Special Article

Objective Structured Clinical Exams: A Critical Review

John L. Turner, MD; Mary E. Dankoski, PhD

Since their inception in the 1970s, objective structured clinical examinations (OSCEs) have become popular and now are part of the US Medical Licensing Examination for all US medical graduates. Despite general acceptance of this method, there is debate over the value of OSCE testing compared to more traditional methods. A review of reliability and validity research does not clearly show superiority of OSCE testing. To use OSCEs in a valid and reliable way, attention must be paid to test content, test design, and implementation factors, especially when the results will be used for high-stakes decision making. While questions remain around the application of OSCE testing, there are also both known and hidden benefits to students, faculty, and organizations that use OSCEs. This paper reviews the pros and cons of the OSCE method and outlines important issues for medical educators to consider when planning to use OSCEs in their programs.

(Fam Med 2008;40(8):574-8.)

Methods for student assessment in medical education has changed over the past 50 years. We have moved from a standard of pen-and-paper tests of knowledge and facts toward a more complex system of evaluation.[1] Medical students today are tested on knowledge, attitudes, and skills across multiple settings and methods, which are often triangulated to reach summative decisions. Current educational and assessment strategies include problem-based learning, computer simulations, faculty global ratings and checklists, standardized patients, and team-based learning.[2]

The standards and processes of licensure and certification have also followed these changes. Current applicants are tested in various proficiencies through innovative methods that were unheard of just a few decades ago. Needless to say, this complex evolution has not proceeded without controversy.

These developments in medical education and assessment must be viewed in the historical context of social changes and advancements in educational theory outside of medicine. The three decades from 1965–1995 saw tremendous growth in innovative educational efforts aimed at new content areas, with a major focus on skill acquisition. When Miller introduced his pyramid of educational objectives,[3] it joined Bloom's cognitive taxonomy,[4] and these stood as the main framework for educational thought in medicine entering the 1990s. Curricular goals and objectives were organized around these hierarchical scaffolds, and student assessment mirrored the proposed developmental steps.

New strategies for assessing learning and competence in these "higher" areas then began to appear in the literature. Societal changes also reinforced the need for real-world testing in medical education, as patients and payers began to question the traditional self-ratification of medical professionals, asking for more evidence of physician training and expertise, and the shift toward patient rights began to emerge. In the midst of this context of medical education transition, Harden first conceptualized the objective structured clinical examination (OSCE) in 1975.[5]

Current State of OSCE Testing

Developed to assess the complex notion of clinical competence, the OSCE uses multiple stations with examinees performing various clinical tasks at each station. Tasks may include test interpretation, history taking, physical examination, patient education, order writing, or other activities. Over time, most OSCEs have come to rely heavily on standardized patients (SPs). SPs are individuals trained to portray a patient with a particular disease or condition, thus affording hands-on testing of students. The OSCE method, as used today, has evolved into a flexible testing approach that can incorporate SPs as well as observer ratings, short written tests, and other methods used to provide a comprehensive clinical evaluation of history taking, physical examination, and communication skills.[1,6]

From the Department of Family Medicine, Indiana University.

Following the development of OSCEs, medical education has continued to evolve, with the move toward competency-based education constituting the most recent sweeping change. The most commonly used model in medical education was largely developed by the Accreditation Council for Graduate Medical Education (ACGME), which categorizes medical competence into six related domains: medical knowledge, patient care, professionalism, communication and interpersonal skills, practice-based learning and improvement, and systems-based practice. The concept of competencies now permeates the culture of medical education, shaping how we currently talk and think about medical student and resident education. The development of quality assessment methods—ie, those that are reliable, valid, practical, generalizable, and replicable—has lagged behind this culture shift, with much of this research still being conducted.

