Improvement of psychometric properties of the objective structured clinical examination when assessing problem solving skills of surgical clerkship

Mohammed Y. Al-Naami, FRCSC, MEd, *Omer F. El-Tinay,* MBBS, FRCS (Glas), *Gamal A. Khairy,* FRCS (Ed), MSc, *Safdar S. Mofti,* MBBS, FCPS, *Muhammad N. Anjum,* FCPS, MRCS (Ed).

ABSTRACT

الأهداف: تحسين الخصائص النفسية المترية للاختبار السريري الموضوعي المركب (الأوسكي)، بالإضافة إلى زيادة تحفيز أعضاء هيئة التدريس على المشاركة.

الطريقة: أُجريت هذه الدراسة الاستفتائية في قسم الجراحة بمستشفى الملك سعود الجامعي، الرياض، المملكة العربية السعودية وذلك خلال الفترة من 6 إلى 7 مايو 2006م، حيث قمنا بعمل الاختبار السريري الموضوعي المركب الذي شمل 95 طالباً تم تقسيمهم على يومين متتاليين، وقد استمر الاختبار في كل يوم 120 دقيقة. وتكون هذا الاختبار من 15 محطة حقيقية، حيث تناولت 10 محطات تصنيف المهارات المتعلقة بحل المشاكل، فيما تناولت 5 محطات اختبار المهارات السريرية.

النتائج: لقد قمنا بقياس مدى استقرار الاختبار بواسطة معامل كرونباخ ألفا ووصلت النتائج في اليوم الأول إلى 0.87، و0.88 في اليوم الثاني، وتم قياس الاتساق الداخلي للاختبار بمقياس ثيتا كارمين حيث كانت النتائج في اليوم الأول 0.76، و0.79 في اليوم الثاني. وكانت درجة ثبات الاختبار عالية على وجه العموم (r=0.8)، وذلك من دون وجود اختلاف واضح في النتائج بين اليوم الأول والثاني. تعتبر درجة مصداقية وشمولية الاختبار جيدة وذلك حسب رأي أعضاء هيئة التدريس والطلبة. وقد تم قياس درجة دقة هذا الاختبار بواسطة معامل ارتباط بيرسون مع الامتحان النظري وكانت النتيجة 0.65. أما عن درجة جدوى الاختبار فقد تحسنت تحسناً ملحوظاً وذلك عند مقارنتها مع درجات الاختبارات السابقة.

خاتمة: أثبتت الدراسة مدى صحة وثبات الاختبار السريري الموضوعي المركب وذلك عند تقييم مهارات طلبة السنة النهائية في الجراحة والمتعلقة بحل المشاكل، كما أن جدوى الاختبار قد تحسنت بشكل ملحوظ بعد استخدام هذه الإستراتيجية التي تميزت بحماس أعضاء هيئة التدريس ومشاركة عدد أكبر منهم.

Objective: To improve the psychometric properties of the objective structured clinical examination (OSCE) and elevate staff motivation.

Methods: The OSCE was conducted in May 2006 at the Department of Surgery, King Saud University, Riyadh, Saudi Arabia as a pilot study for 95 students split over 2 consecutive days lasting 120 minutes each day. There were 15 actual stations on each day consisting of 10 stations that addressed problem solving skills, and 5 addressed clinical skills testing.

Results: The stability of the OSCE measured by Cronbach's alpha on day one was 0.87, and 0.88 on day 2. The internal consistency of the OSCE measured by Carmine's theta on day one was 0.76, and 0.79 on day 2. Overall, OSCE's reliability for each day was high (r=0.8), without a significant difference between the 2 days. Credibility and comprehensiveness of the the OSCE were considered good by both staff and students. Accuracy of the OSCE measured by Pearson's correlation with multiple choice question scores was 0.65. Feasibility of the OSCE has also improved remarkably compared with previous OSCEs.

Conclusion: The OSCE proved to be highly reliable, and a valid format when more problem solving skills testing has been emphasized for the final year surgical clerkship. Feasibility has also improved remarkably after using this strategy, marked by more staff participation and enthusiasm.

