Original Article

Exploring the process of final year objective structured clinical examination for improving the quality of assessment

Nadia Saeed, Tara Jaffery, Khaja Hameeduddin Mujtaba Quadri
Department of Medicine, Shifa College of Medicine, Shifa International Hospital, Islamabad.
Corresponding Author: Nadia Saeed. Email: saeed.nadia@hotmail.com

Abstract

Objective: To explore the process of final year Objective Structured Clinical Examination (OSCE) for improving the quality of assessment.

Methods: The analytical cross-sectional study was conducted with purposive sampling on one-year Medicine Objective Structured Clinical Examination (OSCE) scores of Final Year batch of 2009 at the Shifa College of Medicine, Islamabad. The scores from December 2008 to December 2009 of 77 Final Year students were analysed. The process of examination and the interpretation of the scores was evaluated using the Standards for Educational and Psychological Testing as the conceptual framework for validity testing which identifies five sources of test validity evidence. Internal consistency reliability of the examination was determined by Cronbach's alpha. Comparison and Correlation between students' end-of-clerkship (EOC) and end-of-year (EOY) examination scores were analysed by paired sample t-test and Pearson's Correlation Coefficient respectively.

Results: There was no significant positive correlation between the scores of end-of-clerkship and end-of-year Medicine Objective Structured Clinical Examination. Overall, the students' performance in the former segment was better. Evaluation of exam stations showed that mean scores significantly decreased in almost all end-of-year stations. Reliability decreased from 0.53 in the former to 0.48 in the latter assessment. Validity evidence showed that content validity was established by blueprinting of the objective exam. Response process evidence revealed that checklists, response key and rating scale were discussed with the faculty observing the stations. However, lack of other important sources of validity like standard setting for pass/fail criteria and poor reliability are serious threats to the validity of such exam scores.

Conclusions: Multiple sources of validity evidence are needed to make defensible assumptions from performance scores. Consideration of all sources and threats to validity evidence is important to improve the quality of assessment.

Keywords: OSCE, Quality of assessment, Clinical examination. (JPMA 62: 797; 2012)

Introduction

Since 1970s, Objective Structured Clinical Examination (OSCE) has become more accepted as a performance-assessment tool and is now used in many medical schools. OSCEs can assess students' clinical competencies in a comprehensive, consistent and standardised manner. It is a proven valid and reliable, formative and summative tool for assessing the clinical skills learned by students of health sciences.[1-3] Multiple studies have shown the impact of OSCE on learning of performance skills. OSCEs help students to develop procedural, communication and physical examination skills.[3,4] At Shifa College of Medicine, OSCE is being used since the inaugural batch of 2003 and it is an integral part of continuous assessment. Over the years the process of OSCE has been improved by blueprinting of OSCE for competencies, developing checklists and rating scales, and introducing simulated patients and faculty training.[5] A study from the same institution showed the perceptions of faculty and the students regarding the acceptability of OSCE as a useful strategy for learning and assessment of performance skills.[5] We wanted to explore the process of test validation by considering important sources and threats to validity evidence for our OSCE assessment in order to propose strategies to improve the quality of our OSCE assessments.

In Shifa College of Medicine, students are evaluated at the end of clinical clerkship in Third, Fourth and Final years. Results of these end-of-clerkship (EOC) assessments are incorporated in the Final Professional Medicine Clinical summative assessment. The weightage is 10% each for Third and Fourth years and 20% for the Final year EOC assessment. Out of this 20%, EOC Medicine OSCE contributes 8%. End-of-year (EOY) Medicine Clinical assessment contributes the remaining 60% in the Final Professional Clinical summative assessment and OSCE contributes 10% to this assessment.

