Sharing of Information by Students in an Objective Structured Clinical Examination

Paul J. Rutala, MD; Donald B. Witzke, PhD; Elizabeth O. Leko, MPA; John V. Fulginiti, MA; Preston J. Taylor, MD

• **Increasing numbers of medical schools are using Objective Structured Clinical Examinations (OSCEs) to evaluate students. An Objective Structured Clinical Examination employs a multiple-station format and standardized patients to document students' clinical skills. A lengthy format is necessary; testing an entire class often necessitates multiple repetitions of the same examination. This dictates a need to minimize sharing of information among students. We studied six administrations of an Objective Structured Clinical Examination designed to measure skills. Analyses were conducted to detect changes in scores over the administrations as well as over the 8.5 hours of each day of testing. An increase in either might indicate information sharing had occurred. No significant increase occurred. If information was shared, it had no significant effect on scores. Skills a student uses to approach a patient should not change even if the patient's complaints are known.**

(*Arch Intern Med.* 1991;151:541-544)

There is a growing movement throughout the US and Canadian medical schools to document objectively each student's clinical competence prior to graduation through the use of standardized clinical situations in which students perform specific patient-related tasks and/or demonstrate problem-solving skills. One commonly employed technique for such an evaluation of student competency is the Objective Structured Clinical Examination (OSCE). Two recent comprehensive reviews[1,2] document the increasing interest expressed in the OSCE since its original description by Harden and coworkers[3] and Harden and Gleeson[4] in the 1970s. Consulting the Curriculum Directory of the Association of American Medical Colleges for the past few years, one sees a growing number of schools reporting the development or use of OSCEs.[5-8] Such a trend toward more objective skills documentation is welcome because for too long the major emphasis in student evaluation has been placed on multiple-choice examinations to certify a knowledge base.

In a typical OSCE students are presented a clinical problem (either a patient or patient-related material) and are asked to deal with that, using a requisite set of clinical skills and knowledge. Performance is usually evaluated by a trained standardized patient (SP) or other observer, using checklists and rating forms. Written records generated by the student may also be scored later by faculty members. Questions (usually multiple-choice) presented on paper following the problem situation may give additional evaluative data. All expectations of the student are based on reset criteria developed through faculty consensus.

Students move from room to room encountering a series of evaluations and clinical problems, most often numbering between 15 and 20. Each room or task is referred to as a "station." Although all OSCEs have in common the multiple-station format, other factors such as the number of stations used, length of time spent by a student in each station, duration of overall testing, and inclusion or not of multiple-choice testing within the format vary considerably from school to school.[2] Another variable imposed by logistics is the number of OSCE administrations necessary to examine an entire cohort of students.

The total amount of OSCE evaluation time necessary to obtain reliable SP-generated scores has been estimated to be approximately 5 hours.[2] Orientation and preparation time will increase the actual administration time beyond the 5 hours of evaluation. If the OSCE contains 20 stations, a maximum of 20 students can be evaluated in a single day, since only one student can be in each station at any one time. Unless multiple patients or SPs are available for each station, a summative fourth-year evaluation for even a small graduating class of 60 to 80 students would require three to four OSCE administrations.

If one were to employ an OSCE as part of a clerkship evaluation, it might be decided to conduct an examination at the end of each clerkship block. If the same OSCE is given every 6 to 12 weeks throughout a year, security must be maintained and an increase in scores must reflect only im-

Accepted for publication September 10, 1990.

From the Department of Internal Medicine (Drs Rutala and Taylor), Preparation for Clinical Medicine (Dr Rutala and Ms Leko), and Division of Academic Resources (Dr Witzke and Mr Fulginiti), College of Medicine, University of Arizona, Tucson.

Reprint requests to Preparation for Clinical Medicine, College of Medicine, University of Arizona, Tucson, AZ 85724 (Dr Rutala).

proved student performance and not student information sharing. As considerations of OSCEs for national certification and licensure continue,[1] the potentially widespread ramifications of testing a single cohort in multiple administrations must be studied carefully. Performance comparisons on which major professional decisions are made must reflect actual examinee ability and not be significantly affected by the examination format employed.[9]

In this study we investigated the effect on examinees' scores of repeated, serial administrations of essentially the same precommencement OSCE over a 2-month period in an effort to see if there was any evidence of significant information being shared among examinees (student-student cueing). Little has been written on this subject. In 1985 Gledhill and Capatos[9] reported an OSCE in which some evidence of cueing was seen in scores obtained on written testing portions of the examination that were interspersed among practical patient-related exercises. It has been assumed by other authors that cueing among students would occur in other OSCE formats and this has been cited as a limitation of the technique.[10] Williams and coworkers,[11] however, reported in 1987 indirect evidence of no cueing effect based on stable scores over 5 days of testing in a 2-week period. This OSCE also involved written testing.

