ASSESSMENT METHODS IN MEDICAL EDUCATION: A REVIEW

ISHTIAQ AHMED[1], SUNDAS ISHTIAQ[2]

ABSTRACT

Good assessment is a major challange in medical education. One of the major obstacle to a comprehensive assessment is the lack of familirity on the part of medical educators about proper selection and effective use of different assessment methods. This primer (review) gives an overview of the basic ideas and vocabulary that one should understand in order to evaluate the quality of any assessment tool designed for the purpose of evaluationg the undergraduates, postgraduates or other medical professionals. Applicability and effectiveness of different assessment tools are described along with their limitations and advantages. Inaddition, assessment methods currently in use are reviewed with attention to their psychometric strength and weaknesses. The data was collected from cross sectional studies, review articles, books on medical education and from guidelines for assessment betweem 1956 to 2013. Websites and other online resources of medline, NCBI and medscape were used to extract the data.

KEY WORDS: Medical Education, Assessment, Assessment Tools, Performance, Curriculum, Learning.

INTRODUCTION

During last two decades the art and science behind medical learning and teaching i.e. medical education has progressed remarkably. The curricula are based on sound pedagogical principles. Learning and teaching have become more scientific, rigorous and problem based and other forms of active and self-directed learning have become the mainstream. The role of teachers has progressed to solution provider rather a problem-identifier[1, 2].

During the last quarter of century the medical schools are facing variety of challenges from patients, society, doctors and students. They have responded in different ways in the form of new curricula development, the introduction of new learning methodologies, new methods of assessment and a realization of the importance of professional development of staff. Due to this, many interesting and effective innovations are made and put in practice [2,3].

The efficiency and effectiveness of health care delivery requires not only knowledge and technical skills but also needs good communication, analytical skills, interdisciplinary care, evidence and system based care. This can be achieved only if the assessment system is sound, comprehensive and robust enough to assess the requisite attributes as well as testing of essential knowledge and skills[3]. Realistically, the assessment should be purpose driven because it has a powerful positive steering effect on learning outcome and the curriculum. It serves multiple purposes for example, formative assessment are used for promoting reflection, guidelines for future learning and shaping values. Similarly, the summative assessment is used for judging an individual's cognitive achievements and clinical performance. So, it conveys what we value as important and acts as the most cogent motivator of student learning. It is essential in designing and planning an assessment to identify and recognize the stakes involved in it. The higher the stakes, greater the implications of the outcome of assessment. Moreover, the more sophisticated the assessment strategies, the more appropriate they become for feedback and learning [4,5]. In this era, the assessment is entering in every phase of professional development and considered crucial steps in the educational process. It is now used during the medical school application process, at the start of residency training and as part of the “maintenance of certification” requirements that several medical boards have adopted[2]. Some important questions must be asked before making a choice of assessment method i.e. what should be assessed?, why assess? Similarly, before deciding an assessment instrument one must also ask: is it valid? Is it reliable? is it feasible? What is assessed and which methods are used will play a significant part in what is learnt[6].

1. Professor of Surgery
Al-Nafees Medical College & Hospital
Isra University, Islamabad Campus
2. Women Medical College, Abbottabad

Correspondence to:
Prof Dr Ishtiaq Ahmed
Head of Surgery Department
Al-Nafees Medical College & Hospital
Isra University, Islamabad Campus
E-mail: surgish2000@yahoo.com

ASSESSMENT METHODS

According to the model proposed by Miller, various assessment methods are available to assess clinical competency of students[7]. The choice of assessment method will depend on the purpose of its use: whether it is for formative purposes (i.e. diagnosis, feedback and improvement), summative purposes (e.g. promotion and certification), or for both. The various characteristics of assessment tools are identified i.e. reliability, validity, feasibility, cost effectiveness and educational impact[6]. Moreover, each assessment method has its advantages and disadvantages, so one assessment method will not assess all domains of competency. Therefore, whatever the purpose of assessment is a variety of assessment methods are required so that the shortcomings of one can be addressed appropriately [6,8].

In 1990, Miller proposed a hierarchical model for the assessment of clinical competence. This model starts with the assessment of cognition and ends with the assessment of behavior in practice[7] (Figure-1). According to this, the professional authenticity increases as we move up the hierarchy and as assessment tasks resemble real practice. The assessment of cognition deals with knowledge and its application (knows, knows how) and this could span the levels of Bloom's taxonomy of educational objectives from the level of comprehension to the level of evaluation[9].

