The discriminatory power of examiners’ OSCE global ratings with other areas of a paediatric assessment program

AUTHOR(S):
- Hasantha Gunasekera, University of Sydney, Discipline of Child and Adolescent Health, Australia (Presenter)
- Jennifer Anderson, University of Sydney, Australia
- Deborah O’Mara, University of Sydney, Australia

ABSTRACT

Background: Tests with high statistical reliability may not assess critical clinical skills well, whereas, clinical assessments may not have high reliability statistics. The Sydney Medical Program’s Child and Adolescent Health (CAH) teaching block’s assessment includes an integrated score across a knowledge test, a short answer clinical skills paper and a three-station OSCE. A global rating scale was introduced for the OSCE where examiners (senior clinicians) make an overall judgement of each student’s performance on each station incorporating communication and inter-personal skills. In this study, we analysed the association between the global ratings and student performance in terms of the integrated block score excluding the global ratings.

Summary of Work: Global rating scores and the in-block CAH integrated assessment results, exclusive of the global rating score, were de-identified for 462 students across 18 months (6 rotations) to December 2018. We categorised students into 4 groups by the number of OSCE stations for which a student was given a satisfactory global rating, and compared their integrated CAH score using Analysis of Variance (ANOVA).

Summary of Results: The number of students who were categorised satisfactory in all 3 OSCEs was 265 (57% of students); in 2 OSCEs was 136 (29%); in 1 OSCEs was 52 (11%); and there were 9 students (2%) who were not satisfactory in any OSCE station. As expected, there was an association between the number of satisfactory global ratings and the overall OSCE score (F=132.18, df 3, 461, p<0.001, eta squared 0.465). However, the ANOVA also showed a significant difference between students’ integrated CAH assessment mark for these 4 groups, increasing consistently according to the number of satisfactory global ratings (F=46.14, df 3, 461, p<0.001, eta squared 0.232).

Discussion and Conclusions: Experienced paediatric examiners’ global ratings in an OSCE station are significantly associated with that student’s final integrated assessment mark. The generalisability of a global score by experts should be acknowledged by being part of the summative assessment calculation. Further research is required on how best to use expert judgements in medical education assessment.

Take-home Messages: Examiners global ratings should be used for a range of assessments and feedback.
512 (2714)

Date of Presentation: Monday, 26 August 2019
Time of Presentation: 1615-1630
Location of Presentation: Room L2, Level 1

Linear OSCE progression in a Physician Associate Programme

AUTHOR(S):
- Sharon Hartwell, Swansea University, UK (Presenter)
- Jayne Dawkins, Swansea University, UK
- Steve Capey, Swansea University, UK

ABSTRACT

Background: Swansea University set up a Physician Associate (PA) programme starting in September 2016. We ensured our final OSCE was at a similar/higher standard than the national OSCE and used OSCE blueprint produced by the Faculty of Physician Associates.

Summary of Work: In Year One the 16 station OSCE was split into two parts containing eight stations each. The first part was sat after 19 weeks of the programme and focused on the process of undertaking skills on ‘normal’ patients. After this OSCE the students are given their percentages for each station and written feedback from examiners on their performance. The second part was sat at the end of Year 1 and assessed both process and differential diagnoses. The two parts of the OSCE are standard set and the results are combined for all 16 stations.

In Year 2, there was a 16 station OSCE held at the end of the year and the Angoff calculated across all stations.

The aim of the OSCEs was to ensure a progressive increase in competency and complexity.

Summary of Results: All students who passed our OSCEs, passed the national OSCE, achieving the fourth highest score of all schools at that sitting. Qualitative feedback from PA graduates stated unanimously that the OSCEs and teaching prepared them well for the national OSCE and clinical work. When comparing the performance rank order between all three OSCEs, there is minimal change between students.

Discussion and Conclusions: We feel that the increase in complexity of the OSCEs across Year 1 and Year 2 OSCE prepared the students for both the national OSCE and the work of a PA. We believe our OSCE standards ensure a steady increase in expected skills, rather than a step-wise increase.

