Volume-3, Issue-1, January-2019: 1-12
International Journal of Recent Innovations in Academic Research ISSN: 2635-3040

Utility of Single-Best-Answer-Questions as a Summative Assessment Tool in Medical Education: A Review

Eze Onyegbule Okubuiro (FWACS, FHEA (UK))[a*], Longinus Ndubuisi Ebirim (FWACS)[b], and Chinwe Edith Okoli (FMCA, FWACS)[c]

[a]Centre for Evidence-Based Medicine, University of Oxford, United Kingdom
[b]Department of Anaesthesiology, University of Port Harcourt, Rivers State, Nigeria
[c]Department of Anaesthesia, Federal Medical Centre, Umuahia, Abia State
* Correspondence: ezeokb@gmail.com

Received December 22, 2018; Accepted December 29, 2018; Published January 2, 2019.

Abstract: Achievement of learning may be measured by written, oral or performance-based assessments. Measurement of competence by written assessment is reliable, cost-effective and proffer logistic advantages hence its universal acceptability in education. Single best answer (SBA) method of assessment in medical education provides a quantitative and objective evaluation of clinical knowledge, thinking and comprehension. Although marking of SBAs is relatively easy, setting of SBAs is usually challenging and requires meticulous planning and extensive human and material resources. This review examines single best answer questions (SBAQ) as a tool of summative assessment in medical education using the framework of utility as defined by Van Der Vleuten (1996).
Keywords: Utility, single-best-answer questions, summative, assessment, medical-education.

Citation: Eze Onyegbule Okubuiro, Longinus Ndubuisi Ebirim and Chinwe Edith Okoli. 2019. Utility of Single-Best-Answer-Questions as a Summative Assessment Tool in Medical Education: A Review. International Journal of Recent Innovations in Academic Research, 3(1): 1-12.

Introduction

Assessment refers to the process and instruments applied to measure a learner's achievements (Mohanna *et al.*, 2011). Such achievements may be tested by written, oral or performance-based assessments or their combinations. Written assessments are well established and widely used to measure competence in all spheres of education (Harden *et al.*, 2012, Schuwirth and Van Der Vleuten, 2013). Their popularity is derived from logistical advantages, high reliability and cost-effectiveness (Schuwirth and Van Der Vleuten, 2013). This review aims to critically appraise the Single Best Answer (SBA) format of multiple-choice questions (MCQ) as a summative assessment tool in medical education.

Background

Since introduction of Medicine as a University subject in the middle ages, competence has remained an 'explicit virtue' (Jackson *et al.*, 2007). However, concerns of partiality and

subjectivity of assessment methods prevailed until 1792 when William Farish introduced a quantitative and objective method of evaluation of students' knowledge (Hogan, 1979). This objective assessment focused on specific facts in the evidence instead of theological and philosophical speculations. In early 20th century, Edward Thorndike developed prototypes of MCQs (Goodenough, 1950). Later in 1914 Frederick Kelly first used MCQs to assess knowledge, critical thinking and comprehension in college education. Following invention of scanners and computers, it has been applied to large population of students with minimal human contact thus making it a favourite method of assessment in education till date (Leman, 2000). MCQs are viewed as a panacea to subjectivity and favouritism in educational assessment (Mabry, 2004). Multiple-choice questions can sample objectively a wide range of a learner's knowledge and understanding (Harden *et al.*, 2012). Whereas marking is relatively easy, setting of MCQs may be challenging, requiring robust planning and extensive human and material resources to ensure validity/reliability. Single best answer questions (SBAQs), multiple true/false questions (MTFQs) and extended matching questions (EMQs) are the most popular forms of MCQs (Brian, 2014). EMQ provides superior construct validity, minimises cueing and enables higher level of assessment of reasoning and application of knowledge using the Blooms taxonomy (Bloom *et al.*, 1956) compared to SBAQ (Brian, 2014). However, it has limited application in certain branches of Medicine (for example epidemiology) and, many options may be redundant (Brian, 2014).

