special communication

Objective structured clinical examinations as an assessment method in residency training: practical considerations

Mohammed Hijazi,[a] Steven M. Downing[b]

From the [a]Department of Medicine, King Faisal Specialist Hospital and Research Centre, Riyadh, Saudi Arabia and the [b]Department of Medical Education, University of Illinois, Chicago, USA

Correspondence and reprints: Mohammed Hijazi, MD, FCCP · Director, Quality Resource Management Department, Department of Medicine, MBC 46, King Faisal Specialist Hospital and Research Centre · PO Box 3354, Riyadh 11211, Saudi Arabia · T: +966-1-442-4731 F: +966-1-442-7499 · mhijazi@kfshrc.edu.sa · Accepted for publication October 2007

Ann Saudi Med 2008; 28(3): 192-199

The assessment of residents in training relies mostly on testing their knowledge by written examinations (multiple-choice tests), assessing their competence using clinical examinations (long and short case oral examinations), and on rating their performance using end of rotations rating forms. Unfortunately, the assessment of competence using clinical examinations has its limitations because of low reliability and validity. Because of improved reliability and validity, the objective structured clinical examination (OSCE) is a more accurate means of assessing the clinical competence of residents. The objective of this review is to provide a practical guide to the role and use of the OSCE as an assessment tool and includes a description of its format, advantages, disadvantages, organization, standards setting, reliability, validity, and utility. The OSCE provides a more reliable assessment of trainee competency compared to the traditional clinical examination. It complements other methods commonly used such as multichoice questions, and end of rotation ratings of performance. Ensuring proper sampling of the competencies to be assessed and an adequate number of stations, proper training of examiners and standardized patients is important in obtaining reliable and valid assessments. Setting standards by using the borderline approach is practical and accurate when compared with other absolute methods. The OSCE needs to be made part of the assessment process for residents in training.

Accurate assessment of trainees is essential to provide them with feedback, to make decisions about their advancement from one level of training to the next and about their successful completion of training. It provides important information to improve the training process and to identify poorly performing trainees that need help. Moreover, accurate assessment protects the community from incompetent trainees that do not improve despite proper training. The assessment of trainees clinical competence and performance using the traditional oral clinical examination and rating of directly observed performance during training suffer from subjectivity and low reliability.[1,2] (Competency-based assessment measures what doctors can do in controlled representations of professional practice, while performance-based assessment measures what doctors do in actual professional practice). This subjectivity and low reliability has caused many training bodies to stop using the traditional clinical examination and search for a more reliable and objective assessment method.[1-4] The OSCE is superior to the oral clinical examination because it overcomes the problem of case specificity by sampling a broad area of competency, resulting in better reliability and validity.[3] Acceptance of the OSCE in the assessment of the clinical competence of residents requires a sound understanding of its format, advantages, disadvantages, organization, standard setting, reliability, validity, and utility. The objective of this paper is to provide an overview and practical guide to the use of the OSCE in resident's assessment, including organization, standard setting, reliability, and validity.

Is OSCE needed in the assessment of residents?

Comprehensive assessment of professional competence, defined as "the habitual and judicious use of communication, knowledge, technical skills, clinical reasoning, emotions, values, and reflection in daily practice for the benefit of the individual and the community being served", needs multiple assessment methods.[5] Most training programs rely on multiple-choice question tests, an end of rotation subjective rating by supervising physicians, and an oral clinical examination (short and long cases) in the assessment of their trainees. While well prepared written multiple-choice tests are

useful for assessing trainee's cognitive abilities (knows and knows how) in a reliable and valid way, the results of the end of rotation subjective rating and oral clinical examinations are less reliable and less valid means for the assessment of clinical performance (does) and clinical competence (shows how), respectively.[5,6] On the other hand, the OSCE has been shown to produce more reliable assessment data.[5,6] It assesses physician performance under examination conditions (competency-based assessment, covering the area of shows how of the Miller's pyramid of assessment, Figure 1) which is a prerequisite for physician performance in real life.[7] The OSCE can be a useful component of multi-method assessment, complementing multiple-choice tests and subjective ratings.[5] It obviates the need for the less reliable assessment data obtained by oral clinical examinations.