Our Medicine OSCE consists of 10 stations of 5 minutes each. The interactive stations are graded through preformed checklists and global rating by trained observers. The non-interactive stations test students' clinical reasoning skills and ask students to interpret clinical data and provide diagnostic or management plans in written form. The answers are graded by a performance key. To reduce the examiners' bias, these checklists and keys are developed after discussion and consensus by the faculty. All EOC Medicine OSCEs are followed by a feedback session in which students receive feedback on their performance in the interactive station from raters observing these stations.

EOY Medicine OSCE differs from EOC Medicine OSCE by the presence of an external examiner appointed by the university. In the EOY OSCE, five out of ten stations are conducted under the supervision of the external examiner who provides scenarios and grading guidelines for stations on the day of the exam.

In order to meaningfully interpret scores, it is necessary to provide validity evidence regarding multiple aspects of the assessment. Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests.[6] According to contemporary conceptualisation, validity is a unitary concept which looks to multiple sources of evidence.[6-9] The type of test validity evidence sources being considered depends on the desired types of interpretation or meaning associated with measures. All validity is construct validity in this current framework.[7-9] Construct, in the current validity framework, pertains to intangible collections of abstract concepts and principles which are inferred from behaviour and explained by educational or psychological theory.[9] For example, educational achievement is a construct; decisions regarding students' educational achievement are inferred from performance of students on assessments such as written tests for cognitive (knowledge) domain and assessments like Mini-clinical evaluation exercise (CEX) and OSCE for performance skills. This concept is reflected in the current Standards of Educational and Psychological Measurement.[6,9]

Messick has described five distinct sources of test validity evidence,[7-9] which is also reflected in the Standards of assessments.[6] It is also important to identify possible threats to validity evidence, as these interfere with meaningful interpretation of the assessment of the data. Two major sources of validity threats are construct underrepresentation (CU) and construct-irrelevant variance (CIV).[8-10] CU refers to the under sampling or biased sampling of content domains, while CIV is systematic error introduced in the assessment by variables unrelated to the construct being measured e.g. flawed cases, untrained standardized patients (SPs), indefensible passing score methods.

Good assessment should have relevance with desired outcomes. There are many studies demonstrating OSCE as a valid and reliable tool of performance assessment.[11,12] There is a lot of literature available on different sources and threats to test validity related to CU and CIV. In this study, we evaluated our OSCE assessment for Final Year students, using a one-year data of Medicine OSCE assessment obtained from EOC and EOY scores of students and considering sources and threats to validity of OSCE results using the Standards for Educational and Psychological Testing[6] as the conceptual framework for validity testing.

Subjects and Methods

The analytical cross-sectional study based on purposive sampling was conducted at the Medicine Department of the Shifa College of Medicine on the basis of

Medicine OSCE scores of Final Year batch of 2009 consisting of 77 students. After approval from the institutional review board, data was acquired from the College Examination Department. Data contained Final Year Medicine OSCE results from December 2008 to December 2009. OSCE scores of the whole class were taken and no record was excluded. The Medicine clerkship was of 8 weeks. The students experienced this clerkship in four batches each containing 19-20 students. At the EOC, each batch was assessed with written assessments, long cases, short cases and OSCE. Students were assessed in a similar manner in the EOY Final Professional assessment. A detailed blueprint of the OSCE assessments was developed which was representative of the competencies the study intended to test (Table-1).

Four EOC and one EOY Medicine OSCE results were analysed using SPSS version 16. Mean scores and standard deviation were calculated for each OSCE station. Mean scores of stations testing similar competencies were compared and correlated using paired sample t-test and Pearson's Correlation Coefficient respectively. A p value of less than 0.05 was taken as significant. Internal consistency reliability of both OSCEs was obtained using Cronbach's alpha. Validity evidence for the OSCE assessment scores was collected using the Standards for Educational and Psychological Testing as the conceptual framework for validity testing.[6]

Results

The purpose of the EOY examination was summative; to certify that the student passing the examination is competent in the clinical competencies being assessed. The EOC OSCE provided judgment on students' learning in the

Table-1: OSCE assessment blueprint for EOC and EOY examinations.

