# *ABC of learning and teaching in medicine*
# Skill based assessment

Sydney Smee

Skill based assessments are designed to measure the knowledge, skills, and judgment required for competency in a given domain. Assessment of clinical skills has formed a key part of medical education for hundreds of years. However, the basic requirements for reliability and validity have not always been achieved in traditional "long case" and "short case" assessments. Skill based assessments have to contend with case specificity, which is the variance in performance that occurs over different cases or problems. In other words, case specificity means that performance with one patient related problem does not reliably predict performance with subsequent problems.

For a reliable measure of clinical skills, performance has to be sampled across a range of patient problems. This is the basic principle underlying the development of objective structured clinical examinations (OSCEs). Several other structured clinical examinations have been developed in recent years, including modified OSCEs—such as the Royal College of Physicians' Practical Assessment of Clinical Examination Skills (PACES) and the objective structured long case (OSLER). This article focuses mainly on OSCEs to illustrate the principles of skill based assessment.

Written tests can assess knowledge acquisition and reasoning ability, but they cannot so easily measure skills

## OSCEs

The objective structured clinical examination (OSCE) was introduced over 30 years ago as a reliable approach to assessing basic clinical skills. It is a flexible test format based on a circuit of patient based "stations."

At each station, trainees interact with a patient or a "standardised patient" to demonstrate specified skills. Standardised patients are lay people trained to present patient problems realistically. The validity of interactions with real patients, however, may be higher than that with standardised patients, but standardised patients are particularly valuable when communication skills are being tested.

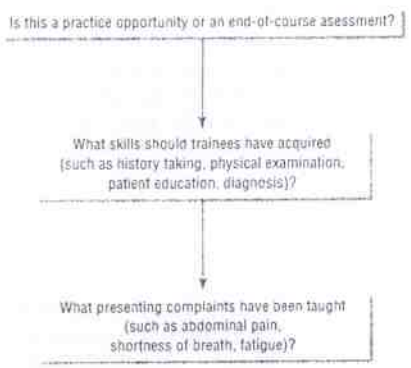
OSCE stations may be short (for example, five minutes) or long (15-30 minutes). There may be as few as eight stations or more than 20. Scoring is done with a task specific checklist or a combination of checklist and rating scale. The scoring of the students or trainees may be done by observers (for example, faculty members) or patients and standardised patients.

Patient-doctor interaction for assessing clinical performance

### Design

The design of an OSCE is usually the result of a compromise between the assessment objectives and logistical constraints; however, the content should always be linked to the curriculum, as this link is essential for validity.

Using many short stations should generate scores that are sufficiently reliable for making pass-fail decisions within a reasonable testing time. (Whether any OSCE is sufficiently reliable for grading decisions is debatable.) Fewer but longer stations maximise learning relative to the selected patient problems, especially when students or trainees receive feedback on their performance. The number of students, time factors, and the availability of appropriate space must also be considered.




Questions to answer when designing an OSCE

### Planning

Planning is critical. Patients and standardised patients can be recruited only after stations are written. Checklists must be reviewed before being printed, and their format must be compatible with the marking method, ideally computerised. OSCEs generate a lot of data—for 120 students in a 20 station OSCE there will be 2400 mark sheets!)

Stations are the backbone of an OSCE, and yet the single most common problem is that station materials are incomplete and subject to last minute changes. The result is increased cost and wasted time.

If OSCE scores are being used for making pass-fail decisions, then it is also necessary to set standards. Several methods for setting standards have been used, with the Angoff method described below being the most commonly used.

Plans should allow sufficient time to process and analyse the scores carefully.

### Costs

OSCE costs vary greatly because the number of stations determines the number of standardised patients, examiners, and staff required. Whether or not faculty members volunteer to write cases, set standards, and examine is also a significant factor.

**Tasks to do ahead**

- Create blueprint
- Set timeline (how long do we need?)
- Get authors for a case-writing workshop
- Review and finalise cases
- Arrange workshop on setting standards
- Recruit standardised patients; recruit faculty members as examiners
- Train standardised patients
- Print marking sheets, make signs
- List all supplies for set-up of OSCE stations
- Remind everyone of date
- Make sure students have all the information
- Plans for the examination day: diagram of station layout; directions for examiners, standardised patients, and staff; possible registration table for students; timing and signals (for example, stopwatch and whistles); procedures for ending the examination
- Anything else?