What is an OSCE?

An OSCE is a series of timed (5 to 10 minutes) stations (ranging from 8 up to more than 20) through which examinees are assessed by one or more examiners while performing a standardized clinical task during a patient examination or standardized patient interaction using a well-defined structured marking sheet.[2,3] The clinical task can be history taking, clinical examination, data interpretation, management, communication skills, counseling, and technical skills.[5] Marking is done using a task-specific checklist, rating scale, or a combination of both.

Other variants of the OSCE exist. Examples are longer stations (15 to 30 minutes), stations with data, multiple choice question response or short written response, and stations were there are no examiners and the marking is made by trained standardized patients. Since its introduction by Harden in the 1970s, the OSCE became widely used in the assessment of medical students.[8] Its use in the assessment of residents and other health care professionals is increasing as well.

Advantages and disadvantages of the OSCE

The main advantage of OSCE is the fact that it allows a sampling of multiple areas of clinical competence compared to the traditional oral clinical examination, overcoming the problem of case specificity and resulting in improved reliability.[2,3] Marking is done in a standardized way to increase interrater agreement. It provides a flexible assessment method through the use of standardized patients.[2]

Logistically, the need for more time, physical space and personnel to organize the OSCE might be an obstacle. It assesses competence (shows how) rather than performance (does).[2] During a short station, trainees perform one aspect of a patient-physician encounter, which fragments the clinical encounter.[2] The use of checklists during marking puts more emphasis on thoroughness, which might become less relevant in advanced level trainees.[2] Moreover, not all clinical situations can be simulated by standardized patients and not all components of a clinical task can be captured by a checklist. Trainee's ethics and behavior need to be observed during training and cannot be reliably assessed using OSCE.

Practical steps for developing the OSCE in residency training

Developing the OSCE for the assessment of clinical competence during residency training is a major task that needs proper planning. It should start a few months before conducting the examination. Practical steps that can be used to guide the process are:

1. Establish support from the department and accrediting body by making the OSCE an essential part of the assessment.
2. Form an organizing team. Ensure an appropriate mix of specialties and expertise.
3. Train the organizing team on the development of the OSCE (workshop).
4. Establish the frequency, timing and purpose of conducting the OSCE during training (end of rotation, annual, end of training, formative assessment, or summative assessment).
5. Develop a blueprint that captures the clinical competencies to be assessed in relation to the objectives of the residency training. This is an important step

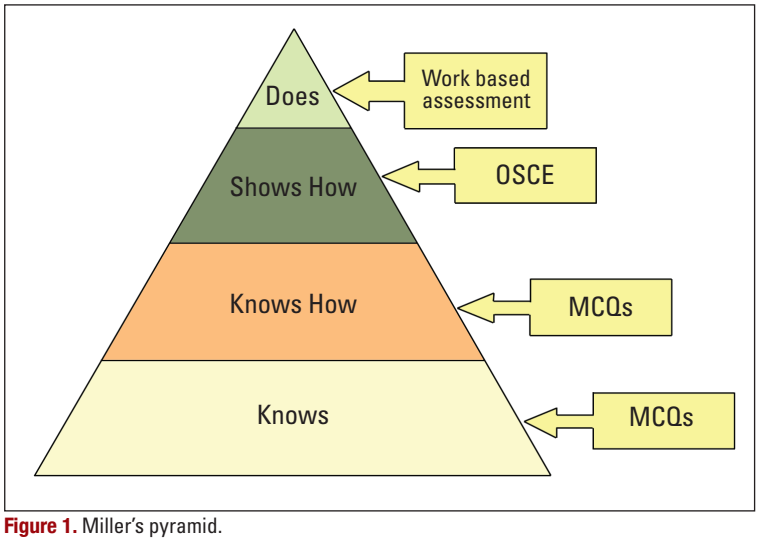

Figure 1. Miller's pyramid.

to ensure adequate sampling of all the essential competencies (content validity). The blueprint can be developed as a simple two-dimensional matrix with the competency (history taking, physical examination, data interpretation, communication, management decisions, patient education, performing procedures) to be assessed on one axis and the clinical situation (patient complaints like chest pain or a clinical problem like anemia) during which the competency is to be demonstrated on the other axis. The clinical situations need to be based on the common problems encountered during the clinical rotations as judged by the training program director and the organizing team (experts).

