Research Paper Medical Science

Looking into the Quality of Our Examinations` Questions: Item Analysis; Do Questions Perform to the Expectations?

Ahmad AbdulAzeem. Abdullah MD, MRCSEd, Department of Surgery Faculty of Medicine, University of Tabuk.

Mohammed Elnibras MD, Department of Surgery Faculty of Medicine, University of Tabuk.

*** Dr. Ibrahim Albalawi** MD, FRCS Ireland, Department of Surgery Faculty of Medicine, University of Tabuk. * Corresponding author

KEYWORDS

Introduction:

Assessment of the quality of examinations is an area of increasing concern in the current era of medical education practice (Epstein 2007, Davies & Howells 2004). Examinations provide measures with which a student`s ability in a specific domain can be estimated (Champlain 2010). In medical education, assessment is used for many purposes which include; provision of information on the student`s progress through the course of his\her education (Champlain 2010), identification of students who need remedial actions, and it is also used for certification and licensure processes. The demand for good quality medical graduates implied by the patient safety concerns and the requirements of different stakeholders necessitates development of high quality examinations, which reflects the true ability of tomorrow doctors (Murdoch-Eaton & Whittle 2012). In view of the aforementioned, it is beyond doubt that important decisions are made out of the examination results, which may have an impact on the students` academic future and career (Tarrant & Ware 2008). However, Downing have shown that there are deficiencies encountered in examinations prepared by classroom teachers and that poorly constructed items are still frequently used in medical education (Downing 2005). Tarrant & Ware have added that these poorly constructed test items can affect students` performance in a given test and in turn distort the validity of the test`s results (Tarrant & Ware 2008).

The construction of good quality MCQ (multiple-choice questions) is known to be tedious and time consuming process imposing excessive demand on teachers (Downing 2005, Epstein 2007). However, the time and effort spent in this process does not necessarily guarantee good quality of those items, which requires, moreover, thorough scrutiny after administration of the items in real examination setting (Osterlind 2002).

Item analysis has been defined by Osterlind as "the process by which test items are examined critically" (Osterlind 2002 p. 257). It is a numerical analysis derived from classical and item response theories (Swanwick 2010, Champlain 2010) and is conducted following test administration. It comprises statistical calculations which, can be applied to test items with dichotomous responses, and offer the opportunity to judge on and improve their quality in order to use them more effectively in future assessment (Tavakol & Dennick 2011, Swanwick 2010, Kapur & Kulenović 2010, Osterlind 2002). It can be used to diagnose structural and other types of errors encountered in MCQ test items and reduce their effectiveness in assessing students` knowledge (Downing 2002).

The difficulty index (*P-value*) is the most popular item analysis factor (Ostrelind 2002). It equates the proportion of the examinees who responded correctly to a particular item (Tavakol & Dennick 2011, Anderson 1983). The ideal range of this value is midway between the maximum score (1) and the chance score of the item (e.g. 0.25 for a 4-option MCQ item). Champlain have indicated that the preferred value of difficulty index is dictated by the goal of the examination; if the students are intended to be ranked according to their level of performance, then the range of 0.3 to 0.7 is ideal (Tavakol & Dennick 2011, Champlain 2010, Lipton & Huxham 1970). He also added that items with difficulty index of 0.5 have the highest ability to discriminate between the students according to their level of abilities (Champlain 2010). The difficulty index can provide inferences about problems of test items that might exist in the content, key and structure of the item and\or teaching quality of the construct being assessed (Osterlind 2002). Flawed MCQ items can affect the difficulty index particularly for high-achieving students (Tarrant & Ware 2008), although other levels of performance are also affected (Downing 2002).

