Techniques for measuring clinical competence: objective structured clinical examinations

David Newble

The traditional clinical examination has been shown to have serious limitations in terms of its validity and reliability. The OSCE provides some answers to these limitations and has become very popular. Many variants on the original OSCE format now exist and much research has been done on various aspects of their use. Issues to be addressed relate to organization matters and to the quality of the assessment. This paper focuses particularly on the latter with respect to ways of ensuring content validity and achieving acceptable levels of reliability. A particular concern has been the demonstrable need for long examinations if high levels of reliability are to be achieved. Strategies for reducing the practical difficulties this raises are discussed. Standard setting methods for use with OSCEs are described.

KEYWORDS clinical competence, *standards; education, medical, undergraduate, *standards; educational measurement; reproducibility of results.

*Medical Education 2004; **38**: 199–203*
doi:10.1046/j.1365-2923.2004.01755.x

BACKGROUND

The main purpose of this paper is to discuss the development of an objective structured clinical examination (OSCE) that meets acceptable standards of validity and reliability. In addressing this issue it is helpful to have an understanding of the background to the rise in popularity of the OSCE as a major tool in the assessment of clinical competence.

Department of Medical Education, The University of Sheffield

Correspondence: David Newble, Department of Medical Education, The University of Sheffield, 1st Floor, Coleridge House, Northern General Hospital, Herries Road, Sheffield S5 7AU, UK. Tel. +44 (0)114-271-5940, Fax: +44 (0)114-242-4896. E-mail: d.newble@sheffield.ac.uk

The traditional method of assessing knowledge until the 1950s and 1960s was the essay. Concerns about the inconsistency of marking, and the inadequate sample of knowledge tested within a given period of time, led to the rapid implementation of objective written tests (e.g. multiple-choice tests), which today have almost entirely replaced essays as the preferred method of assessing the recall and application of knowledge in medical examinations. This trend has been particularly evident in high stakes testing situations where reliability and content validity are essential ingredients in making the results of such assessments defensible to both students and external agencies. In North America, the same concerns were raised about the traditional clinical and oral examinations used for assessing clinical competence in the 1960s. The National Board of Medical Examiners, after discovering low correlations between examiners, discontinued their clinical oral examination on the basis of unacceptable reliability.[1] Such decisions took longer to reach in other parts of the world, partly perhaps as a result of the lack of an alternative approach to the assessment of clinical competence. The advent of the OSCE in the 1970s promised the equivalent advantages in clinical testing to that of objective written examinations in knowledge testing.[2] In other words, the use of checklist based marking would enhance interrater consistency and the testing of students' performance on multiple stations would increase the number and range of competencies that could be sampled. The OSCE has subsequently been subject to a considerable amount of research into its strengths and limitations, the outcomes of which form the basis of generalizations to be made in this paper.[3]

In attempting to make such generalizations about OSCE's it is important to keep in mind various points. The first is that an OSCE is not a test method in the same way as an essay or a multiple-choice question. It is basically an organization framework consisting of multiple stations around which students

Key learning points

Traditional clinical examinations have serious limitations in terms of validity and reliability.

Objective structured clinical examinations (OSCEs) have the capacity to improve the validity and reliability of assessments of many aspects of clinical competence.

To achieve high levels of reliability OSCEs have to be longer than is often practicable.

OSCEs can be combined with other methods of assessment to enhance reliability.

Easy to apply standard setting procedures are now available.

A considerable body of evidence on OSCEs exists to guide decisions and further developments.

rotate and at which students perform and are assessed on specific tasks. The conventional view of an OSCE is of a series of 5–10 minute stations where a standardized clinical task is performed under the observation of one or two examiners who score the performance on a structured marking sheet. However, many variants exist. For example, stations may be much longer and examiners may not be present, with the marking being undertaken by the simulated patients on whom the task was performed.[4] In other OSCE's there may be stations at which multiple-choice questions are asked or at which other forms of written responses are required. This makes a discussion about OSCEs difficult if the format is not fully described. In this paper I will refer to the conventional short station format as this is the approach being used in most medical schools and by many licensing bodies such as the General Medical Council[5] and the Medical Council of Canada.[6]

Broadly speaking, the issues to be addressed in regard to the OSCE revolve around those to do with organization and those to do with the quality of the assessment. This paper focuses on the latter, though organizational issues, such as the numbers of examinees, location and resources, may have a major impact on the technical quality that can be achieved.

