The Objective Structured Clinical Examination
A Step in the Direction of Competency-Based Evaluation

Carol Carraccio, MD; Robert Englander, MD, MPH

Background: The Accreditation Council for Graduate Medical Education is embarking on the major task of a paradigm shift in graduate education in the direction of competency-based medical education and evaluation of outcomes. The Objective Structured Clinical Examination (OSCE), a measure of clinical competence that focuses on outcomes via observable behaviors, is gaining national recognition.

Objective: To review the pediatric literature relevant to the OSCE.

Method: A MEDLINE search from the date of the original report of the OSCE (1975) to the present was performed. All English-language studies regarding the use of the OSCE in pediatric education published in the United States and Great Britain were reviewed.

Main Outcome Measures: Reliability and validity of the OSCE were examined. Use of standardized pediatric patients was discussed.

Results: A greater number of stations and similarity between tasks at different stations increased the reliability of the OSCE. A greater number of stations increased sampling of material and content validity. Correlation between the OSCE and precertification examinations ranged between 0.59 and 0.71, with \( P \leq .01 \). Correlation between the OSCE and monthly clinical evaluations was much lower (0.39-0.57), but still statistically significant at \( P \leq .05 \). Gaps between expected and actual performance were documented. Overall, the experience of being a standardized patient was viewed as positive by children and their parents.

Conclusions: With appropriate attention to design, acceptable reliability and validity can be achieved for the OSCE. Significant correlations between the OSCE and precertification examinations as well as monthly clinical evaluations were found, the former being stronger than the latter. We conclude that the combination of the OSCE, standardized board examinations, and direct observation in the clinical setting has the potential to become the "gold standard" for measuring physician competence.

patients into the US Medical Licensing Step 3 Examination within the next several years.

We reviewed the existing literature published in the United States and Great Britain relevant to the use of standardized patients in medical student and resident education within the field of pediatrics. For comparison, we reviewed recent key literature using adult standardized patients. Our intent is to raise awareness regarding the effect of study design on the reliability of the OSCE and to highlight the relationship between performance on the OSCE as a measure of competence and other, more traditional, measures of clinical competence.

EVOLUTION OF THE OSCE

The OSCE was introduced as a way of measuring clinical competence that allowed for control of many of the biases of conventional methods. Traditional clinical examinations are given by a single faculty member who observes a student or resident taking a history and performing a physical examination. The learner must be prepared to discuss the case and answer questions posed by the examiner. The student's or resident's evaluation is subject to the whims of both the examiner and the patient. The first description of the OSCE in the medical literature appeared in 1975.1 The purpose of the article was to describe a test of clinical competence that avoided many of the disadvantages of the traditional clinical examination. Thirty-three students spent 5 minutes at each of 16 stations, either procedure stations or question-and-answer stations. Each procedure station was followed by a question-and-answer station that pertained to the previous procedure station. At procedure stations learners were asked to perform a focused history and/or physical examination on standardized patients or to perform other focused tasks such as interpretation of x-rays, microscopic slides, or electrocardiograms. The use of simulated patients helps spare faculty time commitment.

The OSCE was blinded to the results of the OSCE. The authors, however, emphasized an important practical point about the OSCE: their examination was used as a criterion-referenced assessment as opposed to the norm-referenced assessment of a traditional examination. Intuitively, this makes sense for evaluating a candidate's competency to practice in the real world. It makes little difference to have one candidate score better than another, and potentially allow the better-scoring one to pass, when neither reach baseline competency for independent practice.

The other studies from the early 1980s begin to address the question of how the OSCE compares with traditional tests in a more systematic way. Watson et al.3 used an OSCE of 20 stations for 67 students, allowing them 4 minutes per station. They compared the results of student scores on this assessment with their traditional grade, composed of clinical and senior tutor reports, project marks, multiple-choice test results, and a 10-minute oral examination. The faculty member assigning these grades was blinded to the results of the OSCE. The authors took this criterion-referenced assessment (the OSCE) and converted it back to a norm-referenced test to determine whether the OSCE results sorted the students into the same grade categories as did the traditional measures. They did so by taking the mean score of the 67 students and empirically delineating a grade of A for students scoring 1 SD greater than the mean, a grade of B for those scoring between the mean and 1 SD above, a grade of C for scores between the mean and 1 SD below, and a grade of D for scores greater than 1 SD below the mean. Percentages were close for all grades except C and D, where 49% and 6% obtained a C and D, respectively, with the traditional grading system as compared with 30% and 13.5%, respectively, with the OSCE. There was a 55% concordance in the final grading system, with the discordance favoring upgrading by the traditional grading system. Despite the imperfections, the authors felt that the OSCE served to highlight the relationship between performance on the OSCE as a measure of competence and other, more traditional, measures of clinical competence.