Mastery testing (criterion-reflected tests) requires that 100% of the items are measured correctly to determine whether students have attained a mastery level of achievements. In non-mastery testing attainment of 65% of a tested material is considered sufficient[8].

Currently, a wide range of assessment methods are available, which include long essay questions, modified essay questions (MEQs) ,short essay questions (SEQs), oral examination/viva, OSCE, MCQs, extended matching items, Constructed Response Questions (CRQs),checklists, critical reading papers, rating scales, student projects, patient management problems, tutor reports, portfolios, short case assessment and long case assessment, log book, trainer's report, audit, simulated patient surgeries, video assessment, simulators, self-assessment, peer assessment, standardized patients etc[1-3].

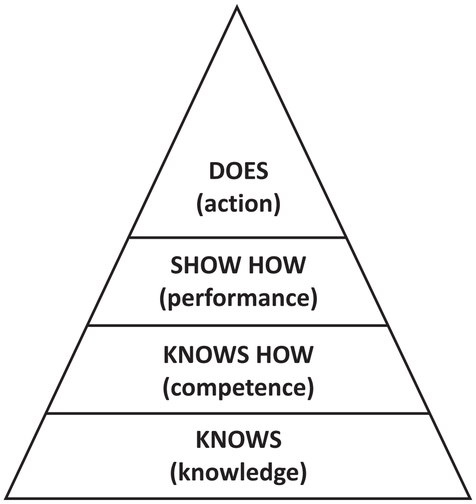

Fig 1: Miller hierarchical model for the assessment of clinical competence

SELF ASSESSMENT

Self-assessment (self-regulation) is a vital aspect of the continuous life long performance of physicians. Self-monitoring requires that individuals are able not only to work independently but also to assess their own performance and progress. All form of assessment can be used as a self-assessment exercise as long as students are provided with 'gold standard' criteria for comparing their own performance against an external reliable measure. Self-assessment methods include written exams like MCQs, MEQs, essay, True/False, modified CRQs and performance exams which comprises of portfolio, student log book, checklists, global rating, video etc[10,11].

ORAL EXAMINATIONS AND VIVA

Oral examinations are also commonly used for assessment. Different studies show that the oral examination/viva has poor content validity, higher inter-rater variability and inconsistency in marking. The instrument is prone to biases and is inherently unreliable. Its validity and reliability can be improved by making it more structured and objective[12].

LONG ESSAY QUESTIONS

This method was the most commonly used in past for the assessment of knowledge. Long essay questions are used when candidates are required to process, evaluate, summarize, supply or apply information to new situations. Long Essay Questions can be used for assessment of complex learning situations that cannot be assessed by other means (writing skills, ability to present arguments). Much more time is required to answer these questions than short answer or multiple choice questions.Therefore a limited number of questions can be used per hour of testing and hence they have lower reliability [5].

Several formats are used for assessment butit should be noted that in choosing any format, the question that is asked is more important than the format in which it is to be answered. In other words, it is the content of the question that determines what the question tests[5,13].

Structuring the marking process and using a correction scheme similar to the one used for short answer questions can improve their reliability. The guidelines for writing short answer questions can also apply to the long essay questions[13,14].

MODIFIED ESSAY QUESTIONS (MEQs)

Modified essay questions are a special type of essay questions which consists of a case summary followed by a series of questions related to the case and that must be answered in the sequence asked. This leads to question interdependency and a student answering the first question incorrectly is most likely to incorrectly answer the subsequent questions too. Therefore, in this assessment no review or possibility of correcting previous answers is allowed and the case is reformulated as the reporting process progresses[15]. A well-written MEQ assesses the approach of students to problem solving, understanding of concept, their reasoning skills, rather than recall of factual knowledge[13]. Due to psychometric problems associated with question interdependency, MEQs are not being used commonly for assessment and replaced by the key feature questions[13-15].