MTFQs require options to be absolutely true or false. This adversely affects its inter-rater and test-retest reliability therefore suggestive of poor validity. Other intrinsic drawbacks include difficulty in formulating questions, cueing effect and guessing. Also, requirement of complicated marking scheme (including negative marking) further diminished its attraction (Harden *et al.*, 2012). Also, this format suffered further rejection on account of its negative educational impact as it encouraged students' studying of the examination itself rather than the subject (Gunderman, 2001). SBAQ is more salient to most areas of Medicine (Brian 2014). Its efficient sampling of knowledge confers a better reliability and validity compared to multiple true/false questions (Tan and McAleer, 2008). It is widely used as a summative assessment tool in undergraduate medical education (Mohanna *et al.*, 2011). SBAQ is cost efficient and can reflect basic science/clinical decisions and hypothesis generation relevant to undergraduate medical studies. A standard SBAQ assesses a range of skills like interpretation, synthesis and application of knowledge rather than mere recall of facts (McCoubre and McKnight, 2008). However, it can only assess student's ability up to level 2 (Knows how) of the Miller's pyramid (Miller, 1990). SBAQ is further limited by inability to assess affective and psychomotor domains of learning (Bloom, 1956). To enhance its utility as a summative tool, SBAQ should be complemented by practical assessments as the former emphasizes learning from written sources (McCoubrie, 2004).

Figure 1. Bloom's Taxonomy (1956)

This review will examine SBAQ using the framework of utility as defined by Van Der Vleuten (1996). This concept describes utility as a product of Reliability (R), Validity (V), Educational impact (E), Acceptability (A) and Cost effectiveness (C). Whereas the individual components have differential weights, neglect of any can adversely affect overall utility.

Utility Index of SBAQ

Reliability

Reliability is defined as reproducibility of results (Schuwirth and Van Der Vleuten, 2014). It demonstrates ability to reproduce differentiation between candidates when utilized in comparable settings. Such differentiation of students' competence is important to ensure progression, patients' safety and achievement of required standards. A tool that ensures that such differentiation is not due to errors is required to discriminate students with different levels of competence (Norman and Eva, 2010). Results from different assessment tools may have some degree of systematic or random error and, this is reflected in classical test theory definition of reliability coefficient as ratio of true and observed variance. The value of reliability coefficient ranges from 0 to 1 and, values ≥0.8 are acceptable for high stake examinations (Gravetter *et al.*, 2000; Norman and Eva, 2010).

Haladyna (1994) attributes high degree of reliability of MCQs to their intrinsic objective scoring process. Veloski *et al.*, (1993) conducted a study on 34 medical students; comparing assessment of their abilities using SBAQ versus un-cued MCQ (28 questions each). They reported Kuder-Richardson 20 (KR-20) reliability coefficient of 0.74 and 0.69 for un-cued MCQ and SBAQ, respectively. In that study tedious clerical work in searching for answers among 564 possible options and need for numerous coding using the un-cued format made it less preferable to students. Sampling size significantly affects reliability therefore, the small number of questions used could have adversely affected (under-estimated) reliability given the low KR-20 obtained with SBAQ format in that study. McCoubrie *et al.*, (2008) reported highly reliable examination (reliability coefficient >0.8) when large samples of SBAQ were conducted over 4 hours.

In a retrospective analysis of 320 SBA and 46 OSCE questions given to 823 sixth year medical students in 2012 and 2013, Fallatah *et al.*, (2015) noticed a reliability coefficient (Cronbach Alpha) of 0.82 and 0.93 for SBAQ and OSCE respectively. Although these values are acceptable for such high-stake examinations, KR-20 would have been more appropriate for SBAQ as, Cronbach Alpha used in analysis of both assessment tools might have exaggerated the values of reliability coefficients obtained. A comparative study examined the psychometric characteristics of MCQs and Patient management Questions (PMPs) over 3 consecutive years of the American Board of Internal Medicine certifying examination (Norcini and Ben-David, 1985).

It revealed a reliability coefficient of 0.74, 0.82 and 0.80 taken over a period of 2.7, 2.8 and 2.8 hours, respectively using the SBAQ format. The values for composite MCQs (SBAQ + Multiple true/false + EMQ) in the same study were 0.92, 0.91 and 0.92 when examinations were taken over 7.4, 7.1 and 6.3 hours respectively. When all MCQ formats were corrected to 4-hour test time they all showed reliability coefficient >0.8 over the consecutive 3-years period. Whereas, reliability coefficient of patient management questions over 4-hour period involving, ≤12 items each were <0.7 over the same period. This confirms that reliability of item format is influenced by number of items/ testing time devoted to each (McCoubrie *et al.*, 2008). Examiner training is vital for improved reliability as it ensures standardization of the processes, unfortunately the studies above did not report the effect of such on reliability of

SBAQs. Authors have noted that examiner training directly affects reliability of SBAQs (McCoubrie, 2004; Considine *et al.*, 2005). Also, test-retest reliability was not assessed in any of the studies. Such omission may suggest a potential for reduced reliability.

