Applying the utility index to review single best answer questions in medical education assessment

Mirbahai, L. & Adie, J.

DOI 10.15761/AEPH.1000113
ESSN 2633-1411

Publisher: AEPH

Archives of Epidemiology and Public Health

Research Article ISSN: 2633-1411

Applying the utility index to review single best answer questions in medical education assessment

Leda Mirbahai[1]* and James W Adie[2]

[1]Warwick Medical School, University of Warwick, CV 7AL, UK

[2]School of Psychological, Social and Behavioural Sciences, Coventry University, CV1 5FB, UK

Abstract

In professional training programmes it is vital to ensure individuals have accomplished all required competencies before qualifying, otherwise patient safety could be placed at risk. This increased emphasis on patient safety and accountability has heightened the need for reliable, valid and suitable methods of assessment that not only can inform if learning outcomes have been achieved but can also promote and encourage learning. One method of assessment that has been traditionally applied in clinical education assessment is the Single Best Answer (SBA) question approach. In this review paper, the benefits and limitations associated with using SBA questions as a method of assessment were critically evaluated. The review clearly highlighted that emphasis should be mainly placed upon the design, coverage and content of SBA questions rather than evaluating the concept of SBA questions as a method of assessment. More specifically, the paper pointed towards the complex skills required for developing a set of SBA questions that can both promote learning as well as evaluate learning. To summarise, the need for defining the necessary skills and criteria required for the careful design and successful application of SBA exam paper is an important avenue to investigate.

Introduction

Assessment is a key component of any educational programme [1] and if used appropriately, assessment can promote learning and quality assurance [1-4]. Assessments can inform educators of the quality of their teaching, areas for improvement and if learning outcomes have been achieved. In professional training programmes, it is vital to ensure individuals have accomplished all required competencies before qualifying, otherwise patient safety can be placed at risk. This increased emphasis on patient safety and accountability has heightened the need for reliable, valid and suitable methods of assessment [1,2].

Written assessments are widely used to assess certain competencies in educational programmes, including medical education [5]. In particular, written exams are divided into constructive, selected, combined constructive and selected response categories; with SBA questions categorised as selected response type [6]. Kelly used Multiple Choice Questions (MCQs) as a method of assessment for the first time in 1914 [5]. Since then different subtypes of MCQs have been developed, including single best answer (SBA) questions, extended matching questions, script concordance and multiple true/false questions [6]. Current practice discourages the use of true/false questions [4], as they are more liable to writing errors, such as inaccurate terminology [7] and cuing effect as the answer needs to be unambiguously correct or wrong [5]. This will unintentionally guide the student to the correct answer [8]. This review will focus upon the use of SBA questions as a subtype of MCQs.

One important consideration determining the selection of a particular method of assessment is what levels of competence can be assessed. According to Miller [4,9-11], SBA questions are commonly used to assess factual knowledge (i.e., the 'knows' level in figure 1). However, a well-written SBA question can promote problem-solving and require students to apply their knowledge to clinical case scenarios (i.e., 'know how' in figure 1). As such, a well-written SBA exam paper can assess student's ability at the first two lower levels of Miller's pyramid [4,5,10,11]. Furthermore, well-written SBA questions can be used to assess both lower and higher cognitive taxonomic levels as described by Bloom's model. As show in figure 2, the highest cognitive taxonomic levels captured by well-written SBA questions are 'analyse' and 'evaluate'. Therefore, the cognitive complexity of the SBA questions can be applied to reflect the cognitive level of the learner [12,13]; (Figure 2). As SBA questions are used extensively in medical education, the aim of this paper is to critically evaluate the suitability of SBA questions as a method of assessment by investigating the utility index of SBA questions.

Utility index

When deciding over and designing an assessment strategy it is essential to consider several factors. These factors are summarised in a conceptual framework referred to as utility index for assessment. The utility index was first described by [2] and it still serves as a framework during assessment design and evaluation. The framework is not a formula and there is no perfect utility index score. The weighting for each component can differ based on the purpose of the assessment (formative versus summative, evaluating knowledge versus change in behaviour) [1,2].