SHORT ANSWER QUESTIONS (SAQs)

The Short Answer Question (SAQ) are semi-structured, open ended question format which can incorporate a clinical scenarios.These questions require students to generate an answer of no more than one, two or few words, rather than to select from a fixed number of options. Many SAQs cannot be asked in an hour of testing time because they require time to answer. This limited sampling leads to less reliable tests but the SAQs have a better content coverage as compared to long essay question. Moreover, the equal or higher test reliabilities can be achieved with fewer SEQs as compared to true/false items. If a large amount of knowledge is required to be tested, then MCQs should be used. It is very important that the questions should be phrased unambiguously and a well-defined answer key is written before marking these questions[13].

A structured predetermined marking scheme is essential to improve the objectivity of SAQs. Moreover, their requirement to be marked by a content expert makes them more costly and time consuming; therefore, they should only be used when closed formats are excluded. In case of the availability of multiple examiners, double marking is preferred and more reliable. For more efficiency and reliability, each marker should assess the same question for all candidates which leads to more reliable scores than if each marker assess all the questions of one group of candidates while another marker assess all questions for another group[5,14]. A similar format is also known as Modified Essay Question (MEQ) or Constructed Response Question (CRQ) can be used[13].

MULTIPLE CHOICE QUESTIONS (A Type)

Single best response MCQs are the most commonly used question type in which students are required to select the single best response from three or more options. They are relatively easy to construct and due to their broad content domain they have high reliability per hour of testing. This assumption is not correct that multiple choice questions (MCQ) are unsuitable for assessing problem solving ability because they require candidates to simply recognize the correct answer, while they have to generate the answer in open ended questions[5,13,14]. Properly constructed MCQs can test the application of knowledge and problem solving skills. Context-free questions can almost exclusively test factual knowledge only and the thought process involved is of simple level (C1)[13]. By contextualizing the questions i.e. by including laboratory findings or clinical scenarios questions are made more authentic and reliable. There is more likely that the student will focus on important information rather than trivia. Moreover, more complex thought process is involved in which the candidates are analyzing different information when making a decision[13,14].

This does not exclude the importance of other question formats which are more suitable than MCQs for asking certain types questions. For example, an essay question will be more suitable than an MCQ when an explanation is required[5].

EXTENDED MATCHING QUESTIONS (R Type)

Extended matching questions or extended matching items (EMQs or EMIs) are context-rich questions having a practical alternative to MCQ, while maintaining consistency and objectivity. The extended Matching Item is based on a single theme and has a long option list to avoid cueing. These questions can be used for the assessment of clinical scenarios with fewer indications they can be used in both basic and clinical sciences[14]. EMQs are organized into sets of short clinical vignettes or scenarios that use one list of options that are aimed at one aspect for example all diagnoses, all laboratory investigations etc. These options can range from 5 to 26, although 8 options have been recommended to make more efficient use of testing time. Some options may apply to more than one theme while others may not apply at all. A well-constructed extended matching set includes four components: theme, options list, lead-in statement, and at least two item stems[14,16].

KEY FEATURE TESTING

Key Feature Test is a clinical scenario-based paper and pencil test. In this assessment, problem description is followed by a limited number of questions which focus on critical, challenging actions or decisions. It has higher content validity with proper blueprinting. Key features questions are short clinical scenarios or cases which are followed by questions aimed at key features or essential decisions of the case. These questions can either be open ended or multiple choice questions. More than one correct answer can be given. When these questions are constructed according to certain guidelines, they can effectively test clinical decision-making skills with a significant validity and reliability[17]. Limitation with this type of questions are that their construction is time consuming, especially if teachers are inexperienced question writers that's why they are less well known than the other types[13,17].

LONG CASE

The long case has traditionally been used to assess clinical competence. In long case usually a non-standardized real patient is used and students interview and examine a patient and then summarize their findings to one or two examiners who question the students by an unstructured oral examination on the patient problem and other relevant topics. The student's interaction with the patient is usually unobserved. Long case may provide a unique opportunity to test the physician's tasks and interaction with a real patient. Different studies show that this assessment has poor content validity, lacks consistency and is less reliable. Moreover, the reproducibility of the score is 0.39 which means 39% of the variability of the score and it is due to actual performance of students and the remaining 61% of the variability is due to errors in measurement[18,19]. In contrary, the long case has face validity and authenticity because the undertaken task almost resembles what the doctor does in real practice. However, it is usually recommended that long case should be avoided in high stake summative assessment[18] and, in fact, it has been discontinued in North America, due to its low reliability. On the other hand, its use in formative examinations is encouraged because of its perceived educational impact[20]. The validity and reliability of long cases can be increased by several modifications e.g. by observing the candidates while interacting with the patient[19], (although this is not a major contributor to reliability);examiners training to a structured examination process[21], and increasing the number of cases[21,22].

