Impact of Observers on Clinical Student Evaluations: Reliability and Variation in DMU OSCEs

Martin Schmidt, Maurice Blodgett, Sarah Parrott
Des Moines University | mschmidt@dmu.edu

DES MOINES UNIVERSITY
MEDICINE & HEALTH SCIENCES

Objective

The assessment of clinical skills is a key component of undergraduate medical education. With the discontinuation of clinical assessment as part of the licensing examinations, the task of certifying competence now falls to the individual medical schools – a development that invites increased scrutiny of the schools' clinical assessment programs. Objective Standardized Clinical Examinations (OSCEs) are an important part of the clinical skills assessment system [1]. In these encounters, students complete a series of tasks in a simulated patient encounter and are being evaluated by trained observers [2]. Assessment of clinical skills via OSCE is a complex task, and an assessment of the quality of the OSCE requires the careful monitoring of multiple performance measures [3]. The present study was undertaken to determine the reliability and generalizability of the type of OSCE conducted at Des Moines University to justify the utilization of this assessment in the certification of clinical skills.

Objective

To establish the reliability and identify the sources of variation of the DMU OSCE assessments 2021-2024.

Methods – The DMU OSCE

OSCE format: In the DMU preclinical DO curriculum, students are assessed with 7 OSCEs featuring four distinct tasks: elicit a patient history, conduct a physical examination, deliver an oral presentation, and document the encounter in a note. Student performance is graded by the standardized patient (SP) and a faculty observer (FO). SPs rate students' interpersonal and physical examination skills using a 5-point rubric. In addition, SPs grade the technical elements of the physical examination with an objective checklist. The FOs rate physical examination, oral presentation and clinical note (SOAP, abbreviated for Subjective, Objective, Assessment and Plan) using a checklist. Depending on the complexity of the physical examination, the SP fraction of the grade ranges from 2.3 to 23% (average 7.3%).

Methods – Data Analysis

Data analysis – Reliability: Data from the 2021-2024 OSCEs were tabulated to track the longitudinal performance of the DO24-DO26 classes in the categories of 1. Interpersonal skills 2. History and physical Examination, 3. Professionalism, 4. SOAP note, 5. the likelihood of the patient to return to the provider and 6. the final grade. After removal of incomplete records, Cronbach's alpha was calculated on a matrix of 594 students x 7 OSCEs (SPSS, IBM).

Data analysis – Sources of variance: For each of the 4450 encounters in the study period, SPs and students were assigned a unique identifier (103 SPs), and the sex of students and SPs were coded. Variance of student grades was analyzed for outcomes in interpersonal skills, professionalism, history/physical rubric and on the likelihood of return to the provider (SPSS, IBM).

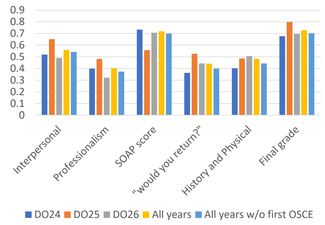

Fig 1: Inter-OSCE reliability of scores and sub-scores DO24-DO26. With the SOAP note score showing the best reliability, the overall reliability of the grade is 0.702.

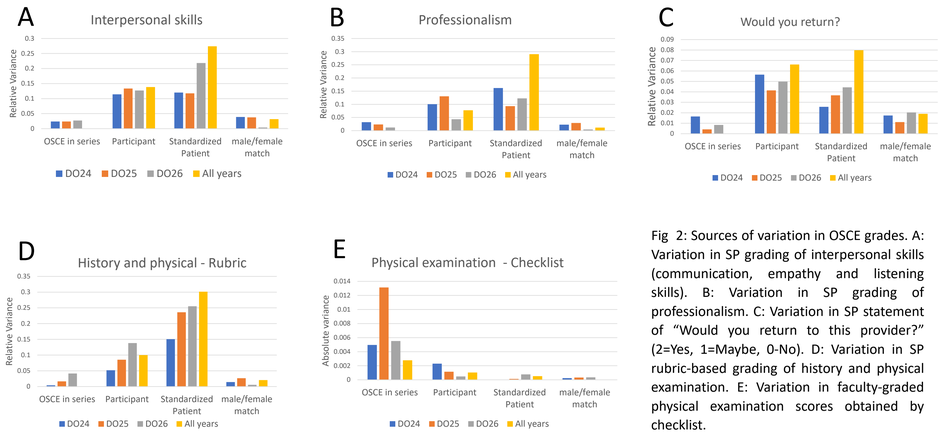

Fig 2: Sources of variation in OSCE grades. A: Variation in SP grading of interpersonal skills (communication, empathy and listening skills). B: Variation in SP grading of professionalism. C: Variation in SP statement of "Would you return to this provider?" (2=Yes, 1=Maybe, 0-No). D: Variation in SP rubric-based grading of history and physical examination. E: Variation in faculty-graded physical examination scores obtained by checklist.

Results

Reliability: The intra-OSCE reliability (i.e., correlation of the sub-scores within one OSCE) varied between 0.22 and 0.76, with an average of 0.5). The inter-OSCE reliability of the final grade for the 7-assessements on the preclinical curriculum showed a reliability of 0.7, with the most reliable grading component being the FO-scored SOAP note. On 14 occasions, SP and FO both evaluated student performance in physical examination using a checklist. The inter-rater correlation of these assessments varied between -.04 and 0.51.

Sources of variance: While most of the variance remains unaccounted for with the 4-factor model (SP, Student, number of OSCE and SP/Student sex match), in most cases the SP is the largest source of variance in the rubric-graded criteria of interpersonal skills, professionalism, history and physical examination. The SPs willingness to return depends most strongly on the student, and the variation on the checklist-graded assessment of the physical examination depends on the particular OSCE.

Conclusions

Final OSCE grades show good reliability (>0.7) over the course of the preclinical years, with the most reliable grading component being the SOAP note score. Ratings of physical examination do not correlate well between faculty observer and standardized patient, reflecting the different vantage points of the examiner and highlighting the need for the inclusion of both components into the final grade. We conclude that the DMU OSCE (with its unique combination of a small number of stations and two-observer scoring of the technical component of the encounter) produces a reliable longitudinal assessment of students' clinical skills over the preclinical phase of medical education, even if each individual assessment does not meet the high reliability threshold of the more commonly practiced 12-20 station OSCE.

References

[1] K. Boursicot *et al.* (2011) Performance in assessment: consensus statement and recommendations from the Ottawa conference, Med Teach. 33, 370-383.

[2] R. M. Harden, M. Stevenson, W. W. Downie, and G. M. Wilson (1975) Assessment of clinical competence using objective structured examination, Br Med J. 1, 447-451.

[3] C. P. Van Der Vleuten (1996) The assessment of professional competence: Developments, research and practical implications, Adv Health Sci Educ Theory Pract. 1, 41-67.