6. Develop the stations: Decide on the number and duration of stations per examination. Sample and select the stations from the blueprint (Tables 2 and 3), train standardized patients as needed, and design the station sheets (stem, instructions to examiners, instructions to examinees, and marking sheet). In general, the higher the number of stations, the better is reliability. Deciding on the number of stations remains a matter of balancing logistic constraints and reliability. An average of 20 stations is likely to result in reasonable reliability.[2] Using 5-minute stations will allow more stations per examination and a sampling of broader areas of competence, thus improving reliability.[2] Sampling from the blueprint should consider covering all the competencies in the setting of diverse common clinical situations. Systems that are more represented in the training might be assessed by multiple stations. Once the number and the type of stations are decided, each station should be developed with particular attention on: 1) providing a clear definition of the task (stem of the station) to be performed by the student, 2) providing clear instructions to the examiner, standardized patient and examinees; 3) designing the marking sheet, and 4) providing a list of requirements (X ray viewer, ophthalmoscope, tendon hammer, etc.). An example of a station instruction sheet is shown in Table 4. Using a standardized format for the stem that states the patient name and age, presenting problem, clinical setting (emergency department, inpatient, ICU,

Table 1. Source of and threats to validity and methods to minimize threats.[16,17]

Sources of validity evidence	Threats to validity	Methods to minimize threats
Content: - Examination blueprint - Representativeness of blueprint to achievement domain - Representativeness of OSCE to domain - Quality of stations - Qualification of OSCE developer - Sensitivity analysis **Response process:** - Format familiarity - Accuracy of scoring, combining scores, and final score - Standard setting - Accuracy of reporting final results - Accurate description of final scores **Internal structure:** - Reliability **Relationship to other variables:** - Correlation with other variable (similar of dissimilar). - Generalizability **Consequences:** - Impact of results on trainee and society - Methods for establishing pass/fail scores - Instructional/learner consequences	**Construct under representation (CU):** - Using few cases or - Unrepresentative cases of the domain of achievement to be tested - Unstandardized patients and rating - Low reliability of rating **Construct irrelevant variance (CIV):** - Flawed cases - Flawed check lists - Flawed rating scale - Poorly trained standardized patient - In appropriate case difficulty - Rater bias - Bluffing of SP - Inaccurate standard setting	- Adequate development of blueprint by experts - Use OSCE stations the adequately sample the domain of interest - Use adequate number of OSCE stations to ensure generalizability of the results (possibly > = 12) - Ensure adequate quality of each station and development by an experienced person - Use well trained SP - Use well developed and clear check lists - Train raters well - Pretest OSCE stations before its use - Ensure reliability (generalizability coefficient > 0.8 for high stake OSCE) - Ensure high quality of data collection and processing - Use defensible standard setting - Ensure clear reporting mechanism

Table 2. Sampling stations in the OSCE.

	History	Examination	Date interpretation/ Workup	Differential Diagnosis	Manage-ment	Procedures skills	Com-munication	Patient education	EBM	System based practice	Others
Card	X		X		X						
Pul		X		X							
ID	X									X	
End					X						
ICU					X				X		
Neur		X									
Hem							X				
ER						X		X			
Rh				X							
GIT		X	X								
Nep			X								
GIM	X										
Onc									X		
All											
Der											
Mis											

Card=Cardiology, Pul=Pulmonary, ID=Infectious Diseases, End=Endocrinology, ICU=Intensive care, Neur=Neurology, Hem=Hematology, ER=Emergency Medicine, Rh=Rheumatology, GIT=Gastroenterology, Nep=Nephrology, GIM=General Internal Medicine, Onc=Oncology, All=Allergy, Der=Dermatology, Mis=Miscellaneous, EBM=Evidence-based Medicine

Table 3. Selecting stations in the OSCE.