In this study, each of 76 examinees was to pass through 16 OSCE stations. A full administration took 8.5 hours. A minimum of 5 days would be required if each examinee completed all 16 stations in 1 day. In fact, six separate administrations were conducted extended over an 8-week period. We hypothesized that in our OSCE specifically designed to measure skill rather than knowledge, ie, one that included no formal written testing, little or no cueing would take place.

METHODS

In August and September 1989, the 76 fourth-year students of the class of 1990 at the University of Arizona (Tucson) College of Medicine (UA) took the OSCE. Students self-selected for one of six Saturday administrations over 8 weeks. Students tended to schedule themselves toward the end of the administration period, as can be seen in Table 1. No OSCE was held in week 4 (Labor Day weekend) nor in week 7 (the Saturday prior to many of the students' taking part two of the National Board Examination).

Students were required to participate by vote of the general faculty but no passing score was required. Each student was instructed that the OSCE was subject to the UA honor code as would be any other examination. Information sharing was discouraged particularly since a student might conclude what the intent of a station was only to be incorrect; sharing of information might in fact jeopardize another student's chance of performing well.

This was the first OSCE experience for each student; however, 73 of the examinees had completed a minimum of 13 interactions with SPs during their second year of medical school in the UA patient instructor program. The other three students were transfer students who entered the UA at the start of the third year. These three had undergone eight diagnostic sessions each with patient instructors at the start of their training at the UA. Therefore, although no student had OSCE experience, all had experience with SPs.

In the OSCE, performance of each student was assessed on 16 clinical problems. Because of scheduling conflicts, 18 stations were actually used during the 6 weeks. The two backup stations were used only if absolutely needed—one for nine students only, the other for 20 students.

Sixteen of the 18 stations involved SPs. Only one SP was trained for each role, ie, every student encountering any particular complaint would always see the same SP. All physical findings were real. All

Table 1.—Number of Students Tested by Week*

Week	No. of Students
1	9
2	11
3	14
4	No OSCE given
5	10
6	16
7	No OSCE given
8	16
All weeks	76

*OSCE indicates Objective Structured Clinical Examination.

SPs had been trained to complete checklist evaluations describing the quality and completeness of their interaction with each student. Each checklist item was counted as one possible point. No faculty observers were used. The two nonpatient stations involved only interpretation of patient material, eg, electrocardiograms, arterial blood gases, urinalyses, Gram's stains, and roentgenograms.

Each station was 18 minutes in length. Between stations there was a 6-minute period. During this time, the student could prepare for the next station, eg, review the patient's chart or organize an approach. During the same time, the SP was completing the checklist for the student just seen. Preparation time and time allowed for the interaction were seen to be adequate or even too much by the vast majority of students (97% for preparation time; 90% for the time allowed for the interaction).

Although only a maximum of 4.8 hours were spent in the stations, each student was required to devote 8.5 hours to the OSCE. The first half hour was orientation. The preparation periods totaled 1.6 hours. There were midmorning and midafternoon breaks (0.6 hours total). One hour was given for lunch. The format allowed no more than four consecutive evaluation sessions for any student or SP without a break, ie, no more than 1.6 hours in evaluation before a 0.3- to 1-hour break.

Students who did not use the full 18 minutes of SP contact time were asked to fill out a brief station evaluation and to have a seat in the clinic lobby. The OSCE staff occasionally passed into the lobby to engage students in conversation and to decrease the opportunity for information sharing.

The clinical problems presented were common and were submitted by 10 different authors representing the disciplines of all the basic, required, third-year clerkships. Problems were chosen for use from among 63 station blueprints submitted to the coordinators. Final selection of stations was made by a 14-member case-development group (composed largely of station authors and required clerkship coordinators) based on importance of the clinical problem and its common occurrence in clinical practice. None of us participated in the case selection process. One third of the stations included in the OSCE were authored by internal medicine faculty.

All patient stations required the student to interview the SP; in two stations there were two SP evaluators present (in one a husband-patient and his wife, in the other a patient and an interpreter). Seven of the stations involved doing directed physical examinations. Seven stations included the requirement of a written progress note.

The number of points possible for each station was not equalized and varied from 38 to 95 on the patient stations (mean, 66 points) and from 22 to 58 on the nonpatient stations (mean, 40 points). The mean number of total points possible for each student was 1060. Station scores are the percent of possible items obtained for that station. Total scores are the unweighted arithmetic mean of individual station percent scores obtained on all stations completed by the student.