SHORT CASE

Short Case assessment involves use of three to four non-standardized real patients with one to two examiners. It provides opportunity for assessment with real patients and allows greater sampling than single long case. This assessment is commonly used in to assess clinical competence of candidate[23,24]. Students are asked to perform a supervised focused clinical examination of a real patient, and are then evaluated on the ability to elicit physical signs, examination technique and to summarize and interpret these findings correctly. To increase the sample size, several cases are used in any one assessment. However, the studies on the validity and reliability of short case assessment are scarce and it is advocated that their empirical validation must be done before promoting their use[25].

OBJECTIVE STRUCTURED CLINICAL EXAMINATION (OSCE)

Objective Structured Clinical examination (OSCE) comprises of different stations where candidate is asked to perform a defined task such as performing focused clinical examination, focused information gathering or to perform some skill activity. For each station, a standardized marking scheme is used. It is an effective alternative to unstructured short case assessment.

The OSCE is primarily used to assess basic clinical skills in which the students are assessed on different discrete focused activities that simulate different aspects of clinical competence at a number of "stations". Each student is exposed to the same stations and assessment and scoring is done with a task. At each station real patients, standardized patients (SPs), or simulators may be used, and demonstration of specific skills can be observed and measured[26,27]. OSCE stations may also incorporate the assessment of interpretation, technical and non-patient skills. Depending upon the complexity of task and assessment, OSCE stations may be short or long (5–30 minutes). The number of stations may vary from as few as eight to more than 20 although an OSCE with 14–18 stations is recommended to obtain a reliable measure of performance[28]. Reliability mainly depends upon sampling, number of stations and competences tested. For assessment, specific checklist or a combination of a checklist and a rating scale can be used. Global ratings produce equivalent results as compared to checklists[27,28]. The scoring of the students or trainees may be done by observers which may be the faculty members, patients, or standardized patients[29].

Mini-Clinical Evaluation Exercise (MINI-CEX)

Mini-CEX is a rating scale developed by American Board of Internal Medicine to assess six core competencies of residents which includes medical interviewing skills, physical examination skills, professionalism, counseling, clinical judgment and other humanistic/generic qualities[30]. Mini-CEX is based on tutor observations of routine interactions that supervising trainee or clinicians have on a daily basis[30,31]. These trainee-patient encounters occur with different evaluators at multiple occasionsin different settings. These encounters are on relatively short observations of 15–20 minutes duration during which performance is assessed on a four point scale i.e. unacceptable, below expectation, met expectations, and exceeded expectations. There is an option for reporting that a particular behavior was unobserved and additional space is provided to record details about the context of the encounter. The mini-CEX is mostly used for formative assessment and incorporates an opportunity for feedback from the evaluator. Evaluators mostly consist of tutors whose primary role is to teach students[30].

DIRECT OBSERVATION OF PROCEDURAL SKILLS (DOPS)

DOPS is a structured rating scale for assessing and providing feedback on practical procedures. The competencies that are commonly assessed include general knowledge about the procedure, informed consent, counseling, communication, pre-procedure preparation, analgesia, technical ability, aseptic technique and post-procedure management[32].

CLINICAL WORK SAMPLING

Clinical Work Sampling is an in-trainee evaluation method that addresses the issue of system and rater biases by collecting data on observed behavior at the same time of actual performance and by using multiple observers and occasions[33, 34].

CHECK LISTS

Checklists are used to capture an observed behavior or action of a student. Checklists are useful for assessing any competence or a component of the competency that can be broken down into specific actions orbehaviors which can be either done or not done. To avoid trivializing the task and to enhance the validity, it is recommended that over-detailed checklists should be avoided[5]. Global ratings (a rating scale which is used in a single encounter, for example in an OSCE, in addition to or instead of a checklist, to provide an overall or "global" rating of performance across a number of tasks) provide a better reflection of expertise than detailed checklists[35].

