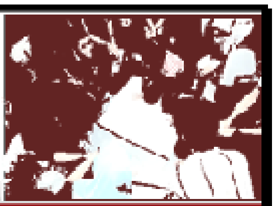
Education in Medicine Journal ISSN 2180-1932

ORIGINAL ARTICLE
Volume 7 Issue 2 2015
DOI: 10.5959/eimj.v7i2.341
www.eduimed.com



A Closer Look at Checklist Scoring and Global Rating for Four OSCE Stations: Do the Scores Correlate Well?

Joong Hiong Sim[1], Yang Faridah Abdul Aziz[2], Anushya Vijayananthan[2], Azura Mansor[3], Jamuna Vadivelu[1, 4], Hamimah Hassan[4]

[1]Medical Education & Research Development Unit, [2]Department of Biomedical Imaging, [3]Department of Orthopaedic Surgery, [4]Department of Medical Microbiology; Faculty of Medicine, University of Malaya, Kuala Lumpur, Malaysia.

ARTICLE INFO

Received	: 14/01/2015
Accepted	: 13/04/2015
Published	: 10/06/2015

KEYWORD

Checklist scoring
Global rating
Correlations,
OSCE stations

ABSTRACT

Introduction: In the marking of objective structured clinical examination (OSCE), checklist scoring and global rating are two commonly used scoring systems. **Objective:** The purpose of this study was to examine correlations between checklist scores and global ratings for four OSCE stations of different station type. **Method:** Data for this study was obtained from the Final Year OSCE (n=185). Each station's score sheet consisted of a detailed checklist of items examined. A global rating scale was also included for the examiner to indicate the global assessment for the station. Spearman's rho correlation coefficients between checklist scores and global ratings were computed for four stations of different station type. For each station, correlations between checklist scores and global ratings were also checked across the three parallel circuits running concurrently and throughout the four rounds. **Result:** Spearman's rho correlation coefficients (ρ) between checklist scores and global ratings for the four stations ranged between 0.62 to 0.88, at p<0.01. Correlation for communication skills station was the highest while correlation for procedural skills station was the lowest. For all stations, ρ ranged between 0.50 to 0.92, at p<0.01 across the circuits and between 0.57 to 0.89, at p<0.01 throughout the rounds. **Conclusion:** Checklist scores and global ratings correlated well for the station as a whole, as well as across the circuits and also throughout the rounds. Although findings of the study showed both checklist and global rating scale could be used as assessment tools in OSCE, it is suggested that for procedural skills station, checklist is preferred.

CORRESPONDING AUTHOR: Joong Hiong Sim, Medical Education and Research Development Unit (MERDU), Faculty of Medicine, University of Malaya, 50603 Kuala Lumpur, Malaysia. Email: simjhjp@um.edu.my

Introduction

First described by Harden and Gleeson [1], objective structured clinical examination (OSCE) has become one of the most widely used assessment methods of clinical competence in medical education [2-4]. The success of OSCE is partly dependent on the method of scoring. Scoring methods for OSCEs vary widely, and they influence reliability [4]. McIlroy [5] found students change behaviors depending on their perceptions of how they are being scored. However, there are still no clearly defined standards for passing an OSCE [6].

Checklists were designed and incorporated into OSCE to increase the objectivity and reliability of marking by different examiners. However, scoring against a checklist may not be as effective as it was thought to be [7]. There is abundant evidence that global rating by an experienced physician is as reliable as the standardised checklist [8, 9]. Compared to checklists, global rating scales administered by experts are a more appropriate summative measure when assessing candidates on performance-based examinations [10]. Hodges and McIlroy [11] found global rating showed substantially higher internal consistency than did the checklists. The study by Malau-Aduli et al. [12] revealed inter-rater reliability was higher for global ratings than for checklist scores.

Checklists may have limits when testing skilled practitioners, who are not as thorough in questioning or examination due to fast pattern recognition and other expert skills. According to Heldi [13], standardised patient OSCEs that are graded with checklists probably do not effectively measure knowledge, clinical skill, or reasoning. Heidi also pointed out several problems with OSCE checklists and recommend abandoning checklists or at least rethinking the approach to creating checklists and supplementing checklists with other measures. Regehr et al. [10] found supplementation with global assessment by a physician has improved testing characteristics as experts generally score low on standardised patient OSCE checklists because they are able to reach decisions with fewer steps [14], thus not completing all the items on the checklist. A higher objectivity does not imply higher reliablity and that global ratings by experienced examiners are a superior tool for assessment. An agreement has to be reached whether replacing the checklists by global rating on particular stations would improve the overall reliability, and then the OSCE can include both types of assessment tools. A balanced approach is suggested by Newble [3] wherein checklists may be used for procedural skills stations and global rating scales are employed for stations pertaining to communication skills and diagnostic tasks. However, Mash [15] cautioned that examiners and standardised patients may need prior training as reliability is increased by performing and assessing the station in the same way with each candidate.