Conceptualizing the acquisition of knowledge, skills, and attitudes as competencies is important because it implies a developmental progression of a medical student from a novice to, ultimately, a proficient and expert clinician.[7] Competency evaluations, then, should include both formative and summative assessments to assess an individual's stage of development. The OSCE method is often used for both types of assessment, and a balanced review of the strengths and weaknesses of OSCEs are included in the ACGME Toolbox of Assessment Methods. Since OSCEs were specifically developed with the intention to evaluate a learner's clinical competence, this method has been heralded by many as the competency evaluation method of choice.[1,2]

Indeed, even though research to investigate the reliability, validity, objectivity, and feasibility of OSCE testing is still ongoing, OSCEs quickly became established as a regular method of learner assessment. As of 2004, 94 of the 126 accredited US medical schools require a comprehensive OSCE test, compared to only 49 schools with such a requirement in 1998.[8] Currently, 107 schools require students to pass the US Medical Licensing Examination (USMLE) Step I, and 83 require passage of USMLE Step II for graduation.[8] Moreover, the USMLE Step II incorporated a Clinical Skills Examination (CSE) in 2004, and this CSE is in essence a national OSCE for all US medical students. Similarly, the Educational Commission for Foreign Medical Graduates (ECFMG) has administered the Clinical Skills Assessment (CSA), also an OSCE-like evaluation, for many years, with the results used to determine future practice preparedness for many foreign medical trainees coming to the United States. Additionally, national certification programs utilizing SPs exist in Canada and the United Kingdom.

Development of scenarios and cases for SPs to use in OSCEs has now grown into a science with accepted guidelines and standards.[24] The frequent networking and sharing of resources among OSCE developers has led to a rapid implementation process in many institutions. National organizations like the Association of Standardized Patient Educators have emerged to provide leadership, education, and structure to the science of SP training and assessment. Continuing education courses, textbooks, and many "how to" articles facilitate faculty development and promote more rigorous evidence behind OSCE testing.

Clearly, the OSCE and the use of SPs to support it have developed into a mainstream method in the education and licensing of physicians, with many high-stakes decisions and potential consequences hinging on their results. Given the central role that OSCEs have assumed in the evaluation of physicians in training, one might assume that OSCEs have excellent psychometric properties. To more fully appreciate how this unique assessment method has impacted medical education, the general psychometric qualities of OSCE testing and the importance of OSCE testing beyond student assessment must be reviewed. As described below, the results are mixed, and educators need a deeper understanding of the research if they are to apply appropriate meaning to OSCE results.

Testing Characteristics

Despite the general acceptance of the OSCE, there has been recent concern over the heavy reliance on this particular format above other assessment methods. For example, Norman challenged the idea that OSCEs provide better assessments than other traditional methods.[9] He noted a lack of evidence to support the superiority of such "high fidelity" testing, which is expensive and resource intensive, over other more manageable evaluation systems. Barman has also challenged the psychometric qualities of OSCE testing in a review of selected publications.[10] Both authors question the validity and reliability of OSCE evaluation.

Reliability

Reliability refers to the consistency of examinee scores over time, test administrations, and sampling.[11] Many studies have reported less than ideal, but generally acceptable reliability scores for OSCEs. In fact, one early study reported no correlation between individual student performance over two similar OSCEs given by the same institution.[12] Various methods have been found to increase reliability of OSCEs. For example, van der Vleuten and Slawson systematically reviewed 10 of the earlier OSCE studies and found that the major source of measurement error was accounted for by variation in student performance from station to station.[6] This "content specificity" is minimized, and OSCE scoring becomes more reliable with large numbers of stations, raters, and good standardization of patients. Other factors influencing reliability include student fatigue,

personal bias, relatively high anxiety about the testing method, and memory lapse.[13,14] Even when efforts are made to control for these sources of error, reliability coefficients generally range from 0.41 to 0.88.[6,15]