Checklist development requires consensus by several experts on the essential behavior's, actions, and criteria for evaluating performance. This is important to ensure validity of content and scoring rules. Moreover, in order to obtain consistent scores and satisfactory reliability, trained evaluators should be used for this assessment[29].

360 DEGREE EVALUATIONS

360° evaluation is a multi-source feedback assessment system which consists of measurement tools that evaluates an individual's competence from multiple perspectives within their sphere of influence. Assessment or feedback collected objectively and systematically through multiple evaluators like peers, students, members of the clinical team, staff, administrative staff, patients and families can provide insight into trainees' work habits, capacity for team work, and interpersonal sensitivity in addition to trainee doing a self assessment. The rating scales vary with the assessment context[31,34]. Their use in formative evaluations might be more appropriate since evaluators provide more balanced and honest feedback when the evaluation is formative and used for developmental purposes rather than for pass/fail decisions[36]. The use of 360° evaluations in summative assessment is not advocated until further studies are conducted to establish their reliability and validity[34]. Limitations with this type of evaluation are that it is time consuming and administratively demanding[11,37].

LOG BOOK

Log books are commonly used in training evaluation or by the clinicians for their personal record. In the logbook students or a clinician can keep a record of the patients seen or procedures performed either electronically or in a book. It documents the range of patient care, complications and learning experience. Logbook is very useful in focusing students on important objectives that must be fulfilled within a specified period of time[38].

Logbooks facilitate and monitor students learning, provide a reward system based on competition among peers, encourage immediate and ongoing interaction between the tutors and the students, provide continuous and objective assessment, provide a feedback loop for the evaluation of learning activities, validate the procedural experience at advanced training levels, and involve training centres[38-40].

PORTFOLIOS

Portfolio refers to a collection of one's professional and personal goals, work, achievements, and methods of achieving these goals. Portfolios demonstrate trainees' development and technical capacity and provide evidence that the learning has taken place. It includes documentation of learning and progression, but most importantly a reflection on these learning experiences[41].

Portfolios documentation may include case reports, record of practical procedures performed, videotapes of consultations, project reports, samples of performance evaluations, learning plans, and written reflection about the evidence provided. Scoring methods include checklists and rating scales which are developed for a specific learning and assessment context and are usually carried out by several examiners who probe students regarding portfolio contents and decide whether the student has reached the required standard or not[41,42].

Portfolio assessment is considered a valid way of assessing outcomes. However, due to its wide variability in the way the portfolios is structured and assessed, it has low to moderate reliability. In addition, due to the time and effort involved in its compilation and evaluation this assessment is not considered very practical. Due to these reasons, portfolios are commonly used for formative assessment and less commonly for summative assessment[42]. Due to these reasons, the strength and extent of the evidence base for the educational effects of portfolios in the undergraduate setting is considered limited[43,44,].

RATING SCALES

To assess performance or behavior of a student or clinician rating scales are widely used. These are particularly useful for assessing personal and professional attributes, generic competencies and attitudes. The observer is required to make a judgment along a scale that may be continuous or intermittent. A limitation or problem of rating scales is the low reliability and subjectivity of the judgments. To get more fair results, multiple independent ratings of the same student undertaking the same activity are necessary. It is also important that before conducting assessment the observers should be trained to use the rating forms[26]. Global rating scales are measurement tool for quantifying behaviors. Raters use the scale either by directly observing students or by recalling student performance. Raters judge a global domain of ability for example: clinical skills, problem solving, etc [10].

SCRIPT CONCORDANCE TEST (SCT)

Script Concordance Test (SCT) is a new format which is slowly gaining acceptance in health professions education. This format is designed to test clinical reasoning in uncertain situations[45] and is based on the principle that the multiple judgments made in these clinical reasoning processes can be probed and their concordance with those of a panel of reference experts can be measured[46]. SCTs are based on short case scenarios followed by related questions that are presented in three parts. The first part contains a relevant diagnostic or management option, second part presents a new clinical finding, and third part is a five point Likert scale that captures examinees' decisions as to what effect the new finding has on the status of the option The test has face validity because its content resembles the tasks that clinicians do every day [47].

HOW TO DO ASSESSMENT?