**The fixed costs of running an OSCE remain much the same regardless of the number of examination candidates. Administering an OSCE twice in one day only slightly increases the fixed costs.**

### Developing the stations

OSCE stations have three components.

[illegible] example, providing the patient's name, age, presenting complaint, and the setting (such as clinic, emergency, or ward) for all stations. The stem must clearly state the task—for example, "in the next eight minutes, conduct a relevant physical examination."

*Checklist*

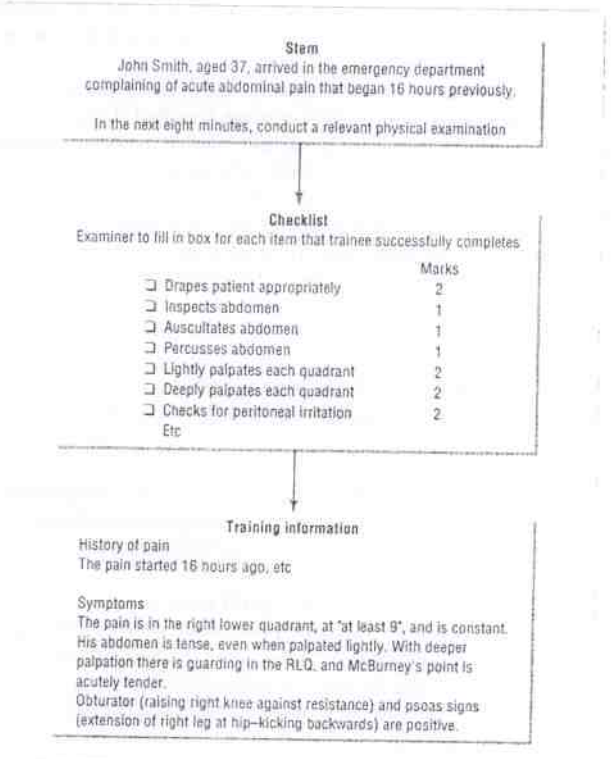
The checklist items are the actions that should be taken in response to the information in the stem. These items should be reviewed and edited to ensure that *(a)* they are appropriate for the level of training being assessed, *(b)* they are task based, and *(c)* they are observable (so the observer can score them).

The length of the checklist depends on the clinical task, the time allowed, and who is scoring. A checklist for a five minute station that is testing history taking may have up to 25 items if a faculty observer is doing the scoring. If a patient or standardised patient is doing the scoring, then fewer items should be used. Use of detailed items will guide scorers: for example, "examines the abdomen" is a general item that might better be separated into a series of items such as "inspects the abdomen," "auscultates the abdomen," "lightly palpates all four quadrants," and so on.

A score must be assigned to every item. Items may be scored 1 or 0, or relative weights may be assigned, with more critical items being worth more. Weights may not change the overall pass-fail rate of an OSCE, but they may improve the validity of a checklist and can affect which trainees pass or fail.

*Training information*

For standardised patients, directions should use patient based language, specify the patient's perception of the problem (for example, serious, not serious), provide only relevant information, and specify pertinent negatives. Responses to all checklist items should be included. The patient's behaviour and affect should be described in terms of body language, verbal tone, and pace. Symptoms to be simulated need to be described.




Components of OSCE station

### Limitations

Skill based assessments are based on tasks that approximate performance in the area of interest. The assumption is that the closer the tasks are to "real world" tasks, the more valid the assessment.

Three aspects of an OSCE limit how closely the stations approximate clinical practice. Firstly, time limited stations often require trainees to perform isolated aspects of the clinical encounter. This deconstructs the doctor-patient encounter and may be unacceptable for formative assessments. The trade-off is that limiting the time allows for more stations, which can provide performance snapshots that allow for reliable, summative decision making.

Secondly, OSCEs rely on task specific checklists, which assume that the doctor-patient interaction can be described as a list of actions. As a result, checklists tend to emphasise thoroughness, and this may become a less relevant criterion as the clinical experience of candidates increases. Thirdly, there are limits to what can be simulated, and this constrains the nature of the patient problems that can be sampled. Again, this becomes more of an issue as candidates' level of training and clinical experience increases.