OSCE Stations	Station type	History taking	Physical Examination	Communication Skills	Clinical reasoning	Procedural skills	Professionalism
Viva 1 (case based)	Interactive				✓		
Viva 2(case based)	Interactive				✓		
Physical Examination 1	Interactive	✓	✓	✓			✓
Physical Examination 2	Interactive	✓	✓	✓			✓
Procedural skills on mannequins	Interactive			✓		✓	✓
Counseling/Communication skills	Interactive				✓		✓
EKG Interpretation	Non-interactive				✓		
Chest X ray Interpretation	Non-interactive				✓		
Pulmonary Auscultation on model	Non-interactive		✓		✓		
Cardiac Auscultation on model	Non-interactive		✓		✓		

✓ Competencies assessed. OSCE: Objective Structured Clinical Exam. EOC: End of Clerkship. EOY: End of Year. EKG: Electrocardiogram.

Table-2: Five Major Sources of Test Validity*.

Source of Validity	Details	Examples
Content	Independent assessment of match between content sampled and the domain of interest.	Examination of blueprint Representativeness of blueprint to domain of assessment Quality of test questions.
Response Process	Quality assurance and quality control of assessment data; Debriefing of examinees.	Student format familiarity Key validation Quality control/accuracy of final scores.
Internal Structure	Data internal to assessment.	Item analysis data Score scale reliability Generalizability coefficient
Relations to Other Variables	Correlation of multiple measures of the same trait with each other and with other measures of same trait to triangulate interpretation of scores	Correlation with other relevant variables Test criterion correlation Generalizability of evidence
Evidence based on consequences of testing	Impact of assessment scores, on decisions, outcomes of assessment on examinee, teaching and learning	Pass/fail score determination, Process used to determine cut score, Impact of test scores / results on students/society

*Adapted from American Educational Research Association, American Psychological Association, National Council on Measurement in Education. Standards for Educational and Psychological Testing. Washington, DC: American Educational Research Association 1999.

Table-3: Correlation of EOC and EOY Medicine OSCE scores (n=77).

OSCE Stations	Mean Scores of EOC OSCE with SD (out of 10)	Mean Scores of EOY OSCE with SD (out of 10)	Difference in Mean Scores of EOC and EOY OSCE	Paired sample t-test P value	Pearson's Correlation Coefficients (r)	Correlation Significance (2-tailed) P value
Viva 1	8.20± 1.39	5.0 ±0.92*	3.20	0.00	-0.066	0.56
Viva 2	7.08± 1.41	4.55±0.97**	2.53	0.00	-0.065	0.57
Physical Examination 1	7.22± 1.30	4.78±0.55	2.44	0.00	0.06	0.60
Physical Examination 2	6.44± 1.31	4.47±0.74	1.97	0.00	-0.22	0.04
Procedural skills	6.19± 1.45	4.82±0.72	1.37	0.00	0.04	0.71
Counseling/ Communication skill	7.07± 1.27	3.75±0.84	3.318	0.00	0.11	0.30
EKG Interpretation	7.12±3.45	3.86±2.37	3.26	0.00	0.10	0.35
Chest X ray Interpretation	7.61±2.99	5.24±1.48	2.3	0.00	0.00	0.98
Pulmonary Auscultation	5.40±3.80	4.90±1.86	0.50	0.29	0.03	0.77
Cardiac Auscultation	4.11±3.57	4.38±2.28	0.272	0.53	0.21	0.06

* Viva taken by external examiner.
** Viva taken by internal examiner.
OSCE: Objective Structured Clinical Exam. EOC: End of Clerkship. EOY: End of Year. EKG: Electrocardiogram.

Medicine clerkship. In addition, it also provided formative feedback to students on their performance.

Test validity evidence (Table-2) revealed that content validity was ensured in both OSCEs. The evidence on validity source related to the process showed that the concerned Medicine faculty was involved in the development of checklists and response keys. Students were familiar with the OSCE format since they had participated in similar OSCE assessments in Third and Fourth years also.