Since this original description of the OSCE, a growing body of literature has evolved to address permutations of the test and describe findings specific to a variety of disciplines. Pediatrics has lagged behind some other disciplines, likely owing to the real or perceived intrinsic difficulties of using children as standardized patients. It was not until the early 1980s that 3 articles highlighted the use of the OSCE in pediatrics.4,5 The first article was merely descriptive.6 Similar to the initial OSCE, the test included 20 stations, allowing students 4 minutes per station. The set-up was essentially identical to that described in the original article, with the exception that the mother of the simulated patient was also asked to comment on student performance. Test reliability and validity were not addressed. The authors, however, emphasized an important practical point about the OSCE: their examination was used as a criterion-referenced assessment as opposed to the norm-referenced assessment of a traditional examination. Intuitively, this makes sense for evaluating a candidate's competency to practice in the real world. It makes little difference to have one candidate score better than another, and potentially allow the better-scoring one to pass, when neither reach baseline competency for independent practice.
they designated as "overall performance." When data were normally distributed, Pearson correlation coefficients were used, and for nonparametric data Kendall correlation coefficients were used. The correlations ranged from 0.11 to 0.32 and all were statistically significant; \( P < .01 \) for the OSCE with the multiple-choice test and \( P < .001 \) for the comparable clinical examinations in other subjects, the in-case assessment, and the overall performance.

**RELIABILITY OF THE OSCE**

It wasn't until the 1990s that investigators began to look more critically at the OSCE as a test. A study by Matsell et al.8 incorporated 10 stations for 77 students. The learners were divided into 4 groups, each taking a different but similar OSCE. In this study, each station evaluated one or several clinical skills, including history-taking, physical examination, interpersonal attributes, ability to generate a differential diagnosis, management, and interpretation of laboratory findings. Interstation reliability, calculated using the Cronbach \( \alpha \), showed a wide range (0.34, 0.12, 0.54, and 0.69). Although one may include the skills listed above all under the rubric of clinical competence, they are different skills. This diversity seems to be a logical explanation for the wide range as well as the low reliability of the scores. The caveat is to ensure a large enough number of stations to maintain an acceptable degree of reliability. The positive relationship between number of stations and reliability was demonstrated in the work by Joorabchi.6 This study introduced pediatric residents as learners and examination participants in addition to medical students completing clinical clerkships in pediatrics. Twenty-nine residents and 6 students completing a 34-station OSCE, with 5 minutes of testing time at each, resulted in a coefficient \( \alpha = 0.80 \) when each station was used as a test item. Dividing stations into 2 equivalent halves based on what was being measured showed a split half reliability index of 0.83. Even correlations between scores derived for the various subsections such as history stations, physical examination stations, and laboratory stations were in the 0.76 to 0.88 range, all statistically significant with \( P < .001 \). In addition to the number of stations used in the OSCE, interstation reliability seems to be dependent on the type of task assignment per station.7 In the study by Hillard and Tallett,7 despite the fact that only 5 stations were used in the OSCE, the Cronbach \( \alpha \) was measured at 0.69. When one takes a critical look at the description of the stations, one sees that all stations focused on interviewing and history-taking skills with a standardized patient. Thus, homogeneity of tasks at different stations increases the reliability.