**Limitations of OSCEs**

- Stations often require trainees to perform isolated aspects of the clinical encounter, which "deconstructs" the doctor-patient encounter
- OSCEs rely on task specific checklists, which tend to emphasise thoroughness. But with increasing experience, thoroughness becomes less relevant
- The limitations on what can be simulated constrain the type of patient problems that can be used

**None of these limitations is prohibitive, but they should be considered when selecting an OSCE as an assessment tool and when making inferences from OSCE scores**

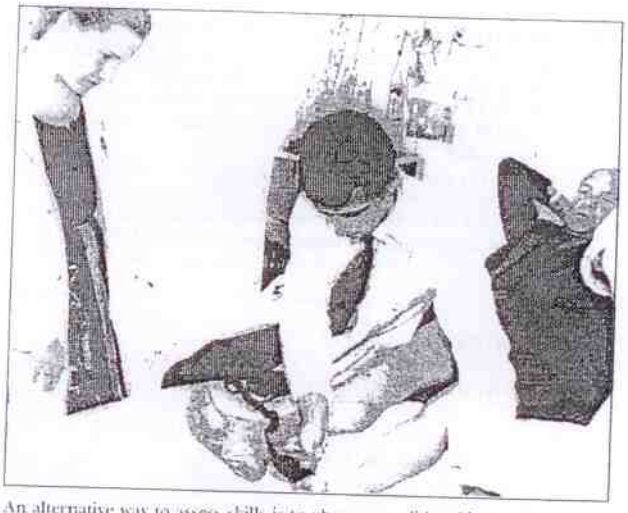
## Other approaches to skill based assessment

### Traditional approaches

The oral examination (also known as the "viva") and the "long case" have long been used for assessing clinical competence. The oral examination is traditionally an unstructured face to face session with the examiners. This allows them to explore the trainee's understanding of topics deemed relevant to clinical practice. The long case is patient based, but the interaction with the patient is usually not observed. Instead, trainees summarise the patient problem for the examiners and respond to examiners' questions about findings, diagnosis or management, and other topics deemed relevant by examiners. The strength of the long case is the validity that comes from the complexities of a complete encounter with a real patient. However, the difficulty and relevance of these assessments varies greatly as the content is limited to one or two patient problems (selected from the available patients), and decisions are made according to unknown criteria, as examiners make holistic judgments. For this reason traditional unstructured orals and long cases have largely been discontinued in North America.

### Alternative formats

Alternative formats tackle the problems associated with traditional orals and long cases by *(a)* having examiners observe the candidate's complete interaction with the patient, *(b)* training examiners to a structured assessment process, and/or *(c)* increasing the number of patient problems. For a short case assessment, for example, one or two examiners may direct a trainee through a series of five or six encounters with real patients. They observe, ask questions, and make a judgment based on the candidate's performance with all the patients. Similarly, a structured oral examination is still a face to face session with examiners, but guidelines for the topics to be covered are provided. Alternatively, a series of patient scenarios and agreed questions may be used so that the content and difficulty of the assessment is standardised across the trainees. Each of these adaptations is aimed at improving reliability, but the most important improvement comes from greatly increasing the number of patient problems, which may well cause an impractical increased testing time.



An alternative way to assess skills is to observe candidates' interaction with patients

## Reliability and validity

The reliability of a test describes the degree to which the test consistently measures what it is supposed to measure. The more reliable a test, the more likely it is that a similar result will be obtained if the test is readministered. Reliability is sensitive to the length of the test, the station or item discrimination, and the heterogeneity of the cohort of candidates. Standardised patients' portrayals, patients' behaviour, examiners' behaviour, and administrative variables also affect reliability.

The validity of a test is a measure of the degree to which the test actually measures what it is supposed to measure. Validity is a property of test scores and justifies their interpretation for a specific purpose. The most basic evidence of validity comes from documenting the links between the content of the assessment and the curriculum's objectives and from the qualifications of those who develop the assessment.

