# The development, implementation and evaluation of a short course in objective structured clinical examination (OSCE) skills

**De Villiers A,** BCur, BCur(Hons), MPhil(Health Sc Edu)
**Archer E,** BCur, BCur(Hons), MPhil(Higher Edu)
Centre for Health Sciences Education, Faculty of Health Sciences, Stellenbosch University, Parow
**Correspondence to:** Adéle de Villiers, e-mail: adeledev@sun.ac.za
**Keywords:** OSCE, Objective Structured Clinical Examination, examiner training, examiner conduct, inter-rater reliability, inter-rater agreement

## Abstract

**Background:** Objective structured clinical examination (OSCE) examiner training is widely employed to address some of the reliability and validity issues that accompany the use of this assessment tool. An OSCE skills course was developed and implemented at the Stellenbosch Faculty of Health Sciences and its influence on participants (clinicians) evaluated.

**Method:** Participants attended the OSCE skills course, which included theoretical sessions concerning topics such as standard setting, examiner influence and assessment instruments, as well as two staged OSCEs, one at the beginning and the other at the end of the course. During the latter, each participant examined a student role-player performing a technical skill while being video recorded. Participants' behaviour and assessment results from the two OSCEs were evaluated, as well as the feedback from participants regarding the course and group interviews with student role players.

**Results:** There was a significant improvement in inter-rater reliability as well as a slight decrease in inappropriate examiner behaviour, such as teaching and prompting during assessment of students. Furthermore, overall feedback from participants and perceptions of student role players was positive.

**Conclusions:** In this study, examiner conduct and inter-rater reliability was positively influenced by the following interventions: examiner briefing, involvement of examiners in constructing assessment instruments, as well as examiners viewing (on DVD) and reflecting on their assessment behaviour. This study proposes that the development and implementation of an OSCE skills course is a worthwhile endeavour in improving validity and reliability of the OSCE as an assessment tool.

Ⓟ **Peer reviewed. (Submitted: 2011-03-02. Accepted: 2011-06-03.)**  **S Afr Fam Pract 2012;54(1):50-54**

## Background

Since the objective structured clinical examination (OSCE) was first described by Harden and Gleeson[1] in 1979, it has become a valuable tool that is increasingly used in the assessment of clinical skills at health sciences faculties. Its popularity can be attributed partly to the fact that the OSCE is one of the few available options for the assessment of "shows how" (performance) as categorised by Miller[2] in his framework for clinical assessment.

However, with increased use many reliability and validity issues have emerged. Some of the factors that influence the validity and reliability of the OSCE are examiner conduct, the scoring method, and the content and number of OSCE stations. In order to address a few of these issues, OSCE examiner training has become mandatory at many leading health sciences faculties.[3]

At Stellenbosch University (SU) OSCE has been utilised for many years by departments including Obstetrics and Gynaecology, Family Medicine and Paediatrics. Examiner briefing often takes place prior to these OSCEs, but by 2009 SU had not yet implemented a structured examiner training programme. In this paper the process is described that was followed in developing and implementing examiner training in the format of an OSCE skills course. The influence of the course on participants' conduct as well as their general perceptions of the course will be discussed.

## Method

A purposive sample of doctors and registered nurses involved in clinical assessment of junior medical students (MBChB III) were invited to participate in an OSCE skills course. Twelve participants, seven registered nurses and five medical doctors, volunteered to take part in the study. One participant did not return for the second session of the course and therefore had to be excluded from the study.

Ethics approval was obtained from the Committee of Human Research at SU and consent forms were signed by all participants prior to taking part in the study. Anonymity of participants was upheld and all collected data have been destroyed since the completion of the study.

Five third-year medical students were recruited to take part in the study. One week prior to the OSCE skills course, they were informed of the procedural skills they would be performing for assessment purposes in two staged OSCEs. Each student received the assessment instrument which would be used to measure their performance, indicating at which competence level they should perform the skill. The four procedural skills were selected to correspond, as far as possible, to participants' fields of expertise: performing an abdominal examination on a pregnant patient, performing a digital rectal examination, taking a 12-lead electrocardiogram (ECG) and performing adult cardiopulmonary resuscitation.