Three recent studies have broken the OSCE down into components and looked at the reliability of the components.8,10 Joorabchi and Devries8 administered the OSCE in 1990, 1991, and 1993 to 29, 32, and 65 residents, respectively, using 30 to 34 stations. The authors found overall generalizability coefficients to be 0.80, 0.81, and 0.86 for each of the 3 years.8 The coefficients for the 3 components—data gathering, applications, and communication—ranged from 0.63 to 0.82 in each of the 3 study years. Rosebraugh et al.9 used a 10-station OSCE and tested 196 students; their results showed a reproducibility coefficient greater than 0.80 in the 2 components of the examination that were studied (presentation and problem solving).9 Lane et al.10 tested 56 residents, using a clinical skills assessment examination at each of 10 stations and allotting 22 minutes per station. Breaking the test down into the 4 component parts of history, physical examination, documentation, and interpersonal skills showed reliability across components with coefficient \( \alpha = 0.69, 0.64, 0.81, \) and 0.76, respectively. These 3 studies lend further credence to the reliability of the OSCE when number and content of stations are taken into account in the design.

Studies using adult standardized patients support the reliability of the OSCE, with reliability coefficients in the 0.40 to 0.91 range, with the majority in the 0.60 to 0.91 range.11-15 In a 1990 review of adult literature assessing use of standardized patients, van der Vleuten and Swanson16 found the following factors to improve reliability: use of checklists as opposed to rating scales, standardized training of patients to maximize reproducibility of individual station performances, minimum of 3 to 4 hours of testing time, stations assessing hands-on clinical skills with patients as opposed to stations using written items, and use of norm-referenced scores (although one may argue that criterion-referenced scores are a more appropriate, if less reliable, measure of the examinee’s competence).

**VALIDITY OF THE OSCE**

As with reliability, the validity of the OSCE was not critically appraised until the early part of this decade. The 2 studies that used more than 30 testing stations in their OSCE attest to the content validity of their studies as acceptable based on extensive faculty review to ensure that all common problems were addressed.6,8 One questions the content validity of those studies with fewer than 10 stations in terms of the ability to incorporate all the necessary material to cover, even superficially, such an extensive body of knowledge.7,17

Several studies addressed construct validity by comparing scores on the OSCE with level of training. Two studies found statistically significant \( P \) values of <.01 when they compared scores on the OSCE with level of training.6,8 Of the 2 studies that did not achieve statistical significance for this measure of their OSCE, one had pediatric residents at only 2 different levels of experience mixed with family practice residents and the other had small numbers with 43 residents distributed unevenly over 4 experience levels.7,10 A 1990 review of tests using adult standardized patients by van der Vleuten and Swanson16 likewise found that groups at different stages of training obtained appropriately different scores. Since that time further studies using adult standardized patients have also showed modest construct validity as measured by the OSCE’s ability to differentiate by level of training.7,11,13-15

The studies which addressed concurrent validity did so by correlating scores on the OSCE with scores on board-sponsored precertification examinations and with monthly clinical evaluations of residents by faculty. Correlations
between the OSCE and the precertification examinations ranged from 0.59 to 0.71, all being statistically significant at the P<.01 level or lower.2,6,8 The correlations in the adult literature are similar and range from 0.37 to 0.68.12,13,18 Correlations of the OSCE with monthly clinical evaluations were typically lower than with the precertification examinations but did reach statistical significance. In these studies the correlations ranged from 0.39 to 0.57 with P<.05 or lower.2,6 Similar correlations are found in the adult literature.12,19 In their review of tests using standardized patients, van der Vleuten and Swanson16 statistically correct for measurement error by making the tests perfectly reliable. In all cases the true correlations between other measures of competence, both written tests and clinical evaluations, increased greatly over the observed correlations, with the majority of coefficients greater than 0.40. However, no “gold standard” for measuring clinical competence against the OSCE has been established. The American Board of Pediatric Certification Examination is the accepted standard for documenting that “competency” to practice pediatrics has been achieved. Despite the reliability and the content validity of this examination, criterion validity suffers from the same lack of appropriate comparisons as the OSCE. One may speculate as to the possibility of a study comparing the OSCE and the board certification examination. Unfortunately, the financial and personnel resources that would be required to use an OSCE to measure a representative sample of the entire content domain of the written certification examination would be prohibitive.

