EDITORIAL

Multiple Choice Questions Revisited: Improvement of Validity for Fair Tests

Mohamed Elhassan Abdalla

Corresponding Author: Dr. Mohamed Elhassan Abdalla MB.BS, Msc-HPE, PhD-HPE. Medical Education Unit . Faculty of Medicine- Jazan University. Gizan-Jazan, KSA. P.O Box 837. Postal Code 45142. Telephone: +966555773042. E.mail hason75@yahoo.com

Abstract:

In the last two decades there is a vast increase in the number of medical schools all over the world, especially in the Eastern Mediterranean Region (EMRO). This necessitates the existence of faculty development programme for the increasing staff.

The objective of this communication is to revisit the multiple choice question elaborate on factors that will affect its validity, aiming at having acceptable fair test for the students.

This communication describe the major threats of validity of multiple choice questions , namely the construct under representation (CU) and the construct irrelevant (CIV) threats and it describe how can they be overcome.

The communication concluded that faculty development programmes and the use of guidelines and blueprint will lead to improvement in the tests validity in medical schools.

Background:

In the last two decades there is a vast increase in the number of medical schools all over the world, especially in the Eastern Mediterranean Region (EMRO) where the number of medical schools is estimated in the range of 10-49 medical schools per country [(1)]. This increase results in increase of the teaching staff to face the increased number of students. This necessitates the existence of faculty development programme by all available means.

The objective of this communication is to revisit the multiple choice question as a widely used examination tool and elaborate on factors that will affect its validity, aiming at having acceptable fair test for the students.

Introduction:

Assessment is one of the important components of the teaching and learning cycle, it is the major drive of student's learning, and it is said that (the assessment tail wags the curriculum dog). A very important rule has been stated by Shwarts that for every evaluative action, there is an equal or (greater) (and sometimes opposite) educational Reaction [(2)], this reaction is mainly affecting the expected impact from the educational process.

In the field of Medical Education, many assessment tools are used to evaluate the achievement of students, the tools are intended to assess the three domains in Bloom's taxonomy of educational objectives.

Among those tools, the Multiple Choice Questions (MCQs) is the most widely used assessment methods in medical schools; it is used in assessment for more than 50 years [(2, 3)]. Many factors have contributed to its popularity; they are easy to administer, easy to be marked even to large number of students, they have testing efficiency and its objectivity is a major cause for its use popularly [(2, 3, 4)]. Theoretically the MCQs can serve the concepts of organization of data and answers to questions as the answers (options) are already there, so students do not need to supply words or phrases and this will contribute to the objectivity of

EDITORIAL

answers and marking. It supports also the concept of regression and summarization of data for example a long case history can be summarized into few words or phrases to fit the exam purpose, yet having the important information, both organization and regression are important concepts supporting educational achievement. [4]

MCQS can test well in the cognitive domain at all levels; although there is considerable criticisms for its use in professional education that , it tends to assess only recall of knowledge , but, it can test any higher level of the cognitive domain, and discriminate well between students, provided that it is well constructed. There is a great mass of research which showes that, the test of knowledge can best determine the expertise [2, 3] that is why many efforts are made to care for the good construction of questions.

There are many types of MCQs described in the literature [5], the most commonly used is the one best answer question, which usually has 4-5 options for the student to choose from them. There is no psychometrical law behind the number of options; one cannot make a judgment on the characteristics qualities upon the number of options [6], and there is great evidence which suggests that, more options do not improve psychometrics and the three options have more distractibility and more discrimination, and it saves time to teachers that can be used for adding questions and hence improving the validity of tests. [7]

Validity and Validity Threats

The utility of any assessment method -including MCQs- depends on its validity, reliability, its educational impact and its feasibility [6]. In this communication we are going to concentrate on some factors that lead to improvement of the validity of MCQs.