Take-home Messages: A two stage OSCE, with increasing complexity across a year of study, can ensure an increase in expected knowledge and skills between the first and second parts of the OSCE and can ensure that students meet expected standards early in the two year course.
Impact of external quality assurance on quality improvement to Objective Structured Clinical Exams (OSCEs) at Australian medical schools

AUTHOR(S):
- Shannon Saad, The University of Notre Dame, Australia (Presenter)
- Cassandra Richmond, The University of Notre Dame, Australia
- Karina Jones, James Cook University, Australia
- Karen D'Souza, Deakin University, Australia
- Clare Heal, James Cook University, Australia
- Bunmi Malau-Aduli, James Cook University, Australia

ABSTRACT

Background: Australian Collaboration for Clinical Assessment in Medicine (ACCLAiM) is a collaborative venture between medical schools that focuses on benchmarking student outcomes in the clinical domain. This is achieved by the collective development and use of Objective Structured Clinical Exam (OSCE) stations with benchmarking of student performances between de-identified institutions. In addition, ACCLAiM members partake in a quality assurance (QA) arrangement, whereby a QA visitor from one institution observes the OSCE at another institution, and then completes a structured feedback report spanning the broad range of processes that contribute to an OSCE. The QA report is submitted to ACCLAiM and the host medical school. Schools involved in the collaboration have access to reports that have been conducted by their member academics, or relating to their own OSCE.

Summary of Work: Perceptions of how the QA process informs practice were obtained by online survey of ACCLAiM-participating schools. Invited participants have roles in medical education and/or assessment, with experience in the QA process. Survey questions addressed issues relating to the use of the QA report, including informing future OSCE practice at each school.

Summary of Results: All respondents reported usefulness of the QA report, primarily citing the ability to use the feedback to drive changes and/or improvements in OSCE processes at their School. It was noted that feedback frequently instigated improvements to OSCE case content, the structure of marking sheets, the processes for standardising SP performances, and examiner training and calibration processes. Improvement of OSCE stations was ranked as the most important outcome consequence. Most respondents (83.3%) indicated additional benefits of having a QA examiner, including the establishment of collegial networks and strengthening of collaborations.

Discussion and Conclusions: Consistent with the aims of ACCLAiM, feedback reports serve to drive QA improvement processes in Australian medical schools. Incidental benefits include establishing a medical education community of practice between member schools.

Take-home Messages: 1. The ACCLAiM QA process has informed quality improvement both within and between schools. 2. Improvement of OSCE station quality is the most valued resulting outcome. 3. Quality assurance of clinical assessment has facilitated a community of practice in medical schools within Australia.
Does Confidence about Competence Decisions Increase as Assessment Data Accumulate? It depends on the OSCE Domain

AUTHOR(S):
- Jorie Colbert-Getz, University of Utah School of Medicine, USA (Presenter)
- Katie Lappe, University of Utah School of Medicine, USA
- Tiffany Weber, University of Utah School of Medicine, USA
- Danielle Roussel, University of Utah School of Medicine, USA

ABSTRACT

Background: Student assessment has traditionally been course-based with assessment scores combined to determine a final grade. Students are motivated to pass a course, but it is unclear if they achieve program competencies. Alternatively, a programmatic assessment model captures performance by competency across courses, which enables identification of students' strengths and weaknesses, with opportunities for growth in all areas. Few studies have measured the impact of switching from a course-based to a programmatic assessment model and no studies have investigated if there is more confidence in competence decisions as assessment data accumulate.

Summary of Work: In AY2017-2018, the University of Utah SOM moved from a course-based to a programmatic assessment model for medical students' year 3 Objective Structured Clinical Exams (OSCEs). Students completed 2-station OSCEs at end of 7 clerkships and received cumulative OSCE domain reports 3 times during year 3 in the areas of history taking, physical examination, clinical reasoning, testing, and documentation. Students also completed a must pass 7-station OSCE at the end of year 3 (EOY3OSCE), but could opt-out if their cumulative OSCE performance met a threshold. Cumulative clerkship OSCE domain scores at times 1 and 3 were correlated with EOY3OSCE scores for 110 students to see if the strength of relationships increased as assessment data accumulated.

Summary of Results: Correlations between clerkship OSCE and EOY3OSCE scores were small (< 0.30) for all domains at time 1. Correlations were moderate for history taking (0.43), physical examination (0.35), and documentation (0.43) and small for clinical reasoning and testing at time 3.

Discussion and Conclusions: As OSCE data accumulate we can be moderately confident in our ability to make competence decisions about students' history taking, physical exam and documentation skills. More data points or additional types of assessment are needed for students' clinical reasoning and ability to interpret tests.

Take-home Messages: As medical schools move to programmatic assessment, it is important to understand how many data points are needed to determine competence. The results of this study suggest the number of assessment data points needed to determine competence may vary depending on the OSCE domain.
Video-based examiner score comparison and adjustment (VESCA): measuring the influence of different examiner-cohorts in fully-nested OSCEs

AUTHOR(S):
- Peter Yeates, Keele University, UK (Presenter)
- Alice Moult, Keele University, UK
- Natalie Cope, Keele University, UK
- Gareth McCray, Keele University, UK
- Bob McKinley, Keele University, UK
- Richard Fuller, University of Leeds, UK

ABSTRACT

Background: Whilst ensuring equivalence of examiners’ judgements in Objective Structured Clinical Exams (OSCE) is critical to assessment validity and fairness, most OSCEs involve multiple different groups of examiners, either in parallel circuits or distributed locations. Despite prior research suggesting different 'examiner-cohorts' may importantly influence OSCE outcomes, the potential for these effects are rarely considered chiefly due to the limitations imposed by fully 'nested' designs (no crossover between student performances seen by different examiner-cohorts). As these effects could importantly bias distributed or national exams we report the development of an intervention to measure and adjust for examiner-cohort effects in fully nested OSCEs.