Validity

This examines the degree to which a test measures what it claims to measure (Norman and Eva, 2010). The different types are; Face, content, concurrent, predictive and construct validity.

Face Validity

It refers to the subjective view of the trainer and trainee regarding the assessment tool. Arguably, this lack of objective assessment limits its scientific value, as it may be prone to bias. However, some authors believe it is a measure of credibility as it reflects perception of students, examiners and experts on test items (Tweed and Cookson, 2001). An online survey analysed responses of 1000 Dental examiners (including 890 course directors) in the United States of America (Albino *et al.*, 2008).

Variability coefficient of 0.84 was obtained in a test-retest analysis of the pilot study of that questionnaire. In that study, forty-five per cent of responders preferred SBAQ for knowledge-based assessment and MCQs were the commonest method of assessment of students' knowledge (28%) in practice. However, there was consensus opinion that MCQs alone were insufficient to determine a students' competence/ fitness to practice. A similar questionnaire-based study involving 340 undergraduate medical students reported that 60% of respondents believed that MCQs adequately evaluated their knowledge base (Chandrasekhar *et al.*, 2010). Although well-designed SBAQs could assess taxonomically higher order processes like interpretation, synthesis and application of knowledge, it does not replace practice-based assessment tools that assess clinical skills, attitudes and communication skill necessary for successful clinical practice (General Medical Council, 2013).

Content Validity

This determines whether the SBAQs are relevant, appropriate and representative of construct being assessed and/ or the cognitive processes that they intend to test (Considine *et al.*, 2005). Content review by experts in the relevant domains is used to ensure content validity of SBAQs (Haladyna, 1999). Content validity relies on personal judgement as; there are no absolutely objective means of establishing it (Considine *et al.*, 2015). As examination, it must be optimally representative of the whole testable domain, blue printing is one of the elements required to demonstrate content validity of SBAQs (Schuwirth and Van Der Vleuten, 2014).

Blue printing of SBAQs is based on pre-set learning outcomes, which is in concordance with a prevailing curriculum. It is against this matrix that an examiner determines how many items per topic or category to be assessed. Relevance of the items is another important element that directly correlates with content validity of an examination (Schuwirth and Van Der Vleuten, 2014).

In the study by Fallatah *et al.*, (2015) (earlier mentioned), four experts (Consultants) with various levels of experience established the content validity of the MCQs and OSCE. Similarly, Velsoki *et al.*, (1993) used a panel of three experts to establish content validity of the two MCQ formats studied. The number and quality of panellists enhanced the credibility of that study as, it achieved the recommended minimum of 3 experts required to establish content validity (Polit and Hungler, 1999).

Concurrent Validity

The change from MTFQ to SBAQ across various spheres of medical education was sequential. Most studies compared validity of these two MCQ formats in different sets of students at different points in time (Tan *et al.,* 2008). There is limited data in literature to directly compare validity of the two formats when administered concurrently to similar distribution of students.

Predictive Validity

Tan *et al.,* (2008) compared the predictive values of MTFQ and SBAQ over 4 series of Final FRCR examination sessions involving 2 previous and 2 subsequent MTFQ and SBAQ segments of the Final FRCR examination; against the overall result in each of these examination periods. Significantly higher predictive values were observed with SBAQ (80%) compared to MTFQ (46%), $P<0.001$. This result suggest that SBAQs had better predictive validity in differentiating knowledgeable from unknowledgeable candidates in that examination. High risk of guessing, mere recall of facts without indebt knowledge, effects of cueing and studying the examination instead of studying for the examination might be responsible for the poor predictive validity of the MTFQ. The study by Fallatah *et al.,* (2015) found good correlation ($r= 0.824$) between students' scores in the SBAQ and the overall result of the final examination in internal Medicine. These reports on high predictive validity of SBAQ may be attributed to its ability to assess and differentiate high order cognitive processes of students. As knowledge drives performance (McCoubrie, 2004), these values suggest that students with higher order reasoning and knowledge base were likely to equally excel in other areas of clinical performance.

