emerald PUBLISHING

Journal of Applied Research in Higher Edu

Does following an "excellent" candidate in the Objective Structured Clinical Examination affect your checklist score?

Journal:	*Journal of Applied Research in Higher Education*
Manuscript ID	JARHE-02-2023-0079.R1
Manuscript Type:	Research Paper
Keywords:	Assessment, OSCE, Medical, Clinical

Title Page:

Does following an "excellent" candidate in the Objective Structured Clinical Examination affect your checklist score?

Abstract

Purpose

The OSCE is regarded as the gold standard of competence assessment in many healthcare programs, however, there are numerous internal and external sources of variation contributing to checklist marks. There is concern amongst organisers that candidates may be unfairly disadvantaged if they follow an 'excellent' preceding candidate. In this study, we assessed if average checklist scores differed depending on who a candidate follows accounted for different sources of variation.

Methods

We examined assessment data from final year MBChB OSCEs at the University of Aberdeen and categorised candidates into three levels dependent on examiner awarded global scores of preceding candidates for each station. We modelled the data using a linear mixed model incorporating fixed and random effects.

Findings

A total of 349 candidates sat the OSCEs. The predicted mean (95% CI) score for students following an 'excellent' candidate was 21.6 (20.6, 22.6), followed 'others' was 21.5 (20.5, 22.4), and followed an 'unsatisfactory' student was 22.2 (21.1, 23.3). When accounted for individual, examiner and station levels variabilities, students following an 'excellent' candidate did not have different mean scores compared to those who followed 'other' (p=0.829) or 'unsatisfactory' candidates (p=0.162), however, students who followed an 'unsatisfactory' student scored slightly higher on average compared to those who followed 'other' (p=0.038).

Originality

There was weak evidence that candidate's checklist variations could be attributed to who they followed, particularly those following unsatisfactory students; the difference in predicted mean scores may be of little practical relevance. Further studies with multiple centres may be warranted assuring perceived fairness of the OSCE to candidates and educators.

Keywords

OSCE, Medicine, Assessment, Practical assessment.

Introduction

The Objective Structured Clinical Examination (OSCE) has been advocated as the 'gold standard' of competence assessment in healthcare programmes.(Sloan *et al.*, 1995) Since first being described by Harden in the 1970's (Harden *et al.*, 1975), its use has become ubiquitous in healthcare education assessment around the globe. The principle of this assessment design is that it is an objective method of performance against a structured marking checklist of clinical encounters in an examination format by multiple assessors. The OSCE was designed to combat the unstandardised, uncontrolled and subjective methods of evaluation used in traditional methods of assessment and subsequently improve the psychometric principles of performance assessment including validity and reliability.(Violato, 2018)

Practically, the OSCE involves candidates rotating around a number of timed 'stations' performing a task or skill, often with a patient, actor, or mannequin present. The candidate will be marked by an examiner against a checklist of observed behaviors or skills and awarded a 'global score' by the examiner rating the candidates' overall performance within the station. See Pell et al & Ilgen et al, regarding the relationship between checklist and global scores.(Ilgen *et al.*, 2015; Pell *et al.*, 2015) The candidate will then move on to the next station, with a different examiner, and so on until the cycle is complete. Due to the numbers of students being examined, institutions will often have multiple 'sites' where the exam is being held simultaneously and have multiple 'runs' or sittings of the exam on each site over the course of the exam period. For further detail on OSCEs see Harden et al.(Harden *et al.*, 2015, 1975; Khan *et al.*, 2013) Figure 1 shows the typical setup of a large scale OSCE.

In addition to the practicalities of the OSCE it is worthwhile considering the 'checklist' itself. The OSCE checklist is the method by which examiners award scores for each station within the exam. Whilst the source of much debate in the literature, checklists can range from a list of actions that an examiner is able to observe non-judgmentally to differentially weighted 'key-features' approaches which provide more of a judgement on how well a particular aspect of the task has been performed. (Homer *et al.*, 2020; Regehr *et al.*, 1999)

Whilst early studies of the OSCE focused on concepts of assessment reliability, more recently the understanding of how raters make their judgements has come under scrutiny. (Chahine *et al.*, 2016; Fuller *et al.*, 2013; Gingerich *et al.*, 2014) Assessment resulting in the judgement of performance through direct observation, either in the OSCE or through work-place-based assessment is vital however, Yeates *et al*, state that these forms of assessment are susceptible to a raft of psychometric weaknesses. Literature supports that whilst examiners are instructed to judge against a behavioural standard, they tend to make judgements by comparing candidates against other candidates. This comparison can lead to assimilation or contrast effects (for example in contrast effects a preceding candidate's good performance reduces the scores given to the current candidate's performance making the current performance appear poor 'by contrast').(Yeates, Cardell, *et al.*, 2015; Yeates, Moreau, *et al.*, 2015)