Reliability is also influenced by whether one or more SPs portray a case, as well as test length. When assessing communication skills, for example, only 2 hours of total testing time is needed to obtain reliability coefficients above 0.7. In comparison, not until test length approaches 6 hours does one find reliability coefficients above 0.7 for data gathering and history taking.[15] An early OSCE study suggested the need for 10 stations and 3–4 hours of testing to obtain reliability coefficients of 0.85–0.90.[16] Other studies have found test lengths upward of 12 hours are required to yield reliability coefficients at the 0.7 or higher level.[17, 18]

Thus, OSCEs yield wide variation in reliability scores. The ECFMG reports reliability scores of the CSA, the longest standing high-stakes OSCE, at only 0.64.[19] Although there is disagreement and some consider these reliability findings adequate,[20,21] many consider them below the acceptable threshold for high-stakes testing.[11,22] In 2004, US medical schools differed widely in the number of stations used in the final comprehensive OSCE: five stations or less in eight schools, six to 10 stations in 61 schools, 11 to 15 stations in 20 schools, and 15 or more stations in 15 schools.[8] A general rule is that seven cases in any domain are required to reach acceptable levels of reliability (Yudkowsky R, personal communication, 2005). Unfortunately, many institutions may thus be inadvertently using OSCEs without achieving adequate reliability.

To ensure high reliability, one must attend to multiple factors when designing, implementing, and scoring any OSCE. Recent promising results came from a study of SP encounters given in small sets of two–four cases during each clinical clerkship at a US medical school. When cases were pooled and examined as a collective OSCE, the reliability was found to be 0.63.[23] Further investigation is needed to see if this practical method of distributing cases over time provides sound data upon which to make high-stakes decisions.

Validity

Highly valid results are critical when using the OSCE method for significant decision making. If a test has low measures of validity, it is doubtful that the test truly measures what it is intended to measure. There are two types of validity to consider.

Content validity is critical to any test and particularly one that can, at best, sample only a small portion of the domain being tested. Early pioneers in the use of OSCEs claim high content validity is obtainable in this format,[24] especially with the application of a "test blueprint" (a framework for content areas of the test).[25]

Concurrent validity is measured by comparing one testing method to another that aims to measure the same construct. OSCE testing, however, measures multiple discreet as well as comprehensive skills and knowledge in a manner unlike other assessment formats. Comparison to other measures of clinical competence, including multiple choice questionnaires, clinical ratings, National Board of Medical Examiners subtests, duration of training, non-SP skills tests, other course grades, and self ratings have produced mixed results.[6,26] Correlation coefficients for these comparisons range from 0.10 to 1.00 (in only nine of 33 studies was the correlation coefficient above 0.70). Another study reported that performance on a multiple choice written test was a better predictor of clinical performance of family physicians than the use of unannounced standardized patients.[27] In contrast, in a more recent study of a comprehensive grading system for internal medicine students, Auewarakul et al found OSCEs to be one of the evaluation methods with the most validity evidence and concluded that "There is clearly sufficient validity evidence to support the utilization of the… OSCE format…"[28]

It is thus difficult to make conclusive statements about the validity of the OSCE method. There are two major explanations offered regarding low validity scores: (1) OSCEs measure different constructs and, therefore, should not be expected to correlate well with standard testing and (2) validity evidence for OSCE testing is hard to determine because OSCEs do not measure what they are intended to measure. There is no clear evidence to support one theory over the other, but emerging research may help clarify the validity concerns. One recent study utilizing SPs to evaluate communication skills concluded that "Scores from OSCE communication checklists may not predict patients' perceptions of communication."[29] In this study, performance on a well-designed checklist did not correlate well with patient perception of effective communication. The many variables that factor into the design and completion of an OSCE will clearly influence the validity; the degree to which low validity is tolerable depends on the application and what is at stake for the learner.