Utility Index = Reliability × Validity × Cost × Feasibility × Educational Impact × Acceptability

Correspondence to: Leda Mirbahai, Warwick Medical School, University of Warwick, Coventry, CV 7AL, UK, E-mail: Leda.Mirbahai@Warwick.ac.uk

Key words: *single best answer questions, multiple choice questions, utility index, assessment, medical education*

Received: February 25, 2020; **Accepted:** March 17, 2020; **Published:** March 25, 2020

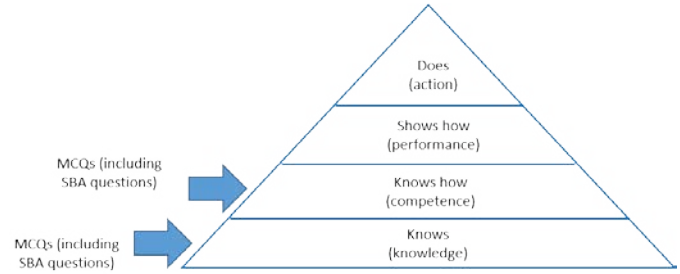

Figure 1. Miller's pyramid of assessment of clinical competence. The two lower levels indicate knowledge (cognition) with the first corresponds to basic facts ('knows') and the second corresponds to applied knowledge ('knows how'). The 'show how' level moves beyond acquisition and application of knowledge. It requires the learner to 'show' an acquired competency (behavior). The highest level of competency truly represents what a competent doctor 'does' in a workplace. This figure is adapted from Miller [9] and Hift [4]

Figure 2. Bloom's modified cognitive taxonomy. The depth of the acquired and retained knowledge can vary. The varying levels of depth of knowledge have been captured in Bloom's cognitive taxonomy. According to Bloom's modified taxonomy, cognitive domain can be divided into six levels with the level of complexity increasing as progressing from remembering facts towards creation and synthesis of knowledge. A well-written MCQ not only can target lower levels of learning, it can also target higher order learning levels, including application, analysis and evaluation. This figure is taken and modified from Hift [4]

Reliability

The outcome of an assessment is only defendable if the results are reliable [14]. Reliability is an indicator of reproducibility of the scores of an assessment [2,3,15,16]. Internal consistency, a type of reliability, is an important consideration for written assessments, including SBA questions. It measures correlation between scores of different items within an SBA examination. It is reliant on all items within an SBA examination measuring the same construct (knowledge) [16,17]. Internal reliability for SBA questions are quantified using Cronbach's coefficient alpha (α) as only one construct is assessed, with reliability coefficient of ≥ 0.80 deemed as acceptable for high-stake exams [3,14,17]. Multiple factors can affect the reliability of an assessment score, including examiner-introduced bias during marking and insufficient number of items [3,14]. SBA questions are marked objectively, and this removes examiner-introduced bias, resulting in increased inter-rater reliability [4]. However, objectivity does not automatically equal reliability. For an assessment score to be reliable sufficient sampling is required [11]. Generally, more sampling will enable a more accurate assessment of the competencies of a learner by reducing the effect of differences in quality of the questions and examinee's characteristics, leading to improved reliability and true evaluation of a student's abilities [4,5]. For good sampling, the selected items should represent the entire content and the ability of the student [8].

SBA questions in general are efficient as high sampling/hour/number of students can be achieved and thus have a high reliability per hour [3,4,15]. For example, Norcini et al. [18] investigated the reliability of the scores for three types of MCQs for three years of certifying exam (n=7000-8000 participants). The study demonstrated that 82-85 SBA questions assessed in 2.8 hours have a coefficient alpha of 0.74, 0.82 and 0.80, for each of the three years. Scores for two of the years are highly reliable (α ≥ 0.8). However, for one-year Cronbach's alpha was below the recommended value of 0.8. Duration of the assessment, homogeneity of the construct, interrelatedness between items and quality of the discriminatory questions can all impact internal reliability scores. The latter will help to discriminate between high and low achievers [16]. In this example, assessment time was consistent between years. Therefore, the low value for alpha (0.74) could be due to poor interrelatedness between items (measuring different constructs) and/or poor-quality discriminating items. This can be evaluated by conducting inter-item and item-total correlation analysis and calculating the item discrimination index [16,19]. However, no item analysis was available for this study. Therefore, it is difficult to analyse why Cronbach's alpha was lower in one particular year. Nevertheless, Norcini et al. [18] demonstrated that by increasing total number of SBA questions and assessment time, the reliability of the scores significantly improves (Table 1). This demonstrates a positive relationship between increased sampling/hour and increased reliability.