Indeed there are many aspects of examination metrics which affect the variability of candidate scores within the OSCE that have been studied including; the performance of simulated patients (Pell *et al.*, 2010); order effects (Burt *et al.*, 2016), examiner effects including the halo effect whereby candidates are scored higher in numerous aspects of their performance because of excellent performance in one area of their assessment (Chong *et al.*, 2017) , examiner leniency or stringency (Finn *et al.*, 2014; McManus *et al.*, 2006) , when during the assessment period a candidate sits the exam (Hope and Cameron, 2015). One aspect of the OSCE which remains under-assessed is that of the effect of the order in which a candidate rotates through the OSCE process with respect to who they follow in their OSCE stations and how this may affect how observers rate. There is concern from OSCE assessors that if a candidate follows a candidate who is excellent within a station then they are immediately compared against that candidate and may be unfairly disadvantaged or vice versa with those who follow ‘unsatisfactory’ candidates.(Gingerich *et al.*, 2014; Yeates, Cardell, *et al.*, 2015; Yeates *et al.*, 2012; Yeates, Moreau, *et al.*, 2015) In one example, a video-based internet experimental study, Yeates *et al* found that when a good performance was preceded by poor performance, candidates’ global scores were higher (on a 6-point scale) when compared with an unbiased prior performance. (Yeates, Cardell, et al., 2015) One aspect of any assessment process which must be considered is that of fairness, described by Harden as “*the quality of making judgements that are free from bias and discrimination*”.(Harden *et al.*, 2015) Whilst the OSCE process is designed to ensure fairness in terms of having different examiners for each station (McManus *et al.*, 2006)), standardisation of stations within the exam which are the same for each candidate (Harden *et al.*, 2015)) and standardisation of patients within the stations (Plaksin *et al.*, 2016) one aspect that is difficult to control for is biases in terms of contrast or assimilation effects whereby scores are biased unfairly away or towards a candidate based on a preceding candidate’s performance (Yeates, Cardell, *et al.*, 2015)

An influential paper by Yeates *et al* sought to examine relationships between scores of successive performances in two high-stakes assessments, the 2011 United Kingdom Foundation Programme Office clinical assessment and the University of Alberta medical school 2008 Multiple Mini Interview. (Yeates, Moreau, *et al.*, 2015)This study compared behavioural scores (how completely candidates performed against listed criteria based on a five-point rating scale) and global scores (an examiner’s holistic judgement on a five-point ordinal Likert scale) for the UKFPO clinical assessment and global score marking for the MMI. They found that both forms of assessment demonstrated evidence of contrast effects when the average of the three preceding candidates was considered. (Yeates, Moreau, *et al.*, 2015)This study however did not examine checklist scores which are still used frequently within many OSCE systems.(Homer *et al.*, 2020) Whilst global scores are used in borderline regression methodology to determine the station’s pass mark, it is the checklist mark when compared to a station’s pass mark that is used to determine whether a student passes or fails the station and exam therefore understanding the effects on a candidate's checklist score based on who they follow in a station seems prudent.(Pell *et al.*, 2015) Chong *et al* concluded in their review of items influencing OSCE examiner's assessment scores that whilst the psychology and impact of various biases such as the halo effect and the hawk/dove effect is well understood further research is required into the influence of the contrast effect when we consider the 'black box' of decision making within the OSCE. (Chong *et al.*, 2017)

Aims and Objectives

- To assess the effect of following on from an 'excellent' or 'unsatisfactory' candidate with regards to the subsequent candidate's checklist scores in a high-stakes MBChB OSCE examination.
- To assess whether a candidate's position in any OSCE station, with respect to the preceding candidate, affects their checklist scores. In particular, whether following a high-achieving candidate (global score of 'excellent') has a detrimental effect on their score compared to the other candidates.
- We hypothesised that following an excellent candidate (global score rating of 'excellent' for that station) leads to an overall difference (detrimental) in the following candidate's score.

Methods

Background & Context

The University of Aberdeen, Scotland, administers a 5-year Bachelor of Medicine and Bachelor of Surgery degree programme (MBChB). Summative clinical examinations, in the form of the OSCE, are conducted each year and candidates are required to pass for progression to the next stage of the programme or graduation. Since the 2017-2018 academic year, a 'sequential' OSCE has been conducted whereby all candidates sit part 1 of the exam, a 'screening test' of 12 x 8-minute OSCE stations. (Duncumb and Cleland, 2019; Pell *et al.*, 2013) In the OSCE, candidates are expected to perform clinical tasks such as history taking, practical procedures or clinical examinations. Most of the stations are manned by an examiner who awards marks based on a structured checklist, standardised out of 30 marks. The examiner also awards an overall 'global score' rated on a 5-point Likert scale against descriptors. (See Appendix 1 for global score descriptors) Each station has a pass mark calculated using the borderline regression method which plots checklist scores against global scores. Practically, this standard setting method means that each station will have a different pass mark depending on the overall cohort's performance each time the question is used.(Kaufman *et al.*, 2000) Since 2015 the final year OSCE examiners have used an iPAD-hosted bespoke app to record their results.(Brown, 2016) Each data entry submission is time-stamped, therefore, the order in which candidates rotate through a station with respect to each other can be analysed.

The 2018 OSCE contained a total of 12 stations held on one day whereas the 2019 OSCE contained 12 stations over 2 days (6 stations per day, all candidates sat both days). There were 11 manned stations each year resulting in 3838 examination encounters comprising 1694 in 2018 and 2144 in 2019.

In 2018, the OSCE was held simultaneously over 5 sites (distinct geographical locations within the same building which allows for many students to be examined at the same time) with 2 or 3 runs (the number of times the whole exam (numbers of stations) is performed) at each site. In 2019, the OSCE was held simultaneously over 7 sites with 3-5 runs at each site over the two days. Each manned station within a site had a different examiner and therefore with multiple sites running concurrently there were multiple examiners examining each question. Over the two-year study period each examiner may have examined a number of different questions. Figure 1 shows the structure of the OSCE.