Each student was randomly assigned to the first station of the day and then rotated counterclockwise through the circuit of examining rooms in an outpatient clinic. There were 1214 student evaluation sessions conducted (two students had to leave early, completing only 15 stations each). Of those 1214, 1185 sessions (98%) were conducted

Table 2.—Range of Scores and Mean Score by Week of the Objective Structured Clinical Examination Administration

Week	Range of Scores, %	Mean, %
1	57.0-72.0	64.5
2	58.2-68.0	63.5
3	49.5-67.3	63.1
5	53.1-70.4	61.1
6	56.8-74.8	64.9
8	56.2-69.8	63.7
All weeks	49.5-74.8	63.6

in the core 16 stations. Although some students passed through a slightly different form of the OSCE, there were enough scores derived from the core stations that any effect of significant sharing of information should have been discernible.

Analyses were conducted comparing the relationship of week of administration with scores on each individual station as well as with total examination score. An increase in scores later in the administration period might be taken to indicate that significant sharing of information about the stations had occurred among examinees or at least that examinees in later weeks had been given some significant idea of what to expect from the OSCE.

In further analyses, scores of morning sessions were compared with those of afternoon sessions, for each individual station as well as across stations. If scores in the afternoon were higher, it might be inferred that during the lunch break sharing of information had occurred (all students ate together, with lunch provided by the college).

RESULTS

Of the 18 stations used for evaluation during the study, 14 were used all 6 days, one was used 5.5 days, and one was used 5 days. The two backup stations were used only 1.5 days or 1 day. None of the core 16 stations used in a minimum of five administrations showed any significant change in scores over the testing period.

Table 2 shows the mean total score by week of administration and the range of scores for each week. It can be seen that there was no systematic change over the six administrations in either parameter.

No station showed any significant difference when its morning and afternoon scores were compared: eight showed slightly lower scores in the afternoon; for 10 stations, afternoon scores were higher.

As might be expected from the above, total scores calculated for the morning and the afternoon were also the same. A mean of all eight morning percent scores was calculated for each student. The mean of these 76 means was 63.8% (SD, 6.8%). A mean of all afternoon percent scores was also calculated for each student. This represented the mean of eight scores for each of 74 students, and the mean of only seven scores for the other two. The mean of these 76 means was 63.3% (SD, 6.9%). The difference was not significant.

COMMENT

Swanson and Stillman[1] have recently admonished all workers investigating the OSCE technique to heed Kane's[12] argument for analyzing all "threats to validity" that might interfere with measurement of the intended domain. We have studied a potentially serious problem and have found no evidence of sharing of significant information among examinees in a format employing multiple administrations of an OSCE in a single testing site.

It was necessary to conduct repeated administrations of the OSCE to examine the entire class. The facility chosen was an outpatient, internal medicine clinic, unavailable to us for a full day except on weekends. An effort to provide maximum flexibility to students in scheduling as well as the occurrence during the administration period of the Labor Day holiday and part two of the National Board Examination led to six administrations of the OSCE extended over an 8-week period.

No significant change of individual station scores or overall OSCE scores was seen over the 2 months. This is taken as indication that no *significant* information was shared during the time and that security was maintained. In fact, no change was seen within any one day of administration either. Although students had ample opportunity to share information regarding the stations, the effect of such, if any, was insignificant.

The OSCE reported by Gledhill and Capatos[9] had a format different from the current one; it included objective written testing in addition to practical (patient-based) testing. A serial increase in mean score on written questions included in the OSCE was seen on all but one of 5 days of administration, suggesting intraclass information sharing. Williams and coworkers[11] also included written testing in their OSCE but they do not report scores for that separate from practical testing. An effect similar to that described by Gledhill and Capatos[9] could have been operative. As the currently described OSCE included no written testing, any such effect becomes irrelevant and this may have contributed to the lack of any cueing being demonstrated.

Gledhill and Capatos[9] further concluded that their use of several parallel patients for each station in an attempt to minimize patient fatigue and the use of several faculty observers introduced factors of variability that precluded identification of student-student cueing when looking at scores obtained in the practical portions of their OSCE. Our study used only the SPs as scorers and only a single SP for each complaint. With this reduced station variability, again no cueing effect was seen.

We would not be so naive as to claim that all students abided strictly by the honor code and did not speak to each other at all of their earlier encounters; this OSCE, though, was not designed to measure knowledge. Its major thrust was to measure skills that, in fact, the students began to acquire 2 to 3 years prior to the OSCE and that had been honed and polished throughout medical school. Thus, we were attempting to measure those enduring skills, developed over time, that with instruction and practice, become stylistic. Even if the patient's complaint is known, the student's style and manner of performance should not change materially.

