Evaluation of Interactive OSCE for Medical Students in the Subject of Medicine; Reliability and Validity in the Setting of Internal vs. External Examiners

Moeen-uz-zafar*, Owayed Al Shammari and Badr Aljarallah
Department of Medicine, Qassim University, Saudi Arabia

Abstract

Objective: To evaluate the reliability and validity of the interactive OSCE in the subject of Medicine for medical students and how it is affected by the presence of internal and external examiners and to identify the appropriate duration of OSCE station.

Methodology: We used the following parameters to evaluate the interactive OSCE process in internal medicine:

• Feedback from the candidate students (n=42 with 28 males and 14 females)
• Feedback from the external examiners (n=11)
• Feedback from the internal examiners (n=8)
• Determination of inter-rater variability at various stations
• Correlation between the OSCE and MCQ results

Results: A high proportion of the student examinees found OSCE to be fair, addressing the course objectives (33/42; 78.6%) and covering a wide range of the skills (35/42; 83.3%).

46% (n=13) of male candidates and 78% (n=11) of female candidates found OSCE to be very stressful (p=0.005). The duration of each station was found insufficient by 15/42 (35.7%) of students. 31/42 (73.8%) of students reported that language barrier seemingly affected their performance at the interactive OSCE stations.

84% of the external and internal examiners found OSCE to be satisfactorily covering the skills needed by a junior physician who will work under supervision but they also suggested the duration of station to be increased and few of the examiners mainly external examiners suggested an additional method of evaluation to be included. Minor interruptions were the main concern of the examiners during the exam, which can affect the reliability and validity of the exam. Patient’s related care was the principle reason for interruption. The overall inter-rater variability was found to be low, with inter-rater agreement having kappa value 0.602 [SE 0.024; 95%CI 0.555-0.650].

The correlation between the result of OSCE and the MCQ was found to be weakly positive with a Pearson correlation coefficient ‘r’ of 0.6203 [p<0.001; 95% CI 0.4503 – 0.7703].

Conclusion: The interactive OSCE in Medicine is generally a satisfactory method for the evaluation of medical students with internal and external examiners with inter-rater agreement of kappa value 0.602. Female students are more stressed in OSCE than male. Language barrier seems to affect the performance of candidates. Weak positive correlation was found between OSCE and MCQ exams as they assess different competencies.

ABBREVIATIONS

OSCE: Objective Structured Clinical Examination; MCQ: Multiple Choice Questions; SEQ: Short Essay Questions; SPSS: Statistical Package for Social Sciences

INTRODUCTION

Evaluation of the skills and the knowledge of medical students is of paramount importance as it not only filters the better students but regular evaluation helps in developing better physicians [1,2]. There are many methods of assessing the medical students in the subject of medicine like multiple choice questions, modified essay questions, long essay questions, computer-based patient encounters, viva voce, long case discussions and objective structured clinical examinations [1]. OSCE was introduced in medicine by Harden in 1975 and since then it has attracted popularity by the students and the faculty [3]. Each individual method of evaluation may address the cognitive, psychomotor or behavioral skills of a student but it is prudent to have all these skills gauged in a single encounter and the best equipped for this purpose is an interactive OSCE where a student faces a real or

standardized patient in front of the examiners [4,5].

Although a comprehensive coverage of the curriculum is possible, OSCE has the disadvantage of losing the holistic approach to a patient as it has segregated encounters at different stations [6].

One critical question in each OSCE exam is how to increase the reliability and validity of the exam; one pivotal aspect is the examiner. Majority of the studies addressing examiners aspect usually focused on the process and content. We are aiming in this study to address the evaluation of the students by internal and external examiners and its effect on the reliability and the validity of the OSCE in the setting of 15 minutes duration OSCE station.