Scoring

Scoring methods for OSCEs also vary widely, and they influence reliability and are open to debate. Checklists have been standard in many established OSCE programs and have intuitive value as an assessment tool. But, checklists may have limits when testing skilled practitioners, who are not as thorough in questioning or examination due to fast pattern recognition and other expert skills. Global ratings, however, may be superior. Global rating scales scored by experts show higher inter-station reliability, better construct validity, and better concurrent validity than do checklists.[30, 31]

No matter what version of scoring is used, there is always concern for rater reliability and differences between rater evaluations. A study of one well established OSCE testing center revealed the presence of four common rater errors (leniency, inconsistency, the halo effect, and restriction of range) despite intensive rater training and experience.[32]

Practical Issues and Feasibility

The success of implementing OSCE testing depends to a great extent on addressing feasibility and practicality issues. Depending on the availability of resources, institutions often need to deviate from ideal test design situations. The cost of OSCE implementation is high in terms of personnel, facilities, finances, and time for examinees, SPs, and faculty. While the goal of reliability measures above 0.70 requires at least 6 hours of testing, in reality, it is nearly impossible to run OSCEs that long, even for large, well organized medical schools. Practical test lengths of 2–4 hours are common, and even these are still resource intensive. Direct costs can also be prohibitive. Dollar cost estimates include $200 minimum per student for acceptable reliability[33]and hourly costs of $15 per student.[34] Quebec Medical College, for example, spends $1,080 to assess the performance of each student in a comprehensive OSCE.[35] Financial pressures on most US medical schools make it difficult to consistently invest this amount of money on OSCE testing. Other practical considerations and potential barriers include recruitment, training, and retention of a large volume of SPs, time and training for faculty observers, test development costs in both time and expertise, and maintenance of usable clinical space to administer the test.

Hidden Benefits

Despite such potential barriers and questions about the psychometric properties of OSCEs, recent reports have highlighted the often hidden benefits of long-term, comprehensive OSCEs. For example, educators subjectively believe in high-fidelity assessment, and students and educators generally feel positive about this type of performance testing. Beyond this subjective experience, Duerson et al reported significant student, curricular, and faculty development outcomes after 9 years of OSCE testing.[36] Student performance improved, small-group teaching sessions were standardized, and faculty received feedback that improved instruction and enthusiasm for teaching physical exam skills. Students evaluated the experience positively and perceived the faculty time commitment as an expression of faculty interest in teaching. Moreover, after passing the OSCE, student confidence increased, and anxiety about upcoming clinical rotations decreased.

"Teaching to the test" is a common phenomenon that helps students pass a certain required assessment. In the case of OSCEs, teaching to the test would possibly lead to enhanced physical exam skills training, thus addressing a recognized deficiency in current medical school graduates.[37] The exact curricular content related to skills education is often clarified and standardized through consensus building. In one published report, the implementation of SP-based testing led to dramatic change in student learning activities, with more time spent on ward-based activities and less on preparation for written tests.[38] Also, the use of OSCEs for evaluation reinforces the patient-centered nature of medical practice, often provides timely and specific feedback on clinical performance, and reminds students that they are practitioners, not mere masters of medical knowledge.[39]

Conclusions

Medical education is a public trust. Indeed, medical educators have always needed the best methods for formative and summative evaluation of trainees. The renewed emphasis on patient safety and quality outcomes in the social consciousness and payer system necessitates that medical educators use high-quality, reliable, valid, educationally sound assessment methods. Direct observation in clinical simulations provides many opportunities for assessment and learning that other traditional evaluation methods also do not afford. The benefits of the OSCE method to learners, faculty, institutions, and the public at large are great.

Despite such benefits, care must be taken to maximize these benefits and generate reliable results. The *de facto* value of high-fidelity performance assessment with OSCEs has been long assumed but has yet to be concretely proven. Norman has summarized, "At best, performance assessment is about as good at predicting actual performance as a multiple-choice test based on relevant knowledge, but no better. There is little comfort here for the notion that performance assessment is, by its very nature, higher on [Miller's] pyramid and hence better."[9] The labor- and resource-intensive OSCE has become standard practice in modern assessment of clinical competence, and the results are used for high-stakes decision making at many levels. Many details must be managed to make one feel confident in the results.