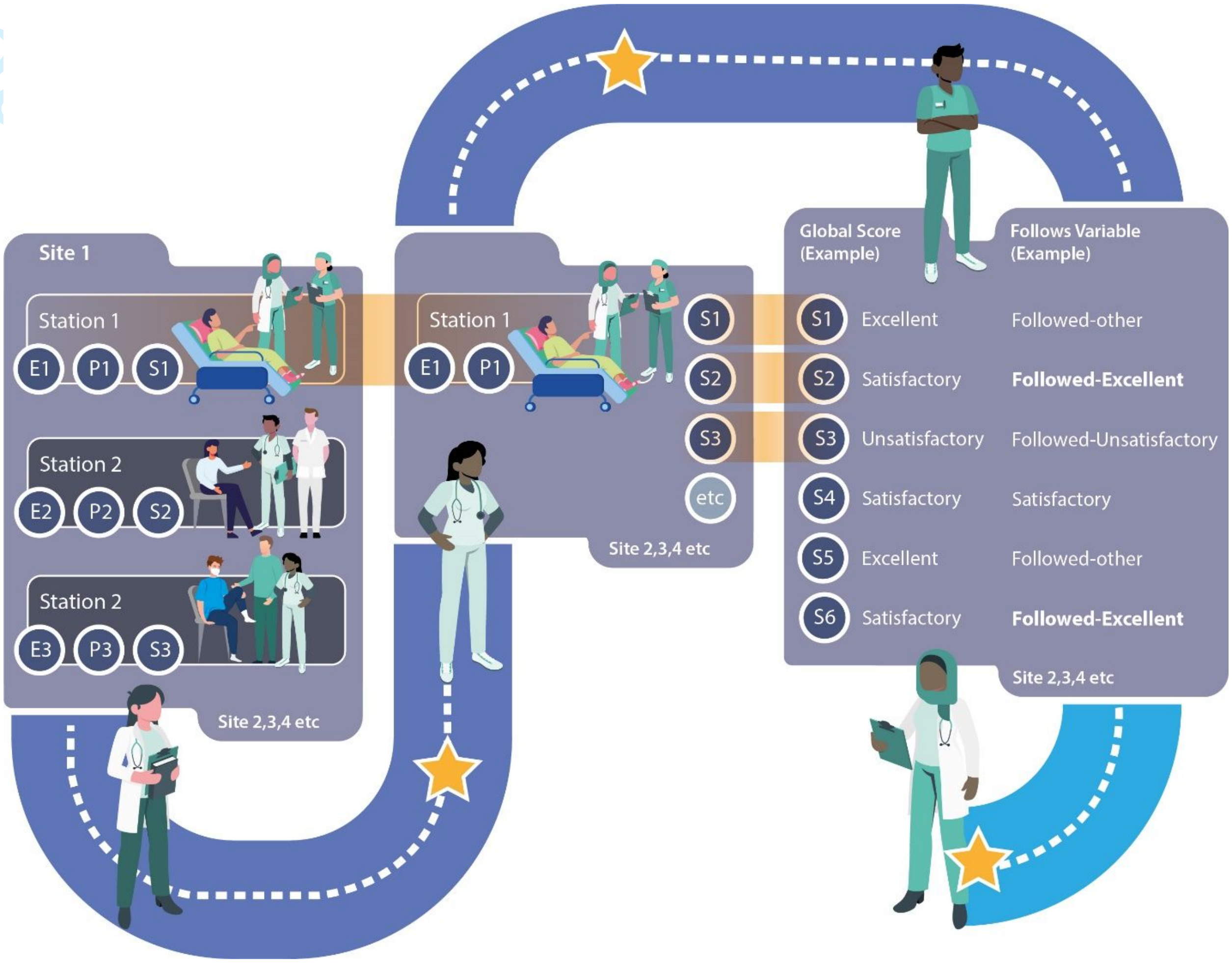

Figure 1: A diagrammatic representation of OSCE structure. The site represents a location where multiple stations are situated. Each station usually has an E-examiner, P-patient partner and S-student. There are usually multiple sites running simultaneously with multiple runs at each site. Within each station (middle portion of above figure) the order in which each S-Student rotates through the station is demonstrated, Student 2 follows Student 1 etc. On the rightmost side, the figure demonstrates how each student is classified according to who they follow within any given station. For example, Student 2 (S2) follows Student 1 (S1) who was marked as 'Excellent' within the station, so S2 is considered to be ‘Follow-Excellent’.

Data Collection

The learning technologies team of the University of Aberdeen extracted the raw data from the OSCE assessment iPADs in Microsoft Excel (Microsoft, Redmund, Washington) format. The data represented the Year 5 MBChB students for academic years 2017-2019 from the University. The data were sorted and organised and candidates within each station were classed against a new variable according to the terms under investigation for each manned station within their OSCE with respect to who they followed within each station based on the preceding student's global score. The new variable 'follow' had three levels: (1) 'follow-excellent' if they followed an 'excellent' candidate, (2) 'follow-unsatisfactory' if they followed a candidate marked as unsatisfactory on the global score and (3) 'follow-other' if they did not follow either an 'excellent' or 'unsatisfactory' candidates. Figure 1 delineates how students were categorized based on who they followed within the station. One unmanned station (prescribing station) in each year was removed from the analysis as a global score is not awarded for this station.

Ethical Considerations

Student identifiable data were removed from the dataset before analysis and the anonymised data were stored, retrieved, and analysed within the University secured computing environment. Ethical approval for this study was granted by the College Ethics Review Board of the University of Aberdeen College of Life Sciences and Medicine. CERB2020/4/1858.

Statistical analysis

We fitted a linear mixed model on the checklist score data including the variable 'follow' as fixed effects and the variables student (n=349), station (n=22) and examiner within station (n=245) as random effects (each examiner could theoretically have examined different stations in combinations of morning or afternoon, day 1 or day 2, 2018 or 2019 but only ever returned one record per candidate when nested within a station). There was no evidence that the mean checklist score was different between the academic years 2018 and 2019 (p=0.192), and therefore, the final model excluded the academic year term. The variances of random effects were estimated using the restricted maximum likelihood method. The comparison of the estimated mean checklist score of the candidate of ‘follow-other’ with ‘follow-excellent’ and ‘follow-unsatisfactory’ was conducted by t-test using Satterthwaite's method and the p-values were adjusted by Tukey-Kramer method to account for multiple comparisons. We checked all model assumptions. Statistical analyses were conducted in the R statistical software environment using the R packages lme4 and lmerTest.

