The use of Borderline Regression Analysis in an online OSCE Marking Tool for improved decisions about pass or fail students

**Background:** Skills examinations use a static or dynamic cut-off score to separate good from bad performing students. A fixed cut-off score is only associated with students ability to pass exams. The difficulty of the exam or variability between examiners in marking student’s performance is not being taken into account. The Online Marking Tool developed within our School of Medicine incorporates various Global Rating Scales (GRS) marking professional competence as Pass, Borderline, Fail, Good and Excellent and incorporates flexible cut-off scores according to Borderline Regression Analysis.

**Summary of Work:** The Online Marking Tool contains a Result Analysis tool including a fully fledged Borderline Regression Analysis pack (Excel export pack). Individual ‘blinded’ scores of examiners at item level are correlated with the GRS of Fail, Borderline, Pass, Good and Excellent or variations of this GRS (Borderline Fail and Borderline Pass, Fail, Borderline, Pass).

**Summary of Results:** Borderline Group Average, the average mean score of those students being marked as ‘Borderline’ is compared with Borderline Regression Method 1 and 2 (Borderline Fail e.g Pass) to determine a flexible cut-off score. Twelve prestigious universities are currently using this unique online marking tool for clinical skills assessments. Over 200 OSCE were successfully administered and analysed using this software solution.

**Discussion and Conclusions:** Approximately 19% more students fail after introducing a flexible cut-off score in clinical skills assessment due to introducing Borderline Regression Analysis. Nevertheless despite the increased percentage of failing students due to their scores, we suggest to introduce a minimum amount of stations that needs to be passed as well.

**Take-home messages:** An online OSCE Management Information System is cheaper, faster and more reliable than a paper based solution.
How students respond to simulated patients’ emotional cues and concerns – Using the Verona Coding Definitions of Emotional Sequences (VR-CoDES) in an OSCE setting

Claudia Kiessling*, Klinikum der Universität München, Institut für Didaktik und Ausbildungsforschung in der Medizin, Munich, Germany/ Brandenburg University of Medicine Theodor Fontane, Neuruppin, Germany
Martin R. Fischer, Klinikum der Universität München, Institut für Didaktik und Ausbildungsforschung in der Medizin, Munich, Germany
Katharina Schäfer, Klinikum der Universität München, Institut für Didaktik und Ausbildungsforschung in der Medizin, Munich, Germany
Clara Wübbolding, Klinikum der Universität München, Institut für Didaktik und Ausbildungsforschung in der Medizin, Munich, Germany
Tanja Pander, Klinikum der Universität München, Institut für Didaktik und Ausbildungsforschung in der Medizin, Munich, Germany

Background: Dealing with patient’s emotions is a challenging task. However, most patients do not express emotions directly but give cues and concerns (CC) as hints to potential underlying emotions. The challenge is to detect these CCs and provide space to elaborate on them if patients wish to. The VR-CoDES were developed to code patient’s CCs and provider responses. Zhou et al. firstly used them for coding student-simulated-patient (SP) encounters in an OSCE setting and concluded that the VR-CoDES can be applied in these settings. In line with these conclusions, we were interested in applying and correlating VR-CoDES with standard OSCE rating scales to collect evidence for the validity of VR-CoDES.

Summary of Work: 88 students from different study years were asked to participate in a voluntary OSCE with four stations. Student-SP encounters were videotaped and then analysed with the VR-CoDES, content-specific checklists, Mini-CEX, and the Berlin Global Rating Scale (BGR) measuring communication skills.

Summary of Results: SP’s gave between one and 15 CCs during the interview. Students’ most popular responses were: ignoring, back channeling, content exploring, and information advice. Of the 1,219 responses, 216 (18%) were empathetic or acknowledging. The ratio between total number of responses and providing-space-responses correlated lowly to moderately with other OSCE ratings (BGR, Mini-CEX, and Checklists).

Discussion and Conclusions: As findings are preliminary, further analysis is needed.

