D:\Biomedica Vol. 33, Issue 4, Oct. – Dec., 2017\Editorial.Doc P. 269 – 272 (KC) I

EDITORIAL

# EFFECTIVENESS OF MCQS IN ASSESSING HIGHER ORDER COGNITION

EHSAN S.B.
*Department of Medical Education, Punjab Medical College, Faisalabad*

"If an assessment asks students to evaluate and create but our instruction asks only that they remember and comprehend, then we have taken a wrong direction". (unknown)

Researchers and educators have agreed upon this fact that there is no one method or tool by which we can assess a person's full ability. Assessment is an important part of curriculum and it needs proper planning and implementation because it not only helps in assessing student's ability, it also drives student's learning[1]. Thus planning, construction and implementation of assessment is not an easy task.

In 1956, Dr Benjamin Bloom a psychologist and a team of educational experts presented three major domains of learning. These are cognitive, psychomotor skills and attitude or behavior. These three domains are present in every educational system but in medical education, it is worth imperative. It is stated that educators have to be very careful while designing curriculum, so that they design teaching methodologies and assessments keeping in mind all the learning domains.[2] Cognition is the major domain which gives foundation for skills and behaviors. Cognitive is further explained by Bloom by dividing it into low order thinking and high order thinking process. It is commonly known as Bloom's taxonomy of understanding or knowledge. There are six categories i.e. recall, understanding, application, analysis, evaluation and synthesis. It is considered to be the stages of difficulty levels. One should master one stage before moving to the next. In short, this taxonomy gives us a panorama of range of educational possibilities against which any curriculum can be contrasted or compared.[3]

Written assessments are designed keeping in mind all the outcomes or objectives based on Bloom's taxonomies which help the assessors to assess different categories/levels of knowledge. In Bloom's Taxonomy first two levels are regarded as lower order learning or recall and rest of the four levels up to creation have been called higher order learning levels.[4] These two broad levels of learning can be tested with written assessment. There are two major forms of written assessments, one is constructed response and other one is selected response format. Multiple choice questions, a selected response format, have been used very successfully for last many years. Selected response item were first used in US army in recruitment test in early twentieth century[5].Major reason of implementing MCQs was to overcome shortcomings of constructed response format.

MCQs can test higher knowledge and can be used on a large group. Time required to make and mark MCQs is generally less than the time required to make and check constructed response tests as computer can be used for MCQ scoring. Making effective MCQs is not an easy task and all the teachers particularly medical teachers are given training in constructing good items.

Many students are test- wise and if questions are not properly constructed then there is a little possibility that exam would differentiate between average and above average students. These are mostly item construction flaws which should be properly taken care of while making MCQs. MCQs despite being very popular due to feasibility and testing efficiency are being criticized for its fairness.[6] Van der Vleuten[7] has criticized it for just focusing on recall of factual knowledge and lack of professional authenticity All these responses and feedback is shared by majority of faculty and it is stated that students are getting more scores in exams due to MCQs and students are not professionally competent. Now if we look at some international exams i.e. PLAB or USMLE, then, these are also based on MCQs and students have to work very hard to achieve high scores. So it can be stated that it is not the poor examination system which allows students to get high scores in MCQs, it is actually its construction that has to addressed properly if educators want to use this tool in a best way.

## Effectiveness of an MCQ Exam

Fairness of an MCQ exam is mainly dependent on proper standard setting, psychometric adequacy, consequential validity and proper construction[8].As far as standard setting is concerned, it is an arbitrary process and it has always been open to criticism, remained controversial, difficult in execution and almost impossible to defend.[9] In examination, standard setting or pass fail decisions are usually based on criterion referenced as compared to norm referenced or holistic method (arbitrary pass marks say, 60%).[10] Norm referenced and holistic methods are commonly used in high

Corresponding Author: Dr. Sumera Badar Ehsan, Assistant Professor
Department of Medical Education, Punjab Medical College, Faisalabad
Email: sumeraehsan@yahoo.co.uk

stakes clinical exams, yet are least defensible.[3] There is no single recommended method of standard setting. In Case of MCQs, Case and Swanson has suggested Modified Angoff's method and Hofstee method. Angoff's method is simple but not widely used. Hofstee method is easy in use and reliable method.[3]

Properly structured, valid and reliable multiple choice questions are vital if educators want a fair and valid examination. A lot of attention is paid to this particular area of research regarding format, design and construction of MCQs.[11] To attain proper psychometric measures it is recommended that each item must measure/test one specific content or mental behavior and all the questions be independent when designing a series of MCQs so that students don't get any cue.[11,12]