The discrimination power index determines differences among individual examinees on the subject matter or psychological construct being tested. It is a relationship between the difficulty of an item and the ability of the examinees (Osterlind 2002). It is based on the assumption that examinees with high mastery of the subject are more likely to answer any particular item about that subject than examinees who exhibit low mastery of the subject (Ostrelind 2002, Champlain 2010, Anderson 1983). The value of the discrimination power is indexed by the point-biserial and the *phi* coefficients (Champlain 2010, Buckley-sharp & Harris 1972). Champlain have indicated that items with discrimination power of < 0.2 should be revised (Champlain 2010, Lipton & Huxham 1970) while it rarely happens that the value of discrimination to exceed 0.5 (Buckley-sharp & Harris 1972, Anderson 1983). Poorly structured (flawed) test items have a low discrimination power and consequently their ability to discriminate between high- and low-achieving students is compromised. This effect increases significantly when the number of these items becomes large in any given examination (Tarrant & Ware 2008, Lipton & Huxham 1970).

The number of distractors that should be developed in an MCQ item is a subject of ongoing debate (Haladyna, Downing & Rodriquez 2010). In their comprehensive account on MCQ writing-guidelines, Haladyna and his colleagues stated "use as many plausible distractors as you can but research suggests three are enough" (Haladyna, Downing & Rodriquez 2010 pp. 312). They have also revised four standardized tests and they found that two thirds of all items have one or two effectively performing distractors and only 1-8% of all questions have three effective distractors (Haladyna, Downing & Rodriquez 2010). They have concluded that MCQ with non-functioning distractors are not desired and have been found less discriminating in comparison to MCQ with plausible distractors (Haladyna, Downing & Rodriquez 2010). In their original study in 1989, Haladyna and Downing have related poorness of MCQ with non-functioning distractors to low "*efficiency*" of these questions in terms of the longer time spent in test development and administration (Haladyna & Downing 1989).

Kapur and Kulenović evaluated a 30-questions best-of-five MCQ paper of anatomy run to 52 students in 2010. They found that for 73.4% of the questions the difficulty index was acceptable (range 0.3 to 0.7) with 16.7% of questions with ideal difficulty index (0.5 to 0.6). 13.3% of the question in their study were too difficult (difficulty index of < 0.3) while another 13.3% of the questions were too easy (difficulty index of > 0.7). The mean discrimination power was 0.4 ± 0.21 with 86.7% of the questions falling in this acceptable range. They have also accounted that 75% of the questions with low discrimination power (< 1.5) were either too easy or too difficult and that 84.6% of the questions with acceptable discrimination index have also reasonable difficulty index (0.3 – 0.7) (Kapur and Kulenović 2010). A direct conclusion that can be drawn here is that questions with acceptable difficulty level will function as good distractors and vice versa.

Our medical School is a newly established one in which item analysis has never been used to judge on the quality of our MCQ test items. We believe that item analysis data will give us a lens through which we can judge the quality of our MCQ test items and improve them in the future.

The objective of the study is to assess:

- The difficulty index and the discrimination power of our MCQ test items.
- The number of questions with one or more non-functioning distractor and their average difficulty and discrimination indices.

Methods:
This is a quantitative descriptive study in which we evaluated two versions of the final written surgical examination administered to the 6th year male and female batch of students (N= 30 and 28 respectively) in 2014. Each examination is composed of a 100 four-option single-best- answer questions. Item analysis was performed with determination of the difficulty index and the discrimination power in addition to identification of questions with one or more non-functioning distractor. Items were categorized according to their difficulty and discrimination values in which we considered the preferable range of difficulty index as (0.3 ≤ *p. value* ≤ 0.7) and the ideal discrimination power as > 0.2. The difficulty and discrimination indices were also calculated for questions with non-functioning distractors. Interesting item analysis data of two selected questions are presented and discussed to highlight usefulness of item analysis in evaluation of the quality of questions and identification of problems with their content and\or teaching.

Results:
Table one shows that the questions which lie within the preferable range of difficulty index ((0.3 ≤ *p. value* ≤ 0.7) accounted for 40.5% and their discrimination power was higher their counterparts which have a difficulty index outside the above-mentioned range; 0.21 and 0.16 respectively, although that association was not statistically significant (p=0.07

Table two compare questions according to their discrimination ability. It shows that questions with discrimination power of ≥ 0.2 accounted for 43.5%. These questions were more difficult than their counterparts with discrimination power less than 0.2 (*P. value* of 0.64 and 0.76 respectively). This relationship was found statistically significant (*P* = 0.00).