Construct Validity

This defines the extent to which an instrument measures a theoretical attribute (Considine *et al.,* 2005). Construct validity of SBAQs examines whether the questions measure the desired domain of knowledge. Construct validity of SBAQs may be established using key check, item discrimination analysis and distractor evaluation (Haladyna, 1999). 'Key check': This ensures that correct answer to a SBAQ is actually correct while confirming that no other item is also correct (Considine *et al.,* 2005). Panel of experts in a particular field perform the key check. And, where there is perceived variation, the SBAQ should be reviewed until consensus is achieved (Haladyna, 1994). Most studies adopt this recommendation in establishing content validity (Venoski *et al.,* 1993; Haladyna, 1999; Palmer and Devitt, 2007).

Item Discrimination Analysis: This examines how each MCQ correlates with overall test performance (Nunnaly and Bernstein, 1994). The underlying principle of this concept suggests that if a question is highly discriminative, the overall test score of those with correct answers will be higher than the overall scores of those with incorrect answers in MCQs (Haladyna, 1999; Masters *et al.,* 2001). This concept is supported by decades of research, which showed that knowledge of a domain is the single best determinant of expertise (Glaser, 1984). SBAQ is a valid tool of competence testing since cognitive knowledge is better assessed using written forms (Downing, 2002).

Pearson correlation coefficient is used to measure relationship between the two variables and values ≥ 0.25 is considered acceptable (Beanland *et al.,* 1999). In the study by Fallatah *et al.,* (2014), Pearson correlation coefficient for the OSCE and SBAQs scores is 0.28. This suggests good construct validity of SBAQs in that study as; OSCE's are also validated for assessment of higher order cognitive processes (Fallatah *et al.,* 2014).

Distractor Evaluation: In Clinical education importance of valid assessment scores is paramount as learner's competence has direct consequence on patient care. A Content validity of SBAQ is enhanced by use of 'functional distractors' (Ali *et al.*, 2016). In SBAQs, a good distractor should be identifiable by a knowledgeable candidate and, should not be discernable by poorly performing students (Linn *et al.*, 2000). Distractors that are not chosen by <5% of respondents are non-functional and those consistently chosen more than the correct answer suggest misleading questions or poor instructions; both should be replaced (Nunnally *et al.*, 1994). Terrant *et al.*, (2010) reported a minimal decrease in mean item difficulty (0.3%) following elimination of a non-functional distractor from 4-or5-option MCQs.

Educational Impact

This describes the effect of an assessment tool on the learning process. Assessment drives learning as, students 'do what you assess' and not 'what you expect' (Schurwirth *et al.*, 2014). Earlier form of MCQ like MTFQ was associated with a negative educational impact because students had a chance of passing by guessing, cueing or/and studying the examination questions only, despite a poor knowledge of the subject (McCoubrie *et al.*, 2008).The wide coverage of learning outcomes covered in SBAQs and higher order cognitive process required to answer the items requires students to conduct in-depth study of the subject in order to succeed in the examination. This provides an incentive for deep learning and utilization of educational resources, which improves the educational impact of SBAQs. Positive correlation of successes in SBAQs with overall summative assessment for progression or qualification motivates students to devote their resources to content of their training assessed in SBAQs (Kim *et al.*, 2012). Unfortunately, educational impact of assessment tools is difficult to explore directly (Schuwirth and Van Der Vleuten, 2014).

Acceptability

Generally, acceptability of SBAQs in undergraduate education is favourably reported (Palmer *et al.*, 2007; Shankar *et al.*, 2010; Brian 2014; Schurwirth *et al.*, 2014). Close scrutiny of results show that majority of students who passed SBAQS also recorded overall pass in their overall summative assessment (Norcini and Ben-David, 1985; Tan *et al.*, 2008; Fallatah *et al.*, 2014). Other stakeholders such as programme directors favourably consider SBAQ as relevant tool for assessment in the cognitive domain (Albino *et al.*, 2008). Many authors accept SBAQs as valid tool for evaluation of factual recall, comprehension, analysis and application of knowledge (Bloom *et al.*, 1956; Considine *et al.*, 2005; Brian, 2014). However, SBAQ is not well received as a sole test of competence by examinees, examiners and regulatory authorities (Albino *et al.*, 2008; Harden *et al.*, 2012; General Medical Council, 2013, Jackson *et al.*, 2013).