Results

A total of 349 students were examined across the OSCEs, 154 in 2018 (72 male and 82 female) and 195 in 2019 (85 male and 110 female). A total of 404 global scores of ‘excellent’ were awarded (10% of global scores). This led to the identification (due to sequencing in runs) of 348 episodes of candidates being considered as following an excellent candidate; 149 in 2018 where the 12 stations were run over one day and 199 in 2019 where 6 stations were held in runs over 2 days (this resulted in a proportionately smaller number of following an excellent candidate in each group as there were only 6 candidates in each run in 2019 as opposed to 12 per run in 2018. A total of 133 global scores of ‘unsatisfactory’ (3.46% of available global scores) were recorded resulting in 118 episodes (50 in 2018, 68 in 2019) of candidates being considered post-unsatisfactory on global scores. (Table I A&B)

2018 Part 1 (Ran over 1 day- 12 stations in one day)								
	Number of candidates				Checklist Scores			
	Global score 'excellent'	Follows global score 'excellent'	Global score 'unsatisfactory'	Follows global score 'unsatisfactory'	Marks Available	Max Obtained	Min Obtained	Passmark
Q1	23	19	7	7	30.0	29.0	14.0	19.7
Q2	12	10	2	2	30.0	28.5	10.0	18.7
Q3	13	11	12	12	30.0	29.5	6.0	17.8
Q4	Unmanned station not included in analysis							
Q5	8	8	8	8	30.0	29.0	8.5	16.2
Q6	13	12	2	2	30.0	30.0	11.5	19.3
Q7	12	12	2	1	30.0	29.5	11.5	19.2
Q8	13	13	6	6	30.0	28.0	10.0	16.2
Q9	23	21	1	1	30.0	30.0	9.5	20.5
Q10	14	13	1	1	30.0	28.0	11.0	17.5
Q11	21	19	2	2	30.0	28.5	13.0	18.9
Q12	13	11	8	8	30.0	28.5	11.0	18.9
TOTALS	165	149	51	50				

2019 Part 1 (Ran over 2 days- 6 stations each day)								
	Number of candidates				Checklist Scores			
	Global score 'excellent'	Follows global score 'excellent'	Global score 'unsatisfactory'	Follows global score 'unsatisfactory'	Marks Available	Max Obtained	Min Obtained	Passmark
Q1	21	19	7	6	30.0	30.0	12.0	19.8
Q2	40	36	16	12	30.0	29.5	14.0	21.1
Q3	18	16	6	5	30.0	28.5	10.5	16.8
Q4	19	15	5	3	30.0	28.0	12.0	17.8
Q5	19	15	2	2	30.0	28.5	10.0	17.9
Q6	37	31	3	3	30.0	29.5	12.0	18.9
Q7	25	20	2	2	30.0	29.5	14.5	18.9
Q8	Unmanned station not included in analysis							
Q9	14	11	6	4	30.0	26.0	12.0	15.7
Q10	22	16	6	5	30.0	29.0	8.0	14.6
Q11	5	4	20	18	30.0	26.0	7.0	13.9
Q12	19	16	9	8	30.0	30.0	14.5	21.9
TOTALS	239	199	82	68				

Table I: A) 2018 and B) 2019 demonstrates the number of candidates receiving an examiner awarded global score of 'excellent' and the number of candidates that, due to the circuitous nature of the assessment process end up following a candidate with a global score of excellent. The same is shown for 'unsatisfactory' and 'follows-unsatisfactory'. In the far right of the table there is a question-by-question breakdown of checklist marks including the minimum and maximum mark achieved and the borderline-regression calculated pass mark for each station.

Among different sources of variability, stations attributed to the highest variability of the checklist score (n=22; variance: 4.8; 95% lower, upper confidence interval: 2.6, 9.1; contribution: 28.8%) followed by examiners within the station (n=245; variance: 1.9; 95% CI: 1.5, 2.4; contribution: 11.3%) and students (n=349; variance: 1.9; 95% CI: 1.5, 2.3; contribution: 11.2%). Figure 2 presents a caterpillar plot of random effects and corresponding 95% confidence interval for all stations, a sample of 50 examiners within stations, and a sample of 50 students. The figure demonstrates the random effects predictions were more extreme for some stations (for example, S21 and S22), resulting wider spread and hence a higher estimate of the variance of the station. In comparison, random effects predictions for examiners within the station and students were very similar, which reflected almost equal estimates of variances for both sources of variations. The minimum and maximum marks and the pass marks in each station confirm the increased station-level variability (Table 1). The intraclass correlation, estimated as the proportion of variation for each source to the total variation, was estimated as 0.28 for the station and 0.40 for the station-examiner clusters. These estimates suggest that the agreements between the global scores of students for the station and station-examiner clusters were moderate.