TECHNICAL ISSUES

The first article in this series addressed the fundamental principles of designing a good assessment.[7] It draws on a set of guidelines for assessing clinical competence which emphasizes the fundamental importance of being clear about the purpose; about defining what is to be tested and using a blueprint to guide the selection of content; about selecting the most appropriate test method and format which should be driven by fidelity to the clinical situation and the task to be posed to the candidate; about issues relating to administration and scoring; and about standard setting procedures.[8]

Purpose

The OSCE provides a test format particularly suitable for assessing many, but certainly not all, components of clinical competence.[9] For example, attitudinal and behavioural aspects are probably better tackled by the use of multiple ratings collected over a period of time during clinical attachments and clerkships. At the other end of the scale, the testing of relevant knowledge required to be competent, including aspects of diagnosis, investigation and management, can be more efficiently and more cheaply tested with written formats. Overall, the OSCE is best suited to testing clinical, technical and practical skills and can do so across a very broad range, often with a high degree of fidelity. These include many skills that were never tested in the traditional clinical examination.

Defining and selecting the content to establish validity

There are different ways of defining the content of a clinical competence examination. Doing so is the basis for establishing the *content validity* of the test, the most fundamental requirement in ensuring the quality of a competency test.[7]

The guidelines referred to previously outline three steps to be taken. The first two are required to define the range of competencies which reflect the 'outcome objectives' for the course or period of training that candidates are to be certified as having achieved. Step one is to identify the problems or conditions that the candidate needs to be competent in dealing with. These may be generated from the opinion of expert groups or by more formal studies based on observation and analysis of what

the student or doctors will have to undertake. Step two is to define the tasks within the problems or conditions in which the candidate is expected to be competent. For example, if the problem was 'Chest Pain' tasks might include taking a history from a patient with angina, performing and interpreting an ECG, demonstrating competence in cardiopulmonary resuscitation and educating a patient about the use of antiangina medication or a diet. While defining the task may be relatively simple, ensuring this is tested at the correct level can be more difficult.

The construction of a *blueprint* or grid is the third step. This is an extremely valuable strategy for enhancing and defending the validity of an examination. It is a way of defining the sample of items to be included in the test. In its simplest form it will consist of a two-dimensional matrix with one axis representing the generic competencies to be tested (e.g. history taking, communication skills, physical examination, investigations, management). The other axis represents the problems or conditions on which the competencies will be demonstrated. An example is provided by the blueprint for the OSCE run by the Professional and Linguistics Assessment Board of the General Medical Council.[5] Research has shown that performance on one problem is a very poor predictor of performance on another, even similar, problem, so wide sampling across problems is required if an adequate level of content validity and reliability (see later) is to be achieved.[10]

Determining and establishing reliability

Other articles in this series and elsewhere deal in detail with the research that has provided us with clear guidance on what we must do if we are to ensure defensible levels of reliability for an OSCE examination.[4] When the OSCE was first devised it was assumed that the main problem undermining reliability in clinical examinations related to the biases introduced by examiners, some of which were personal and some related to the lack of standardization of the tasks and scoring criteria. The 'objective' part of the OSCE referred to the standardization of both the task and the scoring (based on checklist type rating forms).[2] While this did indeed improve interrater reliability, research using generalizability theory showed that the problem of rater consistency paled into insignificance relative to the issue of *case specificity*.[11,12] The bottom line of such studies was that OSCE examinations, and indeed many other test formats used to assess aspects of clinical competence, needed to incorporate measures across a large number of cases or problems. The undeniable fact that emerged was that OSCEs, used alone, would need to be much longer (of the order of 4–8 h) than those in common use and that this potentially made them impractical.[12]

Various strategies have subsequently been adopted to minimize the practical difficulties raised by case specificity. The simplest is to combine the OSCE with other test formats that provide more efficient sampling of content.[13–16] As long as all test components are based on the same blueprint, this is a justifiable approach. One example is provided by our own experience with an undergraduate final examination where the combination of a 90-minute OSCE with an unacceptable reliability of around 0.6 was combined with a 90-minute free-response item written test (reliability 0.8) to produce an overall and acceptable reliability for the clinical competence examination of 0.8.[14]