**MATERIALS AND METHODS**

**Study Setting**

At the department of medicine, Qassim University, OSCE is being used for the formative and summative assessment of students since the year 2006. In the summative examination, the OSCE is divided into two segments: firstly the silent (unmanned) OSCE, that has twelve stations including radiology, laboratory interpretation, and electrocardiogram and picture spot diagnosis. Secondly the interactive OSCE, where the potential of a student is judged by two examiners on real patients. It is composed of eight stations including history taking, emergency case management, cardiovascular, abdomen, general physical, chest, rheumatology and neurology physical examinations. Duration of each station is 15 minutes. Among the two examiners, one is internal examiner (from the faculty of the department and participates in students teaching during the course) and external examiner who is typically a physician, specialized in one of the internal medicine subspecialties. They are provided with pre-designed global rating scale evaluation sheets (Appendix) that have been developed and validated at the department. The student performs the asked skills and answers the questions. Examiners are not allowed to discuss the assigned scores with each other to increase authenticity of each individual’s score.

Each station is evaluated out of 10, where 10 are the highest and 01 is the lowest performance. Student need to get at least 6 out of 10 to pass the station. The final marks of the course is divided between written and OSCE exams equally. Final exam carries 50 % of the total mark, which is also divided equally between written and clinical exams. To pass internal medicine course, student need to get at least 60 % of the total mark as per university rules.

**Study design**

This is a cross sectional study analyzing the interactive OSCE examination stations for the 4th year medical students final examination of internal medicine course at Qassim university, King Fahad specialists hospital in 2013. College of medicine research ethical committee approved this study.

**Participants**

Medical students of the 4th year who were taking internal medicine course final exam were enrolled. All students were included as OSCE exam is mandatory for their assessment. Written invitation to the post exam survey was offered to the whole group (students & examiners) immediately after the exam.

**Data collection**

All OSCE sheets were collected and analyzed. Subsequently, paper-based feedback from students, internal and external examiners were also recorded and analyzed. The questionnaires were developed at the department of Medicine and reviewed by experts in Medical education. Most of the questions were closed ended with few open-ended questions (Figure 1-3).

**Data analysis**

Content validity of our OSCE evaluation sheets was ascertained at a review by internal and external examiners. To gauge the reliability of the OSCE stations, we compared the two examiners at each station and we estimated the inter-rater agreement at each station using the ‘kappa test’. We were interested to find any relation between OSCE and MCQ results of our students to judge their overall performance, for that we used the Pearson correlation coefficient. SPSS 15 and MEdCALC software were used to analyze the data.

**RESULTS**

The total number of students was 46 (31 males; 15 females), 42 responded to the questionnaire (28 males, 14 females). Eight

**Figure 1: Students’ questionnaire.**

| 1. Were the objectives of bedside clinical teaching covered in the OSCE? |
| 2. Did you understand the process of OSCE exam beforehand? |
| 3. Were the commands and the tasks well understood and clear? |
| 4. Tasks during the OSCE were demonstrated during the course? |
| 5. Contents of the OSCE matched the curriculum? |
| 6. Was the time for each station adequate? |
| 7. Were the facilitators co-operative? |
| 8. Was there any interference during the OSCE? |

**• By facilitators**

**• By others**

1. Were the patients co-operative?

2. Which one of the following tools is the most useful in assessing your performance?

   - MCQ
   - SEQ
   - OSCE
   - Daily evaluation

3. Which one of the following tools is the most stressful?

   - MCQ
   - SEQ
   - OSCE
   - Daily evaluation

4. Was the OSCE structured and well-sequenced?

5. Did the OSCE cover wide range of skills?

6. Did the OSCE highlight your weaknesses and strengths?

7. Did language barrier affect your performance?

8. Any other comments?
Figure 2: Internal examiners’ questionnaire.
1. Were the objectives of Bedside teaching covered in the OSCE?
2. Tasks during the OSCE were demonstrated during the course?
3. Contents of the OSCE matched the curriculum?
4. Was the time for each station adequate to evaluate the students?
5. Was there any interference during the OSCE?
6. Which one of the following tools is the most useful in assessing student’s performance?
   - MCQ
   - SEQ
   - OSCE
   - Daily evaluation
7. Was the OSCE structured and well-sequenced?
8. Did the OSCE exam cover wide range of skills?
9. Do you think that the results of this OSCE are valid?
10. Do you suggest any additional method of evaluation?
11. What should be the appropriate response of examiner during the OSCE?
    - Mask face
    - Encouraging
12. Do you think that there is high inter-rater variability between the examiners of internal and external?
13. Any other comments.

