RESEARCH

Open Access

Measuring and correcting staff variability in large-scale OSCEs

Skerdi Haviari[1,2,3], Christian de Tymowski[1,4,5], Nelly Burnichon[1,6,7], Cédric Lemogne[1,8,9], Martin Flamant[1,4,10], Philippe Ruszniewski[1,4,11], Saja Bensaadi[1], Gregory Mercier[1], Hasséne Hamaoui[1], Université Paris Cité OSCE study group, Tristan Mirault[1,6,12], Albert Faye[1,13,14] and Donia Bouzid[1,2,15*]

Abstract

Context Objective Structured Clinical Examinations (OSCEs) are an increasingly popular evaluation modality for medical students. While the face-to-face interaction allows for more in-depth assessment, it may cause standardization problems. Methods to quantify, limit or adjust for examiner effects are needed.

Methods Data originated from 3 OSCEs undergone by 900-student classes of 5th- and 6th-year medical students at Université Paris Cité in the 2022-2023 academic year. Sessions had five stations each, and one of the three sessions was scored by consensus by two raters (rather than one). We report OSCEs' longitudinal consistency for one of the classes and staff-related and student variability by session. We also propose a statistical method to adjust for inter-rater variability by deriving a statistical random student effect that accounts for staff-related and station random effects.

Results From the four sessions, a total of 16,910 station scores were collected from 2615 student sessions, with two of the sessions undergone by the same students, and 36, 36, 35 and 20 distinct staff teams in each station for each session. Scores had staff-related heterogeneity (p<10^{-15}), with staff-level standard errors approximately doubled compared to chance. With mixed models, staff-related heterogeneity explained respectively 11.4%, 11.6%, and 4.7% of station score variance (95% confidence intervals, 9.5-13.8, 9.7-14.1, and 3.9-5.8, respectively) with 1, 1 and 2 raters, suggesting a moderating effect of consensus grading. Student random effects explained a small proportion of variance, respectively 8.8%, 11.3%, and 9.6% (8.0-9.7, 10.3-12.4, and 8.7-10.5), and this low amount of signal resulted in student rankings being no more consistent over time with this metric, rather than with average scores (p=0.45).

Conclusion Staff variability impacts OSCE scores as much as student variability, and the former can be reduced with dual assessment or adjusted for with mixed models. Both are small compared to unmeasured sources of variability, making them difficult to capture consistently.

Keywords OSCE, Score variability, Inter-rater variability

*Correspondence:
Donia Bouzid
donia.bouzid@aphp.fr
Full list of author information is available at the end of the article

Introduction

Since their introduction in the 1970s, Objective Structured Clinical Examinations (OSCEs) have become a widely employed tool for measuring the clinical competence of healthcare students. Indeed, this tool aims to embody the qualities of an ideal assessment approach, emphasizing the enhancement of validity, reliability, objectivity, and practicality [1]. It is extensively used in evaluating the practical skills of undergraduate health students as part of summative assessments. However, OSCEs' ability to discern a general underlying trainee skill or talent level, as understood in generalizability (G) theory, has been questioned by work measuring variance attributable to students and finding values as low as <5% [2]. Various forms of rater bias exist, including the halo effect, the rater's mood, familiarity with candidates, rater's experience, or even gender [3–5], and in contrast with the low contribution of a student-attributable variance, some studies have found that as much as 16% to 34% of score variance may be due to examiners [6, 7].

Furthermore, in 2013, Yeates et al. introduced three concepts to elucidate the origins of rater variability [8]. These concepts were labelled as follows: differential salience, i.e., when the importance attached to a given aspect differs from one rater to another; criterion uncertainty, i.e., varying and uncertain rater's notions of competence; and information integration, i.e., raters typically form global impressions using their unique descriptive language instead of discrete numeric scores. OSCE inter-rater low reliability is resistant to standard improvements, e.g., in rater training [9]. In theory, multiplying OSCE stations can allow a broad sampling, thus a satisfactory measurement of average skill mastery, leading to an increased number of stations, with recommendations in the 10-20 range [10]. However, OSCEs are expensive and time-consuming, which limits the feasible number of stations; for example, in 2012, in the UK, an estimate of the cost for a 15-station OSCE was 23.67 GBP per student and per station and 335 GBP total per student [11], which at the time was close to 4 and 57 times the 6.19 GBP minimum hourly salary for adults in the UK [12]. In addition to practical difficulties, adding stations to a single session to better measure skill mastery may not necessarily help better predict future performances (and predicting future performance is an essential way in which grades are used in practice, e.g., to allocate professional positions). In this context, better designs and/or analysis methods would be helpful to assess the underlying level of students, assuming that such a concept can be operationalized, as educational and early career institutions mostly do.

