The criteria and analysis of good multiple choice questions in a health professional setting

Ahmad A. Abdel-Hameed, *MPH, PhD,* ***Eiad A. Al-Faris,*** *MSc, MRCGP (UK),*
Ibrahim A. Alorainy, *DMRD, FRCPC,* ***Mohammed O. Al-Rukban,*** *ABFM, SBFM.*

ABSTRACT

Assessment of health workers as students and professionals has a profound impact on their learning and is an essential safety valve before certification. It is used for their training, their placement, their certification, and their promotion. The multiple choice question (MCQ) type of tests represents one of the most important examination tools that is commonly used in this assessment. The MCQs can be reliable, valid, and cost-effective in assessing medical knowledge. This paper portrays the different purposes of assessment in the medical field. The paper discusses in detail the criteria of a good assessment tool. Interpretation of MCQ test results is the final section of the paper.

Saudi Med J 2005; Vol. 26 (10): 1505-1510

Multiple choice questions (MCQ) tests are the most commonly used type of tests deployed on their own or in conjunction with other types of test tools for educational assessment. One of the advantages of MCQ tests is that they could be self-administration and could therefore be used for self-assessment. They are good for measuring knowledge, comprehension and could be designed to measure application and analysis. The MCQ tests are quick and simple to mark (electronic marking). This ease of marking permits rapid turn-round and personalized feedback in a very short time. The nature of MCQ tests makes them less likely to be affected by subjective bias from the marker, and therefore more reliable. A great advantage of MCQ tests is that they allow comprehensive coverage of topic area(s). The examiner can focus on detailed parts of the course. The MCQ test results could be statistically analyzed to provide information on facility (difficulty) as well as discrimination power of the test items. Particularly where there is a language problem, MCQ tests reduce reliance on skills of writing and self-expression. For all these reasons, MCQ tests are increasingly being used in the educational assessment of health professionals. The present article reviews how to interpret results obtained by an MCQ and how to evaluate the educational value of an MCQ test.

Interpretation of MCQ test results. For the appropriate analysis of a test, we have to take into consideration the purpose for which the test was made. Tests are used for different purposes during the educational process. At the start of instruction, the trainers or teachers need to decide the student's level in a program or course. For this purpose, they use a placement test (placement assessment). Assessment during instruction could be a diagnostic tool that points to areas of instruction deficiency where there is need of remediation (formative or diagnostic assessment). After completion of instruction, examiners and certifying bodies

From the Departments of Pathology (Abdel-Hameed), Family & Community Medicine (Al-Faris, Al-Rukban), and Radiology (Alorainy), College of Medicine and King Khalid University Hospital, King Saud University, *Kingdom of Saudi Arabia*.

Address correspondence and reprint request to: Dr. Mohammed O. Al-Rukban, PO Box 91678, Riyadh 11643, *Kingdom of Saudi Arabia*. Tel. +966 (1) 4670836. Fax. +966 (1) 4671967. E-mail: mrukban@ksu.edu.sa

use examination results to make promotion or certify candidates (summative assessment). Two broad categories of standards are used in interpretation of test results: norm-referenced and criterion-reference standards. The norm-referenced tests are designed to provide a measure of performance that is interpretable in terms of an individual selective standing in some known group. An example of these is the selection examination for admission to medical school. On the other hand, a criterion-referenced test is designed to provide a measure of performance that is interpretable in terms of clearly defined reference or criterion. An end of course test, which assesses whether the students have mastered specific learning objectives or domains of the course, exemplifies this. Most tests include a combination of the 2 standards.

Evaluation of an MCQ test. Specific criteria are used to evaluate the particular test. The most important of these are the test reliability and its validity, including its educational impact and usability of the test.[1]