Saudi Med J 2011; Vol. 32 (3): 300-304

From the Department of Surgery, College of Medicine, King Saud University, Riyadh, Kingdom of Saudi Arabia.

Received 24th November 2010. Accepted 1st February 2011.

Address correspondence and reprint request to: Dr. Mohammed Y. Al-Naami, Director, Teaching, Learning, and Assessment Program, King Saud University, Head, Gastrointestinal and Bariatric Surgery Unit, Department of Surgery #37, King Saud University, PO Box 7805, Riyadh 11472, Kingdom of Saudi Arabia. Tel. +966 (1) 4679417. Fax. +966 (1) 4679493. E-mail: mohammed_alnaami@yahoo.com / alnaami@ksu.edu.sa

Since its introduction by Harden and colleagues[1] in the 1970s, the objective structured clinical examination (OSCE) gained popularity in many medical schools to assess the clinical competence of medical students with acceptable rates of reliability and validity.[2-4] Furthermore, the OSCE has been used successfully not only in medical schools, but also as a clinical assessment tool in residency training,[5] dental,[6] nursing,[7] physiotherapy,[8] and for medical licensing.[9-11] However, recent reviews criticized OSCEs for inadequate psychometric properties reporting,[12-14] and for being resource-intensive.[4] At the Department of Surgery, King Saud University Medical School, Riyadh, Saudi Arabia, the OSCE has been introduced to assess the clinical competence of the surgical clerkship since 2005. Previously, the long and short cases were the standard clinical examinations used. Initial results of our OSCEs' reliability and validity were reported to be moderate.[15] Feasibility was a problem, especially regarding recruiting and using actual and simulated patients, as well as staff motivation.[15] The main purpose of the current study was to improve the OSCE's psychometric properties by emphasizing more problem solving testing, and to motivate faculty staff to participate efficiently in its development and conduct.

Methods. The surgical staff was introduced initially to the OSCE and its development in a half-day workshop when original stations' scenarios and corresponding checklists were developed and used. Subsequently, additional stations have also been developed throughout the years and added to the OSCE bank. A 2 days OSCE was conducted on May 6-7, 2006 at the Department of Surgery, King Saud University in Riyadh for the final year surgical clerkship. A mixture of stations (blueprint) testing various competencies including history taking and communication skills, physical examination skills, and problem solving skills were recently submitted by the teaching staff from different divisions of the department of surgery. The surgical course consists of tutorials, bedside teaching, early morning ward rounds, rotational exposure to the surgical intensive care unit, operating theater, surgical clinics, and emergency department over a 10-week period, 5 weeks in general surgery, and 5 weeks in other surgical specialties. Submitted stations were reviewed by the course committee, some were approved as such, others were returned for revision, and some were excluded. Two-thirds of the OSCE stations addressed problem-solving skills, and one-third addressed clinical skills testing. The OSCE was conducted over 2 consecutive mornings for the final year clerkship including 48 students on day one, and 47 students on day 2. Each day, the students were split into 2 groups. The examination included 24 stations; the time given for each station was 5 minutes, so each half of the examinees will complete the OSCE in 2 hours, followed by the second half, with a brief break between the 2 rotations. There were 15 actual (rated) stations and the remainder was dispersed rest (unrated) stations. Almost all stations involved an examiner matched to the corresponding station he/she developed as much as possible. Exclusion criteria included: difficult stations for all were removed, patient no show stations due to non-cooperation or exhaustion were removed if no alternative patients were available, and an unclear demonstration of the task (for example, unclear picture, x-rays, and so forth) were removed. Checklists were designed to contain desired competencies to be examined in one column (averaging around 10 items), against asked or not asked or partially asked, carried out or not carried out or partially carried out columns. Rating of the total score for each station was carried out by dividing the total marks achieved by the student over the total mark of all items in the checklist multiplied by one hundred. The global rating contained 3 categories; pass, borderline, and fail against check boxes added at the bottom of each checklist. Each student's total score of the whole OSCE was calculated by taking the average percentage of scores achieved in all rated stations. Rating followed a norm-referenced assessment method. A student has to score 60% and above in the OSCE as well as in written examinations to pass the course. However, if global rating indicates a borderline student's performance in most of the stations (>50%) even if the total score is ≥60%, this candidate will be discussed in the departmental board meeting for possible remedial short course (6 weeks attachment) with another OSCE or to repeat the whole course again. This study was reviewed by our institutional review board and found to be ethically acceptable. There was no need to obtain consent from the department and students, however, informed consent was obtained from the patients participating in the OSCE examination.

