High stakes simulation in anaesthesia

Andrew McIndoe FRCA

Revalidation
FOR ANAESTHETISTS
RCA Revalidation matrix
Matrix reference 1H02, 1103

Key points

Simulation-based assessments can be safely introduced into high stakes examinations if first evaluated for validity and reliability in parallel with established benchmarks of performance.

Test scenarios should be designed and structured to allow every candidate potential access to all the marks available on the station.

Carefully formulated checklist marking schemes applied during simulated critical incidents have been shown to effectively document and measure both technical and behavioural aspects of performance.

The criteria chosen to mark performance in a simulated critical incident are not intended to be all-inclusive but should be items that have been shown to differentiate best between strong and weak candidates.

Rater scoring variability may be minimized by specific examiner training, both in the management of the human patient simulator and in the implementation of the chosen scoring system.

Andrew McIndoe FRCA
Consultant Anaesthetist
University Hospitals Bristol NHS Foundation Trust
Bristol
UK
Co-Director
Bristol Medical Simulation Centre
Bristol
UK
OSCE Section Head
Simulation and Resuscitation
Primary FRCA Examination
E-mail: andrew.mcindoe@uhbristol.nhs.uk (for correspondence)

In 2004, the Examinations Committee of the Royal College of Anaesthetists decided to pilot the use of a commercially available full-body interactive patient simulator as an Objective Structured Clinical Examination (OSCE) assessment station within the Primary Fellowship of the Royal College of Anaesthetists (FRCA) examination.

This article examines the results of that pilot and the subsequent formal inclusion of simulator-based assessment into the FRCA and wider medical curriculum.

Background

Artificial patient simulators have been used for medical teaching in the UK since the 1990s. However, it was not until the publication of the Chief Medical Officer's 2008 annual report *Safer Medical Practice: Machines, Manikins and Polo Mints* that this mode of teaching gained broad national support and widespread popularity.[1] It is now generally accepted that simulated patients allow teaching sessions on specific clinical conditions to be more readily scheduled within a teaching programme. Scenarios may be stopped, discussed, re-started, and re-run at will. Rare events can be experienced, and lessons learned in an environment that is safer both for patients and for trainees. Facilitated discussion or feedback consolidates the learning experience and careful combination of this with formative assessment exercises allows trainees insight into their performance, both as individuals and as members of a wider team, helping to target future learning. It is at this point that opinions start to diverge. Despite recommendations included in the Chief Medical Officer's 2008 report, there remains a general uneasiness, even amongst simulation enthusiasts, to accept the use of simulators as a medium for examining competence. High stakes summative assessments performed on simulators are designed to judge clinical performance involving a range of integrated knowledge and skills. Is a simulated patient capable of delivering an accurate reflection of how a candidate would respond when faced with the real clinical situation?

Drivers promoting the use of simulators within the FRCA examination

Real patient clinical cases made their last appearance in the final component of the FRCA examinations in 1994. Although the inclusion of real patients in the examination paralleled some challenges faced by anaesthetists during their preoperative ward visits, it was never ethical or possible to extend testing to include the more practical aspects of real patient management in the operating theatre. Assembling a cohort of willing and appropriate patients for the exam week was a daunting task, made more difficult by the uncertainty about how patients might actually respond to direct questioning or examination by a candidate. Reliability of the clinical case component of the examination could easily be adversely affected by inter-patient variability between days and sittings. Some might argue that real life is similarly unpredictable, but this does not assist in the creation of a level playing field for candidate assessment. As a possible alternative, an OSCE was devised using a series of test stations to replace the 'clinical' in the final FRCA. Subsequently, the OSCE was moved to the Primary FRCA where it has remained ever since. In addition to providing consistency and reproducible examination conditions for candidates, its power as an examination modality lies in a perceived ability to test the integration of knowledge and skills with elements of actual practice—'show how' as well as 'know how'. Miller[2] describes this aspect as part of a pyramid framework of competence.