Another potential weakness in the literature about OSCEs is the longitudinal variation of scores over time. OSCEs have been found to correlate weakly with other measures of student performance, which has been interpreted to mean that OSCEs measure something different from other student examinations [13]. However, although significant, the low correlation could also be interpreted as OSCEs being a noisy measure, not capturing anything other than motivation and familiarity with the station subjects on a specific day, in addition to a small general talent for OSCEs.

Moving towards a practical measurement of such a relatively stable "student's performance" (akin to general intelligence in psychometrics, but in the limited context of medical OSCEs) would require methods to disentangle many sources of variability, such as station wording, station topic, student' emotional state, student preferences, and inter-rater variability. Inter- teacher variability, in particular, has often been accused of reducing the OSCEs validity and can worry students.

Therefore, in large-scale OSCEs performed in a different cohort of students in the medical school of Université Paris Cité, we aimed (1) to improve OSCE validity with the implementation of a two-rater system during the session and (2) to better assess the staff-related effect when analyzing the results with a refinement of a previously described mixed statistical model [14].

Methods

Study design and populations

We conducted a retrospective analysis based on anonymized students' scores from Université Paris Cité Medical School. The sessions whose data is analyzed are described in Table 1. The medical school of Université Paris Cité conducted on-site OSCEs as mandatory examinations for the the class we refer to as C19 (class of 2019, by year of confirmed admission into the medical curriculum) in Feb 2023 (session OSCE-C19-1) and June 2023 (session OSCE-C19-2). A different class (of 2018, C18) also took a mandatory exam in February 2023 (OSCE-C18). A third class (of 2020, C21) also took a mandatory exam in December 2023 (OSCE-C21).

Teachers from the Université Paris Cité medical school, heterogeneous with respect to OSCE experience, administered the exam and were involved as raters or standardized participants. For all three sessions, teachers administered the same station throughout the event but could swap roles freely between rater and standardized participant. The team assigned to a station is referred to as "station staff" since standardized participants and raters were always linked at each station throughout the day, forming a single unit of statistical analysis.

The study was carried out anonymously as part of routine quality monitoring under the purview of data-processing agreements required from all students and teachers. Our institutional IRB, the "Comité d'Evaluation

Table 1 Description of the OSCE sessions

Session name	OSCE-C19-1	OSCE-C19-2	OSCE-C18	OSCE-C21
Date	Dec 2022	Jun 2023	Feb 2023	Dec 2023
Class (year of admission)	C19 (2019)	C19 (2019)	C18 (2018)	C21 (2021)
Curricular year	5th	5th	6th	4th
Mandatory	Yes	Yes	Yes	Yes
Number of expected students	896	888	869	767
Number of scored students	881	871	863	747
Number of stations per student	5	5	5	5
Number of station staff teams	180	180	175	105
Number of tracks	36	36	35	20
Students per complete track	25	25	25	40
Number of staff per station, of which	1 or 2	2 or 3	1 or 2	2 or 3
• Standardized participant	0 or 1	0 or 1	0 or 1	0 or 1
• Rater(s)	1	2	1	2
Staff role swapping allowed	Yes	Yes	Yes	Yes
2-rater consensus or average grading	NA	Consensus	NA	Average

de l'Ethique des projets de Recherche Biomédicale Paris Nord" (IRB00006477), waived the need for ethical review.

OSCE tracks

In each session, each student participated in all five OSCEs stations. Expert teachers from the Université Paris Cité OSCE group carefully prepared the scripts for these stations. Two other teachers evaluated each script and tested it with volunteer residents.

Each station lasted 8 minutes, with 1 minute to switch between 2 stations, except for OSCE-C19-2, where a 2-minute pause was scheduled to reach a consensus between the two raters. For logistic reasons, OSCEs were organized in tracks. Students were divided into groups of 5, and these groups of 5 underwent the same five stations approximately every hour at different locations in the same building (referred to as tracks), rotating every 10 minutes. Thus, the 5 stations from the same track shared students with each other, but not with the other tracks, throughout the day. This increased the effect of staff-related variability on overall scores ; if a track happened to draw particularly lenient (or severe) raters and helpful (or unhelpful) standardized participants in its 5 stations, no compensation of this stroke of (bad) luck by other stations's staff could be expected for the students affected to this track.

Two groups of students had extra time due to disabilities (+3 min) and did their sessions at the end of the day.

OSCE station scripts

In OSCE-C19-1, the five scripts covered the following topics: etiologic diagnosis of an asthenia secondary to an oedematous-ascitic decompensation; etiologic diagnosis of a chest pain secondary to ischemic cardiopathy; etiologic diagnosis and management of a child with atopic dermatitis; management of the discovery of a pulmonary nodule; hand disinfection and suturing (technical OSCE).