Statistical analysis. Reliability of a test refers to its precision in discriminating students' performance upon repetitions and when examiners are in close agreement in their ratings. Reliability is determined by a correlation coefficient (r) that can be measured by multiple correlations using different methods such as test-re-test, split-halves, and currently by several statistical software programs available commercially. A correlation coefficient (r) <0.6 indicates low reliability, (r) 0.6-0.8 moderate, and (r) >0.8 high reliability. Reliability can be further sub-categorized into stability (stable students' performance upon repetitions) and internal consistency (consistent score correlations with the sum of all other scores), which were measured by Cronbach's alpha and Carmine's theta using the

BMDP® statistical software (Statistical Solution Ltd, Saugus MA, USA). Credibility (face validity) and comprehensiveness (content validity) were judged by faculty and students. Accuracy (concurrent validity) was measured by Pearson's correlation of OSCE scores with one best answer type multiple choice question (MCQ) score. Feasibility issues are discussed in the text.

Results. Day one OSCE stations and results are presented in **Table 1**. All stations scored a high stability coefficient (r=0.87), which indicates that the OSCE is highly stable in discriminating students' performance upon repetitions and when any station has been removed. Internal consistency (correlation of each station's scores with the sum of all other stations' scores) was also high (r>0.8) in 7 stations, moderate (r=6-8) in 3, and low (r<0.6) in 5 stations, with an overall moderate internal consistency (r=0.76). Internal consistency for each individual station, however, is more important than overall internal consistency for purposes of individual item (station) analysis and revision. The 5 stations with low internal consistency should be revised and improved for future use. The overall reliability of day one OSCE was high (r) >0.8 when both stability and internal consistency are combined. Day 2 OSCE stations and results are presented in **Table 2**. All stations scored a high stability coefficient (r=0.88). Internal consistency was high in 6 stations, moderate in 6, and low in 3 stations, with an overall internal consistency (r=0.79). The overall reliability of day 2 OSCE was high (r) >0.8 when both stability and internal consistency are combined. Results of the

Table 1 - Day one objective structured clinical examination stations' content, stability, internal consistency, and surgical specialty.

Stations' content	Cronbach's alpha (stability)	SMC-Carmine's theta (internal consistency)	Surgical specialty
1. A laryngoscope and endotracheal tube (PS)	0.8740*	0.99336**	Anesthesia
2. Thoracostomy tube and drainage system (PS)	0.8735	0.98564	Thoracic surgery
3. Alternating patients with ventral hernia (Phx)	0.8712	0.99448	General surgery
4. Alternating patients with biliary colic (Phx)	0.8703	0.99622	General surgery
5. A CT picture of a subarachnoid hemorrhage (PS)	0.8742	0.83245	Neurosurgery
6. Lower intestinal obstruction x-rays (PS)	0.8907	0.57359	Pediatric surgery
7. Actual case of small bowel obstruction (PS)	0.8948	0.62688	General surgery
8. Actual case of lower limbs ischemia (Hx-CS)	0.8894	0.48465	Vascular surgery
9. A child with a nephrostomy tube (PS)	0.8901	0.67235	Pediatric urology
10. Actual case of hematuria (Hx-CS)	0.8961	0.54126	Adult urology
11. Actual case of end-colostomy (PS)	0.8818	0.60979	General surgery
12. Parotid tumor picture (PS)	0.8861	0.45037	General surgery
13. Triple lumen central venous catheter (PS)	0.8789	0.57638	General surgery
14. A gallbladder with stones specimen - jar (PS)	0.8679	0.99972	General surgery
15. A burn picture (PS)	0.8678	0.99971	Plastic surgery
Average	**0.8700**	**0.76000**	

*Overall Cronbach's alpha (OSCE reliability) with this station being removed. **Station's coefficient to the sum of all stations' coefficient correlation (item-total correlation). SMC - squared multiple correlations, PS - problem solving, Phx - physical examination, Hx-CS - history-communication skills

Table 2 - Day 2 objective structured clinical examination stations' content, stability, internal consistency, and surgical specialty.