Successful OSCEs are often the result of significant planning, coordination of multiple resources, commitment to large-scale testing, and judicious use of assessment data. Care must be taken to minimize the multiple sources of error and find validity evidence to justify OSCE use. Such attention to these issues—to do it right—comes with a hefty price tag. When high-stakes consequences hang in the balance, however, it is essential that these details are not taken lightly.

More research is needed about the best uses of the OSCE method and how to maximize reliability and

validity. Advocates of the OSCE method should continue to produce and disseminate evidence of the far-reaching impact that is seen by students and educators. All institutions should avoid the overreliance on any single evaluation method. Each institution must judge the relative value of comprehensive testing in light of local resources as well as the need to prepare students for the CSE component of the USMLE Step II.

Corresponding Author: Address correspondence to Dr Turner, Indiana University, Department of Family Medicine, 1110 W. Michigan Street, Long Hospital, 2nd Floor, Indianapolis, IN 46202. 317-278-0300. Fax: 317-274-4444. jlturner@clarian.org.

References

1. Howley L. Performance assessment in medical education: where we've been and where we're going. Eval Health Prof 2004;27(3):285-303.
2. Epstein R. Assessment in medical education. N Engl J Med 2007;356(4):387-96.
3. Miller G. The assessment of clinical skills/competence/performance. Acad Med 1990;65:S63–S67.
4. Bloom B. Taxonomy of educational objectives. In: Handbook I: the cognitive domain. New York: David McKay Co Inc, 1956.
5. Harden R, Stevenson M, Downie WW, Wilson GM. Clinical competence in using objective structured examination. Br Med J 1975;1:447-51.
6. van der Vleuten CPM, Swanson DB. Assessment of clinical skills with standardized patients: state of the art. Teach Learn Med 1990;2(2):58-76.
7. Dreyfus H, Dreyfus S. Mind over machine. New York: Free Press, 1986.
8. Barzansky B, Etzel SI. Educational programs in US medical schools, 2003–2004. JAMA 2004;292(9):1025-31.
9. Norman G. Inverting the pyramid. Adv Health Sci Educ 2005; 10:85-8.
10. Barman A. Critiques on the objective structured clinical examination. Ann Acad Med Singapore 2005;34(8):478-82.
11. Buckendahl CW, Plake BS. Evaluating tests. In: Downing SM, ed. Handbook of test development. Mahwah, NJ: Lawrence Erlbaum Associates, 2006:725-38.
12. Gledhill RF, Capatos D. Factors affecting the reliability of an objective structured clinical exaination (OSCE) in neurology. S Afr Med J 1985;67:463-7.
13. Rutala PJ, Witzke DB, Leko EO, Fulginiti JV, Taylor PJ. Student fatigue as a variable affecting perrformance in an objective standardized clinical examination. Acad Med 1990;65:S53-S54.
14. Tamblyn RM, Klass DJ, Schnabl GK, Kopelow ML. Sources of unreliability and bias in standardized patient rating. Teach Learn Med 1991;3:74-85.
15. Swanson DB, Norcini JJ. Factors influencing reliability of tests using standardized patients. Teach Learn Med 1989;1:158-66.
16. Reznick RK, Blackmore D, Cohen R, et al. An objective structured clinical examination for the licentiate of the Medical Council of Canada: from research to reality. Acad Med 1993;68(suppl):S4-S6.
17. Petrusa ER, Blackwell TA, Ainsworth MA. Performance of internal medicine house officers on a short station OSCE. In: Hart I, ed. Further developments in assessing clinical competence. Montreal: Can-Heal, 1987:598-608.