The assessment is an integral component of overall educational activities. Assessment is a comprehensive decision making process having important and broad implications beyond the measure of students' success. Assessment is also related to program or curriculum evaluation because it provides important data to determine the effectiveness of program. It also helps in improvements in teaching program and developing educational concepts[32].

It should be purpose driven and designed prospectively keeping in view the learning outcomes. The assessment methods used must provide a valid and usable data. While devising assessment strategies the principles of assessment were kept clearly in mind. The format, content and frequency of assessment, as well as the timing and format of feedback, should follow from the specific goals of the teaching program of institution. Importantly, the purpose of assessment should direct the choice of instruments used for assessment[3].

It is important that different domains of competence should be assessed in coherent, integrated, and longitudinal fashion with the use of multiple methods with the provision of frequent and constructive feedback. Educators should be aware of the impact of assessment on learning, limitation of each method (including cost), the potential inadvertent effects of assessment and the existing status of the program or institution in which the assessment is occurring.

Needs assessment is the starting point of a good assessment that identifies the current knowledge and skills of the students before the commencement of the actual educational activities. It is used to assess the existing knowledge base, future needs and priority areas that should be addressed[4].

Various assessment tools are available which are appropriate for the different levels of the hierarchy. Van der Vleuten has proposed a conceptual model for defining the utility of an assessment tool[4,48]. In this model several weighted criteria are multiplied conceptually on which the assessment tools can be judged. These criteria are validity (does it measure what it is supposed to be measuring?); reliability (does it consistently measure what it is supposed to be measuring?); educational impact (what are the effects on teaching and learning?); acceptability (is it acceptable to staff, students and other stakeholders?), and cost. So the weighing of the criteria depended on the purpose for which the tool was used[6].

For summative purposes, (i.e. Selection, promotion or certification) the reliability is more important, while for formative purposes, (i.e. diagnosis, feedback and improvement) the educational impact carries more weight in assessment. Similarly, the test of clinical competence, (which allows decisions to be made about medical qualification and fitness to practice) must be designed with respect to key issues including blueprinting, validity, reliability, and standard setting[4,28].

In assessment process, Long essay questions, Short essay questions, MCQs and oral examinations could be used to test applied knowledge, and factual recall. Similarly, to assess clinical performance more sophisticated methods are needed which includes directly observed long and short cases, objective structure clinical examinations (OSCE) and the use of standardized patients. The Objective Structure Clinical examination (OSCE) has been widely adopted as a tool to assess students, or doctor's competences in a range of subjects[27]. It measures outcomes and allows very specific feedback. In this regard, for knowledge, concepts, and application of knowledge ('Knows' and 'Knows How' of Miller's conceptual pyramid for clinical competence context-based MCQ, extended matching item and short answer questions are appropriate. For 'Shows How" multi-station OSCE is feasible. For performance-based assessment ('does') mini-CEX, DOPS is appropriate. Alternatively clinical work sampling and portfolio or log book may be used [33, 48].

ASSESSMENT OF PERFORMANCE

Performance assessment usually divided into two categories; assessment of performance in vitro, i.e. in standardized or simulated conditions, and assessment of performance *in vivo*, i.e. in real conditions. Both categories involve demonstration of a behavior or skill continuously or at a fixed point in time by a student and observation and marking of that demonstration by the examiner. Several tools can be used which comprise of rating scales, checklists, structured and unstructured reports. All these tools can be used to record observations and to assist in the assessment or marking of such demonstrations. Checklists and rating scales are used as scoring methods in various forms of assessments, including Objective Structured Clinical or Practical Examinations (OSCE, OSPE), Direct Observation of Procedural Skills (DOPS), peer assessment, self assessment, and patient surveys[24,28].

The assessment of real performance is that what a doctor do in his real practice i.e. clinical competence, which is the ultimate goal for a valid assessment. The face validity of this "in-training" assessment is excellent but has problems of inadequate reliability which is due to lack of standardization, limited sampling of skills and limited observations. This is a major cause of concern which limits their use as summative "high-stakes" or qualifying examinations. To overcome this issue, assessments in simulated settings which mimic the real conditions should be designed to assess performance such as OSCE/ OSPE[33, 48].