After the successful introduction of the OSCE, in 2004, the Examinations Committee decided to increase the number of test stations used, in order to increase the reliability of this component of the examination. As an intended substitute for the old 'clinical case', the existing OSCE already incorporated a variety of simulation techniques, including the use of

doi:10.1093/bjaceaccp/mks034 Advance Access publication 3 July, 2012

actors to play the part of patients in a history-taking scenario and mock-ups of clinical situations. A decision was made to extend the use of simulation by piloting the inclusion of a commercially available programmable manikin (Laerdal SimMan®)[3] in a newly devised interactive OSCE station aimed primarily at assessing a candidate's ability to recognize and subsequently manage an anaesthetic critical incident. Although this was a new concept in UK high stakes examinations, precedents existed elsewhere. Byrne and Greaves[4] had conducted a review of assessment instruments used during anaesthetic simulation between 1980 and 2000, but had concluded that the lack of evidence supporting validity and reliability meant that 'introduction of simulator-based tests for certification or re-certification of anaesthetists would be premature' at that stage. Subsequent to that review, objective measurement of technical performance using a simulator-based methodology had been shown to successfully document and mirror improvement of technical performance of novice UK anaesthetists in their first 3 months of supernumerary training, and demonstrated an ability to objectively differentiate novice from experienced performance using an observational scoring checklist designed by expert consensus.[5 6] Development of an Anaesthetists' Non-Technical Skills (ANTS) scoring framework in 2003 complemented these findings, providing evidence that key behavioural markers could also be reliably assessed during simulator scenarios.[7]

The simulated patient OSCE station

A variety of short test scenarios were written for the new simulation OSCE station based around incidents taken from the published anaesthetic core curriculum.[8] Conditions were selected that would allow candidates to:

- demonstrate an ability to assimilate clinical information from a variety of sources,
- recognize and evaluate the significance of ongoing but standardized changes in the condition of a patient,
- formulate an appropriate differential diagnosis,
- evolve and demonstrably apply a logical and safe management plan.

Figure 1 describes an example test scenario based around an elderly male patient undergoing a hemiarthroplasty. In this particular case, a cement reaction might be viewed as the most likely cause of a sudden loss of cardiac output, and credit would be given for appropriate clinical management. However, candidates might also receive marks for recognizing the presence and significance of changes in the monitored parameters; inclusion of circuit disconnection, myocardial infarction, pulmonary embolus, and depth of anaesthesia in a differential diagnosis; exclusion of potential causes; and communication of the problem and management plan to senior help and other members of the team. Thus, the emphasis lies more on patient stabilization rather than making a specific diagnosis.

It is important that all candidates are exposed to an equal opportunity to gain the marks that are available on the station. In this sense, the design of simulation-based exam scenarios differs somewhat from the model used to design teaching scenarios. When teaching, an instructor may allow a scenario to develop in a branching pattern towards a number of differing clinical outcomes depending upon the interactions made by the trainee. Alternative outcomes not encountered during the simulation can be discussed later as part of a feedback session, perhaps even allowing the scenario to be re-run to demonstrate the effect of different management strategies. Exam scenarios tend to follow a more linear pathway, so that test items may be presented in turn to every candidate. This does not mean that changes will not occur in the condition of the patient, rather that these changes will occur regardless of whether the candidate has intervened or not (Fig. 2).

Assessing candidate performance

Having selected an element from the curriculum to be tested and having chosen a scenario to act as a vehicle, it is then necessary to define the metrics used to mark the station. A checklist approach is

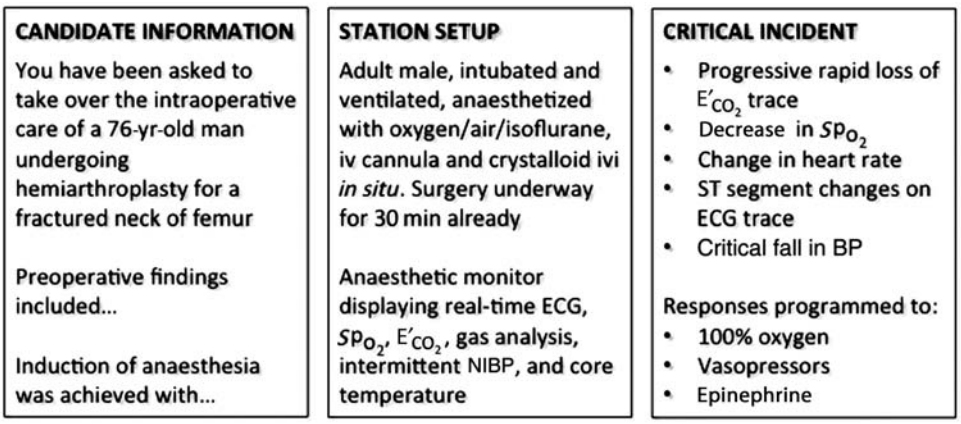

Fig 1 Example of an OSCE simulation station test scenario outline.