	History	Examination	Date interpretation/ Workup	Differential Diagnosis	Manage-ment	Procedures skills	Com-munication	Patient education	EBM	System based practice	Others
Card	Chest pain		ECG		Acute MI						
Pul		Chest (Pneumonia, IPF)		Chronic cough							
ID	Fever										
End					Uncontrolled DM						
ICU					Hypotension						
Neur		Sudden weakness									
Hem							Bad news				
ER						Airway		MDI technique			
Rh				Acute knee arthritis							
GIT		Abdominal distension	Jaundice Hepatitis								
Nep			Oliguria Nephritis								
GIM	Weight loss										
Onc									Prognosis		
All											
Der											
Mis											

Card=Cardiology, Pul=Pulmonary, ID=Infectious Diseases, End=Endocrinology, ICU=Intensive care, Neur=Neurology, Hem=Hematology, ER=Emergency Medicine, Rh=Rheumatology, GIT=Gastroenterology, Nep=Nephrology, GIM=General Internal Medicine, Onc=Oncology, All=Allergy, Der=Dermatology, Mis=Miscellaneous, EBM=Evidence-based Medicine

Table 4. Objective structured clinical examination sheet.

OSCE station number…… **Examiner name** **Examinee name**
STEM Patient name, age, presenting problem (chest paint), clinical setting (ER), duration Task to be performed: obtain history related to chest pain over the coming 5 minutes
CHECK LIST **For each item assign "0" if not done, "1" if done inappropriately, or "2" if done appropriately** ____ Introduced him/her self and asked permission from the patient ____ Asked about the location of the pain ____ Asked about the onset of the pain ____ Asked about the character of the pain ____ Asked about relation to activities ____ Asked about the severity of the pain ____ Asked about radiation ____ Asked about relieving and aggravating factors ____ Asked about associated symptoms (sweating, SOB, palpitation, dizziness, cough, fever) ____ Asked about risk factors for ischemic heart disease
Global rating **(Please circle based on the overall impression about the resident performance)** Pass Borderline Fail

clinic), the task to be performed, and over how long, throughout the stations is to be encouraged.[2] The marking sheet can be a task-specific checklist, global rating scale (pass, borderline, fail), or a combination of both. The number of items per checklist depends on the task performed. It has been shown that fewer items per checklist is associated with better reliability and validity.[9] A score must be assigned for each item on the checklist. The score can be zero for not done, one for inappropriately done, and one for appropriately done.

7. Select the examiners, standardized patients, and supportive staff. Train the examiners on using the marking sheet (training, workshops), and the standardized patient on the specific task to be performed during the station. Examiners need to pay attention to recording the examinees names on each sheet and to complete the checklist and rating scale. Examiner training and commitment to the whole process are important to minimize variation in rating and improve objectivity.[10] Also important is the proper training of standardized patients and piloting the stations before the examination.
8. Select the location and date of the examination and set the OSCE schedule. Communicate the date with examiners, examinees, patients, and organizing staff. Develop a map for the stations to guide the examinees and organizers during the examination. Allow time between stations (one to two minutes).
9. Check that everything is in place before the examination using a checklist and ensure security by restricting access to the examination location. Although ensuring security is important, it has been shown that knowing the OSCE stations before the examination does not result in scores that are different from the scores of those who did not know the stations.[11]
10. Conduct the examination. Time control is very important to avoid disturbing the flow of the examination because of the consecutive nature of the station. Using an alarm or a stopwatch or a timekeeper for every 5 stations are ways of maintaining time control. Collect the entire marking sheet from each station examiner after checking it for completeness (especially examinees name). Enter the data from the marking sheets into an electronic database that will help in performing the needed analysis.

Setting standards

Setting standards (setting defendable passing score and grades) on an examination is a matter of judgment and requires the use of systematic methods. Coming up with defensible passing scores requires the use of an acceptable method and the selection of experienced, qualified, and unbiased judges. The standard can be set using relative (norm-referenced), absolute (criterion referenced), or compromise (Beuk, Hofstee) methods.