Item-writing quality has a significant effect on the reliability of the SBA scores. A flawed item can affect the performance of a student by either making the question too easy or too difficult. If the question is too easy, it cues the student towards the correct answer and thus the scores are not a true reflection of the student's ability [20-22]. Furthermore, inaccurate and vague terminology can cause confusion, resulting in reduced reliability of the data [7]. Therefore, a well-drafted SBA exam will require discriminatory questions and plausible distractors [19]. Item-writing quality can be improved by providing training, quality control and following item-writing guidelines [21,23-25]. Overall, if items are well constructed and appropriate sampling is achieved, SBA examination scores can be highly reliable [3,4,15].

Validity

Validity, alongside reliability, are the two most important components of the utility index for high-stake assessments, with reliability a pre-requisite for validity [16]. Face, content, concurrent, predictive and construct validity have been classified as different types of validity [2,3]. Although Downing [26] has proposed that construct validity is not a subtype of validity but rather it is validity in its entirety and evaluation of construct validity requires evidence from multiple sources, including content, response process and intrinsic flaws and errors associated with a method of assessment [26]. For SBA questions, the errors can relate to the quality of item-writing and non-functioning distractors (i.e. the other least plausible options in an SBA question) [27,28]. Overall, validity is defined as the degree to which an assessment method and its content measure what it is expected to evaluate and at an appropriate level [2,3,5,26].

Content validity ensures that the content coverage, focus and depth is adequate to provide a true representation of the measured

Table 1. Improving the reliability of the MCQ scores by increasing sampling number and assessment time. As shown in this table increasing the sampling number from 82, 45, 308 to total of 435 samples and increasing the assessment time from 2.7h, 1.5h and 3.2h to total of 7.4h increased the Cronbach's coefficient alpha from 0.74, 0.76 and 0.88 to 0.92. The table is extract from the study published by Norcini et al. [18]

	Best single answer	Matching questions	True/false questions	Total MCQ questions
Number of questions	82	45	308	435
Time (h)	2.7	1.5	3.2	7.4
Cronbach's coefficient alpha	0.74	0.76	0.88	0.92

construct [2,26,29]. In SBA exam papers, blueprinting ensures that all items align with the course learning outcomes and are set at an appropriate cognitive level [26,29]. Therefore, blueprinting can prevent construct under-representation (CU) and construct-irrelevant variance (CIV), which are usually caused by either under-sampling or biased sampling [30,31]. Multiple studies have demonstrated that more than one expert is required to evaluate content validity [32,33], as the process of content blueprinting can be subjective and prone to errors [5]. As detailed in the 'reliability section', SBA questions have a high sampling/hour, therefore it is possible to achieve high validity. Conducting item analysis by investigating item discrimination index (DI) and item difficulty can provide valuable information regarding validity of an assessment result. DI is measured by calculating point biserial correlation coefficient, which ranges from -1 to +1 with values above 0.35 deemed acceptable while item difficulty is measured by calculating facility index, range between 0-100 (or 0-1), with higher values indicating easier questions and values ranging from 30-70 (or 0.3-0.7) deemed as acceptable [34]. In SBA questions, the quality of the distractors can significantly affect both reliability and validity of the results [27,35]. For example, Ali et al. [27] demonstrated that replacing non-functioning detractors (options selected by less than 5% of the students in 23 SBA questions) improved the reliability of the data (averaged α improved from 0.62 to 0.72; n=30 first-year medical students). Furthermore, it improved the difficulty of the SBA questions by reducing the gap between expected difficulty index to observed difficulty index from 0.4-0.59 to 0.15. This can subsequently result in improving the quality of the questions by improving the focus of the question and the validity of the question. However, a limitation of this study is small sampling size, which is reflected in the alpha (α is <0.8).