Validity refers to the evidence presented to support or refute the meaning or interpretation assigned to assessment results; without evidence of validity, assessment has little or no meaning [8]. Validity is now believed to be mostly as construct validity, rather than having different types of validity as it has been thought of. The construct validity can have different sources of evidence, either content which means outline and plan of the test described by a detailed test blueprint or response process evidence that refers to evidence of data integrity such that all sources of error associated with the test administration are controlled or eliminated to the maximum extent possible or internal structure which relates to the statistical or psychometric characteristics of the examination questions or performance prompts, the scale properties – such as reproducibility and generalizability and the psychometric model used to score and scale the assessment, the other two sources are the consequences of the test and the relation to other variables [8].The first three are most important in the written test.

There are many threats to validity of any test, they can be grouped under two headings, construct under-representation (CU) which refer to under sampling the content domain or construct irrelevance variance (CIV) which refer to the variables that systematically interfere with the ability to meaningfully interpret scores or ratings (flawed items, inappropriate reading level, statistically biased items, testwiseness, teaching to test and cheating) [9, 10]. The role of teachers is to make sure that, those threats and its related factors are eliminated to the minimum in order to have a valid and fair test.

The content under-representation can mainly be avoided by the use of test blueprint, a term that has been borrowed from architecture. It indicates that a process of assessment needs to be conducted according to a replicable plan [11]. The use of the blueprint will ensure that, the test has been developed and mapped carefully against the educational objectives of the course, ensuring fair representation of objectives. [11, 12]

EDITORIAL

The use of blueprint can contribute to the elimination of some of the Construct irrelevant variance (CIV) also, as it will ensure that, the test is using the appropriate methods to educational objectives and that the students are having un-biased results and score in well represented test. [9] [11]

Despite its importance as it is shown above, the use of blueprint is underestimated by teachers in medical schools; in a study published in 2003, the authors have indicated that very few medical schools in USA are using blueprint for validation of tests [11, 12]. Teachers must be trained in construction and use of blueprints, many publications have elaborated in the steps to develop the exam blueprint. The Association of Medical Education in Europe (AMEE) has published a simple guide for blueprint development under the series of Twelve Tip which can be used for training and discussion [13]. The following is an example of part of a blueprint for the written test in Paediatrics

The above blueprint can be constructed in many ways but the simple steps are summarized below:

1- Determine the content areas and topic and their relative importance according to relative weight of the objectives of the course. In the above example the contents and weights are Cardivascular (30%), Respiratory (30%) and Endocrinology (40%).
2- Determine the weight for each domain you need to assess. In the above example the domains and weights are Knowledge (Recall) (30%), Comprehension (Reasoning) (40%) and Application of knowledge (30%).
3- Determine the time needed by the students for each item; MCQs (1-1.5 min.), Modified Essay Question (MEQs) (5-7 min.), etc.

System (Clinical-Problem)	Basic Science	Patho-physiology	Clinical Presentation	Investigation & Diagnosis	Management	Follow-up	Prognosis	Total Items
Cardiovascular System Heart Failure Infective endocarditis	 1K		 1C	 1Ap				3 (30%) 2 1
Respiratory System Asthma Tuberculosis	 1K			 1K	 1Ap			3 (30%) 2 1

EDITORIAL

Endocrine System Diabetes Mellitus Hypothyroidism		1C		1C	1Ap	1C		4 (40%) 3 1
Total Items Knowledge Comprehension Application	2 (20%) 2 0 0	1(10%) 0 1 0	1(10%) 0 1 0	3(30%) 1 1 1	2(20%) 0 0 2	1(10%) 0 1 0		10 3(30%) 4(40%) 3(30%)

K= Knowledge, C= Comprehension, AP=Application

4- Having the total time for the exam you can determine the number of items in each domain (see step No. 2) and then in each content area (see step No. 1); if you have 120 min., for the application of knowledge domain in this exam the distribution may be as follows;

Total Time for the exam = 120 min

Application of knowledge = 30% = 36 minutes = 10 minutes
(Cardivascular), 10 minutes (Respiratory) & 16 Minutes (Endocrine).
10 minutes (application of knowledge in the Cardiovascular) = 3 MCQs + 1 MEQ.
10 minutes (application of knowledge in the Respiratory) = 3 MCQs + 1 MEQ.
16 minutes (application of knowledge in the Endocrine) = 3 MCQs + 2 MEQ.