Cost Effectiveness

Use of SBAQ format is highly cost effective (Schuwirth and Van Der Vleuten, 2014). They are more difficult to produce but use of optical scanners makes them easier to score. Production of SBAQs involves a rigorous process that consumes significant human and material resources. However, the expense is incurred only once as questions generated can be kept in a confidential "bank" and re-used over time or shared among comparable institutions in a region (Schuwirth and Van Der Vleuten, 2014). SBAQ can be administered to very large population of students over a wide region in unit time (McCourie, 2004). Automation of the scoring process reduces human and material resources required to process SBAQ. This eliminates favouritism and assessor bias and, makes SBAQ a ubiquitous tool for assessment of students' knowledge globally (Haladyna, 1999).

Poor item banking, centralised management, administrative support for logistics and administration of SBAQs are identified as major setbacks to its cost efficiency (Schuwirth and Van Der Vleuten, 2014). Institutions sharing similar curriculum and educational goals may collaborate in production and sharing of SBAQs, this will further improve its cost effectiveness.

Reflective Analysis on Single-Best-Answer-Qestions in Summative Assessment of Final Year Medical Students of the University Of Port Harcourt

High stake (qualifying) examinations must exhibit high levels of reliability and validity (Case *et al.*, 2003). Reports from literature demonstrate high validity, reliability, cost efficiency and educational impact as evidence of the utility of SBAQs in Clinical education (Considine *et al.*, 2005; Mohanna *et al.*, 2011; Schuwirth and Van Der Vleuten, 2014). Final year undergraduate medical school examination is a qualifying examination. It assesses candidates' competence and fitness to practice medicine: a pre-requisite for licensure. It is seen as a means of checking that candidates have learnt the basics as it relates to knowledge, skills and attitudes in a wide range of medical disciplines to a pre-determined level of proficiency and that, the patients feel protected by the transparency of the process (Hawthorne, 2007). Results are final/irreversible and have lasting consequences (Linn *et al.*, 2000). Therefore, any summative tool employed in this process must possess high level of validity and reliability to protect its credibility/acceptability (Considine *et al.*, 2005).

SBAQs are the most objective method of assessment (Norcini *et al.*, 2013). SBAQs are created via a rigorous process involving a panel of experts. This group ensures that items reflect the right context, assesses the right breadth and depth of students' knowledge and differentiates abilities of a cohort of students (Brian 2014). This process of standard setting demands good knowledge of educational processes and in-depth assessor training (Considine *et al.*, 2005). SBAQs as summative assessment tool in final undergraduate medical examination is criterion-referenced and success achieved following a candidate's demonstration of a minimum level of competence in the relevant domains of learning (General Medical Council, 2013; Norcini *et al.*, 2013). Norm-referenced SBAQs are used in ranking of students; this method is more relevant in selection examination where many qualified candidates compete for limited opportunities as seen during Medical specialty selection process (Mohanna *et al.*, 2011). This review will now address setting, conduct, and marking/scoring of SBAQs in final year undergraduate medical examination and feedback to students.

Setting of SBAQs

The purpose of any assessment is to permit inferences to be drawn with regard to candidates' competence (Case *et al.*, 2001). Relevant tools are needed to determine students' competence in the various domains of learning. It is established that SBAQs are most suited for assessment of students' knowledge (Mohanna *et al.*, 2011). Ensuring SBAQs assess the right knowledge base of students is best achieved through the instrument of a blueprint based on existing curriculum (Hawthorne, 2007). Blueprinting ensures that the SBAQs are mapped carefully against learning objectives to produce a 'valid examination' (Hamdy, 2006). About 4 hours of test time using SBAQs is recommended to achieve a reliability coefficient >0.80 (McCoubrie *et al.*, 2008). This period may be divided into two sessions. However, for final year medical students' examination of this University, SBAQs are limited to 2 hours of 150 questions as, other forms of written assessments such as short answer questions and essay questions are also administered to achieve a composite written examination. It is thus difficult to formulate SBAQs to cover all the topics in Anaesthesia for the cohort of students in the

examination. The panel of experts therefore select the most relevant and critical Anaesthesia topics to be assessed. Arguably, this reduces the high reliability associated with SBAQs (McCoubrie, 2004). However, such combinations of assessment tools are thought to enhance the validity of the written examination (Schuwirth and Van Der Vleuten, 2013). Also, relevant stakeholders had found such combinations acceptable in undergraduate medical education (Tweed and Cookson, 2001).