Purpose of the study

The aim of this study was to examine correlations between checklist scores and global ratings for four interactive OSCE stations of different station type (communication skill, procedural skill, examination and history taking) conducted in our institution. A closer look at correlations across the three parallel circuits running concurrently as well as throughout the four rounds, was also attempted.

Method

Data for this study were obtained from the Final Year OSCE. A total of 185 candidates took the examination.

The OSCE

The OSCE comprised 16 work stations from 11 clinical departments (emergency medicine, opthalmology, obstetrics and gynaecology, biomedical imaging, anaesthesiology, otorholaryncology, orthopaedics surgery, primary care medicine, psychological medicine, surgery, paediatrics) and one rest station. The time allocated for each station was five minutes, with a gap of one minute between stations. There were three concurrent sessions or parallel circuits of 17 stations each. Each circuit had identical stations but with different examiners/candidates/standardised patients. The examination was held for four rounds from morning until late afternoon. Each candidate was required to perform a defined clinical task. Only standardised patients were used. A standardised marking scheme specific for each case was prepared. For non-interactive stations, only checklists were used for scoring. For interactive stations, candidates were scored using both checklists and global ratings. Each station's score sheet consisted of a detailed checklist of items examined in that particular station (total score=10 marks). Global rating on a 3-point scale with anchors at (Fail, Borderline, Pass) was also included for the examiner to indicate the global assessment for the station. Examiners had gone through OSCE examiners' training and briefing workshops conducted prior to the OSCE by the OSCE team, Faculty of Medicine. Each examiner rated candidates' performance by first scoring the task-specific checklist and then completing a global rating. The scores for the two components were independent of each other.

Validity and Reliability of the OSCE

For quality assurance and content validity of the OSCE, blueprinting was done and question vetting were conducted both at department and faculty levels. For consistency in marking and to increase reliability of the scores, examiners were provided with standardised marking scheme for each station. Examiners also shared the rationale for awarding a global score to a candidate's performance with other examiners marking the same stations as part of the intellectual discourse during the training and briefing session for examiners two days before the OSCE. Briefing session for standardised patients was also conducted one day before the OSCE, after the stations were set up at the examinaation ward.

Data Analysis

Although the Final Year OSCE comprised 16 work stations, for the purpose of this paper, data analysis was done for four stations. The four stations were a purposive sample from the 16 work stations with two requirements that the station must be (i) interactive, and (ii) of a different station type (see Table 1). For each station, checklist score was marked out of a total of 10 marks to provide a numerical score. Global rating of Fail, Bordeline, Pass were assigned the numerals 1, 2 and 3 respectively. Spearman's rho correlation coefficients between checklist scores (ratio data) and global ratings (ordinal data) were computed for the four stations. For each station, Spearman's rho correlations between checklist scores and global ratings were also checked across the three parallel circuits as well as throughout the four rounds.

Result

Table 1 provides a summary of the correlation coefficients for the four OSCE stations overall (column 3) as well as for the three circuits (columns 4 to 6) and the four rounds (columns 7 to 10).

Spearman's rho correlation coefficients (ρ) between checklist scores and global ratings for the four stations ranged between 0.62 to 0.88, at p<0.01, with a mean correlation coefficient of 0.76. For all stations, ρ ranged between 0.50 to 0.92, at p<0.01 across the three parallel circuits and between 0.57 to 0.89, at p<0.01 throughout the four rounds. Example: for Station 6, ρ=0.88 (p<0.01). ρ across the three circuits were respectively 0.92, 0.86, 0.87 at p<0.01 while ρ throughout the four rounds were 0.86, 0.89, 0.85 and 0.88, at p<0.01 respectively.

Table 1: Correlations between checklist scores and global ratings (n=185)

Station No.	**Station Type**	**Spearman's rho correlation coefficient (ρ)**							
		Overall	**Circuit 1**	**Circuit 2**	**Circuit 3**	**Round 1**	**Round 2**	**Round 3**	**Round 4**
6	Communication Skills	0.88	0.92	0.86	0.87	0.86	0.89	0.85	0.88
9	Procedural Skills	0.62	0.78	0.53	0.74	0.75	0.71	0.57	0.65
10	Physical Examination	0.77	0.84	0.70	0.91	0.64	0.83	0.75	0.77
12	History Taking	0.78	0.81	0.50	0.73	0.79	0.78	0.83	0.78

*correlation is significant at p<0.01

Discussion

With only four stations, we cannot report on the reliability or internal consistency of station score. However, reliability analysis reported an alpha value of 0.68 across the 16 stations for the 185 candidates. This indicated the OSCE as a whole had acceptable reliability.