THE OSCE AS A FORMATIVE EVALUATION TOOL

One of the clear benefits of the OSCE is the stimulus it provides for formative evaluation of both the learner and the teaching program. Feedback to the individual student can be accomplished at all stations with faculty observers. Immediate feedback during the initial phase of the OSCE has been studied and actually was found to both improve competency at subsequent stations and improve the quality of the learning experience for the examinee.20 In addition, review of group performance on the OSCE is helpful in demonstrating to faculty areas of weakness in the educational program. This use of the OSCE has been suggested in both the adult and pediatric literature.1,3,8-10,13,18

Noteworthy from the review of the use of the OSCE in pediatrics is the gap between faculty expectations of the learner’s performance and the actual scores achieved by the learner. A survey was mailed to 59 preceptors who taught students in a community practice setting.11 The survey contained 2 of the same cases given on the OSCE with questions identical or similar to those used in testing. The community preceptors were asked to score the items according to how they thought their own student would perform. There was a 64% response rate. For a case on anemia, the correlation between the preceptor’s proposed score and the student’s actual score was 0.19 (P = .15). For a case dealing with growth, the correlation between the preceptor’s proposed score and the student’s actual score was 0.41 (P = .06). This represents a significant disparity between what the community preceptors predicted their students would know and the knowledge base their students demonstrated. Although not tested, there may also be disparity between what academic faculty feel a student should know as compared with the community faculty, since the former group created the test. Involving the community as well as the academic faculty in the OSCE design may increase the content validity of the test.

A disconcerting study looked at the gap between faculty expectations of resident performance and actual resident performance.6 Before the test administration, the faculty comprising the planning task force agreed on the correct answers and arrived at a consensus score that a minimally competent resident must achieve to pass. A minimum pass level (MPL) was calculated for each station and subsection of the test. Separate MPLs were calculated for the first- and third-year residents and the level for the second-year residents was an average of both. Of the 64 first-year pediatric residents taking the OSCE, 41% scored below the MPL, which was set at 88% of the maximum score for this group. For the 36 second-year residents, 55% scored below the MPL, which was set at 57% of the maximum score for this group. Of the 26 third-year residents taking the examination, 96% scored below the MPL, which was set at 68% of the maximum. Despite similar findings by other investigators, the revelation of such a disturbing discrepancy between faculty expectations and resident performance in one’s own training program compels a comprehensive, critical evaluation of existing educational practices.8 Why the poor performance on the OSCE? The residents functioned well in the clinical setting. Acceptable test reliability and validity were demonstrated in their study. The authors speculated that unrealistic expectations of the faculty, the contrived nature of the test, and faculty equation of cognitive skills with clinical proficiency may have all contributed to the gap between faculty expectations and resident performance. They conclude that we need to move to a competency-based curriculum with competencies expressed in specific observable behaviors.

POTENTIAL PITFALLS OF THE OSCE

One concern in using pediatric standardized patients is the need for large numbers of children to account for a fatigue factor during prolonged testing periods. The author of a recent editorial suggests using a variety of tasks at the stations to minimize the need for large numbers of children as standardized patients. Videotaped recordings of physical findings, microscopic slides, photographs of dermatologic findings, x-rays, and other forms of imaging have potential value in an OSCE.21 These adjuncts minimize the number of children needed and allow for substitutes at the stations with children to alleviate the potential for patient fatigue, which may lead to the child’s unwillingness to cooperate.

Another concern is the potential negative effect on the children who acted as standardized patients. One group of researchers addressed this concern by holding focus-group sessions with the children and their parents. All of the real parents felt that they would allow their
children to be standardized patients again. They observed that their children were proud about participating and that they believed they had worked hard at a “real” job to earn money. Both the real and the standardized parents felt that the children had actually learned a lot from the process. All the children liked the fact that their participation in the OSCE gave them a chance to earn money. In another random sample of 7 standardized patients younger than 18 years, Woodward and Gliva-McConvey found that young children tended to view the simulation as play-acting and found it fun. One negative experience occurred when a 6-year-old, simulating an emergency, overheard the physician talking about death. The concept of death at her age had not occurred to her and this was subsequently discussed with the child at length. The older children felt that the simulation made them realize that learning is a lifelong process, that adults are fallible, and that it is important to understand the experiences and problems of others. Overall, the experience was viewed as positive and the authors underscore the importance of ensuring that the case assignment and thus the learning be commensurate with the child’s developmental age.