CONCLUSION

Good quality assessment not only satisfies the needs of accreditation but also contributes to student's learning. Assessment methods should match the competencies being learnt and the teaching formats being used. Multiple methods of assessment implemented longitudinally can provide the data that are needed to assess trainees' learning needs and to identify and remediate suboptimal performance by clinicians. Decisions about whether to use formative or summative assessment formats, how frequently assessments should be made, and what standards should be in place remain challenging. Educators also face the challenge of developing tools for the assessment of qualities such as professionalism, teamwork and expertise that have been difficult to define and quantify.

REFERENCES

1. Al-Wardy NM. Assessment Methods in Undergraduate Medical Education. Sultan Qaboos Univ Med J. 2010; 10(2): 203–9.
2. Epstein RM. Assessment in medical education. N Engl J Med. 2007;356(4): 387-96.
3. Tabish SA, Assessment Methods in Medical Education. Int J Health Sci (Qassim). 2008; 2(2): 3–7.
4. Van der Vleuten CP, Schuwirth LW. Assessing professional competence: from methods to programmes. Med Educ. 2005;39:309–17.
5. Schuwirth LW, van der Vleuten CP. Different written assessment methods: what can be said about their strengths and weaknesses? Med Educ. 2004;38:974–9.
6. Norman G. Postgraduate assessment – reliability and validity. Trans J. Coll. Med. S. Afri. 2003;47:71–5.
7. Miller GE. The assessment of clinical skills/ competence/performance. Acad Med. 1990;65:S63–7
8. Wass Cees Van der Vleuten, Shatzer John, Jones Roger. Assessment of clinical competence. The Lancet. 2001;357:945–9.
9. Bloom BS. Handbook I: Cognitive domain. Taxonomy of educational objectives. New York: David McKay; 1956:12-56.
10. Swartz M, Colliver J, Bardes C, Charon R, Fried E, Moroff S. Global ratings of videotaped performance versus global ratings of actions recorded on checklists: a criterion for performance assessment with standardized patients. Acad Med. 1999; 74:1028–32.
11. Wood J, Collins J, Burnside ES, Albanese MA, Propeck PA, Kelcz F, et al. Patient, faculty, and self-assessment of radiology resident performance: a 360-egree method of measuring professionalism and interpersonal/ communication skills. Acad Radiol. 2004; 11:931–9.
12. de Silvia V, Hanvella R, Ponnamperuma G. Validity of oral assessment (viva) that assess specific and unique competencies in a postgraduate psychiatry examination. Sri Lanka J of Psychiatry 2012 ;(2):15-19.
13. Schuwirth LWT, van der Vleuten CP. In: Written Assessments. Dent J, Harden R, editors. New York: Elsevier Churchill Livingstone; 2005. pp. 311–22.
14. Case SM, Swanson DB. Constructing written test questions

for the basic and clinical sciences. Website:[http://www.nbme.org/PDF/ItemWriting_2003/2003IWGwhole.pdf].