An OSCE skills course was developed by the researcher. Subject experts were involved in determining and presenting course content, which included an introductory session on general assessment principles; logistics and blueprinting; standard-setting; assessment instruments; the use of standardised patients (SPs), real patients and plastic models/part-task trainers; and examiner influence. After obtaining copyright clearance, a reader was compiled containing articles about these topics along with supplementary reading references and information regarding the assignment required for the course.

The OSCE skills course was presented in two morning sessions which were scheduled approximately one month apart. On the first day of the course, participants took part in a staged OSCE during which they had the opportunity to assess a student performing a procedural skill. There were four OSCE stations and consequently every student role player was assessed by three different participants. Each of the three participants used a different assessment instrument with which to evaluate the student role player, namely a global rating scale, a checklist or a combination of the two. After completion of the staged OSCE, the more theoretical part of the first session followed, covering half of the topics mentioned in the previous paragraph.

On the second day of the course, following the remainder of the theoretical sessions, the OSCE exercise was repeated. The only difference was that course participants were briefed beforehand concerning the individual stations with regard to functionality of the manikins, logistics and acceptable examiner behaviour. Furthermore, prior to the OSCE, the three participants who were assessing the same procedure were grouped together in order to discuss the assessment instrument which they had selected to use and to adapt it to suit their requirements.

Video recordings were taken during both OSCEs. Video material from the first OSCE was written on DVD and each participant received a copy of their interaction with the student role-player to take home. In the month between the two sessions, participants were expected to watch the DVD and write a reflective report about their conduct during the OSCE, including information such as whether they considered themselves to be a "dove" or a "hawk", what they would change about their conduct in subsequent OSCEs and their opinion about prompting and teaching during OSCEs. Participants were further encouraged to exchange DVDs with one another in order to obtain feedback from a peer.

The video recordings were assessed by the researcher and a second independent health sciences educator. Participant conduct was evaluated with regard to whether or not teaching or prompting took place during the assessment. The procedural competence of student role players was assessed, utilising the same assessment instruments as the participants to validate the standard established by the specific student role. Quantitative data obtained from the first and second OSCE assessments were compared and analysed using Spearman rank order correlations in order to measure the effect of the OSCE skills course on inter-rater reliability. Further information was obtained from a questionnaire which participants had completed on the first day of the course, including demographic and background information, as well as participants' perceptions concerning the design and planning of current OSCEs in their workplace. Information gained from questionnaires was included in the statistical analysis where appropriate.

Following each OSCE, student role players participated in a group interview during which they were questioned regarding their perceptions of the participants' conduct in the OSCEs. Interviews were audio recorded and transcribed. Course evaluation forms were completed by all participants at the end of the course. Qualitative data from the focus group interviews and course evaluations were thematically analysed and recurring themes were identified.

## Results and discussion

### Questionnaires

Data obtained from the questionnaires revealed that six participants (50%) were involved exclusively with the assessment of students and setup of OSCE stations, and not with the planning and designing of OSCEs. The latter was mostly done by senior staff members, e.g. consultants. Wilkinson et al. propose that inter-rater reliability is improved when examiners are involved with administration and design of the OSCE and that examiner "ownership"

of the entire assessment is the crucial factor.[4] Ideally all examiners should be involved in the development of the assessment instruments, resulting in a shared definition of good clinical performance.[5]The two participants (17%) who had been involved with the planning and designing of OSCEs also had the most (more than four years) experience with this assessment tool. Four participants had no prior OSCE involvement.

Fifty per cent of participants had health sciences education training, e.g. courses in mentorship, assessment or other. Despite this, certain concepts, such as blueprinting and standard-setting, were unfamiliar to many participants. Furthermore, it became apparent that standard-setting is not often used and most departments make use of an arbitrary value, such as 50%, when determining a pass mark.