18. Dawson-Saunders B, Verhulst S, Marcy M, Steward D. Variability in standardized patients and its effect on student performance. In: Hart I, ed. Further developments in assessing clinical competence. Montreal: Can-Heal, 1987:451-8.
19. Boulet JR, McKinley DW, Whelan GP, Hambleton RK. Quality assurance methods for performance-based assessments. Adv Health Sci Educ 2003;8:27-47.
20. Schuwirth L, van der Vleuten C. The use of clinical simulations in assessment. Med Educ 2003;37(suppl):65–71.
21. Nayer M. An overview of the objective structured clinical examination. Physiother Can 1993;45(3):171-8.
22. Sloan DA, Donnelly MB, Schwartz RW, Strodel WE. The objective structured clinical examination. The new gold standard for evaluating postgraduate clinical performance. Ann Surg 1995;222(6):735-42.
23. Bergus GR, Kreiter CD. The reliability of summative judgements based on objective structured clinical examination cases distributed across the clinical year. Med Educ 2007;41(7):661–6.
24. Harden RM, Gleeson FA. Assessment of clinical competence using an objective structured clinical examination (OSCE). Med Educ 1979;13:41-54.
25. Newble DI, Noare J, Elmslie RG. Validity and relaibility of a new examination of the clinical competence of medical students. Med Educ 1981;15:46-52.
26. Brown R, Roberts J, Rankin J, Stevens B, Tompins C, Patton D. Further developments in assessing clinical competence. In: Hart IR, Walton HJ, eds. Further developments in assessing clinical competence. Montreal: Canadian Health Publications, 1987:563-71.
27. Ram P, van der Vleuten C, Rethans JJ, Schouten B, Homba S, Grol R. Assessment in general practice: the predictive value of written knowledge tests and a multiple-station examination for actual medical performance in daily practice. Med Educ 1999;33:197-203.
28. Auewarakul C, Downing SM, Jaturatamrong U, Praditsuwan R. Sources of validity evidence for an internal medicine student evaluation system: an evaluative study of assessment methods. Med Educ 2005;39:276-83.
29. Mazor KM, Ockene JK, Rogers HJ, Carlin MM, Quirk ME. The relationship between checklist scores on a communication OSCE and analogue patients' perceptions of communication. Adv Health Sci Educ 2005;10(1):37-51.
30. Regehr G, MacRae H, Reznick RK, Szalay D. Comparing the psychometric properties of checklists and global rating scales for assessing performance on an OSCE-format examination. Acad Med 1998;73(9):993-7.
31. Hodges B, McIlroy JH. Analytic global OSCE ratings are sensitive to level of training. Med Educ 2003;37(11):1012-6.
32. Iramaneerat C, Yudkowsky R. Rater errors in a clinical skills assessment of medical students. Eval Health Prof 2007;30(3):266-83.
33. Stillman PL, Swanson DB, Smee S, et al. Assessing clinical skills of residents with standardized patients. Ann Intern Med 1986;105:762-71.
34. Petrusa ER, Blackwell TA, Rogers LP, Saydjari C, Parcel S, Guckian JC. An objective measure of clinical performance. Am J Med 1987;83:34-42.
35. Grand'Mason P, Lescop J, Rainsberry P, Brailovsky CA. Large-scale use of an objective structured clinical examination for licensing family physicians. CMAJ 1992;146:1735-40.
36. Duerson MC, Romrell LJ, Stevens CB. Impacting faculty teaching and student performance: nine years' experience with the objective structured clinical examination. Teach Learn Med 2000;12(4):176-82.
37. Ozuah PO, Curtis J, Dinkevich E. Physical examination skills of US and international medical graduates. JAMA 2001;268(9):1021.
38. Newble D, Jeager K. The effect of assessment and examinations on the learning of medical students. Med Educ 1983;17:165-71.
39. Barrows HS. An overview of the uses of standardized patients for teaching and evaluating clinical skills. Acad Med 1993;68(6):443-53.