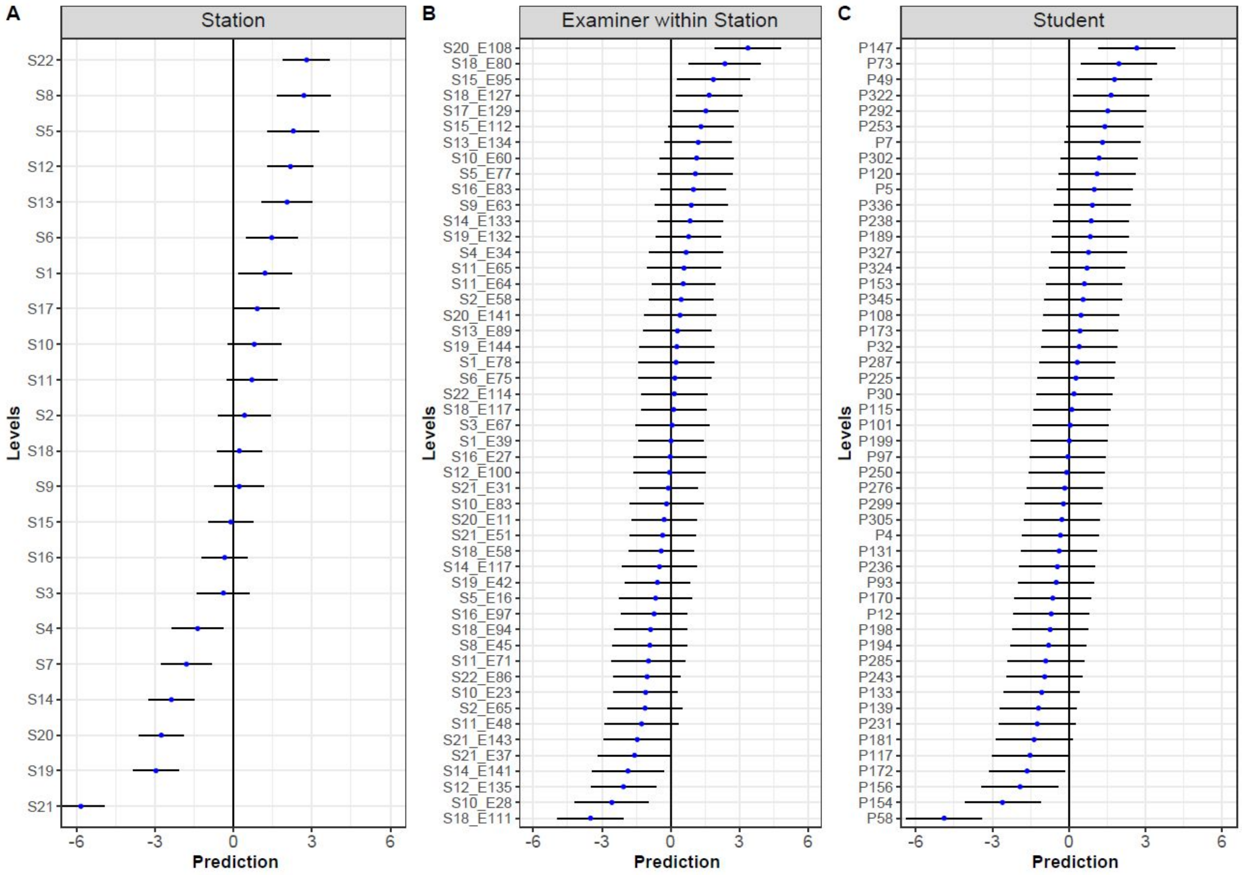

Figure 2: Caterpillar plots of random effects and corresponding 95% confidence interval for all stations, and a sample of 50 examiners within stations and a sample of 50 students based on the fitted linear mixed model. The y-axis labels 'S21' indicates Station 21, 'S18_E111' indicates Examiner 111 in Station 18 and 'P58' indicates Student (pupil) 58. The plot shows a wider spread and hence a higher estimate of the

variance of the station compared to the examiner within the station and student.

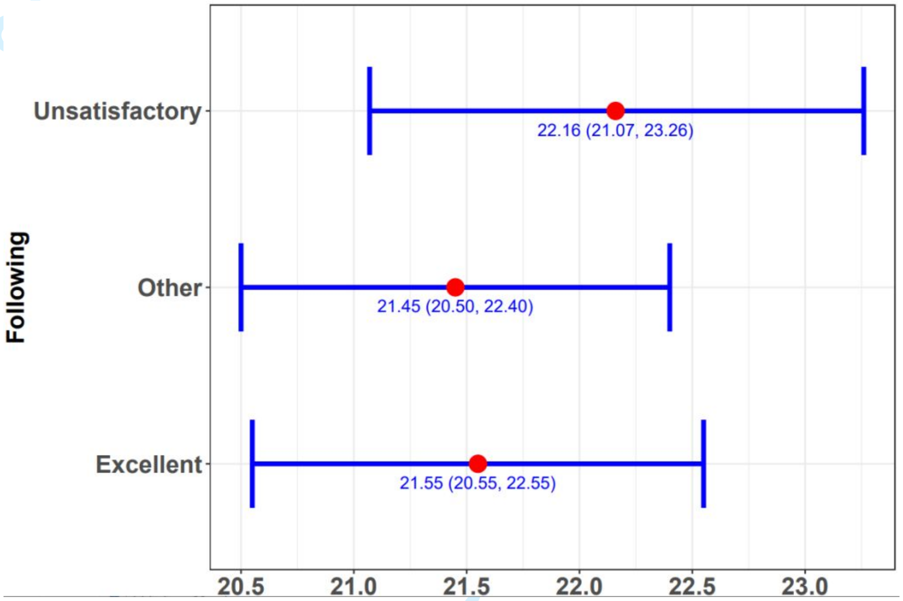

Figure 3: Predicted means and 95% confidence intervals of students who are following 'Excellent', 'Other' and 'Unsatisfactory' groups.

The predicted mean checklist score for students who followed others (i.e. who did not follow an excellent or unsatisfactory candidate on the global score for that question) was 21.5 (95% confidence interval: 20.5, 22.4). The predicted mean checklist score for students who followed an excellent candidate was 21.6 (95% CI: 20.6, 22.6) and those who followed an unsatisfactory student was 22.2 (95% CI: 21.1, 23.3). When accounted for individual, examiner and station levels variabilities, students who followed an 'excellent' candidate did not have different mean scores compared to those who followed 'other' (adjusted p=0.829) or 'unsatisfactory' candidate (p=0.162). However, students who followed an 'unsatisfactory' student scored slightly higher on average compared to those who followed 'other' (p=0.038). The absolute mean difference of checklist score ranged from 0.10 to 0.71 suggesting very small effect sizes.