Construct validity is evaluation of how well a single construct is measured in an assessment [3,5]. As mentioned above, blueprinting can reduce construct-irrelevant variance (CIV) error [30,31]. In general, SBA questions are accepted by examiners and examinees as a suitable method for assessment of knowledge (i.e. face value for assessment of knowledge domain). Therefore, if written well the assessment data can have high construct validity for knowledge. Although it has been shown that SBA questions can assess both factual and applied knowledge, some still believe that SBA questions can only be used for assessment of factual knowledge [4,5]. Furthermore, SBA scores have been shown to demonstrate excellent predictive validity pertaining to final year medical exam results [33]. Fallatah et al. [33] demonstrated a significant correlation (r=0.82, p<0.001) between SBA scores with 320 questions and the final exam results comprised of objective structured clinical examinations (OSCE), SBA questions and long-case presentation (number of students=824). However, the study is not without limitations as Cronbach's alpha was used to also measure the reliability of OSCEs. The authors highlighted that the assessments had a high internal consistency (reliability) by measuring Cronbach's alpha for all assessments, including OSCE. However, as OSCE is a multi-construct assessment method, Cronbach's alpha will overestimate the internal consistency and reliability. Generalisability coefficient is a more suitable method for assessing reliability of OSCEs [14]. Most importantly, the SBA data is part of the final assessment data when they calculated the correlation and subsequently this would have overestimated the positive correlation.

Cost and feasibility

The feasibility of an assessment method depends on the resources required to develop and construct the items within the assessment as well as the actual running cost [8]. Writing a well-constructed SBA question with plausible distractors is challenging, time consuming and requires training [19,36]. However, once a high-quality question bank is created, SBA questions become a highly feasible and effective method of assessment [1,4]. Use of scanners for marking contributes to cost effectiveness of SBA questions [1]. Certain steps can reduce the cost associated with development of SBA items, including dedicated trained staff for generation of high-quality questions and shared question banks between institutes. However, the latter does require initial heavy investments by multiple institutes [1].

Reducing the number of plausible options can also increase feasibility of SBA questions, as writing high quality distractors is time consuming. Currently in medical schools, SBA questions with five choice of responses (5-options) are widely used. However, there is no clear evidence on optimal number of options for a SBA item. The key factor should be the quality of the question and writing plausible distractors, as non-functioning distractors directly affect the reliability and validity of the SBA scores [8,27,35,37,38]. Multiple studies have investigated the optimal number of distractors, with varying level of evidence indicating that using SBA questions with only three choice of responses has no impact on quality of the question while improving efficiency [19,38,39]. For example, Vegada et al. [19] conducted a study whereby 132-second year medical students were divided into three groups with equal distribution of high, mid and low achievers per group. The students undertook a 30-item SBA exam with either five, four or three choice of responses. The authors concluded that SBA questions with three choice of responses are as valid as SBA questions with five choice of responses, as there was no significant difference in the reliability of the scores between groups. However, the study is not without limitations. The scores of the three groups are statistically different (p =0.000). Although within an acceptable range, item difficulty was significantly higher (p =0.004) for SBA questions with three choice of responses (55.45 ± 17.34) versus five choice of responses (39.05 ± 19.09), indicating that higher percentage of students found the SBA questions with three choice of responses easier than SBA questions five choice of responses. Most importantly, the Cronbach's alpha for none of the groups was above 0.8 (Cronbach α for three, four and five-option questions was 0.61, 0.67 and 0.75, respectively). Therefore, the scores are not highly reliable and defendable, and therefore any conclusion from these results should be viewed with caution. Overall, further research on optimal number of distractors is needed. However, the most important factors are the quality of the distractors rather than the quantity [27,38].

Educational impact

Assessments can promote learning. However, it is also widely acknowledged that students are strategic learners and will prioritise and centre their learning around assessment topics. In other words, they will adapt a learning strategy that is suitable for the method of assessment. Therefore, assessments should be used strategically to promote desired learning strategies [2,3,11,40,41], as approaches to learning (superficial, deep and achieving) impacts performance in exams [42]. Poorly constructed SBA questions have low educational impact. This is partly due to students guessing the correct answer (cuing effect) and developing a pattern recognition learning technique rather than learning the content. However, well-written SBA questions with clinical vignettes that require application of knowledge and higher cognitive process (see Figure 2) can promote deep learning and thus have a higher educational impact [5,42,43]. In general, the learner's method of preparation is different for SBA questions compared to open-ended written questions and methods of assessment that correspond to higher

