ORIGINAL ARTICLE

INTER-RATER RELIABILITY OF OBJECTIVE STRUCTURED LONG EXAMINATION RECORD

Syed Inamullah Shah, Mehreen Baig, Sajida Shah*, Eitezaz Ahmad Bashir, Hajira Sarwar, Jamil Ahmad Shah

Department of Surgery, Foundation University Medical College, Islamabad, *Shifa International Hospital, Islamabad-Pakistan

Background: Objective Structured Long Examination Record (OSLER) scale was introduced in 1997 by Gleeson to improve the long case examination. There is no psychometric evidence to support reliability of OSLER. This study was done to analyse inter-rater reliability of OSLER. **Methods:** Two groups of examiners assessed 105 students in long case examination of their final professional examination, using OSLER scale. Group 1 was composed of actual examiners while Group 2 was mock examiners. Kappa statistic and intraclass correlation coefficient (ICC) were used on SPSS 23 to calculate reliability. **Results:** Mean score awarded by actual examiners was 55.36 (SD=11.2) whereas mean score by mock examiners was 57.74 (SD=14.1). Cronbach's alpha was 0.586, Kappa was 0.019 whereas inter-rater reliability on ICC was 0.413. **Conclusion:** Although OSLER is a practical modification of long case examination with good validity, the scale needs to be more structured to improve its reliability.

Keywords: Long case; OSLER; Reliability

Citation: Shah SI, Baig M, Shah S, Bashir EA, Sarwar H, Shah JA. Inter-rater reliability of Objective Structured Long Examination Record. J Ayub Med Coll Abbottabad 2018;30(2):180–3.

INTRODUCTION

In nineteenth century, Cambridge University introduced the long case examination as an assessment tool for clinical acumen.[1] Since then, long case examination has been used as one of the tests of clinical competency for certification of undergraduate medical students.[2] The traditional procedure of long case examination requires the candidate to take history and examine a real patient without being observed by the examiner, before presenting the findings in an unstructured manner to the examiner.[3] This is followed by questions by the examiner to assess clinical reasoning of the student.[4] Like most medical colleges in Pakistan, in our institution the certification requirement for final year MBBS includes one long case examination in addition to other assessments including OSCE, MCQs and SAQs. It forms 10% of the total assessment. The examination is criterion referenced, with a pass value at 50%.

The long case replicates the actual daily practice of patient encounters of a candidate.[5] Nevertheless, as an examination tool, it is perceived to have poor reliability[3,6] and has largely been omitted from the repertoire of assessment tools in North America.[7] However, it is still an integral part of undergraduate certification examination in the subcontinent.[8] The main objection raised against the long case is that its outcome depends on many inconsistent variables including the difficulty level of the patient and biases of the examiner.[9,10]

In order to make the long examination more objective, valid and reliable, many modifications in the original format were suggested.[11–13] One of these modifications[13] standardized the long examination by making it structured. It consists of a ten-item analytic scale to score the performance of each candidate on a long case. This is called Objective Structured Long Examination Record (OSLER).[9,13] Literature shows that comparisons of OSLER with traditional long case examination found it to be student and examiner-friendly as well as offering an improvement in the format of long examination.[9]

Although reported to have adequate face validity, there is no evidence to support the reliability of OSLER.[7] Reliability refers to the ability of a test or instrument to consistently measure what it is supposed to measure. It includes internal consistency of the instrument, test-retest reliability (inter-case reliability in long examination), and inter-rater reliability to assess the difference in score awarded by different examiners on the same student and patient.[2] This study was carried out to assess inter-rater reliability of Objective Structured Long Examination Record

MATERIAL AND METHODS

This was a non-randomized study carried out in Department of Surgery of Foundation University Medical College, Islamabad. The study included 105 students appearing in the final professional examination of surgery in 2014. The students were not informed of their participation in the study to maintain decorum and their undivided attention towards the exam. Ethical approval was obtained

from Ethics Review Committee of Foundation University Medical College.