Unlike some workers in this field, we provided no feedback to students until after all had been tested. Although some students asked questions about particular stations, no "right answers" were shared with them by the OSCE staff. This, too, minimized the chance of significant cueing.

The exit questionnaire completed by all students included one item for them to indicate if they thought their performance would change were they to take the OSCE again the next day. Many (47%) believed it would not, since studying would not change their skills substantially. The thing they most frequently cited as potentially influencing a reexamination (22) was studying electrocardiography in the interim.

Thus, over two thirds of the examinees either believed a repeated examination would yield the same results or that their scores would change only on those items dealing with their ability to interpret electrocardiograms. Only 31% believed that studying topics covered in the OSCE or increased familiarity with the format would aid them on a repeated examination. This impression on the part of the students helps substantiate our above-mentioned conclusions.

In 1989 Roberts and Norman[18] reported an OSCE in which 71 students were asked to repeat one of five stations they had already been through at their completion of the five-station circuit. Although students achieved slightly higher mean scores on three of the five stations when they tookit a second time, there was no statistically significant pattern of increase. This lends objective credibility to the UA students' impression.

At the UA, many class members spend large amounts of their fourth year away from the main teaching hospitals in Tucson. This, plus the year's curriculum being totally comprised of electives, disperses students such that one might not expect frequent opportunity for sharing of information between the weekly administration. But even when the students were known to be together, ie, the nine to 16 students who took the OSCE together, no detectable information sharing occurred.

In the program reported, it would appear that student scores are not materially affected by time of day or by week of administration when the student-SP interaction takes place. Other logistic factors that might detract from a reported score's accurately reflecting examinee performance must be investigated carefully. Fatigue on the part of SPs as the weeks of administration pass might cause an increase in scores in later weeks if SPs score less stringently as they tire. Conversely, a decrease in scores over the weeks of administration might indicate that SPs score more critically as they gain more experience and have more interactions to use for comparison. However, the stability of scores we have reported suggests neither of these factors significantly alters OSCE evaluations over time.

We are confident that the format described can be continued at the UA and believe that our results can encourage other OSCE coordinators to employ similar designs. The class of 1991 at the UA will be examined in the same format, but each student will have to achieve a preset overall score to pass. Any change in the results presented here, brought about by the requirement to pass, remains to be investigated.

References

1. Swanson DB, Stillman PL. Use of standardized patients for teaching and assessing clinical skills. *Eval Health Prof.* 1990;13:79-103.
2. Van der Vleuten CPM, Swanson DB. Assessment of clinical skills with standardized patients: state of the art. *Teach Learn Med.* 1990;2:58-76.
3. Harden RM, Stevenson M, Downie WW, Wilson GM. Assessment of clinical competence using objective structured examination. *BMJ.* 1975;1:447-451.
4. Harden RM, Gleeson FA. Assessment of clinical competence using an Objective Structured Clinical Examination (OSCE). *Med Educ.* 1979;13:41-54.
5. Wilson V, ed. *1986-87 AAMC Curriculum Directory.* 15th ed. Washington, DC: Association of American Medical Colleges; 1986.
6. Wilson V, ed. *1987-88 AAMC Curriculum Directory.* 16th ed. Washington, DC: Association of American Medical Colleges; 1987.
7. Ahari V, ed. *1988-89 AAMC Curriculum Directory.* 17th ed. Washington, DC: Association of American Medical Colleges; 1988.
8. Turner C, ed. *1989-90 AAMC Curriculum Directory.* 18th ed. Washington, DC: Association of American Medical Colleges; 1989.
9. Gledhill RF, Capatos D. Factors affecting the reliability of an Objective Structured Clinical Examination (OSCE) test in neurology. *S Afr Med J.* 1985;67:463-467.
10. Sachdeva AK. Objective Structured Clinical Examination (OSCE). In: *Manual of Evaluation Techniques for Medical Students.* Springfield, Ill: Committee on Testing and Evaluation, Association for Surgical Education; 1987:91-103.
11. Williams RG, Barrows HS, Vu NV, et al. Direct, standardized assessment of clinical competence. *Med Educ.* 1987;21:482-489.
12. Kane M. The validity of licensure examination. *Am Psychol.* 1982;37:911-918.
13. Roberts J, Norman G. Reliability and learning from the OSCE. Proceedings of the 28th Annual Conference on Research in Medical Education, Association of American Medical Colleges; Washington, DC; October 27-November 2, 1989.