Discussion

Main findings

The OSCE examination is considered by many as the 'gold standard' of competence-based assessment (Sloan *et al.*, 1995) however concerns have been raised that candidates may be unfairly disadvantaged based on who they follow within the OSCE.(Gauthier *et al.*, 2016; Hyde *et al.*, 2022; Yeates, Cardell, *et al.*, 2015; Yeates, Moreau, *et al.*, 2015) . Out results indicate that there was weak evidence, within the framework of our study, that the mean checklist scores were different between candidates who followed an 'excellent', or 'unsatisfactory' candidate compared to 'other' global score candidates (global p=0.043). When accounted for individual, examiner and station levels variabilities, students who followed an 'excellent' candidate did not have different mean scores compared to those who followed 'other' (adjusted p=0.829) or 'unsatisfactory' candidate (p=0.162), however, students who followed an 'unsatisfactory' student scored slightly higher on average compared to those who followed 'other' (p=0.038). A comparison of the mean scores between three groups suggests that the observed effect size is small with little practical relevance (the absolute mean difference of checklist score between different categories of following ranged from 0.10 to 0.71). It was interesting to note that a large proportion of candidates (139 candidates) did not achieve a global score of excellent in any station, and 'excellent' was maximally achieved by one single candidate in six out of the 11 stations. This shows that whilst a candidate may be excellent in one station, this does not equate to them being considered excellent throughout all stations. Combined with the observed variances of stations (indicated by the differing pass marks within each station) and examiners within each station, this again demonstrates the importance of having an adequate number and variety of stations and examiners within the OSCE setup.

Comparison with previous literature

Rater cognition has become a topic of interest in the healthcare education literature as we seek to understand the complexities of how rater's make their decisions. (Hyde *et al.*, 2022; Yeates, Cardell, *et* *al.*, 2015; Yeates *et al.*, 2012) Whilst previously conceptualized as a mostly passive process of observation, it has now been theorized as a complex active cognitive process. (Gauthier *et al.*, 2016; Hyde *et al.*, 2022)

Assessor's scores of performances are highly variable; one study attributes this variability due to differential salience whereby different assessors valued different aspects of performance to varying degrees, and criterion uncertainty where assessors constructed criteria differently and were influenced by recent exemplars.(Yeates *et al.*, 2013) In this study we wished to assess whether examiner's were influenced by the preceding candidate who they thought was 'excellent' on the global score and if a student who followed an 'excellent' candidate within a question was unfairly disadvantaged. Articles in the literature mainly focus on contrast effects in domain-based behavioral scores in the MMI or work-place-based assessments and experimental studies as opposed to the effect on real-life checklist scores with which our interest lies.(Chong *et al.*, 2017; Yeates, Cardell, *et al.*, 2015; Yeates *et al.*, 2012)

Recent debate within healthcare professions education concerns whether the OSCE marking scheme should be domain-based, checklist-based (including differentially weighted checklist marks) or something other such as entrustment measures as opposed to performance measures.(Homer *et al.*, 2020; Pinilla *et al.*, 2023; Regehr *et al.*, 1998) Our results indicated a low absolute mean difference of checklist scores between different categories of following, although statistically significant (for one comparison) also testifies that there is little practical difference in the mean scores that comes from who one follows.

Whilst checklist marking schemes are considered by some to be reductionist, with some research interest on checklist/global score alignment being conducted when investigating station-related metrics.(Pell *et al.*, 2015) Our study shows that by using a checklist marking scheme candidates can be assured that their performance is based on what is actually observed despite who they follow. In our institution, the OSCE process also involved examiner and patient partner training and calibration alongside lead examiners, considered important steps with regards to the standard setting, reliability and quality assurance of assessment procedures.

Our study findings contrast with an experimental study whereby assessors were randomised to viewing poor postgraduate candidate performance of a mini clinical examination (miniCEX) encounter (similar to an OSCE station but usually conducted in the real-world clinical environment) or good candidate performance, then each assessor being given a borderline performances to rate.(Yeates *et al.*, 2012) Their results showed a significant contrast bias effect where assessors rated performance of candidates following a good candidate lower than those who were exposed to a poor candidate on subsequent candidates' global scores.(Yeates *et al.*, 2012) These differences may be attributable to the wider range of candidates often seen in the postgraduate context as compared to those who had been trained in one institution and were used to the assessment process as well as this experiment focusing on global/behavioural scores as opposed to checklist scores similar to the findings of Yeates, Cardell *et al*. and Yeates, Moreau, et al. (Yeates, Cardell, *et al.*, 2015; Yeates, Moreau, *et al.*, 2015)

Strengths and Limitations

The main strengths of our study include that empirical 'real world' OSCE data were used as opposed to experimental methods or non-OSCE studies of examiner contrast effects when studying rater behaviour on candidate's checklist scores which are still used in many healthcare programmes.(Homer *et al.*, 2020; Pell *et al.*, 2015; Yeates *et al.*, 2012; Yeates, Moreau, *et al.*, 2015) We also chose to consider candidates as 'excellent' or following on from 'excellent' within each station as opposed to considering the top performing candidates in the exam as a whole. We chose this method as we wanted to assess the contrast effects of each examiner, nested within each question, on the subsequent student.

Our study has some limitations. We conducted this study on two years of final year undergraduate MBChB assessment data, therefore, results may not be generalisable to other OSCE situations such as postgraduate settings or in other professional groups. In this study, we chose not to examine data from the sequential days (Duncumb and Cleland, 2019) as this cohort was less likely to be representative of the general student population in terms of its self-selectiveness.