Stations' content	Cronbach's alpha (stability)	SMC-Carmine's theta (internal consistency)	Surgical specialty
1. Spinal anesthesia picture (PS)	0.8818*	0.99287**	Anesthesia
2. Pneumothorax chest x-ray (PS)	0.8818	0.98492	Thoracic surgery
3. Alternating cases of inguinal hernia (Phx)	0.8768	0.99445	General surgery
4. Alternating cases of biliary colic (Phx)	0.8771	0.99592	General surgery
5. An MRI picture of multiple brain lesions (PS)	0.8805	0.72231	Neurosurgery
6. Upper intestinal obstruction x-rays (PS)	0.8817	0.73407	Pediatric surgery
7. Barium enema picture (PS)	0.9023	0.47435	General surgery
8. Actual case of lower limbs ischemia (Hx-CS)	0.8874	0.56768	Vascular surgery
9. A nasogastric tube (PS)	0.8848	0.71059	General surgery
10. Actual case of hematuria (Hx-CS)	0.8932	0.64324	Adult urology
11. Cleft lip and palate picture (PS)	0.8855	0.58038	Plastic surgery
12. Lateral neck mass picture (PS)	0.8764	0.74320	General surgery
13. Lower urinary tract syndrome (Hx-CS)	0.8860	0.70075	Adult urology
14. Foley's catheter (PS)	0.8705	0.99967	General surgery
15. Appendix specimen - jar (PS)	0.8704	0.99979	General surgery
Average	0.8800	0.79000	

*Overall Cronbach's alpha (OSCE reliability) with this station being removed. **Station's coefficient to the sum of all stations' coefficient correlation (item-total correlation). SMC - squared multiple correlations, PS - problem solving, Phx - physical examination, Hx-CS - history-communication skills

2 days were comparably stable and consistent without significant statistical difference (p>0.05). Credibility (face validity) and comprehensiveness (content validity) were judged by examiners and students for both days as very good. Accuracy (concurrent validity) measured by Pearson's correlation of OSCE scores with the written one best answer MCQ scores of all students involved in the course was 0.65. Feasibility was much improved compared to previous OSCEs.