Table three shows that questions which contain one or more non-functioning distractor represented 56% of all questions. These questions are less discriminating and easier than their counterparts in which all distractors are functioning (discrimination power of 0.14 & 0.22, and *p. value* of 0.89 & 0.57 respectively). These relations were found statistically significant (*P* = 0.003 & 0.004).

Table four and five present item analysis data of two interesting examples of questions on both sides of an extreme. Table four shows a very difficult question which was not answered by all candidates (P. value = 0.0), while table five present a very easy question which was answered correctly by candidates (p. value of 1.0). The discrimination power of both questions was zero.

Table No. (1): Comparison of questions according to their difficulty index.

Category of difficulty	Number	Percentage	Discrimination power (Mean)	Significance (*t-test*)
Within Preferable range of difficulty (0.3 ≤ *p. value* ≤ 0.7)	81	40.5%	0.21	0.07
Outside preferable range of difficulty (0.3 > *p. value* > 0.7)	119	59.5%	0.16	

N=200.

Table No. (2): Comparison of questions according to their discrimination power.

Category of discrimination	Number	Percentage	Difficulty index (Mean)	Significance (*t-test*)
Within Preferable range of discrimination (> 0.2)	87	43.5%	0.64	0.00
Outside preferable range of discrimination (< 0.2)	113	56.5%	0.76	

N=200. P < 0.05

Table No. (3): Comparison of questions with and without non-functioning distractors.

Parameter/Questions	Questions with Functioning distractors	Questions with Non-functioning distractors	Significance (*t-test*)
Number and percentage	88 (44%)	112 (56%)	-
Mean difficulty index	0.57	0.89	0.003
Mean discrimination power	0.22	0.14	0.004

N=200. P < 0.05

Table No. (4): An interesting example of item analysis data of a selected question.

Difficulty index: 0.0					
Group	Number	A	B*	C	D
Total	28	0	0	0	28
High performers	13	0.0	0.0	0.0	1.0
Low performers	15	0.0	0.0	0.0	1.0
Discrimination power		0.0	0.0	0.0	0.0

(*) Correct answer.

Table No. (5): An interesting example of item analysis data of a selected question.

Difficulty index: 1.0					
Group	Number	A	B*	C	D
Total	28	0	28	0	0
High performers	13	0.0	1.0	0.0	0.0
Low performers	15	0.0	1.0	0.0	0.0
Discrimination power		0.0	0.0	0.0	0.0

(*) Correct answer.

Discussion:

The ability of a question to discriminate between different levels of performance is related to its difficulty and the ability of the test's candidates. In line with the findings of Kapur & Kulenovic, it was shown that the questions with an appropriate level of difficulty are more discriminating than questions that are either too difficult or too easy. This is an expected and logical finding since too difficult questions are answered by few students while too easy questions are answered by most of the candidates and in both cases the question loses its ability to differentiate between high- and low-achieving students. This effect is most clear in questions which are either answered correctly by all examinee or, on the other extreme, not answered correctly by any examinee; the discrimination power in both situations is equal to zero as shown in table 4 & 5. Thus, questions with an inappropriate level of difficulty may warrant revision because one of the essential purposes of any examination is to discriminate between different levels of ability of examinees. This is particularly important when the number of those questions amounts to considerable portion of the examination as the case in this study.

The aforementioned result is also supported by the finding that questions with an appropriate discrimination power tend to have also a more appropriate difficulty level than questions which are less discriminating, highlighting more on the close relationship between the difficulty of a question and its ability to differentiate between high- and low-achieving students. This relationship is proved statistically significant in this study. However, the number of questions with undesired discrimination power exceeded their counterparts with an appropriate discrimination ability.

Questions that contain one or more non-functioning distractor accounted for almost two thirds of the questions in this study. This result coincides with the finding of Haladyna etal who indicated similar result in their series. The quality of those questions does not seem good, as they are found easier and less discriminating than their counterparts with all functioning distractors and this association is proved statistically significant. This result coincides with the finding of Haladyna and his colleagues (Haladyna, Downing & Rodriquez 2010). Those questions are also assigned in the literature as being "low-efficient" in terms of the time wasted in preparation and administration of such test items. Consequently, those questions worth revision in order to improve their quality in the future and there are number of strategies which have been proposed in the literature on how to deal with non-functioning distractors which is out of the scope of this study (Haladyna & Downing 1989).