Although the OSCEs were run concurrently in three parallel circuits and continuously for four rounds from morning until late afternoon, checklist score and global rating score correlated well for the station as a whole, as well as across the circuits and also throughout the rounds.

Comparing the four stations as a whole, it can be seen that Station 6 (communication skill station) has the highest correlation coefficient (ρ=0.88) while Station 9 (procedural skill station) has the lowest correlation coefficient (ρ=0.62). See column 3 Table 1. With correlation coefficients ranging from 0.62 (moderately good) to 0.88 (very good), it appears that both checklist scoring and global rating can be used as assessment tools for all station types. Nonetheless, global rating is perhaps better suited for communication skill station as a more holistic assessment of a candidate's competence while checklist scoring is more suitable for procedural skill station which requires a more detailed assessment of the skills. Such findings support Newble's [3] suggestion.

A closer look at Table 1 shows that of the three circuits and the four rounds, correlation coefficients for Station 6 (communication) was ranked the highest among the four station type in two out of the three circuits (Circuit1, Circuit2) but in all the four rounds. For Station 9 (procedural skill), correlation coefficient was ranked the lowest in one of the circuit (Circuit1) and second lowest in the other two circuits but in three out of the four rounds. Hence, there was more consistenccy throughout the rounds as compared to across the circuits. This findings could be due to variations such as examiners and/standardised patients. This is because for the same station, examiners and standardised patients were different for each circuit as the examination ran concurrently across the circuits. However, throughout the rounds, examiners and standardised patients were the same.

The use of global rating on a 3-point scale with anchors at (Fail, Bordeline, Pass) could possibly cause the correlation coefficient of the procedural skill station (Station 6) to be the lowest. This is because for procedural skill station, a more detailed assessment of the candidate's competency is required and a finer rating scale is needed to discriminate between the candidates with different levels of competency. A detailed checklist can cater to this need but not a global rating on a 3-point scale. Examination of the raw scores showed candidates with checklist score of 6, 7, 8, 9 or 10 were awarded the same global score of "Pass". Perhaps higher correlation coefficients between checklist scores and global rating scores could be obtained with the use of a finer rating scale. Hence, it is suggested that global rating on a 4-point scale with anchors at (Unsatisfactory/Fail, Bordeline, Satisfactory/Pass, Very Good) or 5-point global score rating scale with anchors at (Poor/Fail, Borderline, Average/Pass, Very Good, Outstanding) be used.

Although global ratings are an important element of OSCE measurement and can have good psychometric properties, OSCE administrators and researchers should clearly define or describe the type of global ratings they use. Since global rating is a rating scale, global score descriptors should be made available to examiners as guidelines as how to award a score. In our institution, global assessment using global rating scales was included in the training of OSCE marking for examiners involved in scoring each OSCE station.

Furthermore, use of global ratings mandates that only people with subject expertise can be used as examiners [8-10]. Other limitations of using global rating scale in OSCE includes difficulty in defending the marking in case of an appeal especially for high-stakes exit examinations. Hence, it is suggested to have an open-ended section to include comments by examiners, especially for borderline candidates.

On the educational impact of assessment tools, there is a possibility that examinees' awareness of marking in OSCE may affect the way they learn medicine. McIlroy et al. [5] found students change behaviors depending on their perceptions of how they are being scored. Boursicot et al. [16] also commented OSCE can promote students to learn the checklist rather than having a deeper understanding of the skills assessed. Given these concerns there is now a trend to group together single `lower-level' checklist items to more `higher-level' items. For example, instead of using separate single marks for hand washing, identification of patient, explaining purpose of encounter – these items are grouped into one rating scale (for example: patient-doctor interaction). Pell et al. [17] found the use of such rating scales can improve the reliability of an OSCE.

Since examiner's global rating score may be influenced to some extent by their knowledge of the checklist score [15], it is advisable not to add up checklist scores prior to giving global rating score to ensure the global score awarded is truly independent of the checklist score.

The authors acknowledged several limitations of this study. These include: (i) the small number of stations analysed, (ii) only a single institution was involved, (iii) there was also a possibility that the rating of the global scales after the checklists could have affected examiners' scoring of students' performance, although the two scores were independent.