Summary of Work: Across two studies we developed 'Video-based Examiner Score Comparison and Adjustment' (VESCA): 1. Volunteer students were unobtrusively filmed on most stations during their year 3 OSCE. 2. Additionally to live examining, all examiners (8 cohorts of 12 examiners in each study) were invited to score 2-4 station-specific, comparator videos of students’ performances. This produced partial crossing. 3. Global scores and total percentage scores were respectively analysed using Many Facet Rasch Modelling (MFRM) and linear mixed modelling (LMM) to estimate the influence of examiner-cohorts on students’ scores. Adjusted ‘fair scores’ were calculated for each student, and influence on students’ categorisation examined.

Summary of Results: Live and video scores were equivalent. Student ability varied by 29%, sd=5.4%. Examiner-cohorts varied in their scoring of the same ability of students by up to 5.7% (Cohen’s d=1.06) suggesting a substantial effect. Adjustment of scores around an artificial pass mark produced a change in classification for 6% of students (fail to pass or vice versa), whilst 9.5% of students’ scores changed by at least 0.5 standard deviations of student ability.

Discussion and Conclusions: Larger effects could potentially be seen between sites in distributed OSCEs, challenging the assumption of equivalence of assessments. Further research should determine: sampling requirements; the influence of other potential biases (contrast effects, rater-drift, idiosyncrasy); and the accuracy of estimates.

Take-home Messages: VESCA offers a promising method to measure and adjust for examiner cohort effects. Once validated, VESCA could potentially enhance the equivalence of high-stakes judgements in distributed or even national exams.
Inter-rater agreement between direct examiners and examiners viewing video recordings in veterinary medicine OSCEs

AUTHOR(S):
- Jean-Yin Tan, University of Calgary Faculty of Veterinary Medicine, Canada (Presenter)
- Irene Ma, University of Calgary Cumming School of Medicine, Canada
- Grace Kwong, University of Calgary Faculty of Veterinary Medicine, Canada
- Emma Read, The Ohio State University College of Veterinary Medicine, USA

ABSTRACT

**Background:** Assessments of video recordings of student performances may improve efficiency and inter-rater consistency in Objective Structured Clinical Examinations (OSCEs) compared to live assessments. However, how live assessments compare to video-based assessments is unknown. The study seeks to compare scores from live examiners with those from video examiners in a Doctor of Veterinary Medicine program OSCE. The hypothesis is that there is no significant difference in scoring between the two formats.

**Summary of Work:** Students (n=33) were assessed by live (n=9) examiners in a three-station OSCE. Performances were video-recorded and assessed by video examiners (n=5). The 3 stations assessed the following skills: inverting suturing, elbow arthrocentesis, and thoracocentesis. Binary checklists were used to determine pass/fail results.

**Summary of Results:** Based on modified Fleiss’ Kappa, overall examiner agreement for pass/fail grading was 0.82 for suturing, 0.56 for arthrocentesis, and 0.76 for thoracocentesis. Overall reliability of the stations was moderate for suturing and arthrocentesis (Cronbach's alpha 0.45 and 0.51 respectively) and fair for thoracocentesis (0.34). Using a Chi-squared test, overall there was no significant correlation between pass/fail outcome and live vs video examiners (p=0.55). Using a generalized linear mixed effects model adjusting for stations, stations (p<0.001) and interaction between station and type of examiner (p=0.03) had statistically significant effects on pass/fail outcomes. The odds of passing a given station did not differ between live vs video examiners (p>0.16) on pair-wise comparison. However, amongst video examiners, candidates had increased odds of passing suturing as compared to both arthrocentesis (odds ratio OR=4.87, p=0.01) and thoracocentesis (OR=6.01, p<0.01).

**Discussion and Conclusions:** Overall examiner agreement varied by station, from moderate to almost perfect agreement. Live assessments did not differ significantly from video-based assessments. However, stations and the interaction with live vs video examiner had statistically significant effects on pass/fail outcomes, with increased odds of passing the suturing station within video-based assessments.

**Take-home Messages:** The study concludes that there were no significant differences in the scoring between live and video examiners, although there were statistically significant station-specific differences among video examiner scores in veterinary OSCEs. Reasons for these differences need to be further explored.