In OSCE-C19-2, the scripts covered the following topics: management of a wrist fracture; management of the discovery of a high-level trisomy 21 screening; interview of a patient with a suspected major depressive episode; call to an intensive care physician to refer a patient for septic shock; management of a high level of prostate-specific antigen with simulated digital rectal examination (technical OSCE).

In OSCE-C21, the scripts covered the following topics : hip fracture; electrocardiogram in the context of chest pain; anemia for subacute fatigue; asthma in primary care; fecaloma with simulated digital rectal examination (technical OSCE).

In OSCE-C18, the scripts covered the following topics: etiologic diagnosis of erythema with infectious origin; management of paracetamol overdose; request for a contrast CT scan for suspicion of pulmonary embolism; evaluation of child psychomotor development; etiologic diagnosis of anal bleeding with simulated digital rectal examination (technical OSCE).

OSCE grading system

Raters observed the OSCE station and then completed the evaluation grid online through a dedicated software. For OSCE-C19-2, the two raters filled out a single form after agreeing on grading ; for OSCE-C21, they each filled a separate form and their average was used. For OSCE-C19-1 and OSCE-C18, the single rater filled the form. In all sessions, standardized participants were not involved in student grading.

The scoring system was binary (Fulfilled/Not fulfilled) for 9 to 12 scenario-specific items. In addition, raters evaluated 2 to 5 general skills (such as "ability to listen" or "ability to communicate with the patient") that were relevant to each script. The 2 to 5 general skills were chosen among 11 possibilities and graded on a 5-point Likert scale. The score for each general skill could take the values {0,0.25,0.5,0.75,1} and was added to the score from the scenario-specific items. The total score, with a minimum of 0 and a maximum from 12 to 15, depending on the number of evaluated items and skills, was linearly normalized to the [0-20] range.

Objectives

The primary objective was to measure grading variability due to evaluation teams (teacher pairs or triplets) referred to as staff variability.

The secondary objective was to assess student effects (i.e., differences in skill) and the ability of the OSCE system to reveal it.

Statistical analysis

Separate descriptive analyses were performed for each station in each OSCE session. We report scores, overall and by station, for students. We also report them for staff teams as an average of all given scores (most staff teams evaluated 5 groups of 5 students, with some rounding discrepancies).

We used linear mixed models with the station score as a dependent variable to account for variability in student skill, staff behavior, and station script difficulty. Random student, random staff , and random station effects were used as predictors. This allowed the computation of a student effect capturing each student's skill, a staff effect capturing each staff team's severity (including both standardized patient helpfulness and rater level of expectation), and a script effect capturing each station's difficulty.

We then estimated intraclass correlation coefficients (ICCs, also called variance partition coefficients) for students, staff, and stations. The ICCs for each class represent the score variance due to the class, expressed as a proportion of the total variance (student + staff + script + residual). They are further referred to as the "explained variance" of each factor.

A low explained variance for students means that there is no stable "student skill" relevant to all stations; a low explained variance for staff teams means that there are few conduct or grading differences between them, and a low explained variance for scripts means that all scripts have similar difficulty.

Confidence intervals for explained variances were obtained by approximating the ratio of explained to unexplained variance as a ratio of independent variances, for which a confidence interval formula is readily available. The latter uses quantiles 0.025 and 0.975 of an F distribution (which describes a ratio of means of summed squares) with degrees of freedom corresponding to the factor of interest (students or raters or stations or residual degrees of freedom) in the numerator and the remaining degrees of freedom in the denominator. These quantiles are multiplied by the observed ratio of variances to obtain a confidence interval [15]. Then, the values for the estimate and the confidence interval were converted from a ratio R (explained/unexplained) to a fraction F with the total variance in the denominator (explained/total) using F = 1/(1+1/R).

The analysis was repeated after transforming scores so that each station had the same standard error. For this, the standardized distance of each student's score to the mean was computed for each station and then retransformed into a standardized score by multiplying it by the average standard deviation across stations, keeping the mean constant. In other words, this is equivalent to scaling the data using the pre-existing mean and the average standard deviation. This sensitivity analysis echoed fairness concerns so that some stations did not have a more significant impact than others on the estimation of student effects.

The student random effects were not used directly as scores but instead formed the basis for rankings. Shifts in ranks were computed between the score-based and random-effects-based ranks, the latter considering staff variability.

All models' variance estimates were obtained based on the restricted maximum likelihood (REML) with the lme4 package of R 4.1.2 software. The analysis was limited to available data without imputation.