Test reliability (reproducibility). Reliability is the degree to which a test consistently measures whatever it is supposed to measure. The more reliable the examination, the greater the confidence that the result would be the same if the examination were re-administered.[1,2] For example, if a student scored 50% in an MCQ examination and in 2 subsequent sittings shortly afterwards scored 30% and 90%, then you could interpret this wide variation of score by poor reliability. So, if a student passes a particular test, one has to be sure that they would not have failed a parallel test, and vice versa. Reliability is measured as a correlation with 1.0 being perfect reliability and the lower the figure the lower the reliability. A test represents at best a sample selected from a range of possible questions. It is a function of the number of questions, and a proper sampling of areas covered. Therefore, the reliability of an MCQ paper increases with the number of questions and the proper sampling of important areas of the discipline or the course. A number of factors are known to influence reliability of a test.[3,4] The length of a test affects its reliability; the more items included in an examination, the greater the reliability. Also, the wider the coverage of contents, the higher is the reliability of the test. When the sample is too narrow and does not cover the course content appropriately, the questions focus only on a certain element; hence, the scores cannot be generalized for the whole discipline. Reliability of a test could also be affected by environmental errors during the examination such as excessive heat, noise, and so forth. Performance may be poor in candidates who are required to sit an examination at the end of a long day. Processing errors might also occur and these decrease test reliability. For example, mistakes may be made by choosing the wrong key answer. Reliability of a test could be assessed by different methods that yield different types of reliability. The test-retest reliability or temporal consistency (stability) is the consistency of the same test given over a period to the same group of people. It is calculated by correlating the scores on a test with scores produced by a repeat administration to the same group. A high positive correlation indicates good reliability. One problem is deciding on the appropriate time period between the 2 administrations. If it is too short then the students are likely to remember their previous answers. If the gap is too lengthy then students may have benefited from further learning.[3] This type of reliability is important for tests used as predictors. In the equivalent-forms reliability (equivalency), 2 equivalent forms of the test are administered to the same group of people, (it can be very difficult to develop 2 truly equivalent forms of a test). Assessment of reliability by re-administering a test is practically difficult. Hence, the main method used for determination of test reliability depends on assessment of internal-consistency of the test as a measure of its reliability. These types of measures require only a single administration of the test. With the split half method, reliability of the examination is divided into 2 parts, for example, all the odd numbered questions and all the even numbered questions. The scores for both halves are correlated, and the degree of correlation reflects the internal consistency of the instrument. Only one administration of the test is necessary, and the method is particularly suitable for tests with many items. The more questions in an examination, the greater the likelihood of high reliability. Internal consistency can also be measured by determination of the coefficient of reliability. There are various statistical techniques for determining this coefficient. The KR20 and Cronbach's Alpha are 2 of the most common.[3] There are no absolute standards that can be used to judge whether a reliability coefficient is high enough. The acceptable minimum for the reliability coefficient depends mainly on the purpose of the test. It has been suggested by some experts that a minimum coefficient of 0.85 should be required if the results would be used to make important decisions about individual examinees and if the examination is the only tool available for their assessment. However, if the decision about a group of individuals, for example, about giving more time to teach about a subject then a minimum standard of 0.65 should be acceptable.[5] Low reliability coefficient could be more tolerable if each score is combined with other methods of assessment, for example, clinical examination, objective structured clinical examination, and so forth.

Test validity. The validity of a test is the extent to which it measures what it purports to measure,[6] or the extent to which inferences made from