Conventional format of SBAQs has three parts: the stem, the correct answer (key) and several incorrect but plausible answers (distractors) (Nunnaly and Bernstein, 1994; Haladyna, 1999). All distractors must be plausible to the uninformed and homogenous with the correct answer. These concepts are adopted in setting the SBAQs for the examination. While there is no agreement in literature on the optimal number of options required in SBAQs, there is a consensus that a minimum of 3 is needed to ensure validity (McCoubrie *et al.,* 2008; Viyas *et al.*, 2008; Tarrant and Ware, 2010). Five options model is favoured by most authors as, number of options greater than seven increases the risk of redundancies (Considine *et al.*, 2005; McCoubre *et al.,* 2008; Kim *et al.,* 2012). The five-option model was used in this examination.

Conduct of SBAQs

SBAQs may be conducted as paper-based or computer-based assessments depending on availability (Considine *et al.,* 2005; McCoubrie *et al.,* 2008,). Paper based format is the most common format used for administration of SBAQs (Tan *et al.,* 2008). The ubiquity of this format is underpinned by its cost efficiency as it can be administered to very large population of students within a very short time (McCoubrie *et al.,* 2008; Vyas *et al.,* 2008). Computer based formats are technology driven and may offer better security, reduced chances of cheating/collaboration by candidates and provide faster and more efficient processing of students results with immediate feedback (Brian, 2014). This suggests higher reliability than paper-based format. However, such technology is not readily available in some resource-limited settings.

Marking/Scoring of SBAQs

SBAQs are marked using computer-based scanning equipment (Brian, 2014). Such minimal human contact reduces chance of bias in marking/ scoring of candidates' scripts and enhances the credibility of this assessment tool. As a criterion-referenced assessment, the standard setting process to establish pass/ fail standard was based on the Angoff method (Ben-David, 2000). This decision was made at the examiners panel after the examination. Each examiner's opinion is informed by their perception of performance of borderline students among a given set of examinees. Five consultants formed the panel of examiners in this SBAQs assessment. This group of experts discussed what constituted adequate or inadequate knowledge. However, the average of individually allocated pass marks was chosen as the pass mark for the SBAQs assessment. Another method that could have been adopted to set the pass mark for the SBAQs is the Ebel method (Ben-David, 2000). However, time constraint restricted use of only Angoff method. With hindsight, the Ebel method could have been used for quality assurance in the process but with added incentives to the examiners. These SBAQs model of MCQs constituted 50% of the written examination.

Student Feedback

One of the drawbacks of SBAQs is the limitation in provision of feedback to students. (Brian 2014). It is very difficult to identify individual items that students performed poorly so as to provide focused feedback. Therefore, feedbacks given to students after SBAQs tend to be

non-specific. This could impair the use of SBAQs as formative assessment tool in medical education (Wood, 2014).

Conclusion

The utility of SBAQs is derived from its high reliability, validity, cost-efficiency and feasibility. Evidence in literature support its use for assessment of knowledge relevant to clinical practice such as synthesis and application of knowledge, data interpretation, problem solving and decision making (Case, 2001). Whereas the utility of SBAQs as valid tool of assessment of knowledge is unquestionable, various stakeholders consider it inadequate as a sole tool for assessment of clinical competence (Tweed and Cookson, 2001). Although it has been suggested that knowledge drives practice (Glazer, 1984), evidence abound that knowledge does not guarantee competence as, professional competence integrates knowledge, skills, attitudes and communication skills. It becomes imperative to incorporate other performance-based assessment tools in qualifying medical examinations in order to provide a comprehensive evaluation of competence of graduating medical students to practice medicine.