Eighty-nine per cent of participants were aware of the utilisation of standardised patients (SPs), although this practice had, according to one of the participants, been abandoned "due to logistical problems" and another considered it "not [to be] relevant to a clinical skills OSCE". According to 67% of these participants, SPs receive training/briefing, when utilised. However, none of the participants makes use of real patients during OSCEs in the Clinical Skills Centre, which implies that most OSCEs are conducted by using part-task trainers or plastic models.

According to 78% of participants, writing stations are used during OSCEs, either as preparation for the next station (57%) or as a "stand-alone" station. A study by Newble and Swansons proposes that in order to use the test time optimally in an OSCE, stations should not contain content that is largely theoretical, such as the interpretation of ECGs or laboratory data. [6] These could be assessed in different formats, whereas technical and practical skills are more difficult to test in formats other than OSCEs.

With regard to assessment instruments, most departments utilise checklists rather than global rating scales. Wilkinson et al. found that, when used by experienced examiners in controlled contexts, global rating scales can be as reliable if not more reliable than checklists.[4] However, at some OSCE stations checklists may be more appropriate (e.g. technical and practical skills) while elsewhere it may be more appropriate to use global rating scales (e.g. communication skills).[7] A few authors suggest using a combination of the two methods.[7,8]

In terms of logistics, one examiner is used per OSCE station. Most of the examiners are briefed beforehand (89%) with regard to acceptable examiner conduct (75%) and the assessment instrument (75%). OSCE stations are usually between five and 10 minutes in duration, with the total number of stations ranging between four and 24. OSCEs are employed for both summative (33%) and formative (44%) purposes.

### Video recordings

Evaluation of the video recordings from the first OSCE revealed the following inappropriate participant behaviour:

- Asking theoretical questions of students when such questions were not part of the assessment
- Allocating marks for "knows how" instead of "shows how"
- Intimidating verbal (e.g. sarcasm) and nonverbal communication
- A complete lack of communication and distancing themselves from students
- Prompting
- Giving feedback and teaching to students in a summative OSCE
- Ignoring items on assessment sheets when such items did not line up with personal preference
- Subjective interpretation of items on the assessment sheet

### Assessment instruments

Analysis of the results from the assessment instruments, utilising Spearman rank order correlations, revealed a significant correlation between the standard and the marks allocated to students by participants in the second OSCE, following the OSCE skills course. The standard was set by predetermining the competency level at which each student had to perform their procedure. The marks from the second OSCE, therefore, reflect the students' performance more accurately than the marks from the first OSCE (Figure 1).

Another interesting finding was that participants who had previously received health sciences education training generally deviated less from the standard than those participants who had not (p=0.01; Figure 2).