15. Knox JD. How to use modified essay questions. Med Teach. 1980; 2:20–4.
16. Swanson DB, Holtzman KZ, Allbee K. Measurement characteristics of Content-Parallel Single-Best-Answer and Extended-Matching Questions in relation to number and source of options. Acad Med. 2008; 83:S21–4.
17. Farmer E, Page G. A practical guide to assessing clinical decision-making skills using the key features approach. Med Educ. 2005; 39:1188–94.
18. Ponnamperuma GG, Karunathilake IM, McAleer S, Davis MH. The long case and its modifications: a literature review. Med Educ. 2009; 43:936–41.
19. Wass V, Jolly B. Does observation add to the validity of the long case? Med Educ. 2001; 35:729–34.
20. Wass V, van der Vleuten C. The long case. Med Educ. 2004; 38:1176–80.
21. Wilkinson TJ, Campbell PJ, Judd SJ. Reliability of the long case. Med Educ. 2008; 42:887–93.
22. Hamdy H, Prasad K, Williams R, Salih FA. Reliability and validity of the direct observation clinical encounter examination (DOCEE) Med Educ. 2003; 37:205–12.
23. Hijazi Z, Premadasa IG, Moussa MA. Performance of students in the final examination in pediatrics: importance of the "short cases." Arch Dis Child. 2002; 86:57–8.
24. Wass V, van der Vleuten C, Shatzer J, Jones R. Assessment of clinical competence. Lancet. 2001; 357:945–9.
25. Epstein RM. Assessment in medical education. Author's reply. New Eng J Med. 2007;356:2108–10.
26. Davis MH, Ponnamperuma GG. In: Work-based Assessment. Dent J, Harden R, editors. New York: Elsevier Churchill Livingstone; 2005. pp. 336–45.
27. Newble D. Techniques for measuring clinical competence: objective structured clinical examinations. Med Educ. 2004; 38(2):199-203.
28. ACGME Outcome Project, Accreditation Council for Graduate Medical Education (ACGME) and American Board of Medical Specialist (ABMS) Toolbox of assessment methods. 2000 website [http://www.acgme.org/outcome/assess/toolbox.asp].
29. Marks M, Humphrey-Murto S. In: Performance Assessment. Dent J, Harden R, editors. New York: Elsevier Churchill Livingstone; 2005. pp. 323–35.
30. Norcini JJ, Blank LL, Duffy FD, Fortna GS. The mini-CEX: a method for assessing clinical skills. Annals Internal Med. 2003;138:476–83.
31. Rees C, Shepherd M. The acceptability of 360-degree judgments as a method of assessing undergraduate medical students' personal and professional behaviors. Med Educ. 2005;39:49–57.
32. Saedon H, Salleh S, Balakrishnan A, Imray CHE, Saedon M. The role of feedback in improving the effectiveness of workplace based assessments: a systematic review. BMC Medical Education 2012;12:25.
33. Noreini JJ, McKinley DW. Assessment methods in medical education. Teachers & Teaching Educ 2009;23(3):239-50.
34. Office of Postgraduate Medical Education. Review of work-based assessment methods. Sydney: University of Sydney, NSW, Australia; 2008.
35. Regehr G, MacRae H, Reznick R, Szalay D. Comparing the psychometric properties of checklists and global rating scales for assessing performance on an OSCE format examination. Acad Med. 1998;73:993–7.
36. Higgins RS, Bridges J, Burke JM, O'Donnell MA, Cohen NM, Wilkes SB. Implementing the ACGME general competencies in a cardiothoracic surgery residency program using 360-degree feedback. Annals Thoracic Surg. 2004;77:12–17.
37. Joshi R, Ling FW, Jaeger J. Assessment of a 360-degree instrument to evaluate residents' competency in interpersonal and communication skills. Acad Med. 2004;79:458–63.
38. Fatemeh K, Alavinia SM. Students perception about logbook: advantages, limitations and recommendations: a qualitative study. JPMA 2012;1184(62):324-9.
39. Patil NG, Lee P. Interactive logbooks for medical students: are they useful? Med Educ 2002; 36: 672-7.
40. Mohammadi A, Khaghanizadeh M, Ebadi A, Mohammadi A, Amiri F, Raeisifar A. Log book; A method of evaluating education and feedback strategy in nursing. Iran J Educ Strateg Spring 2010;3:41-5.
41. Davis MH, Ponnamperuma GG. In: Portfolios, projects and dissertations. Dent J, Harden R, editors. New York: Elsevier Churchill Livingstone; 2005. pp. 346–56.
42. Rees C, Sheard C. The reliability of assessment criteria for undergraduate medical students' communication skills portfolios: The Nottingham experience. Med Educ. 2004;38:138–44.
43. Buckley S, Coleman J, Davison I, Khan KS, Zamora J, Malick S, et al. The educational effects of portfolios on undergraduate student learning: a Best Evidence Medical Education (BEME) systematic review. BEME Guide No. 11. Med Teach. 2009;31:279–81.
44. Thistleth waite J. How to keep a portfolio. Clin Teach. 2006;3:118–23.
45. Charlin B, van der Vleuten CP. Standardized assessment of reasoning in context of uncertainty. The Script Concordance Test approach. Eval Health Profess. 2004;27:304–19.
46. Hall KH. Reviewing intuitive decision-making and uncertainty: the implications for medical education. Med Educ. 2002;36:216–24.
47. Fournier JP, Demeester A, Charlin B. Script Concordance Tests: Guidelines for Construction. BMC Med Inform Decis Mak. 2008;8:18.
48. Dijkstra J, Van der Vleuten CPM, Schuwirth LWT. A new framework for designing programmes of assessment. Adv Health Sci Educ Theory Pract.2010; 15(3): 379-93.