While a norm-referenced standard setting (for example, the fixed percentage method by passing the highest 90% of examinees) might be used for low-stakes OSCE intended for formative assessment (end of rotation), criterion-referenced methods are more appropriate for competency-based assessment, particularly high-stakes examinations.[3,12,13]

Absolute standard setting approaches are either examination centered (like the Nedelsky, Eble, Jaeger, Angoff and the modified Angoff methods) or examinee centered (like the contrasting groups and borderline group methods), all of which can be used for standard setting in an OSCE.[14] In the Angoff method, a number of judges review all the checklist items of OSCE stations and decide on the score of the borderline examinees on each item (similar to the contrasting groups method).[14] The average score of the judges is used as the pass score. In the modified Angoff method, the judges decide on the performance of the borderline examinees at the OSCE station level rather than the item level.[14] While supported by a lot of research, this method is time consuming and difficult.[14]

The contrasting groups and borderline methods are widely used for deciding the pass score on both small-scale and large-scale OSCE conducted by medical schools and by the Medical Council of Canada.[13] It provides results similar to the Angoff method while being simpler.[3] In the contrasting groups and borderline approaches, the examiners rate the examinees as pass, borderline, or fail (other 4- to 6-point scale could be used as well: outstanding, excellent, borderline pass, borderline fail, poor, and inadequate). This is done on each station in addition and independent of the checklist scoring. The intersection of the scores of the borderline and pass groups constitutes the pass mark in the contrasting groups methods, while the mean of all borderline scores on a given station will be the pass mark for the station. The sum of the pass marks for all the stations will be the pass mark for the examination in the borderline group method.[3,13,14] Such methods have been shown to produce reliable and valid passing scores and have the advantage of being easier to conduct compared to the Angoff method.[13,14] It is essential, although frequently difficult, to have enough borderline examinees.

Using a fully compensatory method (passing the examination depends on the final total score regardless of the number of stations passed) versus passing a certain number of stations to make pass/fail decisions is controversial.[3] Using a fully compensatory mechanism is more practical and has been shown to produce more reliable results.[13]

Standards need to be set by experienced, unbiased, and qualified judges. Using two or more groups of judges for the rating allows for measuring the agreement between their ratings (reliability) and increases the credibility of the standards. The pass/fail scores need to be acceptable to stakeholders. Comparing the OSCE assessment results based on the level of training and with the results of other assessment methods during training helps to support defensibility of the standard.

Using the results

The results of an OSCE can be used for a formative assessment during training or as a component of a summative assessment. As a formative assessment, it will help to provide feedback for trainees and to the training program director about the areas that need to be improved. As a summative assessment, it helps to decide the successful completion of a rotation or a training level, and in making decisions about board certification. It can be used to complement other assessment methods like written multiple choice tests and global ratings during training.

Reliability

Reliability refers to the reproducibility of assessment data or scores over time. It is a quality of the outcome or results and not the assessment instrument itself. Reliability estimates the random error of measurement.[15] Low reliability indicates that large variation in results is expected upon retesting (unacceptably large number of trainees passed or failed might get different results upon retesting without much change in their knowledge or skills), making the results of the assessment useless. It is an important source of validity evidence and allows for generalization of the assessment results to the domain of the competency assessed.[15]

For the OSCE, reliability can be assessed using generalizability theory to account for all measurement errors (examiners, cases, examinees, and standardized patients). The OSCE stations should be used for the assessment of reliability and not the checklist items in the stations.[15] The number of stations needed depends on how much reliability is acceptable based on the intended use of the assessment results. A higher number of OSCE stations is required to achieve a higher level of reliability. In general, acceptable reliability for high-stakes (board certification), moderate-stake (end of course, end of year summative assessment), and low-stakes (in training, formative assessment) OSCE are more than 0.9, from 0.8 to 0.89, and 0.7 to 0.79, respectively.[15]

The reliability coefficient can be used to estimate the standard error of measurement (SEM), to create confi-

dence bands around the scores, and to calculate the reliability of the pass-fail decision. Reliability of the OSCE can be improved by using a higher number of stations, using stations of medium difficulty, and testing examinees with heterogeneous abilities.[15]

Validity

Validity refers to the accumulation of evidence that supports meaningful interpretation of the assessment results.[16] Without evidence of validity, OSCE results cannot be interpreted. Validity is a unitary concept that requires multiple source of supporting evidence.[16] Common sources of validity evidence (Table 1) are content, response, relationship to other variables, internal structure, and consequences. Attention needs to be given especially to collecting such evidence during the process of designing and conducting the OSCE. In general, the higher-stake OSCE needs the collection of more validity evidence to support interpretation (board certification).