The internal consistency reliability of EOC Medicine OSCE was 0.53 which dropped to 0.48 in EOY Medicine OSCE. Therefore, the validity related to internal structure of the assessment was poor, because the internal consistency reliability of both OSCE assessments was <0.70, which is the minimal accepted reliability for any assessment. Generalisability coefficient was not estimated and factor analysis for variance was not done. Evidence for other validity sources was not available, particularly the pass/fail score determination by using a standard setting process was not done.

Analysis of the EOC and EOY Medicine OSCE scores of 77 students revealed that except for the station of cardiac auscultation, mean scores significantly decreased in most of EOY OSCE stations (Table-3). Improvement in mean scores was noticed in cardiac auscultation station only which was from 4.11±3.57 to 4.38±2.28. However, this improvement was statistically insignificant. The decrease in mean scores was the most for counseling/communication skill (from7.07± 1.27 to 3.75±0.84), EKG interpretation (from7.12±3.45 to 3.86±2.37) and viva station 1 (from8.20± 1.39 to 5.0 ±0.92) with significant p values.

Scores of stations testing similar competencies did not reveal any significant positive correlation between EOC and EOY Medicine OSCE. Pearson Correlation Coefficient (r) for viva stations 1 and 2, and physical examination stations 1 and 2 was -0.06, -0.06, 0.06 and -0.22 respectively. The correlation coefficient for Procedural skills and counselling/communication skills was 0.04 and 0.1, while it was 0.1, 0, 0.03 and 0.21 for EKG interpretation, chest X-ray interpretation, pulmonary auscultation and cardiac auscultation stations respectively.

Discussion

OSCE is an important assessment format of performance, demonstrating multiple competencies by using multiple stations to assess those competencies. This assessment is important because performance of students on one case does not predict their performance on another case i.e. their performance is case-specific.[13] A large number of stations allow for better sampling of domains of interest, thereby improving the validity and reliability of the assessment. It is also important to appreciate the impact of the assessment on learning.[14] Assessing clinical skills leads to increased attention by students to learning and practising those skills.[15] OSCE is an important resource for defining clerkship-related learning needs of students.[16] The use of OSCE results for performance assessment is important because the OSCE process serves to identify areas of weaknesses in the curriculum and/or teaching methods, and thus serves as a mechanism to improve educational effectiveness.[17] However, OSCE is an expensive tool for assessment.[18] In view of its importance, it is essential to ascertain the validity of the OSCE assessment by validity evidence review in order to make meaningful interpretations of students' performance. The current Standards of Educational and Psychological Measurement[6] provide a useful framework for collection of validity evidence for any assessment.

Our EOC and EOY OSCE assessment results demonstrated areas of strength as well as areas requiring additional attention for improvement in future examinations. The content validity of the examination can be improved by systematic sampling across desired domains and ensuring that stations are assessing multiple competencies. This can be achieved by detailed blueprinting of the examination according to the clerkship objectives.

The second evidence source for test validity relates to data integrity which shows that all sources of error associated with the test administration have been controlled or eliminated to the maximum possible extent.[9] Stress of examination affects a student's performance and may relate to lack of familiarity with the assessment format. OSCE is a strong anxiety-producing experience.[19] This effect is usually more pronounced in EOY examinations as assessing students for multiple disciplines and the long duration of exams are major contributors to exam anxiety.[20] The presence of external examiner in EOY OSCE may add to the stress. Our students were familiar with the OSCE format, having participated in OSCE assessments in their Third and Fourth years also. Our process of grading the student responses of the non-interactive stations was acceptable. A response key was generated at the time of developing the station. This was discussed with the faculty which was grading the responses. In addition, all the responses were graded by the same faculty members who sat together during the process. Any confusion with the student responses and clarification of key was done by discussion within the group immediately.