**Reliability**

Reliability is the degree to which an instrument produces the same results on repeated testing.[13-15] MCQs are considered to have a high degree of reliability of objective scoring method.[11] Main concepts related with reliability are precision, consistency, stability, equivalence and internal consistency.[13] Reliability is measured by reliability coefficients or co-relation coefficients and in case of MCQs it should be positive or strong (usually more than 7).[16] Reliability and validity are usually tested in pilot studies and conditions are kept close to the actual exam. Students or participants who contribute in pilot study should be representative of the target population in terms of their age, ability and educational level. Stability of a single MCQ test is based on test retest correlation. Time interval between test and retest has a major effect on scores and stability of test. Ideal time interval between test and retest is a debatable issue.[15] Issues related with test retest can be minimized by using two alternative forms of MCQs. This equivalence will measure weather two sets are measuring the same attributes and these MCQs are usually administered in succession and in random order.[14]

Another concept related with reliability is internal consistency i.e. reliability based on average correlation among items within a test and examines the degree to which MCQs measure the same domain of knowledge.[13,14] Internal consistency is routinely measured by calculating reliability coefficient.[13,14,17] There are different statistical ways to calculate internal consistency but commonly used method is 'coefficient alpha.[17] However in case of MCQs many research reports refer to the Kuder-Richardson coefficient (KR-20). Kuder-Richardson (KR-20) is a specific form of the coefficient alpha and is usually used in dichotomous data.[13,14] In MCQs dichotomous can be described in the form of correct and incorrect options. This coefficient ranges from zero to 1. Closer to 1 indicates high reliability. Usually a value of 0.7 or more is considered significant. In case of MCQs if the alternative form correlation is considerably lower than the reliability coefficient (0.2 or greater) then it is indicated to change the variation and content of exam in future.[18]

**Validity**

Validity is generally defined as the degree to which an instrument measures what it is supposed to measure[19]. It has been explained as "To validate a proposed interpretation or use of test scores is to evaluate the claims being based on the test scores. The specific mix of evidence needed for validation depends on the inferences being drawn and the assumptions being made" **(p. 131)**.[20] So it is evident that validity is an important concept in assessment and in constructing MCQs, one has to take a special care about validity. It helps in interpretation and gives legitimate meanings to data. No assessment can be said valid or invalid, it is actually the scores of that assessment that validity is linked with. In case of MCQs validity may be affected by many factors that may increase or decrease its difficulty. These factors could be vague statements, grammar mistakes, poor instructions, improperly constructed MCQs and in appropriate distracters.[21] Validity is dependent of different elements and these include construct, face and content validity. Each element will ultimately contribute to the overall validity.

Construct validity is an important concept and was explained as a hypothesized attributes to be tested by an assessment.[22] In case of MCQs, it will be explained as whether or not it measures the domain of knowledge that is being examined. Construct validity is routinely established by using a "Key Check" and item analysis that includes difficulty index, discrimination index and distracters evaluation.[11,23,24]

Key check confirms whether the answer is correct. It helps in removing any confusion between right answer and distracters and it is usually done by a team of experts and if any item has any ambiguity in key then it should be revisited. Item discrimination is an index that helps in discriminating scores of high achievers and low achievers for an MCQ item. In simple words it is stated that it explains, how an MCQ is linked with overall performance.[18] This can be measured easily. However each item response in comparison of total performance is statistically measured. Nunnally and Bernstein[18] have recommend this method to examine the item discrimination in which relationship of one item with the total test scores is analysized. Point biserial correlation (rpb) use the Pearson correlation coefficient (r) to measure the correlation between two variables. In MCQs we will take a dichotomous ratio i.e. overall test score and correct or incorrect response. The point biserial correlation coefficient (rpb) will measure the relationship of correct or incorrect answer with overall score and it ranges from 0 to 0.4. An uncorrected item to total score of 0.25 or greater is consi-

dered acceptable.[13] MCQs having high positive discrimination values are considered good. Zero value indicates that equal number of students from high achievers and low achievers select the right or wrong answer. Negative value shows that students getting high scores have selected wrong options. MCQs with zero and negative values should be revised, corrected or discarded after experts' opinion. MCQ with item to total score correlation greater than 0.7 indicates that it has so much similarity or overlapping, and it can be removed to reduce similarities.[13]

Similarly distracters evaluation also plays an important role in overall validity. A good distracter must be chosen by those who have not prepared well for the test and must be ignored by all those students who perform well. If some distracter is selected by majority of the students instead of correct answer then that distracter must be revisited, or changed. In addition to all above mentioned elements, consequential validity is also important. It is assessment that drives learning and a fair exam will certainly motivate students learning.

**Construction of MCQs**

Any kind of cognitive knowledge learned can be tested by written assessment. It is stated that any aspect of cognitive achievement can be best tested either by multiple choice or true false form[1].In health professions education all kind of learning can be tested through written assessments.

Writing good MCQs is not an easy task and it is considered one of the limitations of selected response examination. Writers of effective MCQs are trained not born[25].It is an art and science to make effective MCQs. Art is linked with writing skills, editing, reviewing, training in writing MCQs, practice and feedback skill etc. while science is associated with structuring and constructing items based on evidence based principles of writing stems, including content, format and options. It is an open fact that any health care expert cannot be a best writer of an effective MCQ until he/she gets a proper training in it. This world is drenched in poorly written MCQs.[25] So writing effective, creative and fair MCQs which could test higher levels of cognition is really a big challenge for busy faculty. This is a main reason for poorly constructed MCQs routinely used in class tests in health education.