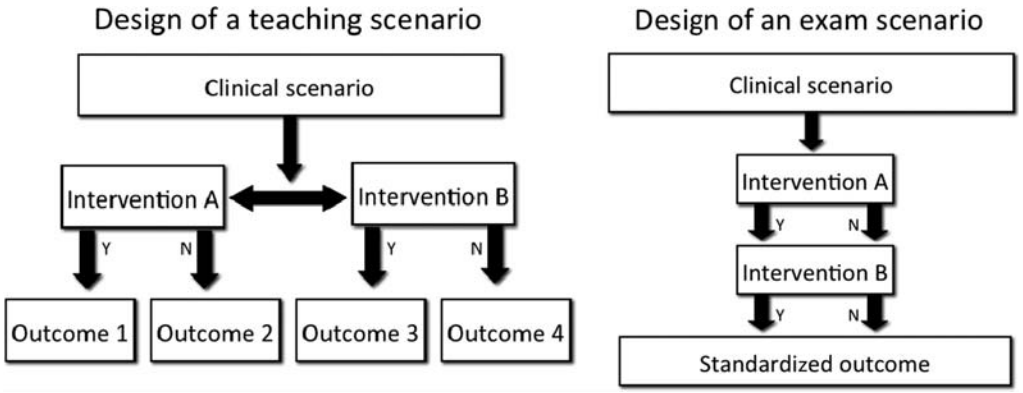

Fig 2 Teaching scenarios are often designed incorporating a branching plan that may encompass multiple outcomes. However this design is less likely to provide a standardized examination test for all candidates. An exam scenario generally includes opportunities for interventions in series rather than in parallel.

used in the OSCE with a maximum of 20 marks available on each station. Elements of the checklist may be related to recognition or interpretation of information that can be gleaned from the simulated patient and monitors, correct prioritization of actions, adherence to published management guidelines, or may simply be direct questions. There is no 'negative marking' of answers or actions that are wrong, but the opportunity to gain marks by indiscriminately listing numerous potential answers in the hope that one may randomly be right is limited by phrasing questions so that candidates may only suggest the best or a small number of most likely options. Candidates are expected to back their assertion that they would 'adopt an ABC approach' with actions that clearly demonstrate this. In real life, situational awareness frequently dictates timing of responses, so candidates may not score for an intervention that does not occur within a defined timeframe as the scenario unfolds. The order of responses or actions is often important, as the possible diagnoses are sometimes made more apparent by later developments or questions from the examiner. Candidates are not allowed to 'go back' and change earlier responses with the benefit of later knowledge. It is important to remember that the overall purpose of a test station is to correctly differentiate between the performance of strong and weak candidates. For this reason, certain basic but essential actions may not actually appear in the scoring matrix if they are found always to be performed by the entire cohort.

Incorporation of the simulated patient station into the OSCE matrix

Initially, the newly designed simulation station was introduced to the 2004 OSCEs as a supernumerary station and was evaluated against existing OSCE stations. Marks were still collected, but neither the candidates nor the examiners assigned to the station were aware as to whether it was contributing as a scoring station in that particular examination. This mechanism allowed adjustments to be made to new questions and to the format of the new station in an entirely realistic exam environment without disadvantage to any candidates before 'going live'. Later, in 2006, a second simulator-based station (interactive resuscitation) was introduced, which included appropriate application of the UK Resuscitation Council guidelines in the scenario as well as overall management.

Evolution of a new question
1. A paper version of a question is prepared and peer-reviewed
2. Scenario software is scripted to strictly control the changes within a 5 min slot
3. The question is field-tested at the next examiners training course
4. The question is submitted for pass-mark setting by modified Angoff process
5. The question is run as a non-scoring station in a real examination (1–3 sittings)
6. Review: the question either goes live, is abandoned, or is further adjusted and re-tested