levels of Miller's pyramid [8,42]. For example, the educational impact of SBA questions is less than methods of assessment that correspond to higher levels of Miller's pyramid, such as Direct Observation of Procedural (DOP) skills as demonstrated by Cobb et al. [42]. In the study conduct by Cobb et al. [42], they analysed the result of a shorten version of the Study Process Questionnaire (SPQ) collected from 70 final year medical students from one institute that had completed 10 DOP assessments throughout a year as well as an end of year SBA exam. The result showed a statistically significant difference ($p<0.001$) in learning approach adapted by the students for the two assessment methods with a more superficial method of learning adapted for SBA questions than DOP. However, the study is not without limitations. For example, the assessment times for the two methods were different. It is well-established that time of assessment (end of year versus during the year) does impact learning strategy [2,42]. Furthermore, the sample size is small with only data collected from one institute.

It is also accepted that SBA questions can promote learning as both summative and formative assessment tools [35]. The key to promote learning is provision of feedback [4,11,41]. For examples, encouraging students to take part in SBA item-writing, answering and provision of peer feedback can promote learning as demonstrated by Walsh et al. [44]. Walsh et al. [44] identified a statistically significant ($p<0.001$) positive correlation between item-writing (r=0.24), answering MCQs (r=0.13) and peer feedback provision (r=0.15) with final summative scores. The study was conducted on two cohorts of first year medical students at Cardiff University (n=297 for 2013/2014 entry and n=306 for 2014/2015) and one cohort of second year students (n=273). Although the results clearly indicate educational impact of this process, it does require repeating at more than one institute.

Overall, as reported by van der Vleuten [2] and others, it is difficult to predict the learning strategy and behaviour that an assessment method will provoke in a learner and thus it is challenging to predict the exact educational impact of an assessment method. However, steps can be taken to minimise superficial learning strategies and improve the educational impact of SBA exams.

Acceptability

Acceptability, although closely linked with face validity, is a broader concept. It encompasses acceptability of an assessment method not only by the educator and trainee but also by all stakeholders, including general public [3]. Acceptability of an assessment method is influenced by stakeholder's values, beliefs and experiences [2]. Therefore, in our view it is extremely challenging to alter a stakeholder's perception of an acceptable method of an assessment, as at times it requires transformation of an individual's core values and beliefs.

The interest of stakeholders in high-stake medical education has resulted in the need for exams, including SBA exams to be fair (high validity and reliability). The fairness of an assessment relates to its construction process, content, quality and standard setting procedure. In general, by improving the fairness of SBA exams both face validity (the degree to which a procedure appears effective in terms of its stated aims) and acceptability of SBA exams can improve [40]. The acceptability of SBA exams has always been a point of discussion and has led to continuous improvement of the structure and content of SBA questions [2]. The high susceptibility of SBA questions to item-writing flaws, lack of resemblance to real-life practice and the incorrect presumption that SBA questions can only assess recall of information has partly resulted in negative view regarding SBA questions [22]. However, despite negative views regarding SBA exams, they are widely used and are accepted. This is partly due to high validity and reliability of a well-written SBA exam and their cost effectiveness. Furthermore, it is widely accepted that knowledge underpins the higher-level competencies of a doctor, including being able to understand, conduct and demonstrate a task with high level of efficiency [5,13,22,40]. Multiple studies, including a study conducted by Pham et al. [22] has demonstrated that SBA questions have the same potential as short answer questions (SAQs) to assess higher cognitive abilities of a student (n=136 final year medical students, number of matching SAQ and SBA=40, statistically significant association with interclass correlation coefficient of 0.77). Therefore, overall SBA questions are accepted as a method of assessment for knowledge construct. However, it is required that SBA exams are combined with other methods of assessment that can evaluate higher-level competencies [5,40].