Discussion. In the Department of Surgery at King Saud University we have been using the OSCE for testing the final year surgical clerkship clinical skills in history taking and communication, physical examination, and problem solving with an almost equal weight for each competency. However, with time we have realized that history taking and physical examination competencies, although very important skills to test, have been overemphasized. Also, our teaching staff felt that their role in such an OSCE is passive when they are given stations with checklists to conduct examinations when they were not involved in the stations development, and the station did not reflect what they have taught their students during the course. Also, most of our students have memorized almost all the different types of surgical history and physical examination skills checklists, as they do not change much with time. Moreover, these skills are more emphasized and tested during another surgical course on basic clinical skills throughout the third academic year of the medical curriculum. Despite moderate scores of reliability and validity in our previous OSCEs,[15] there was a need to improve reliability, validity, and feasibility of the examination. In this OSCE, almost two-thirds of the stations addressed problem solving skills testing. In this way, reliability, validity, and feasibility of the OSCE have improved, and the role of the teaching staff was better achieved as their objectives were better met (namely, emphasizing more management and problem solving skills over basic clinical skills). History taking and physical examination skills, however, were not completely neglected. Townsend et al[16] found better reliability and performance of OSCEs when problem solving and physical examination were more emphasized over other competencies. Dennehy et al[17] also found that OSCEs were more accurate for problem-solving ability, critical thinking, and communication skills over history taking and physical examinations skills. We tried to have a mirror image OSCE format for the 2 days as much as possible. However, the contents of the 2 days OSCE were almost completely different. It would have been more interesting to present actual OSCE results in addition to psychometric properties of this OSCE, but they contain too much detail, and are difficult to tabulate in this paper. Overall reliability of the OSCE was high (r>0.8) without a significant difference between the 2 days. Ideally, the internal consistency coefficient should exceed 0.8.[18] Low internal consistency coefficients (r<0.6) were documented in 5 stations one on day one, and 2 stations on day 2. This could have been attributed to many factors including station difficulty, inadequate clinical exposure, inappropriate checklist design, inconsistent examiner, or combinations of these and possibly other factors. All stations with low internal consistency coefficients should be completely revised or excluded, moderate coefficients to be reviewed for further improvements, and high coefficient stations stored in a secure OSCE stations bank. Examiners who are consistently matched to low internal consistency coefficient stations, after ensuring adequacy of other factors, will be scrutinized further by having them trained again on writing and rating checklists, or have another examiner to increase inter-rater reliability. Checklists were marked using the global rating method described by Newble.[3] Global rating scoring has been advocated also by some authors for marking checklist of OSCEs.[19,20] Comprehensiveness (content validity) as well as credibility (face validity) of this OSCE was considered adequate and good by expert faculty staff. Accuracy (concurrent validity) of the OSCE scores when correlated to MCQ scores of the same cohort group indicated good Pearson's correlation coefficient (r=0.65). However, this reflects only our local experience as results of OSCEs ideally would be correlated to students' performance in national or board MCQs examinations. Although a "gold standard" clinical test is still lacking, the reported accuracies were mostly carried out by correlations with written tests. Reported accuracies are generally low.[21,22] However, moderate as well as high accuracy coefficients have also been reported.[6,23,24] Predictive and construct validity aspects of the OSCE, although important to reflect a high fidelity test, were not addressed in this OSCE due to some logistics and difficulties to apply, which make it a limitation of this study. Feasibility was a problem that we initially faced due to limited resources. To find an adequate number of cases and to teach them to be consistent in their performance was not an easy task. Dropouts, absence, and lack of cooperation by some patients during the examinations were additional problems. Also, interviewing and examining actual patients by many medical students is becoming an ethical dilemma in our institute especially with the increasing number of admitted students. A structured simulated patient training program is still lacking in our institute. In this OSCE, most of these difficulties were sorted out especially when more problem solving stations were developed and used. These stations were

found easy to prepare and conduct by teaching staff. Also, their enthusiasm and interest in participating in this OSCE have improved remarkably. Friendly and transparent feedback from both parties on the OSCE was found very useful. Constructive feedback sessions after OSCEs were also found very useful by other authors.[25,26]

In conclusion, the OSCE proved to be a highly reliable and valid format when problem solving skills testing is emphasized for the final year surgical clerkship. Feasibility has also improved remarkably after using this strategy, marked by more staff participation and enthusiasm.