Figure 3: External examiners’ Questionnaire.
1. Do you think the OSCE covered a satisfactory level of skills and knowledge for a medical student?
2. Contents of the OSCE matched the extent of clinical skills expected from a fresh medical graduate?
3. Was the time for station adequate to evaluate the skills of a student?
4. Was there any interference during the OSCE?
5. Which one of the following tools is the most useful in assessing student’s performance?
   - MCQ
   - SEQ
   - OSCE
   - Daily evaluation
6. Was the OSCE structured and well-sequenced?
7. Do you suggest any additional method of evaluation?
8. What should be the appropriate response of examiner during the OSCE?
   - Mask face
   - Encouraging
9. Do you think that there is high inter-rater variability between the examiners of internal and external?
10. Any other comments.

Table 1:
<table>
<thead>
<tr>
<th>STATION</th>
<th>SPECIALTY</th>
<th>KAPPA</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>EMERGENCY</td>
<td>0.732</td>
</tr>
<tr>
<td>2</td>
<td>CHEST</td>
<td>0.806</td>
</tr>
<tr>
<td>3</td>
<td>CVS</td>
<td>0.416</td>
</tr>
<tr>
<td>4</td>
<td>ABDOMEN</td>
<td>0.589</td>
</tr>
<tr>
<td>5</td>
<td>GENERAL PHYSICAL</td>
<td>0.680</td>
</tr>
<tr>
<td>6</td>
<td>RHEUMATOLOGY</td>
<td>0.471</td>
</tr>
<tr>
<td>7</td>
<td>HISTORY</td>
<td>0.448</td>
</tr>
<tr>
<td>8</td>
<td>NEUROLOGY</td>
<td>0.604</td>
</tr>
</tbody>
</table>
internal examiners and eleven external examiners were involved in the OSCE. Post exam survey was given to all of them; response rate was 91% (42/46) among students, 87% (7/8) and 100% (11/11) among internal and external examiners respectively. The reason for not responding to questionnaire was mainly delaying it to a different time for other commitments. 33/42 (78.6%) students were generally satisfied by the OSCE process reporting that the course objectives were made clear to them beforehand and the examination covered them comprehensively, a wide range of the skills were covered was reported by 35/42 (83.3%). OSCE was found very stressful by 13/28 (46.4%) of male candidates and 11/14 (78%) of female candidates (p=0.005). In their opinion among the various assessment methods the most useful tool in assessing and developing their skills was daily evaluation 19/42 (45.2%) and OSCE by 12/42 (28.6%). Duration of each station was found adequate by 27/42 (64.3%). Some patient related interference was reported by 08 (44.4%) of the examiners. As the examination is conducted in English language and the native language is Arabic, a large proportion of students 31/42 (73.8%) reported that language barrier affected their performance.

A vast majority of internal examiners 06/7 (85%) reported patient related interference during the exam, like changing an un-cooperative patient or bringing in a standardized case. 3/7 (43 %) of them raised the concern that this interruption may affect the validity and reliability of the results. Most of them 06/7 (85%) suggested that the duration of OSCE stations has to be increased. Few of them 2/7 (28%) suggested an additional method of evaluation like a long case. In the opinion of 9/11 (81.8%) of external examiners, OSCE was satisfactorily covering the skills needed by a junior physician who will work under supervision, 7/11 (63.6%) found the duration of each station to be insufficient. A small proportion 2/11 (18.2 %) suggested making long case & short case a part of our evaluation. Kappa values for the inter-examiner agreement can be seen in table 1, overall kappa was 0.602 [SE 0.024; 95%CI 0.555-0.650]. The correlation coefficient between the results of OSCE and the MCQ was 0.6203 (P<0.0001 95% CI 0.555-0.650]. The correlation coefficient between the results of OSCE and the MCQ was 0.6203 (P<0.0001 95% CI 0.555-0.650]. The correlation coefficient between the results of OSCE and the MCQ was 0.6203 (P<0.0001 95% CI 0.555-0.650].