One issue which has been given some prominence in recent years is that of the approach to rating. The original description of OSCE's anticipated that the use of checklists would enhance interrater reliability and would solve the problem created by global ratings used in traditional examinations. One problem that emerged from the checklist approach was the phenomenon of trivialization.[8] Unfortunately, it is easy to fall into the trap of developing detailed checklists that produce reliable scores but which do not truly reflect the examinee's performance of the task. Only criteria that are easy to define may be included on the marking sheet at the expense of equally or more important criteria that are more difficult to define and measure. Trivialization of the scoring may also be apparent if appropriate weightings within the marking schedule are not made. A related problem is possible unintended effects on student learning. If checklists are made available to students, they will inevitably use them to guide their learning. If they are not well constructed this may lead students to practise the wrong approach simply to enhance their chances in the OSCE.[17]

More recently, the issue of global vs. checklist ratings has been investigated in more depth. It is becoming apparent that global ratings, within the framework of structured tasks and used by informed or trained assessors, may be as reliable or even more reliable than checklists.[18] However, a balanced approach is probably best. There are OSCE stations where checklists may be more appropriate (e.g. some

practical and technical skills stations) and others where global ratings may be more appropriate (e.g. communication skills stations and some diagnostic task stations where there may be alternative routes to the same outcome). Our own preference is to use a combination of both, with checklists used to identify specific elements of content or skill that must be demonstrated and global ratings used for providing a measure of process aspects (e.g. patient education skills, general approach to a task). In the end the most important thing to evaluate is whether the final score truly reflects the level of competence of the examinees on the task they were asked to perform.

Standard setting

Another major issue which has achieved recent prominence in the literature is standard setting. This is dealt with in greater depth in another article in this series.[19,20] In general, the standard setting procedure uses either a relative (or norm-referenced) approach or an absolute (or criterion-referenced) approach. In testing for competence an absolute method is usually going to be the most appropriate. One broad approach is to use expert judges prospectively to estimate the probability that a borderline candidate will succeed on each item in the test. An example of such a method is the Angoff procedure.[21] The alternative but simpler approach is the borderline group method, which provides similar results to the Angoff method.[22] This is becoming more popular both for large scale OSCEs conducted by national licensing bodies, such as the Medical Council for Canada, and for small-scale OSCEs conducted by medical schools.[23,24] Such methods involve examiners giving a global rating of each student's overall performance independent of the mark they award as a result of completing the station scoring sheet. In our own experience we have used the categories pass/borderline/fail. The mean of all borderline scores becomes the pass mark for the station and the pass mark for the whole OSCE is calculated by adding the mean borderline scores of all stations. Examiners find this process easier than the Angoff procedure, it is less time consuming and has the added credibility associated with being based on direct observation rather than on a hypothetical student's performance. There are other variants on the borderline approach that are beyond the scope of this article.

One other issue that is sometimes debated is whether final decision-making should be based on the overall score across all stations – a fully compensatory model – or on passing a defined proportion of stations. Some organizations use a combination of both. There is no right or wrong answer as to which is the more valid approach. As a result, our preference is to use the simplest, which is the overall mark, to which additional statistical indices can then be applied such as the standard error of measurement, the educational measurement equivalent of a confidence interval.[25]

OTHER ISSUES

There are many other issues that could have been addressed in this article but space precludes dealing with them in any detail. For instance, a considerable amount of work has been done on the use of simulated and standarized patients. Generally speaking, it has been demonstrated that, when well trained, they cannot be distinguished from real patients, are stable over time, and can provide accurate feedback and assessments.[26,27] Other examples include station length and effects of the order in which students take the stations. A recent review provides a useful starting point for those interested in such issues.[4]

REFERENCES

1 Hubbard JP. *Measuring medical education.* Philadelphia: Lea & Febiger; 1971.

2 Harden RM, Gleeson FA. Assessment of clinical competence using an objective structured clinical examination (OSCE). *Med Educ* 1979;**13**:41–54.

3 Epstein RM, Hundert EM. Defining and assessing professional competence. *JAMA* 2002;**287**:226–35.

4 Petrusa ER. Clinical performance assessment. In: Norman GR, Van der Vleuten CPM, Newble DI, eds. *International handbook of research in medical education.* Dordrecht: Klruwer Academic Publications; 2002.

