (options other than the key answer chosen by less than 5% of the students or have positive discrimination) was a minimum of 46 per exam and a maximum of 253 per exam, and the mode number of nonfunctioning options was 105. When considering the percentage of the nonfunctioning options the minimum was 27.6 % of the total exam options and the maximum was 70.5% of the total exam options. The mode was 52.5% of the options , while the mean was 44.5%. Within the total number of options in the 51 exams, which was 12240 options, there were 5597 (45.7%) options.

When considering the reliability co-efficient; that was calculated using the KR-20 co-efficient we found that the minimum KR-20 within the 51 exams was 0.42 and the maximum was 0.91. The mode was 0.70. The reliability between 0.40-0.69 was found in 14 (27.5%), between 0.70-0.89 was found in 34 (66.7%) and 0.90 and above were three (5.8%).

When considering the percentage of items with facility index <35% per exam, the minimum percentage was 0% and the maximum was 18.3% of the items. In total, there were just eight out of 42 exams (19.05%) without items with facility index <35%. Table (1) shows the number of exams in relation to the percentage of items with facility index <35%.

Table (1): Number of exams in relation to the percentage of items with facility index <35%.

Percentage of items with facility index <35% per exam	Number of Exams	Percentage of Exams
Zero%	8	19.05%
1-4.9%	8	19.05%
5-9.9%	15	35.7%
10-14.9%	9	21.4%
15-19.9%	2	4.8%
Total	42	100%

When considering the percentage of items with facility index >90% per exam, the minimum percentage was 2.9% and the maximum was 75% of the items. Table (2) shows the number of exams in relation to the percentage of items with facility index >90%.

Table (2): Number of exams in relation to the percentage of items with facility index >90%

EDITORIAL

Percentage of items with facility index >90% per exam	Number of Exams	Percentage of Exams
0-9.9%	7	16.7%
10-19.9%	6	14.3%
20-29.9%	12	28.8%
30-39.9%	5	12%
40-49.9%	5	12%
50-59.9%	3	7.2%
60-69.9%	3	7.2%
70-79.9%	1	2.4%
Total	42	100%

When considering the percentage of items with discrimination index <19 % per exam, the minimum percentage was 4% and the maximum was 45% of the items. Table (3) shows the number of exams in relation to the percentage of items with discrimination index <19%.

Table (3): Number of exams in relation to the percentage of items with discrimination index <19 %.

Percentage of items with discrimination index <19% per exam	Number of Exams	Percentage of Exams
0-9.9%	5	11.9%
10-19.9%	20	47.6%
20-29.9%	10	24.3%
30-39.9%	5	11.9%
40-49.9%	2	4.8%
Total	42	100%

When considering the percentage of items with discrimination index >60 % per exam, the minimum percentage was 0% and the maximum was 17.7% of the items. Table (4) shows the number of exams in relation to the percentage of items with discrimination index >60 %.

Table (4): Number of exams in relation to the percentage of items with discrimination index >60 %.

Percentage of items with discrimination index >60% per exam	Number of Exams	Percentage of Exams
Zero%	11	26.2%
1-4.9%	21	50%
5-9.9%	6	24.3%

10-14.9%	**2**	**4.8%**
15-19.9%	**2**	**4.8%**
Total	**42**	**100%**

A correlation between the reliability coefficient (KR-20) and the other variables is shown in the table below. There is a positive correlation between KR-20 and both the number of items in the exam and the percentage of items with discrimination index >60%. And the table shows a negative correlation with percentage of the nonfunctioning options, percentage of items with facility index <35%, percentage of items with facility index >90% and percentage of items with discrimination index <19%.

Table (5): Correlation between reliability coefficient (KR20) and the other variable

		KR 20
No. of Items	**Pearson Correlation**	**.420[a]**
	Sig. (2-tailed)	**.002**
	N	**51**
Percent of Non-functioning Options	**Pearson Correlation**	**-.371-[a]**
	Sig. (2-tailed)	**.007**
	N	**51**
KR 20	**Pearson Correlation**	**1**
	Sig. (2-tailed)	
	N	**51**
Facility Index< 35%	**Pearson Correlation**	**-.019-**
	Sig. (2-tailed)	**.907**
	N	**42**
Facility Index>90%	**Pearson Correlation**	**-.457-[a]**
	Sig. (2-tailed)	**.002**
	N	**42**
Discrimination Index <19%	**Pearson Correlation**	**-.501-[a]**
	Sig. (2-tailed)	**.001**
	N	**42**
Discrimination Index> 60%	**Pearson Correlation**	**.485[a]**
	Sig. (2-tailed)	**.001**
	N	**42**

a. Correlation is significant at the 0.01 level (2-tailed).