Validation and standard setting for the simulation station

From May 2004, candidates' cohort performance data from the simulation station were collected in parallel and reviewed against performance on existing scoring stations (Fig. 3). Further analysis of the score profile obtained from the new simulation station appeared to demonstrate closest agreement to scores obtained in the resuscitation, X-ray, and measurement stations. The greatest difference in station scores was apparent when compared with scores obtained on the history and communication stations, suggesting perhaps that the simulation stations, as devised then, were providing assessment of data interpretation and implementation of protocols rather than interpersonal skills. Did the examiners assigned to the station agree with the outcomes of the scoring system? During the trial period, examiners were also asked to give a subjective opinion of each candidate's overall performance on the simulation station using one of three terms 'pass', 'borderline', or 'fail'. These subjective ratings were later compared with candidate outcomes based on the objective scoring matrix (Table 1). Variability was assessed by the Pearson product moment correlation calculated from OSCE simulation station score and subjective examiner rating for each candidate. Correlation was high ($r=0.85$, $P<0.001$),[9 10] suggesting that the objective rating achieved by

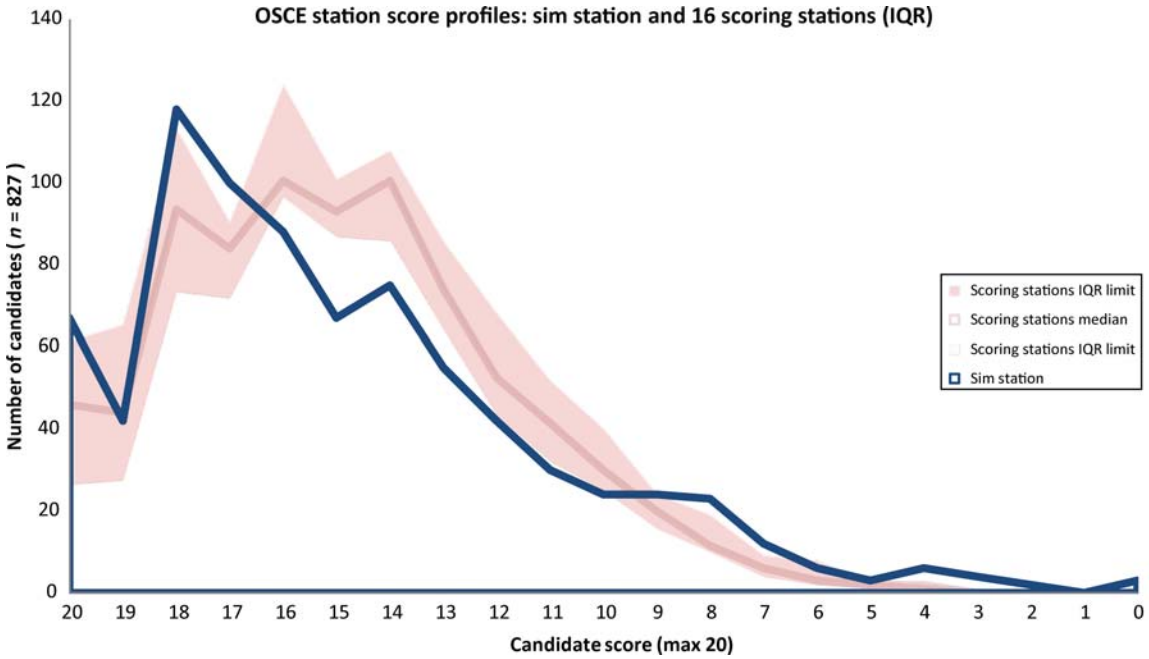

Fig 3 OSCE simulation station candidate score profile (blue) compared with scores from existing live stations (red).

Table 1 Simulation station scores demonstrated high correlation with examiners' subjective rating of each candidate and statistically significant correlation with the overall OSCE score. No correlation was shown with scores obtained in the basic science SOE1 examination. *Pearson's product moment correlation coefficient. †SOE 'Viva' 1, pharmacology and physiology. ‡SOE 'Viva' 2, clinical case and physics

Non-scoring trial comparison 2004	Correlation value (r)*	P-value
Simulation station score *vs* examiner subjective rating	0.852	<0.001
Simulation station score *vs* overall OSCE result	0.133	0.01
Simulation station score *vs* SOE1† result	−0.026	0.33
Simulation station score *vs* SOE2‡ result	0.049	0.21
Simulation station score *vs* overall Primary FRCA result	0.065	0.14

candidates using the checklist system on the simulation station closely matched the overall subjective opinion of the examiner.