Sponsors: None

Conflict of interest: None

References

1. Ali, S.H., Carr, P.A. and Ruit, K.G. 2016. Validity and Reliability of Scores Obtained on Multiple-Choice Questions: Why Functioning Distractors Matter. Journal of the Scholarship of Teaching and Learning, 16(1): 1-14.

2. Beanland, C., Schneider, Z., Lo-Biondo-Wood, G. and Haber, J. 1999. Nursing Research: Methods, Critical Appraisal and Utilisation. Sydney: Mosbey.

3. Ben-David, M.F. 2000. AMEE Guide No. 18: Standard setting in student assessment. Medical Teacher, 22(2): 120-130.

4. Bloom, B.S. 1956. Taxonomy of Educational Objectives. Handbook I: The Cognitive domain, 2nd Edition, New York: David McKay Co Inc.

5. Brian J. 2014. Written Assessment. In: Swanwick, T., (Ed.), Understanding Medical Education: Evidence, Theory and Practice. 2nd Edition, Chichester, Wiley Blackwell.

6. Case, S.M., Holtzman, K. and Ripkey, D.R. 2001. Developing an item pool for CBT: A practical comparison of three models of item writing. Academic Medicine, 76(10): S111-S113.

7. Chandrasekhar, T.S., Subish, P., Mohan, L., Upadhyay, D.K. and Mishra, P. 2010. The Views of Medical Students about the Purpose and Objectivity of Assessment in a Medical College in Western Nepal. Journal of Clinical and Diagnostic Research, 4: 2271-2278.

8. Considine, J., Botti, M. and Thomas, S. 2005. Design, format, validity and reliability of multiple choice questions for use in nursing research and education. Collegian, 12(1): 19-24.

9. Downing, S.M. 2002. Assessment of knowledge with written test forms. In: Norman, G., Van Der Vleuten, C. and Newble, D. (Eds.), International Handbook of research in medical education, Springer, Dordrecht. 647-672 pp.

10. Fallatah, H. I., Tekian, A., Park, Y.S. and Al Shawa, L. 2015. The validity and reliability of the sixth-year internal medical examination administered at the King Abdulaziz University Medical College. BMC Medical Education, 15(1): 10.

11. General Medical Council (Great Britain). 2013. Good medical Practice. London, General Medical Council, 16-17 pp.

12. Glaser, R. 1984. Education and thinking: The role of knowledge. American Psychologist, 39(2): 193-202.

13. Goodenough F.L. 1950. Edward Lee Thorndike, 1874-1949. The American Journal of Psychology, 63: 291-301.

14. Gravetter, F.J. and Wallnau, L.B. 2000. Statistics for the behavioral sciences Stamford. CT: Wadsworth.

15. Gunderman, R.B. 2001. The perils of testing. Academic Radiology, 8(12): 1257-1259.

16. Haladyna, T.M. 1994. Developing and validating multiple-choice test items. Lawrence Er1baum, New Jersey.

17. Hamdy, H. 2006. Blueprinting for the assessment of health care professionals. The Clinical Teacher, 3(3): 175-179.

18. Harden R.M. and Laidlaw, J.M. 2012. Written and Computer-Based Assessment. In: Essential Skills for a Medical Teacher: An introduction to teaching and learning in medicine. London, Churchill Livingstone.

19. Hawthorne, K. 2007. Assessment in the undergraduate curriculum. In: Jackson, N., Jamieson, A. and Khan, A., (Eds.), Assessment in Medical education and Training: A practical Guide. Oxford, Radcliffe Publishing.

20. Hogan, D.B. 1979. The Regulation of Psychotherapists: A study in the philosophy and practice of professional regulation (Vol. 1). Ballinger Publishing Company.

21. Hogan, R.L. 2007. The historical development of program evaluation: Exploring past and present. Online Journal for Workforce Education and Development, 2(4): 5.

22. Jackson, N., Jamieson, A. and Khan, A. (Eds.). 2007. Assessment in medical education and training: a practical guide. Radcliffe Publishing.

23. Kim, M.K., Patel, R.A., Uchizono, J.A. and Beck, L. 2012. Incorporation of Bloom's Taxonomy into multiple-choice examination questions for a pharmacotherapeutics course. American Journal of Pharmaceutical Education, 76(6): 114.