Threats to validity evidence that can affect the interpretation of the OSCE results are many. The major threat to validity evidence of the OSCE is construct underrepresentation due to undersampling (few OSCE stations) or improper sampling of contents.[17] This can be avoided by carefully developing blueprints that cover all the contents and competency to be assessed and by selecting the OSCE that samples common clinical problems and that covers all such content and competency areas. Involving experts in the blueprint development and sampling process will help to minimize this threat. The other threat to validity evidence is construct irrelevant variance (CIV) due to improper training of standardized patients (SP), flawed SP and checklist, too easy or too difficult cases, bluffing of SP, rater bias (central tendency and halo effect), and indefensible pass criteria.[17]

Conclusion

OSCE enables the assessment of resident's competence in a more reliable and valid way compared to the traditional clinical examination. A multimethod approach to assessment using multichoice questions, OSCE, and performance-based assessment (rating based on observation at work by peers, patients and supervisors) is the way to a more accurate assessment.

REFERENCES

1. Wilkinson J, Benjamin A, Wade W. Assessing the performance of doctors in training. British Medical Journal 2003;327(7416):91-2.
2. Smee S. Skill based assessment. British Medical Journal 2003;326(7391): 703-6.
3. Newble D. Techniques for measuring clinical competence: objective structured clinical examinations. Medical Education 2004;38(2):199-203.
4. Norcini JJ. The death of the long case? British Medical Journal (2002);324(7334):408-9.
5. Epstein RM, Hundert EM. Defining and assessing professional competence. Journal of the American Association 2002;287(2):226-35.
6. Miller G. The assessment of clinical skills/competence/performance. Academic Medicine 1990;65:563-7.
7. Rethans JJ, Norcini JJ, Baron-Maldonado M, Blackmore D, Jolly BC, LaDuca T, et al. The relationship between competence and performance: implications for assessing practice performance. Medical Education 2002;36(10):901-9.
8. Harden RM, Gleeson FA. Assessment of clinical competence using an objective structured clinical examination (OSCE). Medical Education 1979;13(1): 41-54.
9. Wilkinson TJ, Frampton CM, Thompson-Fawcett M, Egan T. Objectivity in Objective Structured Clinical Examinations: Checklists Are No Substitute for Examiner Commitment. Academic Medicine 2003;78(2):219-23.
10. Wilkinson TJ, Frampton CM, Thompson-Fawcett M, Egan T. Objectivity in objective structured clinical examinations: checklists are no substitute for examiner commitment. Academic Medicine 2003;78(2):219-23.
11. Wilkinson TJ, Fontaine S, Egan T. Was a breach of examination security unfair in an objective structured clinical examination? A critical incident. Medical Teacher 2003:25(1)42-6.
12. Boulet JR, De Champlain AF, McKinley DW. Setting defensible performance standards on the OSCE and standardized patient examinations. Medical Teacher 2003;25(3):245-9.
13. Wilkinson TJ, Newble DI, Frampton CM. Standard setting in an objective structured clinical examination: use of global ratings of borderline performance to determine the passing score. Medical Education 2001;35(11):1043-9.
14. Norcini JJ. Setting standards on educational tests. Medical Education 2003; 37(5):464-9.
15. Downing SM. Reliability: on the reproducibility of assessment data. Med Edu. 2004;38:1006-12.
16. Downing SM. Validity: on meaningful interpretation of assessment data. Medical Education 2003;37(9):830-7.
17. Downing SM, Haladyna TM. Validity threats: overcoming interference with proposed interpretations of assessment data. Medical Education 2004:38(3); 327-33.