Comparisons were also made between scores achieved by candidates on the simulation station and their performance in other components of the Primary FRCA examination. A statistically significant correlation was found to exist between simulation station score and overall OSCE result (P=0.01). In this component of the examination, a single station contributes a potential 20 marks to the total 320 marks available in the OSCE. Correlation with overall Primary FRCA outcome was weak and statistically insignificant (P=0.14). An overall 'pass' in the entire FRCA examination is dependent upon satisfactory performance in the multiple choice question paper, OSCE, and two separate structured oral examinations (SOEs). The influence of the simulation station alone is limited to 20 data points out of a possible 818 in the entire examination. Further analysis revealed no statistically significant correlation between a candidate's performance on the simulation station and performance in either of the two SOEs that specifically test the application of basic science knowledge; however, correlation was higher with SOE2, which includes discussion of theoretical management of a clinical case.

Establishing that a test station is able to contribute to the process of differentiating those that perform well or poorly overall in the examination is essential, but it is also necessary to define standards or thresholds for passing successful candidates. Historically, these have been both norm-based and criterion-based in many postgraduate examinations. However, the uncoupling of different components of the Primary FRCA examination allowing candidates to accumulate overall passes over time means that a norm-based approach is not appropriate to the FRCA where the standard of a cohort may vary considerably at different sittings. Therefore, the actual passmark in the OSCE examination is set as an aggregate of individual passmarks for each OSCE question used in that particular examination. These individual question passmarks are each established by a modified Angoff technique to moderate for the level of difficulty. Experienced examiners are asked independently to estimate a mean score that a minimally competent or borderline trainee might achieve on the specified question, and the average of these scores is then accepted as that question's passmark contribution to the overall OSCE passmark.

Potential error, reliability, and internal consistency

'In the clinical examination there are three variables—the student, the examiner, and the patient. The aim should be to

Table 2 Simulation and interactive resuscitation (IR) station scores from the current OSCE format sampled in 2010 both now show closer correlation with the OSCE and overall Primary FRCA result. *Pearson's product moment correlation coefficient

Active station comparisons 2010	Correlation value (r)*	P-value
Simulation station score *vs* overall OSCE result	0.5	<0.001
Simulation station score *vs* overall Primary FRCA result	0.2	<0.001
Interactive resuscitation score *vs* overall OSCE result	0.5	<0.001
IR station score *vs* overall Primary FRCA result	0.4	<0.001

> standardize the examiner and the patient so that the student's performance can be seen as a measure of his/her clinical competence.' (Collins & Harden)[11]

In the context of a simulator-based assessment, the interaction between the examiner and the simulator is of vital importance if the patient presented to the trainee is to be perfectly standardized for every candidate. Similarly, an examiner's interpretation of responses and interactions made by a candidate may theoretically affect the score recorded for the station.[12] Examiner-based error in the FRCA was minimized by the creation of a 2 day examiner training course. This, in addition to the introductory examiner training, served the dual purpose of orientating new examiners to the hardware and concept of the simulation stations while allowing them to calibrate their assessment and scoring skills appropriately. Re-evaluation of the scores obtained from the simulation and interactive resuscitation stations in a 2010 sample examination sitting confirmed closer correlation with overall OSCE and Primary FRCA outcomes (Table 2). Internal consistency or reliability of candidate scores and outcome obtained in the OSCE was assessed by the calculation of Cronbach's α.[6 13] The OSCE examination itself was found to achieve α-value of 0.762. Generally, a value >0.6 and <0.9 indicates acceptable internal consistency. The value of α can be expected to increase with the number of effective test stations included, so removal of the data from the simulation or interactive resuscitation stations gives an indication of each station's contribution to the overall reliability of the examination. Removal of the simulation station scores reduced α-value to 0.749; independent removal of the interactive resuscitation station scores reduced α to 0.743, indicating that both stations contribute to the reliability of the OSCE examination.