24. Leman, N. 2000. The Big Test. New York: Favar, Strauss and Giroux.

25. Linn, R.L. Gronlund, N.E. and Davis, K.M. 2000. Measurement and assessment in teaching. 8th Edition, Merrill; London, Prentice-Hall International (UK), Upper Saddle River, N.J.

26. Mabry, L. 2004. Strange, yet familiar: Assessment driven education. Holding accountability accountable: What ought to matter in public education, 116-134 pp.

27. Masters, J.C., Hulsmeyer, B.S., Pike, M.E., Leichty, K., Miller, M.T. and Verst, A.L. 2001. Assessment of multiple-choice questions in selected test banks accompanying text books used in nursing education. Journal of Nursing Education, 40(1): 25-32.

28. McCoubrie, P. 2004. Improving the fairness of multiple-choice questions: a literature review. Medical Teacher, 26(8): 709-712.

29. McCoubrie, P. and McKnight, L. 2008. Single best answer MCQs: a new format for the FRCR part 2a exam. Clinical Radiology, 63(5): 506-510.

30. Mohanna, K., Cottrell, E., Chambers, R. and Wall, D. 2011. Teaching made easy: a manual for health professionals. Radcliffe Publishing.

31. Norcini, J. and Ben-David, F. 2013. Concepts in Assessment. In: Dent, A., Harden R.M. (Eds.). A Practical Guide for Medical teachers. 4th Edition, London, Churchill Livingstone.

32. Norcini, J.J., Swanson, D.B., Grosso, L.J. and Webster, G.D. 1985. Reliability, validity and efficiency of multiple choice question and patient management problem item formats in assessment of clinical competence. Medical Education, 19(3): 238-247.

33. Norman, G. and Eva, K.W. 2010. Quantitative research methods in Medical Education. In: Swanwick, T. (Ed.), Understanding medical Education: Evidence, Theory and Practice. 2nd Edition, Chichester, Wiley Blackwell.

34. Nunnally, J.C. and Bernstein, I.H. 1967. Psychometric theory (Vol. 226). New York: McGraw-Hill.

35. Palmer, E.J. and Devitt, P.G. 2007. Assessment of higher order cognitive skills in undergraduate education: modified essay or multiple choice questions? Research paper. BMC Medical Education, 7(1): 49.

36. Polit, O.F. and Hungler, B.P. 199. Nursing research: principles and methods. Lippincott Williams and Wilkins, Philadelphia.

37. Schuwirth L.W.T. and Van Der Vleuten C.P.M. 2013. Written Assessments. In: Dent, A., Harden, R.M., (Eds.), A Practical Guide for Medical teachers. 4th Edition, London, Churchill Livingstone.

38. Schuwirth L.W.T. and Van Der Vleuten C.P.M. 2014. How to Design a Useful Test: The principles of assessment. In: Swanwick, T. (Ed.), Understanding Medical Education: Evidence, Theory and Practice. 2nd Edition, Chichester, Wiley Blackwell.

39. Tan, L.T. and McAleer, J.J.A. 2008. The introduction of single best answer questions as a test of knowledge in the final examination for the fellowship of the Royal College of Radiologists in Clinical Oncology. Clinical Oncology, 20(8): 571-576.

40. Tarrant, M. and Ware, J. 2010. A comparison of the psychometric properties of three-and four-option multiple-choice questions in nursing assessments. Nurse Education Today, 30(6): 539-543.

41. Tweed, M. and Cookson, J. 2001. The face validity of a final professional clinical examination. Medical Education, 35(5): 465-473.

42. Van Der Vleuten, C.P. 1996. The assessment of professional competence: developments, research and practical implications. Advances in Health Sciences Education, 1(1): 41-67.

43. Veloski, J.J., Rabinowitz, H.K. and Robeson, M.R. 1993. A solution to the cueing effects of multiple choice questions: the Un-Q format. Medical Education, 27(4): 371-375.

44. Vyas, R. and Supe, A. 2008. Multiple choice questions: a literature review on the optimal number of options. National Medical Journal of India, 21(3): 130-3.

45. Wood D.F. 2014. Formative assessment. In: Swanwick, T., (Ed.), Understanding Medical Education: Evidence, Theory and Practice. 2nd Edition. Chichester, Wiley Blackwell.