Regarding the validity evidence related to the internal structure of OSCE assessments, the reliability of our OSCE assessments was low for both EOC and EOY examinations. This is of major concern because unless assessment scores are reliable and reproducible, it becomes almost impossible to interpret the meaning of those scores - thus, validity evidence is lacking.[9] Possible reasons for this low reliability are the number of stations as well as the time allocated for each station.[21,22] We recommend increasing the number of OSCE stations to 14 and allocating 10 minutes for each station. Longer duration for each station will allow for increasing the complexity of the task and making it more real-life.

In our setup, although the basic format, checklist and grading system of stations were consistent across all EOC OSCE assessments, but there were variations in the examiners. Gledhill and Capatos[23] found that in spite of efforts to control patients and examiner variability, inaccuracies due to these effects remain in judgments. Inter-rater variability further increased in EOY Medicine OSCE as half of the OSCE stations were influenced by external examiner. We recommend doing generalisability study to determine the G coefficient, which demonstrates how well the student behaviour in a station can be generalised to similar cases and variance analysis to estimate the factors contributing to source error.

The fourth source of validity evidence relates to relationship with other variables. We hypothesised that there would be a positive correlation between students' scores in EOC and EOY OSCE assessments. However, our results did not show any correlation between the two OSCEs. Possible reasons for this lack of correlation can be our small sample size and limited data of only one EOC and EOY Medicine OSCE results.[24] The reduction in mean scores demonstrated poor performance of students in EOY OSCE assessment and, hence, it is essential to explore the deficiencies which may have contributed to these results. Studies have shown that strong long-term memory is associated with over-learning in the initial phase and the spacing of learning over time.[25] The duration of clerkship plays a vital role in a student's performance and reduction in clerkship duration can lower subject examination scores.[26] Final Year students were rotated for eight weeks in Medicine clerkship which seems an insufficient duration to acquire and retain adequate exposure. After the medicine clerkship, the students were rotated to other disciplines and their knowledge could not be reinforced over time.

All EOC OSCEs were followed by timely feedback sessions in which students were guided about their deficiencies. Feedback about OSCE was also taken from the students. Failure in improvement may occur if this one time feedback was inadequate or if it was not valued by the students. Lack of longitudinal assessments and informal feedback also play significant role in performance reduction at the end of the year.

The last source of test validity relates to the consequences of the assessment. This implies consequences of the examination results on students, institution and society. Review of consequences is also important in guiding future curriculum changes in view of the needs of society and students as they begin independent practice of Medicine. The process of developing examination blueprint should consider the impact of the OSCE assessment on student learning and future role as a doctor. In addition, determination of pass/fail cutoff score through established and defensible methods of standard setting for assessment is essential. Validity evidence related to this source was lacking in both OSCE assessments.

Limitations of the study were the poor reliability of OSCE assessments and lack of correlation of student performance which precludes any meaningful analysis of the OSCE scores. Reliability can be improved by increasing

the number of OSCE stations and time allocated to each station. Increasing the sample size and assessment data was needed to perform correlational and predictive studies of OSCE assessment.

Conclusion

The process of test validation is very similar to the scientific method of theory development, hypothesis generation, collection of data for the purpose of hypothesis testing and forming inferences regarding the accuracy of interpretations. The study brought to light areas of concern in the OSCE process, and opened avenues for developing mechanisms to improve the quality of OSCE assessments.

Acknowledgements

We are grateful to Dr. Ali Yawar Alam, Head of Department, Community Medicine, and Mr. Sabir Tabassum, Assistant, Examination Department, Shifa College of Medicine, for their help.