assessment results are appropriate, meaningful, and useful in terms of the purpose for the assessment. A particular examination might be valid for one purpose but invalid for another. For example, a series of MCQs that test factual recall may be a valid measure of whether a student has read a textbook on diabetes but invalid as an indicator of whether that same student can actually manage a patient suffering from diabetes. Validity is a unitary concept that depends on a variety of types of evidence, is expressed by degree (high, low), and not number, and refers to the inferences drawn (not the instrument itself). One simple piece of evidence could be, for example, that experts score higher than students do on the test. Alternative approaches include an assessment of the soundness of individual test items.[6] The measure of validity is not a straightforward process as different types of evidence of validity are described. Previously, we used to talk about "types of validity" but now we consider these as categories for accumulating evidence of validity.[1,7] Thus, we speak of content-related evidence, construct-related evidence, criterion-related evidence and consequences of using the test as different types of evidence to establish validity of a test. Content-related evidence is the most important and feasible type of evidence to be assessed in an MCQ test. It relates to the sampling of the course topics within test elements (table of specification) (**Tables 1 & 2**). Therefore, the MCQ paper for a final medicine examination that does not have questions on the respiratory, or renal system, has a poor content validity. The examination committee or course teacher should form the table of specification. This type of grid should identify the content areas, for example, cardiology, and nephrology. It should also specify learning outcomes; for example, ability to make diagnosis or management. The number of test items for each content area and learning objective should be clearly specified, ensuring that the number of items in each cell is in proportion to the time spent in teaching and learning. Content validity is based on expert judgment, and the assessor should compare the course objectives and its' contents on one hand with what is measured by the examination, on the other hand. Construct related evidence is the extent to which a test measures hypothetical construct. If the aim of the MCQ paper as stated by the examination committee is to test the candidate's problem solving skills and it contained recall of knowledge (context-free) questions with no or little application (context-rich) questions, this means the test has low construct validity. Criterion-related evidence includes concurrent studies for a parallel criterion, for example, another test or predictive study relating the test to an event in the future, for example, relating admission tests with future performance. Thus, the students' scores on one MCQ paper could be compared with their scores on another established MCQ test performed at the same time or another test that tests knowledge. This would be achieved by correlating the 2 sets of scores and computing the correlation coefficient. The greater the positive coefficient, the greater the validity. Predictive studies relate to the certainty with which a test can predict future performance. It is particularly important if you are using your assessment for selection purposes. No test will have perfect predictability, so it is wise to base any decision on more than one predictor. It could be achieved by correlating the students' scores with their future performance. The magnitude of the correlation coefficient will determine the predictive validity. There are varieties of factors that may influence the validity of an assessment instrument. Vague or misleading instructions to candidates decrease the test validity. The language of the test might also affect its validity for example using inappropriate vocabulary or overcomplicated wording. If there are too few test items, this might lead to poor sampling and lower test validity. Duration of the test could be crucial. If insufficient time is allowed for answers then the test turns into one based on speed. The items in the test should be appropriate for the outcomes being measured, and the item should be of moderate difficulty. Too easy and too difficult items will fail to discriminate. One

Table 1 - Example of a table of specifications based on the stimulus (for family medicine examination).

Branches	Context free	Context rich	Total
Pediatrics	4	12	**16**
Medicine	3	11	**14**
Surgery	2	6	**8**
Obstetrics/Gynecology	3	9	**12**
Emergency Medicine	2	8	**10**
Orthopedics	1	3	**4**
Therapeutics	4	2	**6**
Primary Health Care	3	8	**11**
Dermatology	1	3	**4**
Psychiatry	2	7	**9**
ENT	1	2	**3**
Ophthalmology	1	2	**3**
Total	**27**	**73**	**100**

important consideration in assessing test validity is to evaluate its possible consequences or educational impact. This is important because students tend to focus strongly on what they believe will be in the examinations. Most of them strategically prepare for the exams depending on the question types expected. For examples: the application (context–rich) questions that start with scenario and test the candidates' problem solving skills, particularly the questions of higher cognitive levels, have a positive impact. They encourage the residents to think deeply and facilitate the development of clinical reasoning skills. The residents will be encouraged to go to work and practice, as it is similar to the examination context. On the other hand, the (context-free) type recall questions test memory, so students will concentrate on memorizing facts, which is not a good preparation for future practice and clinical work.

Test usability. This refers to the practical requirements of the test, such as cost of the test and its acceptability to the examinees and the examiners. The test should be cost-effective and easy to administer and mark.[1]

Item analysis in MCQs. In MCQ tests, item analysis provides a way of measuring the quality of questions - seeing how appropriate they are for the candidates, and how well they measure their ability. It also provides a pool of evaluated items that could be re-used repeatedly in different tests with prior knowledge of how they are going to perform. Now, a number of software computer programs provide a report on item analysis together with the MCQ test results. Two important indices provided by item analysis are the difficulty factor and the discrimination index.

Difficulty factor. This is essentially the proportion or percentage of students who answered the question correctly. It is based on either the total number of students (the percentage who answered correctly) or it could be based on a sample of upper and lower scores as follows:

$$\text{Difficulty factor} = \frac{R_u + R_L}{T}$$

Where: R_u = number selecting the correct option in the upper scoring group, R_L = number selecting the correct option in the lower group, T = total of examinees in upper and lower groups.

It is rather confusing because the higher the difficulty factor the easier the question.