Table four presents an interesting item analysis data of a selected question in which all examinees have chosen only one option; the incorrect one. Such data may indicate problem with the question content or structure but may also point to poor teaching. However, a similar condition is sometime encountered in questions that assess information related to guidelines because these tend to change from time to time. On the other hand, table five shows an item analysis data of another selected question in which all examinees have chosen one option; the correct one. While this question does not serve any discrimination value, in a criterion-reference test it may reflect learning achievement of important knowledge domain particularly if it is about emergency and life-saving conditions. In this circumstance, such item result should in fact be celebrated rather than being dismissed. Nevertheless, such questions should always be revised an inspected vigilantly and they demonstrate the usefulness of item analysis in provision of essential feedback regarding the quality of item construction and/or teaching.

Conclusion:

It has been shown that there is a close relationship between the difficulty of an MCQ item and its ability to discriminate among different levels of students` achievement. Items, which have an appropriate difficulty level, tend to be strong discriminators and vice versa. Questions with one or more non-functioning distractor are still frequently encountered in our examinations, they have also been found easier and less discriminating than their counterparts in which all distractors are functioning. Item analysis provide an essential tool through which problems in MCQ questions can be detected and treated following test administration.

Limitations:

The results of this study cannot be generalized to other contexts owing to the small number of students undertaking the tests.

References:

1. Bhakta B., Tennant A., Horton M., Lawton G., Andrich D. Using item response theory to explore the psychometric properties of extended matching questions examination in undergraduate medical education.BMC Medical Education, 2005; 5: pp. 5-9.
2. Champlain A.A primer on classical test theory and item response theory for assessments in medical education.Medical Education, 2010; 44(1): pp. 109-117.
3. Downing S.Effects of violating standards item writing principles on tests and students: the consequences of using flawed test items on achievement examinations in medical education.Advances in health sciences education. 2005; 10: pp. 133-143.
4. Downing S.Item response theory: Applications of modern test theory in medical education.Medical Education, 2003; 37: pp. 739-745.
5. Downing S.Evaluation Methods: What do we know?Academic Medicine, 2002; 77(10): pp. S103-S104.
6. Davies H., Howells R.How to assess your Specialist Registrar.Arch Dis Child, 2004; 89: pp. 1089-1093.
7. Epstein R.Assessment in Medical Education.N Engl J Med, 2007; 356: pp. 387-396.
8. Osterlind J.Constructing test items: Multiple-Choice, constructed-Response. Performance and Other format.Second edition, 2002, Kluwer Academic Publishers, New York. p. 253-263.
9. Swanwick T.Understanding Medical Education: Evidence, Theory and Practice.First edition. Wiley Blackwell, London, 2010; pp. 225-229.
10. Tarrant M., Ware J.Impact of item-writing flaws in multiple-choice questions on student achievement in high-stakes nursing assessments.Medical Education, 2008; 48: pp. 198-206.
11. A. Lipton and G. J. HuxhamComparison of multiple-choice and essay testing in preclinical physiology.British Journal of Medical Education, 1970, 4: pp. 228-238.
12. Deborah Murdoch-Eaton & Sue Whittle.Generic skills in medical education: developing the tools for successful lifelong learning.Medical Education 2012: 46: pp. 120–128.
13. M. D. Buckley-Sharp and F. T. C. Harris.Methods of analysis of multiple-choice examinations and questions.British Journal of Merlica1 Education, 1912, 6: pp. 53-60.
14. J. 0. Nnodim.Multiple-choice testing in anatomy.Medical Education 1992, 26: pp. 301-309.
15. Tavakol M & Dennick R.Post-examination analysis of objective tests.Medial Teacher, 2011; 33: pp. 447–458.
16. E. Kapur, A. Kulenović.Analysis of Difficulty and Discrimination Indices in One-Best Answer Multiple Choice Questions of an Anatomy Paper.Journal of Medical Faculty University of Sarajevo, Bosnia & Herzegovina, 2010, 45(1): pp. 14-20.