**DISCUSSION**

Our OSCE exam is reliable and valid in the setting of using both internal and external examiners. It is more stressful for female students than male students. Language barrier is playing a role in making the exam more cumbersome for the students. Interruption during the exam is the most reported obstacle facing examiners. Our OSCE exam assesses multiple domains including competency in clinical examination, history gathering, emergency situation management skills and its priorities.

The results of our study are consistent with Elms lie et al and Carraccio et al that OSCE is a valid and reliable method of assessment, as the OSCE evaluation sheets were found satisfactory by the reviewers in addressing what they were designed for, and the inter-examiner variability was low for the same student as the kappa was moderately positive [7,8].

Addressing the effect of internal and external examiners is rarely being commented on in the recent publications. Only one recent paper surveyed 350 final year students and asked them about the possible improvement of the OSCE, involving external examiners was one of these points [9]. One of the points we address in our study was asking both internal and external examiners about the possibilities of inter rater discrepancies, which further highlighted the importance of the concept.

The students found the examination to be satisfactorily covering the objectives of the curriculum provided to them earlier and they believed that their skills were properly evaluated in adequate time. Likewise, both types of examiners found OSCE to be adequate in assessing the skills needed by a junior physician to practice under supervision. Previously, Hodges and his colleagues found OSCE to be powerful in training and practice [10]. Others found OSCE to be acceptable for students and staff [11,12]. Although the examination was stressful for the students more so for the female candidates, results of other studies are nearly consistent which found OSCE to be stressful than other forms of evaluation [13]. There is no clear reason for OSCE being more stressful for female students, but one possible explanation is a cultural reason where segregation between male and female is a common practice. Although, this practice doesn’t apply to the hospital setting where medical students are usually taught by opposite gender.

Our OSCE exam is 15 minutes encounter, which allows more comprehensive and relaxed environment for both students and examiners. A recent systemic review [14] that included 34 studies to assess the reliability and validity of OSCE checklists to assess the communication skills of medical students showed that the duration of OSCE stations is varied so widely between centers and is ranging from 5 – 20 minutes, predominantly in 5-8 minutes range [15-30]. No specific reasons being mentioned for allocating the duration of OSCE station.

Although it is a bit improper to correlate two modalities of evaluation that address different skills of a medical student, Denney et al found a weak correlation between OSCE and MCQ results [31]. In our study we found a weakly positive correlation between the two forms of examination. Similar study in our setting reported moderate correlation [16].

Our OSCE is considered adequate and credible. There is no “gold standard” to correlate with OSCE in assessing physical examination competencies, thus concurrent validity is not feasible [32]. The smooth running of OSCE needs a huge amount of resources and sometimes a standardized patient [33]. Our OSCE examination is conducted on a real patient at the floor of the hospital and smooth running is hampered by the daily routine care of patients and occasionally they refuse, hence our internal examiners reported interference during the exam.

The native language of student is different than OSCE which is conducted in English, which is familiar, but still most of the
Language barrier seems to affect the performance of students in the subject of Medicine in the setting of internal and external examiners if both were external to a mixed internal and external exam including comparing the Kappa between a committee of examiners as the main obstacle. Also comparing different duration of OSCE station and its effect on the reliability. In Conclusion, the results of our study prove that OSCE is a valid and reliable method of evaluation of students on the reliability. In Conclusion, the results of our study prove that OSCE is a valid and reliable method of evaluation of students on the reliability.

Further study is needed to address other aspects of the OSCE exam including comparing the Kappa between a committee of examiners if both were external to a mixed internal and external. Also comparing different duration of OSCE station and its effect on the reliability. In Conclusion, the results of our study prove that OSCE is a valid and reliable method of evaluation of students in the subject of Medicine in the setting of internal and external examiner. It is a source of stress mainly for female students. Language barrier seems to affect the performance of candidates.

ACKNOWLEDGEMENTS

We acknowledge the cooperation of the students, external physicians and the faculty of the department.

REFERENCES

5. Benjamin S Bloom. Taxonomy of educational objectives; the classification of educational goals. 1956.