Conclusion

It is apparent that high validity and reliability can be achieved for SBA exams if questions are constructed well and appropriate sampling is conducted. SBA questions are suitable for assessment of factual and applied knowledge and they can be constructed to assess lower and higher cognitive taxonomic levels, up to analysis and evaluation. Once generated they are cost effective as high number of students can be assessed per time required for examining and marking. Although they can be used to provide instant feedback and promote learning, issues surrounding fairness of SBA questions, misconception regarding suitability of SBA questions to assess applied knowledge, and dissimilarity of SBA exam conditions to clinical setting has always led to questioning the educational impact and acceptability of the SBA questions. Subsequently, this has resulted in continuous improvement of SBA questions, generation of item-writing guides and use of multiple subtypes of MCQs alongside other written formats, such as SAQs. However, their high reliability, validity and cost effeteness has made them a standard component of high-stake medical education assessment programmes. Nevertheless, it is important to mention that no single method of assessment can measure all levels of competency. Therefore, it is recommended to use a suite of assessments to make a sound judgment regarding an individual's competencies as a health care professional and it is important to align the method of assessment with what it is intended to measure.

References

1. Schuwirth LWT, van der Vleuten CPM (2010) How to design a useful test: the principles of assessment. Understanding medical education: evidence, theory and practice. Wiley-Blackwell, Malaysia. pp: 195-207.
2. Van der Vleuten CPM (1996) The assessment of professional competence: developments, research and practical implications. *Adv Health Sci Educ Theory Pract* 1: 41-67. [Crossref]
3. Oliver C (2007) Developing and maintaining an assessment system: a PMETB guide to good practice. [online] Available from: https://www.researchgate.net/publication/264405740_Developing_and_Maintaining_an_Assessment_System-a_PMETB_Guide_to_Good_Practice
4. Hift RJ (2014) Should essays and other open-ended –type questions retain a place in written summative assessment in clinical medicine? *BMC Med Educ* 14: 249. [Crossref]
5. Okubuiro EO, Ebirim LN, Okoli CE (2019) Utility of single best answer questions as a summative assessment tool in medical education: a review. *Int J Recent Innov Acad Res* 3: 1-12.
6. Jolly B (2010) Written exams. Understanding medical education: evidence, theory and practice. Wiley-Blackwell. Malaysia: pp: 208-231.
7. Holsgrove G, Elzubeir M (1998) Imprecise terms in UK medical multiple-choice questions: what examiners think they men. *Medi Educ* 32: 343-350. [Crossref]

8. Schuwirth LW, van der Vleuten CP (2004) Different written assessment methods: what can be said about their strengths and weaknesses? *Medi Educ* 38: 974-979. [Crossref]

9. Miller GE (1990) The assessment of clinical skills/competence/performance. *Acad Med* 65: S63-S67. [Crossref]

10. Tabish SA (2008) Assessment methods in medical education. *Int J Health Sci (Qassim)* 2: 3-7. [Crossref]

11. van der Vleuten CP, Schuwirth LW, Scheele F, Driessen EW, Hodges B (2010) The assessment of professional competence: building clocks for theory development. *Best Pract Res Clin Obstet Gynaecol* 24: 703-719. [Crossref]

12. Krathwohl DR (2002) A revision of Bloom's taxonomy: an overview. *Theory into Practice* 41: 212-218.

13. Javaeed A (2018) Assessment of higher order thinking in medical education: multiple choice questions and modified essay questions. *MedEdPublish* 7: 60.

14. Downing SM (2004) Reliability: on the reproducibility of assessment data. *Med Educ* 38: 1006-1012. [Crossref]

15. Schuwirth LWT, van der Vleuten CPM (2003) ABC of learning and teaching in medicine: written assessment. *BMJ* 326: 643-645. [Crossref]

16. Tavakol M, Dennick R (2011) Making sense of Cronbach's alpha. *Int J Med Educ* 2: 53-55. [Crossref]

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

18. Norcini J, Swanson DB, Grosso LJ, Webster GD (1985) Reliability, validity and efficiency of multiple-choice question and patient management problem item formats in assessment of clinical competence. *Med Educ* 19: 238-247. [Crossref]

19. Vegada B, Shukla A, Khilnani A, Charan J, Desai C (2016) Comparison between three option, four option and five option multiple choice question tests for quality parameters: a randomised study. *Indian J Pharmacol* 48: 571-575. [Crossref]

20. Tarrant M, Ware J (2008) Impact of item-writing flaws in multiple choice questions on student achievement in high-stakes nursing assessments. *Med Educ* 42: 198-206. [Crossref]

21. Omer AA, Abdulrahim ME, Albalawi IA (2016) Flawed multiple-choice questions put on the scale: what is their impact on student's achievement in a final graduate surgical examination? *J Health Specialties* 4: 270-275.