Conclusions and implications for practice and further study

In summary, the OSCE assessment is a complex interaction, conceptualised as a 'black box' of variance between test format design issues, assessor behaviours and candidate performance. (Chong *et al.*, 2017; Pell *et al.*, 2015) This study adds to the literature when considering the interaction between the OSCE design format, assessor behaviours and the candidate's performance based on who they follow in this structured clinical examination. There was weak evidence that the variations in checklist scores of candidates could be attributed to who they followed in a high-stake OSCE particularly those followed unsatisfactory students; the difference in predicted mean scores, however, may be of little practical relevance. A future study with larger sample size and predefined limits of equivalence may be warranted to assess the examiner's contrast effects more rigorously and assuring the perceived fairness of the OSCE examination to candidates and educators alike. Further work is required to assess the effect of following on from excellent candidates in formal postgraduate practical assessments as well as examining the effect of following an 'excellent' candidate in the sequential portion of the OSCE.

References

Brown, C.W. (2016), “Tablet- or iPAD-based marking of OSCEs and MMIs: An imaginative cost-saving approach”, *Medical Teacher*, Vol. 38 No. 2, pp. 211–212, doi: 10.3109/0142159X.2015.1072270.

Burt, J., Abel, G., Barclay, M., Evans, R., Benson, J. and Gurnell, M. (2016), “Order effects in high stakes undergraduate examinations: an analysis of 5 years of administrative data in one UK medical school”, *BMJ Open*, Vol. 6 No. 10, p. e012541, doi: 10.1136/bmjopen-2016-012541.

Chahine, S., Holmes, B. and Kowalewski, Z. (2016), “In the minds of OSCE examiners: uncovering hidden assumptions”, *Advances in Health Sciences Education*, Vol. 21 No. 3, pp. 609–625, doi: 10.1007/s10459-015-9655-4.

Chong, L., Taylor, S., Haywood, M., Adelstein, B.-A. and Shulruf, B. (2017), “The sights and insights of examiners in objective structured clinical examinations”, *Journal of Educational Evaluation for Health Professions*, Vol. 14, p. 34, doi: 10.3352/jeehp.2017.14.34.

Duncumb, M. and Cleland, J. (2019), “Student Perceptions of a Sequential Objective Structured Clinical Examination”, *Journal of the Royal College of Physicians of Edinburgh*, Vol. 49 No. 3, pp. 245–249, doi: 10.4997/jrcpe.2019.315.

Finn, Y., Cantillon, P. and Flaherty, G. (2014), “Exploration of a possible relationship between examiner stringency and personality factors in clinical assessments: a pilot study”, *BMC Medical Education*, Vol. 14 No. 1, p. 1052, doi: 10.1186/s12909-014-0280-3.

Fuller, R., Homer, M. and Pell, G. (2013), “Longitudinal interrelationships of OSCE station level analyses, quality improvement and overall reliability”, *Medical Teacher*, Vol. 35 No. 6, pp. 515–517, doi: 10.3109/0142159X.2013.775415.

Gauthier, G., St-Onge, C. and Tavares, W. (2016), “Rater cognition: review and integration of research findings”, *Medical Education*, Vol. 50 No. 5, pp. 511–522, doi: 10.1111/medu.12973.

Gingerich, A., Kogan, J., Yeates, P., Govaerts, M. and Holmboe, E. (2014), “Seeing the ‘black box’ differently: assessor cognition from three research perspectives”, *Medical Education*, Vol. 48 No. 11, pp. 1055–1068, doi: 10.1111/medu.12546.

Harden, R., Lilley, P. and Patricio, M. (2015), *The Definitive Guide to the OSCE The Objective Structured Clinical Examination as a Performance Assessment.*, 1st ed., Elsevier Health Sciences., Amsterdam.

Harden, R.M., Stevenson, M., Downie, W.W. and Wilson, G.M. (1975), “Assessment of clinical competence using objective structured examination.”, *BMJ*, Vol. 1 No. 5955, pp. 447–451, doi: 10.1136/bmj.1.5955.447.

Homer, M., Fuller, R., Hallam, J. and Pell, G. (2020), “Shining a spotlight on scoring in the OSCE: Checklists and item weighting”, *Medical Teacher*, Vol. 42 No. 9, pp. 1037–1042, doi: 10.1080/0142159X.2020.1781072.

Hope, D. and Cameron, H. (2015), “Examiners are most lenient at the start of a two-day OSCE”, *Medical Teacher*, Vol. 37 No. 1, pp. 81–85, doi: 10.3109/0142159X.2014.947934.

Hyde, S., Fessey, C., Boursicot, K., MacKenzie, R. and McGrath, D. (2022), “OSCE rater cognition – an international multi-centre qualitative study”, *BMC Medical Education*, Vol. 22 No. 1, p. 6, doi: 10.1186/s12909-021-03077-w.

Ilgen, J.S., Ma, I.W.Y., Hatala, R. and Cook, D.A. (2015), “A systematic review of validity evidence for checklists versus global rating scales in simulation-based assessment”, *Medical Education*, Vol. 49 No. 2, pp. 161–173, doi: 10.1111/medu.12621.

Kaufman, D.M., Mann, K. V, Muijtjens, A.M. and van der Vleuten, C.P. (2000), “A comparison of standard-setting procedures for an OSCE in undergraduate medical education.”, *Academic Medicine : Journal of the Association of American Medical Colleges*, Vol. 75 No. 3, pp. 267–71, doi: 10.1097/00001888-200003000-00018.

Khan, K.Z., Ramachandran, S., Gaunt, K. and Pushkar, P. (2013), “The Objective Structured Clinical Examination (OSCE): AMEE Guide No. 81. Part I: An historical and theoretical perspective”, *Medical Teacher*, Vol. 35 No. 9, pp. e1437–e1446, doi: 10.3109/0142159X.2013.818634.