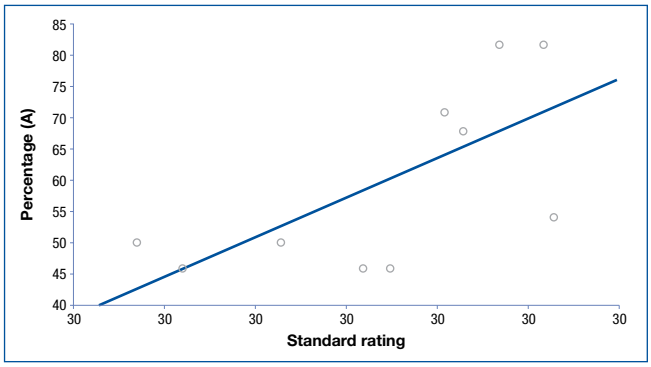


**Figure 1:** Scatter plot illustrating participants' rating (A) against standard rating

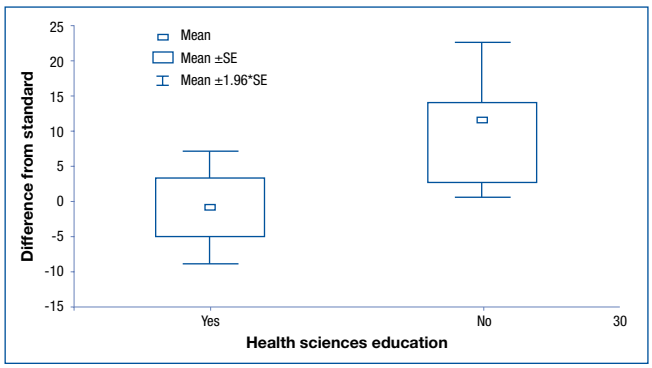


**Figure 2:** Box plot illustrating examiners' deviation from the standard depending on prior health sciences education

### Inter-rater reliability

The marked correlation between the station standard and the participants' marks during the second OSCE demonstrates an improvement in inter-rater reliability. There are a number of factors that may have contributed to this.

Firstly, prior to the second OSCE, examiner briefing and group discussions about the assessment instrument were included. Secondly, during the examiner conduct session, participants received general feedback with regard to, e.g. ensuring that marks are allocated for "shows how" instead of "knows how".[2] Studies have shown that giving feedback to raters about their clinical performance assessments after an examination may help to standardise future evaluations and motivate them to provide accurate ratings, a method used with Olympic skating judges.[5]

Thirdly, prior to the second session of the OSCE skills course, participants had viewed their interaction with the student role player on DVD and reflected on their behaviour, possibly recognising whether they were either too "hawkish" or "dovish". This "hawk-dove" effect is a potential weakness of clinical examinations due to examiners differing with regard to their degree of leniency or stringency when scoring students: hawks tend to fail more candidates as a result of their own very high standards; doves, on the other hand, tend to pass most candidates.[9] This phenomenon was described by Osler[10] as early as 1913. He differentiated between "the two extreme types, the metallic and the molluscoid [which] illustrate inborn defects of character." Studies have shown that extensive training of examiners has a statistically significant effect on the accuracy of examiners' scoring of clinical examinations.[11,12]

A study done at a Malaysian university reported not only an improvement in inconsistent marking, but also a reduction in inappropriate conduct, such as teaching and prompting.[13] With regard to the latter, our study produced similar findings, but not appreciably so. One factor that may have contributed to the insignificant decrease in inappropriate examiner conduct is that participants did not receive individual feedback concerning their conduct during the OSCE. Barth et al.[14] found that, in preparing surgeons for their role as clinical teachers, individualised feedback from video-recorded teaching sessions played a considerable role in their improvement in teaching effectiveness. In isolation, self-study or reflection does not yield the same favourable results.

### Feedback from participants

According to the information gleaned from the course evaluations, participants felt that they benefitted from taking part in the staged OSCE, affording them the opportunity to view their performance on DVD and to reflect on their conduct. The following comments were made in this regard:

"This improved the way I assessed."

"Gives an idea of how students may perceive you as examiner, whether you express yourself clearly when communicating."

Although most participants agreed that the course as a whole was beneficial, many singled out the topic of examiner influence as being pertinent. This partiality may be attributed to the fact that most of the participants function as examiners and are not involved in administration and planning of OSCEs and therefore find topics such as standard-setting and blueprinting irrelevant.

### Feedback from students

Information gained from student role players during the group interviews shows that they appreciated some form of interaction from the participants, even when it was intimidating or inappropriate:

"He [the examiner] started asking me these questions, like blowing it at me ... I think like maybe that's how a real life OSCE is, ... what was nice about him, he didn't make me feel bad that I didn't know the answer… And the other two, they weren't really responsive, they were just like 'Are you finished, ok, I'm leaving now'. So, I didn't get a feedback" [translated].

"But I think what made it easier for me was when he speaks to you during the exam … so, at least he's giving some feedback while you're busy, so you know that he's actually looking or taking interest … whereas, if they're silent, I don't know what they're thinking."

This quote from one of the students suggests dissatisfaction with the decline in prompting and teaching by participants during the second OSCE:

"They were less friendly and stricter" [translated].