Take-home messages: Coding OSCE stations with the VR-Codes was feasible but highly time consuming. This may be an obstacle for using them for routine assessments.
Hawks, Doves and Rasch decisions. Using Many-Facet Rasch Modelling (MFRM) to understand the overall impact of “examiner-groups” on OSCE scores

Peter Yeates*, University of Manchester, Centre for Respiratory Medicine and Allergy, Manchester, UK
Stefanie Sebok, Queen's University, Faculty of Education, Kingston, Canada

Background: Ensuring fairness of Objective Structured Clinical Examinations (OSCEs) is important. Unwanted score variability can arise from examiners, content, examination site and cycle timing. Whilst recommended procedures help to ensure acceptable reliability, it can remain unclear what overall impact such variability has on students’ scores.

Summary of Work: We used Many-Facet Rasch Modelling (MFRM) to analyse undergraduate students’ scores from a 16 station, multi-site, multi-cycle summative OSCE. Facets were: students, stations and site plus a novel facet: “examiner group”: the combined influence of 16 examiners within a given cycle of the OSCE on students’ average scores. Outcomes were “fair-average” scores and logit-scale-based standard errors of measurement (SEM) for each facet.

Summary of Results: Data comprised 235 students, 16 examiner-cycles, 4 sites and 16 stations. “Observed averages” in examiner-cycles suggested potentially unfair differences (i.e. average scores of 5.2 vs. 4.6 out of 7.0 in different examiner-cycles) but “fair averages” indicated differences are substantially attributable to real differences in students’ performance: mean score 5.11, (logit scale=0.10,SEM+/ -.08 logits) vs. 4.80 (logit scale=0.20,SEM+/ -.06). The effect of exam site was trivial. Stations varied moderately in difficulty. Using fair rather than observed averages altered the pass/fail decisions of 2 students(<1%).

Discussion and Conclusions: These OSCE procedures achieved good fairness which can be further enhanced by MFRM derived “fair average” scores. The logit scale-based standard error of measurement (SEM) provides an individualised margin of error for each candidate.

Take-home messages: We use these findings to illustrate potential benefits of MFRM and recommend that the “examiner group” be considered as a means to understand the fairness of OSCE exams.

Fixing the Rubber Band: Calibration of Communication Skills items in OSCE checklists

Winny Setyonugroho*, Universitas Muhammadiyah Yogyakarta, Yogyakarta, Indonesia
Thomas Kropmans, National University of Ireland Galway, Galway, Ireland
Kieran Kennedy, National University of Ireland Galway, Galway, Ireland
Brian Stewart, National University of Ireland Galway, Galway, Ireland
Jan van Dalen, Maastricht University, Netherlands

Background: Communication skills are commonly measured through ‘communication-items’ in OSCE station-checklists. Our aim is to calibrate the communication component of OSCE station checklists according to the MAAS-Global.

Summary of Work: The term ‘calibration’ is used to rate how close items in station checklists correspond to an accepted standard. Three raters independently compared checklists of 4 departments with 17 items of the MAAS-Global. G-theory was used to analyse the reliability of the calibration procedure.

Summary of Results: G-Kappa was 0.8. For two raters G-Kappa is 0.72 and fell to 0.57 for one rater. 46% of the checklist items correspond to section three of the MAAS-Global (i.e. medical content of the consultation), whilst 12% correspond to section two (i.e. general communication skills), and 8.2% to section one (i.e. communication skills for each separate phase of consultation). 34% of the items were not considered to be communication skills.

Discussion and Conclusions: This research confirms a reliable and valid procedure for calibrating OSCE CS item checklists using the MAAS-Global. Such calibration will enable comparison of results of CS assessments between and across students. We strongly suggest that such a procedure is more widely employed to arrive at a stable judgment of the communication component of the doctors’ behaviors.

Take-home messages: Calibration of different OSCE checklist’s is possible and calibrating with 2 raters was found to be reliable. Future research is needed to compare clinical skills assessment outcome.