McManus, I., Thompson, M. and Mollon, J. (2006), “Assessment of examiner leniency and stringency ('hawk-dove effect’) in the MRCP(UK) clinical examination (PACES) using multi-facet Rasch modelling”, *BMC Medical Education*, Vol. 6 No. 1, p. 42, doi: 10.1186/1472-6920-6-42.

Pell, G., Fuller, R., Homer, M. and Roberts, T. (2010), “How to measure the quality of the OSCE: A review of metrics – AMEE guide no. 49”, *Medical Teacher*, Vol. 32 No. 10, pp. 802–811, doi: 10.3109/0142159X.2010.507716.

Pell, G., Fuller, R., Homer, M. and Roberts, T. (2013), “Advancing the objective structured clinical examination: sequential testing in theory and practice”, *Medical Education*, Vol. 47 No. 6, pp. 569–577, doi: 10.1111/medu.12136.

Pell, G., Homer, M. and Fuller, R. (2015), “Investigating disparity between global grades and checklist scores in OSCEs”, *Medical Teacher*, Vol. 37 No. 12, pp. 1106–1113, doi: 10.3109/0142159X.2015.1009425.

Pinilla, S., Lerch, S., Lüdi, R., Neubauer, F., Feller, S., Stricker, D., Berendonk, C., *et al.* (2023), “Entrustment versus performance scale in high-stakes OSCEs: Rater insights and psychometric properties”, *Medical Teacher*, pp. 1–8, doi: 10.1080/0142159X.2023.2187683.

Plaksin, J., Nicholson, J., Kundrod, S., Zabar, S., Kalet, A. and Altshuler, L. (2016), “The Benefits and Risks of Being a Standardized Patient: A Narrative Review of the Literature”, *The Patient - Patient-Centered Outcomes Research*, Vol. 9 No. 1, pp. 15–25, doi: 10.1007/s40271-015-0127-y.

Regehr, G., Freeman, R., N, M. and R, H. (1999), “OSCE performance evaluations made by standardized patients: Comparing checklist and global rating Scores”, *Academic Medicine*, Vol. 74 No. 10, pp. s135–s137.

Regehr, G., MacRae, H., Reznick, R.K. and Szalay, D. (1998), “Comparing the psychometric properties of checklists and global rating scales for assessing performance on an OSCE-format examination”, *Academic Medicine*, Vol. 73 No. 9, pp. 993–7, doi: 10.1097/00001888-199809000-00020.

Sloan, D.A., Donnelly, M.B., Schwartz, R.W. and E, W. (1995), “The Objective Structured Clinical Examination The New Gold Standard for Evaluating Postgraduate Clinical Performance”, *Annals of Surgery*, Vol. 222 No. 6, pp. 735–742, doi: 10.1097/00000658-199512000-00007.

Violato, C. (2018), *Assessing Competence in Medicine and Other Health Professions*, CRC Press, Boca Raton : Florida : CRC Press, [2019], doi: 10.1201/9780429426728.

Yeates, P., Cardell, J., Byrne, G. and Eva, K.W. (2015), “Relatively speaking: contrast effects influence assessors’ scores and narrative feedback”, *Medical Education*, Vol. 49 No. 9, pp. 909–919, doi: 10.1111/medu.12777.

Yeates, P., Moreau, M. and Eva, K. (2015), “Are Examiners’ Judgments in OSCE-Style Assessments Influenced by Contrast Effects?”, *Academic Medicine*, Vol. 90 No. 7, pp. 975–980, doi: 10.1097/ACM.0000000000000650.

Yeates, P., O’Neill, P., Mann, K. and Eva, K. (2013), “Seeing the same thing differently”, *Advances in Health Sciences Education*, Vol. 18 No. 3, pp. 325–341, doi: 10.1007/s10459-012-9372-1.

Yeates, P., O’Neill, P., Mann, K. and Eva, K.W. (2012), “Effect of Exposure to Good vs Poor Medical Trainee Performance on Attending Physician Ratings of Subsequent Performances”, *JAMA*, Vol. 308 No. 21, p. 2226, doi: 10.1001/jama.2012.36515.

Appendix 1: Examiner awarded Global Score descriptors for final year MBChB students (ordinal scale)

Excellent (5)

Excellent demonstration of a cohesive and logical approach. Demonstrated excellent medical knowledge and clinical skills. Uses insightful and adaptive approach to patient with excellent interaction with the patient. Demonstrates an accomplished level of professionalism.

Highly Satisfactory (4)

Demonstrates cohesive and logical approach. Demonstrated thorough medical knowledge and clinical skills. Appropriate and adaptive approach to patient with good interaction with the patient. Demonstrates professionalism.

Satisfactory (3)

Overall reasonably cohesive and logical approach. Demonstrated adequate medical knowledge and clinical skills. Evidence of attempts to adapt approach to patient with reasonable interaction with the patient. Attempted to demonstrate professionalism.

Borderline (2)

Lacks cohesive or logical approach. Demonstrated basic understanding (with some inaccuracies) of required medical knowledge and clinical skills. Limited attempts to adapt to situation and very poor interaction with the patient. Demonstrated little evidence of professionalism.

Unsatisfactory (1)

Disorganized approach with several omissions. Very limited understanding of required medical knowledge and/or clinical skills. Fails to demonstrate logical approach with little flexibility and poor interaction with the patient. Fails to demonstrate professionalism.

Declarations

Ethics approval and consent to participate

Ethical permission for this study was granted by the College Ethics Review Board of the College of Life Sciences and Medicine, University of Aberdeen (CERB2020/4/1858).

Availability of data and materials

The data is available upon reasonable request by contacting the corresponding author.

Competing interests

The authors declare that they have no competing interests.

Funding

Not applicable.

Presentations:

An early version of this study was presented as a poster presentation at the Higher Education Teaching and Learning (HETL) conference in Aberdeen, June 2023.