Conclusions

Simulator-based assessments have been successfully introduced into the Primary FRCA examination over a 6 yr period. The Royal College of Anaesthetists occupies a unique position where it has been able to evaluate the use of simulators against an established gold standard for candidate assessment without potential disadvantage to candidates during the appraisal period. At the time of introduction, face validity was cross-checked against subjective examiner opinion. However, in contrast to the Israeli National Board, the College has adopted a marking system that does not include an examiner's global holistic rating of a candidate's performance, and a trainee cannot fail the examination overall purely on the basis of a single simulator assessment.[14] Content validity was reinforced by mapping of test items to the published FRCA curriculum. Construct and criterion validity were carefully considered by comparison with existing and accepted assessment instruments used within the FRCA examination. Subsequent tests of internal consistency have confirmed the reliability of this examination tool. So should simulator-based assessments be rolled-out on a wider scale? Increasingly, there is evidence that simulation scenarios are being included at a local and regional level in selection processes for specialty doctors. Here, the emphasis lies in the ranking of candidates rather than the establishment of a threshold performance for pass/fail purposes. However, the smaller number of test data points included in appointment procedures affects reliability and increases the risk of types 1 and 2 errors. It is therefore especially important that the scenarios and assessment methods chosen are thoroughly tested in advance and shown to be reproducible and appropriate to the intended context. Careful orientation of the station assessors should help to achieve this. Nationally, the General Medical Council (GMC) has shown interest in the use of RCoA scenarios for fitness to practise assessments involving doctors in difficulty. However, extrapolation of data obtained from FRCA candidates to this context is not without difficulty too. Trainees are expected to be able to operate in a wide range of clinical environments. Further specialization frequently involves a narrowing of breadth of practice in favour of depth of experience in specific sub-specialty areas. Suitable denominator data, while not impossible, become difficult to obtain to support conclusions drawn from simulator-based assessments of established specialists' practice. That said, if suitable safeguards are put in place to acquire relevant data and to compensate for these potential issues, we can probably expect to see simulators play an ever-increasing role not only in education and trainee assessment, but also in the revalidation processes for anaesthetists in the UK.

Acknowledgements

The Royal College of Anaesthetists has received support and advice from Laerdal in developing the use of simulators in the Primary Fellowship examination. I am grateful to Dr Janis Shaw who led the initial pilot, and the RCoA OSCE working party and Examination staff for collection of data published within this article.

Declaration of interest

None declared.

References

1. Annual report of the chief medical officer. *Safer Medical Practice: Machines, Manikins and Polo Mints* (dh_096227.pdf). Department of Health Publications, 2008. Available from www.dh.gov.uk/en/Publicationsandstatistics/Publications/AnnualReports/DH_096206

2. Miller GE. The assessment of clinical skills/competence/performance. *Acad Med* 1990; **65** (Suppl): S63–7

3. www.laerdal.com

4. Byrne AJ, Greaves JD. Assessment instruments used during anaesthetic simulation: review of published studies. *Br J Anaesth* 2001; **86**: 445–50

5. Forrest FC, Taylor MA, Postlethwaite K, Aspinall R. Use of a high-fidelity simulator to develop testing of the technical performance of novice anaesthetists. *Br J Anaesth* 2002; **88**: 338–44

6. Morgan PJ, Cleave-Hogg D, DeSousa S, Tarshis J. High-fidelity simulation: validation of performance checklists. *Br J Anaesth* 2004; **92**: 388–92

7. Fletcher G, Flin R, McGeorge P, Glavin R, Maran N, Patey R. Anaesthetists' Non-Technical Skills (ANTS): evaluation of a behavioural marker system. *Br J Anaesth* 2003; **90**: 580–8

8. *CCT in Anaesthetics II: Competency Based Basic Level (ST Years 1 and 2) Training and Assessment*. Royal College of Anaesthetists. Available from www.rcoa.ac.uk

9. Soper DS. p-value calculator for correlation coefficients (online software), 2012. Available from http://www.danielsoper.com/statcalc3

10. Cohen J, Cohen P, West SG, Aiken LS. *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*, 3rd Edn. Mahwah, NJ: Lawrence Earlbaum Associates, 2003

11. Collins JP, Harden RM. The use of real patients, simulated patients and simulators in clinical examinations. *Association for Medical Education in Europe (AMEE) Guide 13*. ISBN: 978-1-903934-14-1. Available from www.amee.org

12. Boulet JR, Murray DJ. Simulation-based assessment in anesthesiology, requirements for practical implementation. *Anesthesiology* 2010; **112**: 1041–52

13. Pell G, Fuller R, Homer M, Roberts T. How to measure the quality of the OSCE: a review of metrics—AMEE guide no. 49. *Med Teach* 2010; **32**: 802–11

14. Berkenstadt H, Ziv A, Gafni N, Sidi A. Incorporating simulation-based objective structured clinical examination into the Israeli National Board examination in anesthesiology. *Anesth Analg* 2006; **102**: 853–8

Please see multiple choice questions 29–32.