All students appearing in long case were given 60 minutes to take history and perform physical examination of their respective patient. The recording of the difficulty of the case before the start of examination was done by the examiners. Two examiners scored the performance of each student at the same time, using OSLER scale. The ten item scale includes 4 items on history taking, 3 items on physical examination and one item each on formation of appropriate investigation in logical sequence, appropriate management and clinical acumen.[9,13] As the scale was being used for a criterion-referenced examination, we used a modified extended marking scheme awarding a definitive score on each item of the OSLER scale.[13]

There were two groups of examiners. Group 1 included actual examiners who asked questions from the candidates. Group 2 were mock examiners who independently rated the candidates on OSLER scale but were otherwise silent. There were four actual examiners. One mock examiner was attached with each actual examiner. The scores awarded by mock examiners were only used for the purpose of this study. All candidates were assessed on the same 10 item scale in 10 minutes duration. The examiner asked each student to take some part of history and perform a specific part of examination under direct observation, to assess technique and communication skills.

Data of scores by both groups were analysed statistically on SPSS 23. Inter-rater reliability was calculated by using Intraclass Correlation Coefficient (ICC) as well as by Kappa.[14–16]

RESULTS

A total of 105 students took the long case examination. All of them were rated by actual examiners but only 97 were rated by mock examiners. Mean scores of actual examiners and mock examiners are given in Figure 1 whereas correlation between them is shown in figure 2.

Scores awarded by actual examiners and mock examiners are shown as histograms in figure-3.

Reliability statistics included calculation of internal consistency of Gleeson's OSLER scale in this study. It was found to be 0.59. Inter-rater agreement as analysed by Kappa statistic was found to be low (p=.362) as shown in table-1. Inter-rater reliability of OSLER scale, using Intraclass Correlation Coefficient (ICC) was found to be low (0.418) (p=.000) (Table-2).

Paired samples t-test was also carried out to analyse the difference between mean scores of actual and mock examiners. The difference was found to be not significant (Table 3).

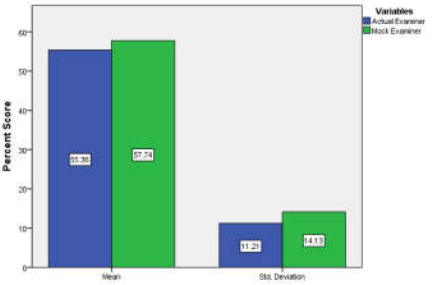

Figure-1: Statistics of actual and mock examiners

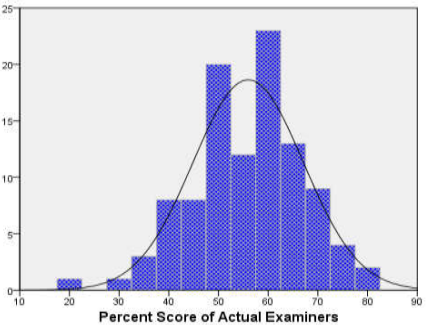

Figure-2: Inter-item correlation matrix

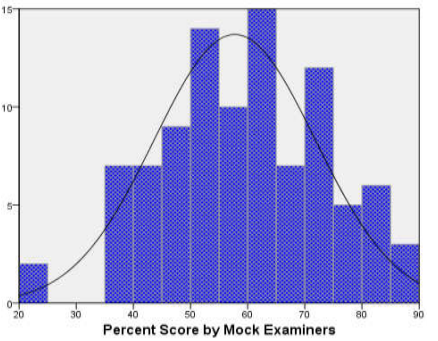

Figure-3: Distribution of scores by examiners

Table-1: Inter-rater agreement on Kappa statistic

Symmetric Measures					
		Value	Asymptotic Standardized Error[a]	Approximate T[b]	Approximate Significance
Measure of Agreement	Kappa	.019	.024	.912	.362
N of Valid Cases		97			
a. Not assuming the null hypothesis.					
b. Using the asymptotic standard error assuming the null hypothesis.					

Table-2: Inter-rater reliability of OSLER

Intraclass Correlation Coefficient							
	Intraclass Correlation[b]	95% Confidence Interval		F Test with True Value 0			
		Lower Bound	Upper Bound	Value	df1	df2	Sig
Single Measures	.413[a]	.236	.564	2.434	96	96	.000
Average Measures	.584[c]	.381	.721	2.434	96	96	.000
Two-way mixed effects model where people effects are random and measures effects are fixed.							
a. The estimator is the same, whether the interaction effect is present or not.							
b. Type A intraclass correlation coefficients using an absolute agreement definition.							
c. This estimate is computed assuming the interaction effect is absent, because it is not estimable otherwise.							