EDITORIAL

Discussion

Item analysis is used to determine that exam scores are able to reflect the intended difficulty and discrimination of constructed items to assess the test taker's true ability.[2] The most important functions of item analysis are that it leads to improvement of teachers' abilities to write items, and to identify items to be eliminated from the exam.[1, 2]

The results showed that all the exams have some nonfunctioning options, some as high as 75% of options. These options add nothing to the items psychometrically.[3] Usually the nonfunctioning options are not plausible distracters, which should be avoided as mentioned in item writing guidelines.[5] All the guidelines on item writing recommend that the number of options should be developed and should be functional and feasible.[6] The presence of nonfunctional options is usually related to the construction abilities of teachers and their capacities in item development.

Reliability provides an indicator of the proportion of error contained in the exam score. There are many ways to define it statistically and to calculate it.[4] In the modern psychometric concepts reliability is considered as an integral part of validity, still it's treated separately because of its importance .[5]

The results showed a significant negative correlation between the reliability (KR-20) and the percentage of nonfunctioning options per exam. This may well affect the standard and fairness of our exams. It is found that removing the nonfiction options will improve the test performance and effectiveness of tests.[6]

To overcome this, some research findings recommend the use of less than five options per item, as this will reduce the number of nonfunctioning options and also will save time that can be used for construction of more items, hence increasing the content validity,[7] specially that their research results showed no change in the reliability, it is noticed that reduction from 4 option to 3 will result in increased reliability[5], but, there is no full agreement on this, as other research have noticed reduction in the reliability with the reduction of options, still, reduction of ineffective options will result in less reduction of reliability than that resulted when deleted random option [5] .

The accepted range of facility index for items in our medical schools is 35%-90%. From the results above it is evident that the majority of exams (81%) have items with facility index <35% and all exams have items with facility index>90%. This may affect the ability of the questions to determine the differences between individuals. It is well known that the maximum determination of these differences will be achieved if the difficulty (facility index) of the items are approaching 50%, and the ability decreases below and above this 50% [3]. From the result above it is evident that there is a negative correlation between reliability (KR20) and the percentage of items with facility index <35% and percentage with facility index > 90%. In the discussion section, this needs to be interpreted.

EDITORIAL

Item discrimination is the most important information about the performance of the item[4], the accept range of discrimination index is 19-60%, the results above shows that, all the exams have items with discrimination index >19% and 89% of the exams have items with discrimination >60%, It well known that the Excellent discrimination can be determined if the index is 40% and up to 75% [3, 4]. This is well supported by the results in this research showed a negative correlation between reliability (KR20) and percentage of items with discrimination index <19%, while it showed positive correlation with the percentage of items with discrimination index >60%.

Conclusion

This work will conclude the importance of conducting item analysis and make use of it to improve the quality of item writing and examination. Focused Faculty development can be arranged because of this item analysis.

Reference

1.Tavakol M, Dennick R. Post-examination analysis of objective tests. Medical Teacher 2011;33:447-58.
2.Hopkins KD. Item Analysis. Educational and Psychological Measurement and Evaluation: Allyn & Bacon; 1998. p. 254-72.
3.Downing S. Statistics of Testing Assessment in Health Professions Education Tylor and Francis 2009.
4.Axelson R, Clarenc K. Reliability. Assessment in Health Professions Education Tylor and Francis; 2009.
5.Rodriguez M. Three Options Are Optimal for Multiple-Choice Items: A Meta-Analysis of 80 Years of Research Educational Measurement: Issues and Practice 2005. p. 4-13.
6.Woddford K, editor. Using multiple choice questions effectively in information. Beyond the Comfort zone: Proceeding of the 21st Ascilite Conference 2004.
7.Rashmi V, Supe A. Multiple Choice Question: A lierature review on optimal number of options The National Medical Journal of India 2008;21(3):130-3.