A major pitfall of the OSCE is the time and financial resources required for administration. Estimated costs vary widely depending on number of examinees and number of testing stations. One study quoted a cost of $6.90 per student per station. Other studies, looking just at per-student costs, found a range from a low of $54 to a high of $496 per student. The discrepancies in cost per-student were related to differences in pay to the standardized patients, honoraria for the faculty, number of stations and examinees, and total time of the examination.

Any examination that involves rotation of examinees through various stations raises the potential for lapses in security through sharing of information. One study addressed this issue specifically and found no difference in scores over the course of several administrations or through the 8.5 hours of each individual administration of the OSCE. The authors conclude that it is unlikely that information was shared or, even if it was, that it had no significant effect on scores.

In those studies that sought learner feedback regarding the test, the results were mixed. One group of 126 residents found the experience challenging but less fair or enjoyable than other forms of testing. In another study of 229 students, four fifths felt the OSCE to be more fair and two thirds felt it was less stressful than their usual methods of testing. They were equally divided as to whether they were actually examined on those aspects of the course that were emphasized in the teaching program.

WHERE HAVE WE COME?

Clinical performance is the outcome of primary concern in medical education, but the consistent measurement of clinical performance remains elusive. It has been 25 years since the first report of the use of the OSCE as a measure of clinical competence appeared in the medical literature. With the exception of a few limited and mostly descriptive studies, the use of the OSCE in pediatrics did not receive much attention until the beginning of this decade. The real or perceived intrinsic difficulties of using children as standardized patients may be responsible for the lag behind other specialties in its more widespread adoption. As researchers took a more critical look at the OSCE, through studies of reliability and validity, important information about the design of the test became evident. Standardized checklists used for grading clinical stations have good inter-rater reliability. Interstation reliability is contingent on both number of stations and diversity of activity across stations. With a small number of stations, tasks across stations must be similar to achieve acceptable reliability. With a large number of stations, one can increase diversity of tasks and still maintain acceptable reliability. Content validity may be achieved provided that there are enough stations to be representative of the content domain being tested. Further study is needed in this area. Construct validity is supported by an increase in scores with an increase in experience. Statistically significant correlations have been achieved between the OSCE and precertification examinations and between the OSCE and clinical evaluations, the former being stronger than the latter. One may question whether these are appropriate comparisons for determining concurrent validity, but a gold standard for comparison does not exist. The gap that has been demonstrated between faculty expectations of student and resident performance is disconcerting and raises questions about the effectiveness of medical education in general and also the methods used to evaluate its effectiveness.

WHERE ARE WE GOING?

The incorporation of the OSCE or some form of clinical skills assessment into the Canadian Licensing Examination, the certification process of the Educational Commission for Foreign Medical Graduates, and the US Medical Licensing Examination is clear evidence of the direction in which medical education is moving. The investment of the Accreditation Council for Graduate Medical Education in national pilot projects for the development of clinical competencies and outcome measures attests to the paradigm shift in education that will need to occur in our residency training programs.

The evidence to date suggests that with appropriate attention to design, the OSCE is a reliable test with modest validity. The challenge that we face as we move into the 21st century is to create a competency-based curriculum for medical education and to further develop and refine measurement tools to evaluate that curriculum. The literature to date supports a strong role for the OSCE in the evaluation process. The combination of the OSCE, the measure of knowledge by standardized board examinations, and direct observation in the clinical setting together has the potential to become the gold standard for measuring physician competence.

Accepted for publication December 17, 1999.

Corresponding author: Carol Carraccio, MD, Department of Pediatrics, Room N5W36, University of Maryland, 22 S Greene St, Baltimore, MD 21201.
REFERENCES


Educational Interventions

Purpose: This section is intended to share information concerning educational efforts in the broad field of pediatrics. We welcome studies on the following topics: undergraduate and graduate education in medicine and allied health occupations; continuing education of health professionals; education of patients and families; and health education for the general public, the community, and organizations that contribute to the promotion and improvement of the health of children and adolescents.