Table-3: Statistical difference in mean scores of actual and mock examiners

Paired Samples Test

Paired Differences							
			95% Confidence Interval of the Difference				
Mean	Std. Deviation	Std. Error Mean	Lower	Upper	t	df	Sig. (2-tailed)
-2.376	13.8	1.398	-5.151	.398	-1.700	96	.092

DISCUSSION

Our study revealed that inter-rater reliability of OSLER is low (0.42) using Intraclass Correlation Coefficient (p=.000), despite the fact that the difference between scores of actual and mock examiners was statistically not significant (*p*=.092) using 95% confidence interval. Internal consistency of the scale was also found to be low (*p*=.362).

Long case examination is a part of assessment in final professional examination in Pakistan. This study is a step towards carrying out meaningful research to assess and improve its reliability. Demise of the long case in North America and elsewhere occurred due to a paucity of evidence based on psychometric research.[5] Presently, there is no evidence to support reliability of OSLER in psychometric terms.[7] As Ponnamperuma *et al* have rightly pointed out that before modifications in long case are adopted for summative assessment, further evidence is required regarding their efficacy and accuracy.[8] This study, in our knowledge, is the first attempt to fill this gap by calculating inter-rater reliability of OSLER.

OSCE has gradually replaced the long case examination to assess psychomotor domains of students in medical schools in North America and much of Europe.[17,18] It has been reported however, that there is poor correlation between long case examination and the OSCE.[19] So a student's performance in one is not a reliable predictor of performance in the other. Hence, it would be unwise to abandon it in medical schools where it is still being used as an assessment tool.[5,8,11]

Unstructured questioning by examiner and single patient encounter are two aspects that raised concerns about the reliability of traditional long case examination.[20] Inter-case reliability is affected because the assessment is based on only one patient encounter whereas inter-rater reliability may be affected because of lack of standardization.[10] This led to modifications in the content and conduct of long examination. One way to improve the long case examination was to increase the number of patient encounters or the number of examiners.[12] Kroboth *et al* showed that even with two patients and two examiners, inter-rater reliability coefficient was 0.40.[21] OSLER makes the questions by examiner more structured as the examiner has to cover all ten items.

Direct observation of history taking and physical examination also affects the scoring by examiners.[22] Although seemingly an ideal solution, it would extend the examination for a very long duration, making it impractical.[11,12] Gleeson's OSLER requires only partial observation of performance and is thus more feasible.[9,13,23]

Case specificity is another matter of concern in long case examination. Some students may get an 'easier' case than others. The OSLER provides for consideration of this factor.[9] The examiner assesses the difficulty level of the case beforehand. The difficulty level is determined by the number of problems that the case presents. If one problem needs to be resolved it represents a standard case, more than three problems would be very difficult.[9,13] Wilkinson et al have reported that case selection has minimal impact on reliability.[2] In our study, patients were not standardized but difficulty level was determined by examiners beforehand.

The findings of this study should be interpreted cautiously as mock examiners used in this study were relatively inexperienced faculty members. Although available evidence in the literature reports that examiner training has very little effect on reliability,[2] their scoring was bound to be different from the actual examiners. Secondly, the mock examiners scored the candidates on questions asked by actual examiners. This could also lead to some misinterpretation. The strength of this study is that it is the first to enquire into the accuracy of OSLER in psychometric terms, in actual examination setting with real patients.

CONCLUSION

Long case examination, with its holistic approach towards patient management, still has a place in

clinical assessment of undergraduate students. The OSLER is a practical modification in its method that makes it more objective and structured. More improvement is needed in OSLER to increase its reliability. An effort should be made to minimize its faults by improving its objectivity and accuracy. Further research is recommended to assess inter-case reliability, make the scale more structured, and to propose improvements in its method.

AUTHORS' CONTRIBUTION

SIS: Conceived the objectives of the research, collected and analysed data, wrote the first manuscript, edited and approved the final draft. MB: Participated in data collection, contributed in writing first manuscript and approved the final draft. SS: Performed literature search, analysed data, compiled results and approved final draft. EAB: Conceived the research question, participated in data collection and approved the final draft of manuscript. HS: Participated in data collection and writing first draft of manuscript. JAS: Participated in data collection, and editing of first draft.