Discrimination index (DI). Item analysis programs provide the numbers and proportions of examinees scoring in the top, middle, and bottom thirds (or the upper quartile versus lower quartile) who select each option. The DI is calculated by subtracting the proportion of students who scored correctly in the lower group from the proportion who scored correctly in the upper group:

$$\text{Discrimination index} = \frac{R_u - R_L}{1/2\ T}$$

Where R_u = number selecting the correct option in the upper scoring group, R_L = number selecting the correct option in the lower group, T = total number of examinees in upper and lower groups.

Table 2 - Example of a table of specifications based on the context (for internal medicine examination).

Branches	Health maintenance	Mechanism	Diagnosis	Management	Total
Cardiovascular	3	4	9	9	**25**
Nephrology	1	3	4	4	**12**
Respiratory	2	3	8	9	**22**
Gastrointestinal tract	1	3	5	4	**13**
Rheumatology	3	2	3	5	**13**
Dermatology	1	0	3	2	**6**
Hematology	1	3	3	4	**11**
Infectious	1	2	9	5	**17**
Neurology	1	5	8	7	**21**
Total	**14**	**25**	**52**	**49**	**140**

It is assumed that persons in the upper group on total scores should have a greater proportion of correct items than the lower group. It follows that DI for correct options should be positive, indicating that students answering correctly tend to have higher scores, whereas DI for the wrong options should be negative, which means that students selecting these options tend to have lower scores. This calculation of the index is an approximation of a correlation between the scores on an item and the total score. Therefore, the DI is a measure of how successfully an item discriminates between students of different abilities on the test as a whole. Any item which did not discriminate between the lower and upper group of students would have a DI=0. An item where the lower group performed better than the upper group would have a negative DI. In general, DI's above +0.30 indicate an item that is working well, but 0.20 is not bad.

Example. This is an example to demonstrate these calculations:

If we have a hypothetical class with 10 examinees in the upper quartile (UQ), and 10 in the lower quartile (LQ). Suppose that Q1, Q2, Q3, Q4 are the correct options for questions 1, 2, 3 and 4. If question 1 were correctly answered by all students in the UQ and none in the LQ then the DI is (10-0)/ 10 which is = 1, difficulty factor is 10/20 which is = 50% or 0.5. Question 2 is correctly answered by all students in the UQ and all in the LQ so the discrimination index is 0 because 10-10 = 10. The difficulty factor is 20/20 which is 1 or 100%. Question 3 is answered by 8 students in UQ and 2 students in LQ, so the DI is 8-2/10 = 0.6, the difficulty factor is 8+2/20 = 50%. Question 4 is correctly answered by more students in the LQ than students in the UQ. This is the odd type of question that would have a negative DI for the correct option. The DI is 4-10/10 = -0.6, difficulty is 14/20 = 0.7 The discrimination index is affected by the difficulty of an item, because by definition, if an item is very easy everyone tends to get it right and it does not discriminate. Likewise, if it is very difficult everyone tends to get it wrong. Such items can be important to have in a test because they help define the range of difficulty of concepts assessed. Overall, students who score well on a particular item tend to score well on all items in the test and students who score poorly on a specific item tend to score poorly across all the items.

Certain questions should be asked concerning DI. 1) Are there any items with a negative DI? These are items where students in the lower scoring group did better than students in the upper scoring group. 2) Was this a deceptively easy item? 3) Was the correct answer key used? 4) Are there any items that do not discriminate between the students, namely, where the DI is 0.0 or very close to 0.0? 5) Are these items which are either very hard or very easy and therefore where you could expect a DI of 0?

Information that could be obtained from item analysis. The main information that could be gained from item analysis includes: 1) Identification of questions which are too easy or too difficult. 2) Identification of questions with wrong key. 3) Identification of questions where alternatives are not performing their proper functions. 4) Showing common misconceptions that the examinees have, as indicated by alternatives, chosen. 5) Identification of examinees who are performing poorly and who require further preparation. 6) Identification of any outstanding examinees who could be extended.