References

1. Newble DI. Assessing clinical competence at the undergraduate level. Med Educ 1992; 26: 504-11.
2. Harden RM, Gleeson FA. Assessment of clinical competence using an objective structured clinical examination (OSCE). Med Educ 1979; 13: 41-54.
3. Cohen R, Reznick RK, Taylor BR, Provan J, Rothman A. Reliability and validity of the objective structured clinical examination in assessing surgical residents. Am J Surg 1990; 160: 302-5.
4. Carraccio C, Englander R. The objective structured clinical examination: a step in the direction of competency-based evaluation. Arch Pediatr Adolesc Med 2000; 154: 736-41.
5. Iqbal M, Khizar B, Zaidi Z. Revising an objective structured clinical examination in a resource-limited Pakistani Medical School. Educ Health (Abingdon) 2009; 22: 209.
6. American Educational Research Association, American Psychological Association, National Council on Measurement in Education. Standards for Educational and Psychological Testing. Washington, DC: American Educational Research Association 1999.
7. Messick S. Validity. In: Linn RL, ed. Educational Measurement, 3rd edn. New York: American Council on Education, Macmillan; 1989; 13-104.
8. Samuel M. Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. Am Psychol 1995; 50: 741-9.
9. Downing SM. Validity: on meaningful interpretation of assessment data. Med Educ 2003; 37: 830-7.
10. Downing SM, Haladyna TM. Validity threats: overcoming interference with proposed interpretations of assessment data. Med Educ 2004; 38: 327-33.
11. Colliver JA, Swartz MH. Assessing clinical performance with standardized patients. JAMA 1997; 278: 790-1.
12. Dupras DM, Li JT. Use of an objective structured clinical examination to determine clinical competence. Acad Med 1995; 70: 1029-34.
13. Elstein AS, Shulman LS, Sprafka SA. Medical Problem Solving: An Analysis of Clinical Reasoning. Cambridge, MA: Harvard University Press, 1978.
14. Van der Vleuten, Schuwirth LW. Assessing professional competence: from methods to programmes. Med Educ 2005; 39: 309-17.
15. Newble DI. Eight years' experience with a structured clinical examination. Med Educ 1988; 22: 200-4.
16. Prislin MD, Fitzpatrick CF, Lie D, Giglio M, Radecki S, Lewis E. Use of an objective structured clinical examination in evaluating student performance. Fam Med 1998; 30: 338-44.
17. Tervo RC, Dimitrievich E, Trujillo AL, Whittle K, Redinius P, Wellman L. The Objective Structured Clinical Examination (OSCE) in the clinical clerkship: an overview. S D J Med 1997; 50: 153-6.
18. Cusimano MD, Cohen R, Tucker W, Murnaghan J, Kodama R, Reznick R. A comparative analysis of the costs of administration of an OSCE (objective structured clinical examination). Acad Med 1994; 69: 571-6.
19. Allen R, Heard J, Savidge M, Bittengle J, Cantrell M, Huffmaster T: Surveying Student's Attitude During the OSCE. Adv Health Sci Educ Theory Pract 1998; 3: 197-206.
20. Hashmat S, Hashmat M, Amanullah F, Aziz S. Factors causing exam anxiety in medical students. J Pak Med Assoc 2008; 58: 167-70.
21. Barman A. Critiques on the Objective Structured Clinical Examination. Ann Acad Med Singapore 2005; 34: 478-82.
22. 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: 266-9.
23. Gledhill RF, Capatos D. Factors affecting reliability of an objective structured clinical examination (OSCE) test in neurology. S Afr Med J 1985; 67: 463-7.
24. Quantitative data analysis. In: Cohen L, Manion L, Morrison K. Research Methods in Education. 6th ed. London and New York: Routledge Taylor and Francis Group; 2007; pp 501-58.
25. Memory. The Acquisition, Retention and Retrieval of Knowledge. In: Halpern FH. Thought and Knowledge: An introduction to critical thinking. 4th ed. Mahwah, New Jersey: Lawrnce Erlbaum Associates; 2003.
26. Edwards RK, Davis JD, Kellner KR. Effect of obstetrics-gynecology clerkship duration on medical student examination performance. Obstet Gynecol 2000; 95: 160-2.