**Factors leading to lower reliability**

- Too few stations or too little testing time
- Checklists or items that don't discriminate (that is, are too easy or too hard)
- Unreliable patients or inconsistent portrayals by standardised patients
- Examiners who score idiosyncratically
- Administrative problems (such as disorganised staff or noisy rooms)

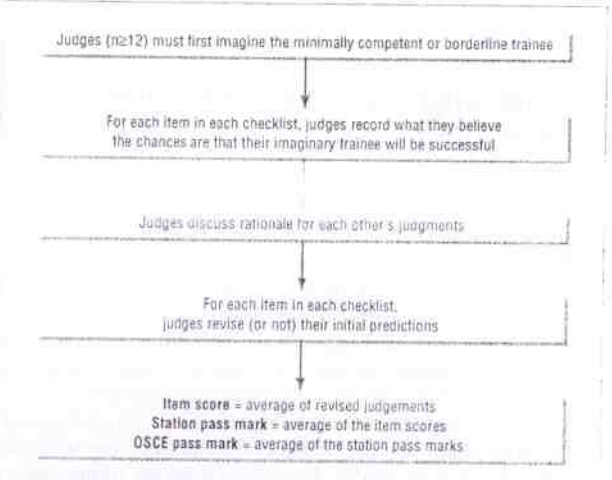
**Questions to ensure validity**

- Are the patient problems relevant and important to the curriculum?
- Will the stations assess skills that have been taught?
- Have content experts (generalists and specialists) reviewed the stations?

## Setting standards

Checklists generate scores; judges set standards. The validity of a standard depends on the judges' qualifications and the reasonableness of the procedure they use to set it. When pass-fail decisions are being made, a skill based assessment should be "criterion referenced" (that is, trainees should be assessed relative to performance standards rather than to each other in a reference group). An Angoff approach is [illegible]

Skill based assessments do not replace knowledge based tests, but they do assess aspects of competence that knowledge based tests cannot assess. Although the use of OSCEs for skill based assessment is increasingly widespread, modifying more traditional formats may be appropriate when they are combined with other forms of assessment or are used to screen trainees. The success of any skill based assessment depends on finding a suitable balance between validity and reliability and between the ideal and the practical.




A modified Angoff procedure for an OSCE

**Further reading**


- Gorter S, Rethans JJ, Scherpbier A, van der Heijde D, Houben H, van der Linden S, et al. Developing case-specific checklists for standardized-patient-based assessments in internal medicine: a review of the literature. *Acad Med* 2000;75:1130-7.
- Hodges B, Regehr G, McNaughton N, Tiberius RG, Hanson M. OSCE checklists do not capture increasing levels of expertise. *Acad Med* 1999;74:1129-34.
- Kaufman DM, Mann KV, Muijtjens AMM, van der Vleuten CPM. A comparison of standard-setting procedures for an OSCE in undergraduate medical education. *Acad Med* 2000;75:267-71.
- Newble DI, Dawson B, Dauphinee WD, Page G, Macdonald M, Swanson DB, et al. Guidelines for assessing clinical competence. *Teach Learn Med* 1994;6:213-20.
- Norcini JJ. The death of the long case? *BMJ* 2002;324:408-9.
- Reznick RK, Smee SM, Baumber JS, Cohen R, Rothman AI, Blackmore DE, et al. Guidelines for estimating the real cost of an objective structured clinical examination. *Acad Med* 1993;67:513-7.


The second picture and the picture showing an oral examination are from Microsoft Clipart.

Sydney Smee is manager of the Medical Council of Canada's qualifying examination part II, in Ottawa, Canada.

The ABC of learning and teaching in medicine is edited by Peter Cantillon, senior lecturer in medical informatics and medical education, National University of Ireland, Galway, Republic of Ireland; Linda Hutchinson, director of education and workforce development and consultant paediatrician, University Hospital Lewisham; and Diana F Wood, deputy dean for education and consultant endocrinologist, Barts and the London, Queen Mary's School of Medicine and Dentistry, Queen Mary, University of London. The series will be published as a book in late spring.

*BMJ* 2003;326:705–6

*Interactive case report*

### A 66 year old woman with a rash

This woman's case was described on 15 and 22 March (*BMJ* 2003;326:588 and 640). Debate on her management continues on bmj.com (http://bmj.com/misc/interactive_case_report.shtml). On 12 April we will publish the outcome of the case together with commentaries on the issues raised by the management, and online discussion from a dermatologist, a vascular surgeon, a general practitioner, and the patient.