**All things being equal? Impact of examiner differences on students’ outcomes in a multi-centre graduation-level OSCE**

Peter Yeates¹
Adriano Maluf², Natalie Cope¹, Gareth McCray¹, Kathy Cullen³, Vikki O'Neill³, Rhian Goodfellow⁴, Rebecca Vallander⁴, Ching-wa Chung⁵ and Richard Fuller⁶

¹ Keele university
² de Montford University
³ Queens University Belfast
⁴ Cardiff University
⁵ University of Aberdeen
⁶ Christie Hospitals NHS Foundation Trust

**Abstract**

**Introduction**

Ensuring inter-institutional equivalence of graduation-level OSCE decisions is critical to fairness and patient safety, however methodological challenges mean this is rarely studied. Recently, an innovation called video-based examiner score comparison and adjustment (VESCA)(1) has enabled linked comparison of examiners within distributed OSCE. Since prior research has hinted at potentially substantial inter-institutional differences(2), we used VESCA to determine the equivalence of different parallel groups (“examiner-cohorts”) within and between UK medical schools, and the impact of adjusting for any differences on students’ pass rate.

**Methods**

We ran the same 6-station formative OSCE at four UK medical schools(3). After examining live performances, examiners additionally scored three station-specific comparison videos which provided 1/ controlled comparison of examiners’ scoring between schools and 2/ data linkage within a linear mixed model. Impact of adjusting for examiner variations on students’ pass/fail and rank were calculated.

**Results**

Controlled comparison of examiners’ scores differed between schools by up to 16.3% from 16.52 (95%CIs 15.52-17.52) out of 27 to 19.96 (95%Cis 18.94-20.97) out of 27, p< 0.001. Examiner-cohorts varied more between schools than within schools (16.3% vs 8.8%). Students’ unadjusted scores suggested inter-school variation in students’ performances of up to 10.8% (17.65(16.87-18.43) to 19.91(19.13-20.69),p<0.001), which was no longer present after adjusting for examiner differences (18.38(17.25-19.52) to 19.14((18.19-20.10), 3.62% difference, p=0.69), thereby suggesting the apparent difference was attributable to examiner, rather than student, variation. Failure rates varied between schools and were substantially
altered by score adjustment (e.g. school 2: observed score failure rate=39.1%; adjusted failure rate=8.7%; school 4 observed=0.0%, adjusted=21.7%).

Discussion and Conclusions:

We found substantial inter-institutional differences in examiner stringency which would challenge the equivalence of outcomes if replicated within a summative setting. These apparent variations in graduation-level expectations warrant prospective investigation in summative settings to safeguard equivalence nationally. VESCA offers a feasible method to perform these comparisons.

References (maximum three)