Conclusions and Implications

With Spearman rho correlation coefficients for the four stations ranging between 0.62 to 0.88 and a mean correlation coefficient of 0.76, checklist score and global rating score correlated well for the station as a whole. Global rating scale could be used as another optional assessment tool in the marking of OSCE. Considering the fact that checklists are more difficult to design compared to rating scales, global rating could also be used to score an OSCE. However, it is suggested that a finer global rating on a 4-point scale be used. It is also crucial to ensure that examiners who score an OSCE station have specialty in the discipline assessed and have undergone training and briefing sessions prior to the OSCE.

Findings of this study on the correlations between checklist scoring and global rating on four OSCE stations of different station type, as well as an exploration of the correlations across the parallel circuits and throughout the rounds, should add to the body of literature on checklist scoring and global rating of OSCEs.

Declaration

An oral presentation related to the content of this work was made during the Australian and New Zealand Association of Health Professional Educators (ANZAHPE) 2014 Conference on 8 July 2014 by Joong Hiong Sim at the Grifith Health Centre, Grifith University Gold Coast Campus. However, the work has not been published elsewhere. An abstract of the work was also included in the Conference Handbook and Programme of ANZAHPE 2014.

Reference

1. Harden RM & Gleeson FA. Assessment of clinical competence using an objective structured clinical examination (OSCE). Medical Education 1979; 13(1): 41-54.
2. Gormley G. Summative OSCEs in undergraduate medical education. Ulster Medical Journal 2011; 80(3): 127-132.
3. Newble D. Techniques for measuring clinical competence: objective structured clinical examination. Medical Education 2004; 38(2): 199-203.
4. Turner JL & Dankoski ME. Objective structured clinical exams: a critical review. Family Medicine 2008; 40(8): 574-578.
5. McIlroy JH, Hodges B, McNaughton N & Regehr G. The effect of candidates' perceptions of the evaluation method on reliability of checklist and global rating scores in an objective structured clinical examination. Academic Medicine 2002; 77(7): 725-8.
6. Gupta P, Dewan P & Singh T. Objective Structured Clinical Examination (OSCE) revisited. Indian Pediatrics 2010; 47(11): 911-920.

7. Reznick RK, Regehr G, Yee G, Rothman A, Blackmore D. & Dauphinee D. Process-rating forms versus task-specific checklists in an OSCE for medical licensure. Academic Medicine 1998; 73: S97-99.
8. Cunnington JPW, Neville AJ & Norman GR. The risk of thoroughness: reliability and validity of global ratings and checklists in an OSCE. Advances in Health Science Education Theory and Practice 1997; 1: 227-33.
9. Wan S, Canalese R, Lam L, Petersen R, Quinlivan J & Frost G. Comparison of criterion-based checklist scoring and global rating scales for the Objective Structured Clinical Examination (OSCE) in pre- clinical year medical students. Medical Education 2011; 45(Supp 3:1). doi:10.1111/j.1365-2923.2011.04089.x
10. Regehr G, MacRae H, Reznick RK & Szalay D. Comparing the psychometric properties of checklists and global rating scales for assessing performance on an OSCE-format examination. Academic Medicine 1998; 73(9): 993-997.
11. Hodges B & McIlroy JH. Analytic global OSCE ratings are sensitive to level of training. Medical Education 2003; 37(11): 1012-1016.
12. Malau-Aduli BS, Mulcahy S, Warnecke E, Otahal P, Teague PA, Turner R &Van der Vleuten C. Inter-rater reliability: comparison of checklist and global scoring for OSCEs. Creative Education 2012; 3(6A): 937-942.
13. Heidi SC. What does an OSCE checklist measure? Family Medicine 2008; 40(8): 589-591.
14. Hodges B, Regehr G, McNaughton N, Tiberius R & Hanson M. OSCE checklists do not capture increasing levels of expertise. Academic Medicine 1999; 74(10): 1129-1134.
15. Mash B. Assessing clinical skills – standard setting in the objective structured clinical exam (OSCE). SA Family Practice 2007; 49(3): 5-7.
16. Boursicot K, Etheridge L, Setna Z, Sturrock A, Ker J, Smee S, & Sambandam E. Performance in assessment: consensus statement and recommendations from the Ottawa conference. Medical Teacher 2011; 33(5): 370-383.
17. Pell G, Fuller R, Homer M & Roberts T. How to measure the quality of the OSCE: A review of metrics – AMEE guide no. 49. Medical Teacher 2010; 32(10): 802-811.