Students indicated an awareness of the subjectivity related to assessment of clinical skills:

"I realised that there were different expectations, like the way they asked the questions, some just looked at me … and the other one was like interacting with me … it's sort of, if you do what they are expecting like from you, then you get good marks, but if you … do it in a different way, then maybe you get lower marks …"

"I think it's a good idea, the whole OSCE thing, to standardise it. Like the current clinical evaluations is, well, very subjective. It matters which doctor you get, which patient you get … so, to sort of, make it fairer for everyone …" [translated].

**Limitations to the study**

This study was characterised by a small sample size (12 participants) which limits generalisations from the results. The use of student role players in a "staged" OSCE removed an important variable, i.e. familiarity with the examinee, which has an influence on examiner conduct. However, according to Jefferies, Simmons and Regehr, this influence might be less significant than previously believed.[15] And lastly, participants' awareness of the video recording could have influenced the way in which they behaved during the OSCE, also known as the Hawthorne or "observer" effect.[16]

## Conclusion

Results from this pilot study suggest that the implementation of an OSCE skills course has a positive contribution to make, especially with regard to increasing inter-rater reliability but also in reducing inappropriate examiner behaviour, such as teaching and prompting. It is recommended that this study be replicated in successive OSCE skills courses in order to reflect the opinions and practices of the larger OSCE examiner population at the Faculty of Health Sciences at Stellenbosch University.

## Declarations

The authors declared no financial or personal conflict of interest.

## References

1. Harden RM, Gleeson FA. Assessment of clinical competence using an objective structured clinical examination (OSCE). Med Educ. 1979;13:41-54.
2. Miller GE. The assessment of clinical skills/competence/performance. Acad Med. 1990;65(9 Suppl):S63-67.
3. Boursicot K, Roberts T. How to set up an OSCE. Clin Teach. 2005;2:16-20.
4. Wilkinson TJ, Frampton CM, Thompson-Fawcett M, Egan T. Objectivity in objective structured clinical examinations: checklists are no substitute for examiner commitment. Acad Med. 2003;78:219-223.
5. Williams RG, Klamen DA, McGaghie WC. Cognitive, social and environmental sources of bias in clinical performance ratings. Teach Learn Med. 2003;15:270-292.
6. Newble DI, Swansons DB. Psychometric characteristics of the objective structured clinical examination. Med Educ. 1988;22:325-334.
7. Newble D. Techniques for measuring clinical competence: objective structured clinical examinations. Med Educ. 2004;38:199-203.
8. Townsend AH, McIlvenny S, Miller CJ, Dunn EV. The use of an objective structured clinical examination (OSCE) for formative and summative assessment in a general practice clinical attachment and its relationship to final medical school examination performance. Med Educ. 2001;35:841-846.
9. McManus IC, Thompson M, Mollon J. Assessment of examiner leniency and stringency ("hawk-dove effect") in the MRCP (UK) clinical examination (PACES) using multi-facet Rasch modelling. BMC Med Educ. 2006;6:1-22.
10. Osler W. Examinations, examiners and examinees. Lancet. 1913;1047-1050.
11. Wass V, Van der Vleuten CPM, Shatzer J, Jones R. Assessment of clinical competence. Lancet. 2001;357:945-949.
12. Downing SM. Threats to the validity of clinical teaching assessments: what about rater error? Med Educ. 2005;39:353-355.
13. Tan CPL, Azila NMA. Improving OSCE examiner skills in a Malaysian setting. Med Educ. 2007;41:517.
14. Barth RJ, Rowland-Morin PA, Mott LA, Burchard KW. Communication effectiveness training improves surgical resident teaching ability. J Am Coll Surg. 1997;185:516-519.
15. Jefferies A, Simmons B, Regehr G. The effect of candidate familiarity on examiner OSCE scores. Med Educ. 2007;41:888-891.
16. Festinger L, Katz D. Research methods in the behavioural sciences. New York: Dryden Press; 1953.