The other threat for validity is the CIV, which can be represented by the ill-structured items; fairness and validity require that the questions should cover the important content and also asking for good construction of questions; that makes it easy to understand (2). One of the major problems in question construction is that very few of staff in medical schools has formal training in question construction [(14)], the other problems in exam construction, may relate to the process of construction of items. In USA, commonly a medical student sits to high stake examination taking at least 3 hours every 4-6 weeks during the college life-time, paradoxically most of the questions are prepared in last minutes, examination are generated by many people who taught the course, and little time is available for review of the questions for overall quality before the examination is provided to students, and the most important factor is that, there may be no agreement on the standard of the question format. [(14)]

In MCQs construction, The questions can be context free or context rich stimulants, the contexts can be a case scenario, and the question to be answered is dependent upon that scenario [(6)], the context free usually test factual knowledge, which is an important part of problem solving, but in the professional life where

EDITORIAL

factual knowledge is usually needed in limited way, so it is recommended that to have more attention for the purpose of using the context free type, to increase the professional authenticity [2]. On the other hand, the context rich question is attempting to test higher levels of knowledge and reflects relevance to the professional life, that is why it is more recommended to be used in medical education [6], The context free question are easy to construct , that is why teachers prefer them mostly in examination.

The following are two examples of context free and context rich stimulant questions, to elicit the difference between the two types:

What is the most important mechanism of Heart Failure in children of one year old?

A. Ventricular weakness
B. Volume overload
C. Pressure overload
D. Ventricular dysfunction
E. Ventricular dilatation

A 1-year-old baby, who is a known case of ventricular septal defect (VSD), was brought to the emergency room by his mother, complaining of cough, shortness of breath which disturbed his feeding for the last 3 days. On examination the baby was ill, his temperature was 38.7 C, his pulse was 110/min which was regular, dyspenic with respiratory rate of 70/min and intercostals recessions, and had a liver which was 4 cm below the costal margin.

What is the most likely mechanism of his symptoms?

A. Ventricular weakness
B. Volume overload
C. Pressure overload
D. Ventricular dysfunction
E. Ventricular dilatation

Although the two questions are apparently the same , but, the first one is totally recall of information , while the second is stimulating students to think , and it resembles the real life situation which the students will face in the future.

The other thing to take care about in construction is to make sure that items are free from flaws. Item-writing flaws can affect student performance on MCQs, making items either more or less difficult to answer, it is found by some authors, that, flaws usually make the item less difficult, but, still my yield some ambiguity, that, 10-25% of failed students would have passed the exam if the flawed items were removed [15]. In another research, the authors found that, Flawed item writing may reach 33% within one test [10], that flawed item tend to be more difficult and fail more students unfairly. [10].

MCQs written at a lower cognitive level were significantly more likely to contain item-writing flaws than those which test higher levels [16], this may be because of the attention and time spent by teachers in construction of those types of items.

Guidelines for item construction are found in many textbooks. The one produced by the National Board of Medical Examiners is an example. It is widely used by medical schools [5], and beside the guidelines it contains examples and templates for good items both in basic and clinical sciences.

EDITORIAL

Terrant et.al has determined the most frequent violations to item construction, they are found to be [16]:

- Ambiguous or unclear information in the stem, such as the use of words like always, frequently, etc.
- Negatively worded stems: such as the use of (Not, except, etc.)
- Implausible distracters: this is usually found when teachers tried to complete the number of options without too much attention to its relation to the right answer.
- Unnecessary or gratuitous information in the stem
- More than one or no correct answer
- The longest option is correct
- Logical cues in the stem
- Word repeats in the stem and correct answer

They have also defined some of the flaws which are presented less frequently, such as fill-in-blank question, complex or K-type MCQs, grammatical cues associated with sentence completions and convergence cues [16].