Open-ended questions as compared to MCQs. For proper educational assessment, it is highly advisable to use more than one method of evaluation, because the different assessment tools tend to complement each other, whereas a single method of assessment may not be sufficient to adequately assess the candidates' different abilities.[8-10] In written examinations, open-ended questions are frequently used to complement MCQs. Open-ended questions differ from MCQs in certain aspects. These differences need to be considered in the different educational contexts. For example, MCQs would be unsuitable if the aim of the question were to establish a diagnosis. A good essay question asks the candidate to process information or knowledge by, for example, requiring the candidates to set up a reasoning process or summarize information, or asking them to apply a known principle in different contexts, and so forth. If such stimuli were the aim of the test, MCQ types would not be applicable. Although cueing clearly exists in the MCQs, the evidence suggests that it does not influence the nature of the thinking processes elicited by the question and the correlation between MCQs and open-ended questions is very high.[9,11] Fewer open-ended questions could be used in one examination because they require more time to answer than an MCQ.[9] Therefore, open-ended questions are associated with lower reliabilities per hour than MCQs. It has been reported that students prepare differently for MCQ tests than for open-ended tests, but this has no demonstrable effect on their performance.[2] Item writers, however, will be influenced in their selection of topics for a test when only a certain format is allowed for. They will then neglect certain important topics because they cannot be asked about easily. Open-ended questions may seem easier to construct, but good open-ended questions require a detailed answer key, which is time-consuming to produce. These criteria need to be balanced against each other, and the outcome of this may vary according to the specific context of the assessment. Different choices need to be made

for high stakes examinations than for formative evaluations. Open-ended questions should be used solely to test aspects that cannot be tested with MCQs. In all other cases, the loss of reliability and the higher cost represent a significant downside. In such cases, MCQs are not less valid than open-ended questions. So, although the essay question type is expensive and less reliable, it can have a benefit in those cases where the particular stimulus cannot be presented in any other question type. Such is not the case with short answer question (SAQs). The stimulus of most SAQs could also be applied with MCQs. These are not only more reliable per hour of testing time, but are also less expensive to produce and to correct. As stated previously, the cueing effect does not influence the type of competence measured. The use of SAQs should therefore be restricted to those situations in which the spontaneous generation of the answer is an essential aspect of the stimulus.

It could be concluded that MCQs can be used in any form of testing, except when the spontaneous generation of an answer is essential, then the SAQs may be appropriate. If the aim is to test complex thinking, processing of information skills such as reasons, construction, comparison or application of knowledge in different tests, then the essay questions are recommended such as in creativity, hypothesizing, reasoning, problem solving and writing skills.[7]

References

1. Linn RL, Gronlund NE. Measurement and Assessment in Teaching. New Jersey: Prentice-Hall Inc; 2000. p. 1-574.
2. Hart I. Objective clinical examinations. In: Dent JA, Harden RM, editors. Edinburgh, London, New York Philadelphia, St. Louis, Sydney, Toronto: Churchill Livingstone; 2001.
3. McAleer S. Choosing assessment instruments. In: Dent JA, Harden RM, editors. A practical guide for medical teachers. Edinburgh, London, New York Philadelphia, St. Louis, Sydney, Toronto: Churchill Livingstone; 2001.
4. Kehoe, Jared. Writing Multiple-Choice Test Items. ***Practical Assessment, Research & Evaluation*** 1995; 4: 47-51.
5. Frisbie DA. Reliability of Scores from Teacher-Made Tests. The Instructional Topics in Educational Measurement Series. Module 3. Washington (DC): National Council on Measurement in Education; 1988.
6. Gronlund NE. Assessment of Student Achievement. 6th ed. USA: Allyn & Bacon Needham Heights; 1998. p. 1-230.
7. Schuwirth LW, Van der Vleuten CP. ABC of Learning and Teaching in medicine, Written Assessment. ***BMJ*** 2003; 326: 643-645.
8. McAleer S, Hesketh EA. Developing the teaching instinct 10: Assessment. ***Med Teach*** 2003; 25: 585-588.
9. Schuwirth LW, Van der Vleuten CP. Different written assessment methods: what can be said about their strengths and weaknesses? ***Med Edu*** 2004; 38: 974-979.
10. Crossley J, Humphris G, Jolly B. Assessing health professionals. ***Med Edu*** 2002; 36: 800-804.
11. Schwartz PL, Loten EG. Brief problem-solving questions in medical school examinations : Is it necessary for students to explain their answers? ***Med Edu*** 1999; 33: 823-827.