Now, if we look at Bloom's Taxonomy then a question is always asked by faculty that at what level of taxonomy MCQs can be constructed. MCQs has always been criticized for testing only superficial knowledge and regurgitation of facts. Some people say that MCQs are not very useful for courses which use problem based learning and self directed approaches. Controversy exists in this aspect that MCQs can only test knowledge, comprehension, application and analysis levels of Bloom's taxonomy, while synthesis and evaluation levels cannot be assessed with MCQs.[24] In contrast to this, " we find that application of knowledge, synthesis and judgment questions can better be assessed by one best answer questions"[1] (p.18). It has been endorsed in another study that MCQs are helpful in enquiry based learning and helped students to be self directed. It is obvious that construction of MCQs is really a challenging task for educators if higher levels of cognition are to be tested.[26]

There is no definite answer to this question, however, it is all mental effort and experience in constructing MCQs that will help the educators to make good questions of higher cognition keeping in mind all the mechanics and psychometrics of an effective exam.

**REFERENCES**


1. Case SM, Swanson DB. 3rd ed. Constructing written test questions for the basic and clinical sciences. Philadelphia: NBME. 1999.
2. Kern, Thomas, Hughes. Curriculum development for medical education A six step approach, second edition, 1998.
3. Krathwohl DR. A revision of bloom's taxonomy, an overview, 2002.
4. Anderson, L.W., & Krathwohl, D. R. (Eds.). A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives. New York: Longman, 2001.
5. Ebel, R. L. Essentials of educational measurement. Englewood Cliffs, NJ: Prentice Hall, 1972.
6. Kaufman, D.M. Assessing medical students: hit or miss, Student BMJ. 2001; 9: pp. 87–88.
7. Van Der Vleuten, C. The assessment of professional competence: developments, research and practical implications, Advances in Health Sciences Education, 1996; 1: pp. 41–67.
8. Paul McCoubrie. Improving the fairness of multiple-choice questions:a literature review, Medical Teacher, 2004; 26 (8): 709-712.
9. Berk R.A.A consumers guide to setting performance standards on criterion referenced tests. Review of educational research, 1986; 56: pp. 137-172.
10. Case, S.M. & Swanson, D.B. Constructing Written Test Questions for the Basic and Clinical Sciences, 3rd ed. (Philadelphia, National Board of Medical Examiners), 2001.
11. Haladyna TM. Developing and validating multiple-choice test items. Lawrence Er1baum, 1991. New Jersey.
12. Haladyna TM. Developing and validating multiple-choice test items. lawrence Er1baum, 1994, New Jersey.
13. Beanland C, Schneider Z. etal. Nursing research: methods, critical appraisal and utilization. Mosby, 1999, Sydney.
14. Polit OF, Hungler BP. Nursing research: principles and methods. Lippincott Williams & Wilkins, 1999, Philadelphia.
15. Considine J., Botti M. & Thomas S. Design, format, validity and reliability of multiple choice questions for use in nursing research and education. Collegian, 2005; Vol. 12 (1).
16. Gravetter F., Wallnau L. Statistics for the behavioral

sciences. Wadsworth/Thomson Learning, 2000, Stanford.
17. Cronbach L. Essentials ofpsychological testing. Harper Collins, 2990, New York.
18. Nunnally J. Bernstein I. Psychometric theory. McGraw Hill, 1994, New York.
19. Downing SM. & Yudkowsky R.. Assessment in health professions education. 1st ed. 2009, Tylor& Francis e-library.
20. Kane, M. Content-related validity evidence in test development. In S.M. Downing & T.M. Haladyna (Eds.), Handbook of test development 2006 (pp. 131–153). Mahwah, NJ: Lawrence Erlbaum Associates.
21. Linn R, Gronlund N. Measurement and assessment in teaching. Prentice Hall, 2000, New Jersey.
22. Cronbach, L.J., & Meehl, P.E. Construct validity in psychological tests. Psychological Bulletin, 1955; 52: p.281–302.
23. Violato C. Item difficulty and discrimination as a function of stem completeness. Psychological Reports; 1991; 69 (3 P11): 739-743.
24. Masters, J., Hulsmeyer, B., Pike, M., Leichty, K., Miller, M. Verst, A. Assessment of multiple-choice questions in selected test banks accompanying test books used in nursing education. Journal of Nursing Education; 2001; 40 (1): 25–32.
25. Downing, S.M. Threats to the validity of locally developed multiple choice tests in medical education: Construct-irrelevant variance and construct underrepresentation. Advances in Health Sciences Education, 2002; 7: 235–241.
26. Honey, M., Marshall, D. The impact of on-line multiple choice questions on undergraduate student nurses' learning. Proceedings of the 20th Annual Conference of the Australasian Society for Computers in Learning in Tertiary Education; 2003, vol. 1. Ascilite, Adelaide, pp. 236–243.