22. Pham H, Trigg M, Wu S, O'Connell A, Harry C, et al. (2019) Choosing medical assessments: does the multiple-choice question make the grade? *Educ Health* 31: 65-71. [Crossref]

23. Case S, Swanson D (2000) Constructing written test questions for the basic and clinical sciences. 3rd ed. Philadelphia: National Board of Medical Examiners.

24. Haladyna T, Downing S, Rodriquez M (2002) A review of multiple-choice item writing guidelines for classroom assessment. *Appl Measurement Educ* 15: 309-334.

25. Reichert TG (2011) Assessing the use of high-quality multiple-choice exam questions in undergraduate nursing education: are educators making the grade?

26. Downing SM (2003) Validity: on the meaningful interpretation of assessment data. *Med Educ* 37: 830-837. [Crossref]

27. Ali SH, Carr PA, Ruit KG (2016) Validity and reliability of scores obtained on multiple-choice questions: why functioning distractors matter. *J Scholarship Teaching Learning* 16: 1-14.

28. Surry LT, Torre D, Durning SJ (2017) Exploring examinee behaviours as validity evidence for multiple-choice question examinations. *Med Educ* 51: 1075-1085. [Crossref]

29. Chandratilake M, Davis M, Ponnamperuma G (2009) Evaluating and designing assessments for medical education: the utility formula. *The Internet J Med Educ* 1: 1-7.

30. Downing SM, Haladyna TM (2004) Validity threats: overcoming interference with proposed interpretations of assessment data. *Med Educ* 38: 327-333. [Crossref]

31. Abdalla ME (2013) Multiple-choice questions revisited: improvement of validity for fair tests. *Gezira J Health Sciences* 9: 27-36.

32. Veloski JJ, Rabinowitz HK, Robeson MR (1993) A solution to the cueing effects of multiple choice questions: the Un□Q format. *Med Educ* 27: 371-375. [Crossref]

33. Fallatah HI, Tekian A, Park YS, Al Shawa L (2015) The validity and reliability of the sixth-year internal medical examination administered at the King Abdulaziz University Medical College. *BMC Med Educ* 15: 10. [Crossref]

34. Chavda P, Misra S, Duttaroy B (2015) Item analysis of multiple-choice questions based undergraduate assessment in community medicine. *South East Asian J Med Educ* 9: 66-68.

35. Sam AH, Hameed S, Harris J, Meeran K (2016) Validity of very short answer versus single best answer questions for undergraduate assessment. *BMC Med Educ* 16: 266.

36. Epstein RM (2007) Assessment in medical education. *N Engl J Med* 356: 387-396. [Crossref]

37. Vyas R, Supe A (2008) Multiple-choice questions: a literature review on the optimal number of options. *Natl Med J India* 21: 130-133. [Crossref]

38. Kilgour JM, Tayyaba S (2016) An investigation into the optimal number of distractors in single-best answer exams. *Adv Health Sci Educ Theory Pract* 21: 571-585. [Crossref]

39. Schneid SD, Armour C, Park YS, Yudkowsky R, Bordage G (2014) Reducing the number of options on multiple-choice questions: response time, psychometrics and standard setting. *Medical Educ* 48: 1020-1027. [Crossref]

40. McCoubrie P (2004) Improving the fairness of multiple-choice questions: a literature review. *Med Teach* 26: 709-712. [Crossref]

41. Schuwirth LWT, van der Vleuten CPM (2018) How 'testing' has become 'programmatic assessment for learning'. *Health Professions Education* 5: 177-184.

42. Cobb KA, Brown G, Jaarsma DA, Hammond RA (2013) The educational impact of assessment: a comparison of DOPS and MCQs. *Med Teach* 35: e1598-e1607. [Crossref]

43. Baig M, Ali SK, Ali S, Huda N (2014) Evaluation of multiple choice and short essay question items in basic medical sciences. *Pak J Med Sci* 30: 3-6. [Crossref]

44. Walsh JL, Harris BHL, Denny P, Smith P (2018) Formative student-authored question bank: perceptions, question quality and association with summative performance. *Postgrad Med J* 94: 97-103. [Crossref]