REFERENCES

1. Dare AJ, Cardinal A, Kolbe J, Bagg W. What can history tell us? An argument for observed history-taking in the trainee intern long case assessment. N Z Med J 2008;121(1282):51–7.
2. Wilkinson TJ, Campbell PJ, Judd SJ. Reliability of the long case. Med Educ 2008;42(9):887–93.
3. Smee S. ABC of learning and teaching in medicine. Skill based assessment. Br Med J 2003;326(7391):703–6.
4. Troncon EA, Fernando ROD, Figueiredo C, Ferriolli E, Moriguti Lio C, Martinelli Ana LC, *et al*. A standardized, structured long-case examination of clinical competence of senior medical students. Med Teach 2000;22(4):380–5.
5. Wass V, Van der Vleuten C. The long case. Med Educ 2004;38(11):1176–80.
6. Teoh NC, Bowden FJ. The case for resurrecting the long case. BMJ 2008;336(7655):1250.
7. Thornton S. A literature review of the long case and its variants as a method of assessment. Educ Med J 2012;4(1):e5–14.
8. Ponnamperuma GG, Karunathilake IM, McAleer S, Davis MH. The long case and its modifications: a literature review. Med Educ. 2009;43(10):936–41.
9. Sood R. Long case examination - Can it be improved? J Indian Acad Clin Med 2001;2(4):252–5.
10. Norcini JJ, Lipner RS, Kimball HR. Certifying examination performance and patient outcomes following acute myocardial infarction. Med Educ 2002;36(9):853–9.
11. Newble DI. The observed long case in clinical assessment. Med Educ 1991;25(5):369–73.
12. Abouna GM. The integrated direct observation clinical encounter examination (IDOCEE) - an objective assessment of students' clinical competence in a problem-based learning curriculum. Med Teach 1999;21(1):67–72.
13. Gleeson F. AMEEN medical education guide no 9: Assessment of clinical competence using the Objective Structured Long Examination Record (OSLER). Med Teach 1997;19(1):7–14.
14. Bannerji M, Capozzoli M, McSweeney L, Sinha D. Beyond Kappa: A review of inter-rater agreement measures. Can J Stat 1999;27(1):3–23.
15. Weir JP. Quantifying test-retest reliability using intraclass correlation coefficient and SEM. J Strength Cond Res 2005;19(1):231–40.
16. Viera JA, Garrett JM. Understanding inter-observer agreement: The Kappa Statistic. Fam Med 2005;37(5):360–3.
17. Barzansky B, Etzel SI. Medical schools in the United States, 2009-2010. JAMA 2010;304(11):1247–54.
18. Bentley BS, Hill RV. Objective and subjective assessment of reciprocal peer teaching in medical gross anatomy laboratory. Anat Sci Educ 2009;2(4):143–9.
19. Kamarudin MA, Mohamad N, Halizah MN, Yaman MN. The Relationship between Modified Long Case and Objective Structured Clinical Examination (OSCE) in final professional examination 2011 held in UKM Medical Centre. Procedia-Soc Behav Sci 2012;60:241–8.
20. Malik A, Bhugra D. Workplace based assessment methods: literature overview. In: Malik A, Bhugra D, Brittlebank A, editor. Workplace-based assessments in psychiatry. 2nd ed. London: RCPsych Publications, 2011; p.14–27.
21. Kroboth FJ, Hansusa BH, Parker S, coulehan JL, Kapoor WN, Brown FH, *et al*. The inter-rater reliability and internal consistency of a clinical evaluation exercise. J Gen Intern Med 1992;7(2):174–9.
22. Wass V, Jolly B. Does observation add to the validity of the long case? Med Educ 2001;35(8):729–34.
23. Nithyanandam S, Joseph M, Vasu U. Can conventional long case examination be improved? Indian J Ophthalmol 2012;60(4):333.

Received: 7 October, 2016	*Revised: 14 January, 2018*	*Accepted: 12 February, 2018*

Address for Correspondence:

Dr Syed Inamullah Shah, Surgical Unit 2, Fauji Foundation Hospital, Rawalpindi-Pakistan
Cell: +92 346 955 9097
Email: smiubk@gmail.com