References

1. Harden RM, Gleeson FA. Assessment of clinical competence using an objective structured clinical examination (OSCE). ***Med Educ*** 1979; 13: 41-54.
2. Carraccio C, Englander R. The objective structured clinical examination: a step in the direction of competency-based evaluation. ***Arch Pediatr Adolesc Med*** 2000; 154: 736-741.
3. Newble D. Techniques for measuring clinical competence: objective structured clinical examinations. ***Med Educ*** 2004; 38: 199-203.
4. Berman A. Critiques on the Objective Structured Clinical Examination. ***Ann Acad Med Singapore*** 2005; 34: 478-482.
5. Pandya JS, Bhagwat SM, Kini SL. Evaluation of clinical skills for first-year surgical residents using orientation programme and objective structured clinical evaluation as a tool of assessment. ***J Postgrad Med*** 2010; 56: 297-300.
6. Gerrow JD, Murphy HJ, Boyd MA, Scott DA. Concurrent validity of written and OSCE components of the Canadian Dental certification examinations. ***J Dent Educ*** 2003; 67: 896-901.
7. Rushforth HE. Objective structured clinical examination (OSCE): review of literature and implications for nursing education. ***Nurse Educ Today*** 2007; 27: 481-490.
8. Wessel J, Williams R, Finch E, Gémus M. Reliability and validity of an objective structured clinical examination for physical therapy students. ***J Allied Health*** 2003; 32: 266-269.
9. Gerrow JD, Murphy HJ, Boyd MA, Scott DA. Concurrent validity of written and OSCE components of the Canadian dental certification examinations. ***J Dent Educ*** 2003; 67: 896-901.
10. Simon SR, Volkan K, Hamann C, Duffey C, Fletcher SW. The relationship between second-year medical students' OSCE scores and USMLE Step 1 scores. ***Med Teach*** 2002; 24: 535-539.
11. Lee YS. OSCE for the Medical Licensing Examination in Korea. ***Kaohsiung J Med Sci*** 2008; 24: 646-650.
12. Volkan K, Simon SR, Baker H, Todres ID. Psychometric structure of a comprehensive objective structured clinical examination: a factor analytic approach. ***Adv Health Sci Educ Theory Pract*** 2004; 9: 83-92.
13. Turner JL, Dankoski ME. Objective structured clinical exams: a critical review. ***Fam Med*** 2008; 40: 574-578.
14. Patricio M, Juliao M, Fareleria F, Young M, Norman G, Vaz Carneiro A. A comprehensive checklist for reporting the use of OSCEs. ***Med Teach*** 2009; 31: 112-124.
15. Al-Naami MY. Reliability, validity, and feasibility of the Objective Structured Clinical Examination in assessing clinical skills of final year surgical clerkship. ***Saudi Med J*** 2008; 29: 1802-1807.
16. Townsend AH, McLlvenny S, Miller CJ, Dunn EV. The use of an objective structured clinical examination (OSCE) for formative and summative assessment in a general practice clinical attachment and its relationship to final medical school examination performance. ***Med Educ*** 2001; 35: 841-846.
17. Dennehy PC, Susarla SM, Karimbux NY. Relationship between dental students' performance on standardized multiple-choice examinations and OSCEs. ***J Dent Educ*** 2008; 72: 585-592.
18. Henson, Robin K. Understanding internal consistency reliability estimates: A conceptual primer on coefficient alpha. ***Measurement and Evaluation on Counseling and Development*** 2001; 34: 177-189.
19. Scheffer S, Muehlinghaus I, Froehmel A, Ortwein H. Assessing students' communication skills: validation of global rating. ***Adv Health Sci Educ Theory Pract*** 2008; 13: 583-592.
20. Hodges B, McIlroy JH. Analytic global OSCE ratings are sensitive to level of training. ***Med Educ*** 2003; 37: 1012-1016.
21. Hatala R, Issenberg SB, Kassen BO, Cole G, Bacchhus CM, Scalese RJ. Assessing the relationship between cardiac examination technique and accurate bedside diagnosis during an objective structured clinical examination (OSCE). ***Acad Med*** 2007; 82: S26-S29.
22. Fitzgerald JT, White CB, Gruppen LD. A longitudinal study of self-assessment accuracy. ***Med Educ*** 2003; 37: 645-649.
23. Shehmar M, Cruikshank M, Finn C, Redman C, Fraser I, Peile E. A validity study of the national UK colposcopy objective structured clinical examination--is it a test fit for purpose? ***BJOG*** 2009; 116: 1796-1799.
24. McLaughlin K, Vitale G, Coderre S, Violato C, Wright B. Clerkship evaluation--what are we measuring? ***Med Teach*** 2009; 311: e36-e39.
25. Khursheed I, Usman Y, Usman J. Students' feedback of objectively structured clinical examination: a private medical college experience. ***J Pak Med Assoc*** 2007; 57: 148-150.
26. Larsen T, Jeppe-Jensen D. The introduction and perception of an OSCE with an element of self-and peer-assessment. ***Eur J Dent Educ*** 2008; 12: 2-7.