5 Tombeson P, Fox RA, Dacre JA. Defining the content for the objective structured clinical examination component of the Professional and Linguistic Assessment Board examination: development of a blueprint. *Med Educ* 2000;**34**:566–72.

6 Reznick RK, Blackmore D, Cohen R, *et al.* An objective structured clinical examination for the licentiate: report of the Medical Council of Canada: from research to reality. *Acad Med* 1992;**67**:487–94.

7 Crossley J, Humphris G, Jolly B. Assessing health professionals. *Med Educ* 2002;**36**:800–4.

8 Newble D, Dauphinee D, Dawson-Saunders B, *et al*. Guidelines for the development of effective and efficient procedures for the assessment of clinical competence. In: Newble DI, Jolly B, Wakeford R, eds. *The certification and recertification of doctors: issues in the assessment of clinical competence*. Cambridge: Cambridge University Press; 1994.

9 Newble DI. Assessing clinical competence at the undergraduate level. In: *Medical Education Booklet*, No. 25. Edinburgh: Association for the Study of Medical Education; 2002.

10 Van der Vleuten C. Validity of final examinations in undergraduate medical training. *BMJ* 2000;**321**: 1217–9.

11 Van der Vleuten CPM. The assessment of professional competence: Developments, research and practical implications. *Adv Health Sci Educ* 1996;**1**: 41–67.

12 Swanson DB, Norman GR, Linn RL. Performance-based assessment: Lessons learnt from the health professions. *Educ Res* 1995;**24**:5–11.

13 Newble DI, Swanson DB. Psychometric characteristics of the objective structured clinical examination. *Med Educ* 1988;**22**:325–34.

14 Newble D, Swanson D. Improving the quality of a multidisciplinary test of clinical competence: A longitudinal study. In. Melnick DE, ed. *Proceedings of the Eighth International Ottawa Conference on Medical Education and Assessment*. Philadelphia: National Board of Medical Examiners; 1998: 376–80.

15 Wass V, McGibbon D, Van der Vleuten C. Composite undergraduate clinical examinations: how should the components be combined to maximize reliability. *Med Educ* 2001;**35**:326–30.

16 Verhoeven BH, Hamers GHC, Scherpbier AJJA, *et al*. The effect on reliability of adding a separate written component to an objective structured clinical examination. *Med Educ* 2000;**34**:525–9.

17 Van Luijk SJ, Van der Vleuten CPM, Van Schelven SM. Observer and student opinion about performance-based tests. In: Bender, ed. *Teaching and assessing clinical competence*. Groningen: Boekwerk Publications; 1990: 497–502.

18 Regehr G, MacRae H, Reznick R, Szaky D. Comparing the psychometric properties of check-lists and global rating scale for assessing performance on an OSCE-format examination. *Acad Med* 1998;**73**:993–7.

19 Norcini J. Setting standards on educational tests. Med Educ 2003;in press (this issue).

20 Cusimano M. Standard setting in medical education. *Acad Med*. 1996;**71**:S112–20.

21 Smee SM. Setting standards for objective structured clinical examination. the borderline group method gains grounds on Angoff. *Med Educ* 2001;**35**:1009–10.

22 Kaufman DM, Mann KV, Miujtjens AMM, Van der Vleuten CPM. A comparison of standard setting procedures for an OSCE in undergraduate medical education. *Acad Med* 2000;**75**:267–71.

23 Rothman AI, Blackmore D, Dauphinee WD, Reznick R. The use of global ratings in OSCE station scores. *Adv Health Sci Educ*. 1997;**1**:215–9.

24 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. *Med Educ* 2001;**35**:1043–9.

25 Streiner DL, Norman GR. *Health measurement scales. A practical guide to their development and use*. Oxford: Oxford University Press; 1995.

26 Van der Vleuten CPM, Swanson DB. Assessment of clinical skills and standarised patients: state of the art. *Teaching Learning Med* 1990;**2**:58–76.

27 Tamblyn RM, Klass DJ, Schnable GK, Koppelow ML. The accuracy of standardised patient presentations. *Med Educ* 1991;**25**:100–9.

Received 6 June 2003; editorial comments to authors 9 July 2003; accepted for publication 18 July 2003