In another research Terrant and ware have indicated a very important conclusion, it is that (Many in-house questions are flawed, because the majority of teachers do not have formal training in item construction). [15]

Conclusion:

In order to help improve the validity of the in-house MCQs, the following must be adopted by the authorities in medical schools:

1- Faculty development programmes that concentrate on assessment validity.
2- Adoption of blueprint process for exam planning
3- Development or adoption of guidelines for item construction.

References:

1. Mapping the World's Medical Schools. *International Medical Education Directory (IMED).* [Online] Foundation for Advancement of International Medical Education and Research. [Cited: May 10th, 2011.] http://www.faimer.org/resources/mapping.html.
2. Mccoubrie, Paul. Improving the fairness of multiple-choice questions:a literature review. *Medical Teacher.* 2004, Vol. 26, 8, pp. 709–712.
3. Anderson, John. For Multiple Choice Questions. *Medical Teacher.* 1979, Vol. 1, 1.
4. Lau, Man Pang. A theory of multiple-choice examination. *British Journal of Medical Education.* 1972, 6.
5. Case, Susan M. *Constructing Written Test Questions For the Basic and Clinical Sciences.* Philadelphia : National Board of Medical Examiners, 2002.
6. Vleuten, Lambert W T Schuwirth & Cees P M van der. Different written assessment methods: what can be said about their strengths and weaknesses? *Medical Education.* 2004, 38, pp. 974–979.
7. Marie Tarrant, James Ware. A comparison of the psychometric properties of three- and four-option multiple-choice questions in nursing assessments. *Nurse Education Today.* August 2010, Vol. 30, 6, pp. 539-543.

EDITORIAL

8. Downing, Steven M. Validity: on the meaningful interpretation of assessment data. *Medical Education* . 2003, 37, pp. 830-837.
9. Steven M Downing, Thomas M Haladyna. Validity threats: overcoming interference with proposed interpretations of assessment data. *Medical Education.* 2004, 38, pp. 327-333.
10. Downing, Steven M. Construct-irrelevant Variance and Flawed Test Questions: Do Multiple-choice Item-writing Principles Make Any Difference? *Academic Medicine*. October 2002, Vol. 77, 10 Supplement , pp. S103-S104.
11. Hamdy, Hossam. Blueprinting for the assessment of health care professionals. *The Clinical Teacher*. 2006, 3, pp. 175-179.
12. Patrick D. Bridge, Joseph Musial, Robert Frank,Thomas Roe,Shlomo Sawilowsky. Measurement practices: methods for developing content-valid student examinations. *Medical Teacher*. July 2003, Vol. 25, 4, pp. 414–421.
13. Sylvian Coderre, Wayne Woloschuk, Kevin Mclaughlin. Twelve tips for blueprinting. *Medical Teacher*. 2009, 31, pp. 322-324.
14. Ralph F. Jozefowicz, MD, Bruce MRalph F. Jozefowicz,Bruce M. Koeppen,Susan Case,Robert Galbraith,David Swanson, Robert H. Glew. The Quality of In-house Medical School Examinations. *Academic Medicine*.. February 2002, Vol. 77, 2, pp. 156-161.
15. Marie Tarrant, James Ware. Impact of item-writing flaws in multiple-choice questions on student achievement in high-stakes nursing assessments. *Medical Education.* 2008, 42, pp. 198–206.
16. Marie Tarrant, Aimee Knierim,Sasha K. Hayes,James Ware. The frequency of item writing flaws in multiple-choice questions used in high stakes nursing assessments. *Nurse Education Today.* 2006, 26, pp. 662-671.