Results

Available data

In OSCE-C19-1, a total of 4480 scores were expected (one score for each of the five stations for each of the 896 students), 75 were missing for 15 absent students; in OSCE-C19-2, 4440 scores were expected (888 students), and 85 were missing for 17 absent students. 864 students underwent both OSCE-C19-1 and OSCE-C19-2 and were available for longitudinal evaluation. In OSCE-C18, 4345 scores were expected (869 students), and 30 were missing for 6 absent students. In OSCE-C21, 3835 scores were expected (767 students) and 100 were missing for 20 absent students.

Evidence for rater variability

The distribution of scores received by each student at each station and their average by student are given in Table 2. The average score given by each rater team is shown in Table 3. A staff effect is visible; for example, for station 1 in OSCE-C19-1, the standard deviation at the student level is 3.9. Given that almost all evaluation teams evaluated 25 students, if picked randomly, the expected standard deviation for these 25-student averages is $3.9/\sqrt{25} = 0.78$. However, the observed value is 1.8 ($p<10^{-15}$ by omnibus ANOVA over staff teams). Similar increased deviations are observed across most stations, indicating that the standardized patient's behavior and/or rater's stringency may differ across staff teams.

Possible scores range from 0 (worst) to 20 (best). OSCE session names are as in Table 1.

For OSCE-C19-1, OSCE-C19-2 and OSCE-C18, all but 3 staff teams per session evaluated 25 students ; the others between 20 and 24. The figure for each staff team is therefore mostly a 25-student average, and there are 36 such figures, whose distribution is presented. Students were not distributed at random from one station to the next : circuits linked together 5 staff teams (one per station), making it meaningful to present the global average per circuit. For OSCE-C21, the 2/3-person staff teams evaluated 40 students each, except for when there were missing students, and one track (5 stations, 5 evaluation teams) had 20 students. The figures are therefore mostly 40-student averages.

The effect is conserved at the overall score effect, where, e.g., the expected standard deviation for these 25-student averages for OSCE-C19-1 is $1.9/\sqrt{25} = 0.38$. In contrast, the observed value is 0.8 ($p<10^{-15}$ by ANOVA over circuits, each linking 5 staff teams). This suggests that averaging over five evaluation teams could not entirely compensate for staff variability. The effect was still present but seemed smaller, for the two-rater session OSCE-C19-2, with an observed standard deviation of 0.6

Table 2 Distribution of student scores overall and by station for the 2022-2023 academic year OSCEs at Université Paris Cité medical school

Scores by student	N	Median	(Interquartile)	[Min-Max]	Mean	+-SD	(SEM)
OSCE-C19-1							
Overall (average)	*N*=881	11.7	(10.4 - 13.1)	[5.7 - 17.2]	11.7	+-1.9	(0.6)
Station 1	*N*=881	9.6	(6.9 - 12.7)	[0.0 - 20.0]	9.7	+-3.9	(1.3)
Station 2	*N*=881	14	(11.7 - 16.3)	[5.3 - 20.0]	13.9	+-3.0	(0.8)
Station 3	*N*=881	8.8	(6.2 - 11.7)	[0.0 - 20.0]	9.1	+-4.0	(1.3)
Station 4	*N*=881	13.0	(11 - 14.7)	[3.0 - 20.0]	12.9	+-2.8	(0.8)
Station 5	*N*=881	13.7	(10.3 - 16)	[0.0 - 20.0]	12.9	+-3.8	(1.1)
OSCE-C19-2							
Overall (average)	*N*=871	10.9	(9.7 - 12.3)	[4.8 - 17.5]	11.0	+-2.0	(0.6)
Station 1	*N*=871	5.7	(3.3 - 8.7)	[0.0 - 17.0]	6.1	+-3.4	(1.4)
Station 2	*N*=871	8.6	(5.4 - 11.8)	[0.0 - 20.0]	8.8	+-4.5	(1.5)
Station 3	*N*=871	15.0	(13.2 - 16.8)	[0.0 - 20.0]	15.0	+-2.6	(0.7)
Station 4	*N*=871	12.1	(9.6 - 14.3)	[0.0 - 18.6]	11.6	+-3.4	(1.0)
Station 5	*N*=871	13.7	(11.3 - 15.7)	[4.3 - 20.0]	13.6	+-3.1	(0.9)
OSCE-C18							
Overall (average)	*N*=863	13.1	(12.0 - 14.1)	[7.1 - 17.1]	13.1	+-1.6	(0.4)
Station 1	*N*=863	11.3	(9.3 - 13.7)	[2.7 - 19.0]	11.4	+-2.9	(0.9)
Station 2	*N*=863	14.3	(12.7 - 16.3)	[5.3 - 20.0]	14.2	+-2.8	(0.8)
Station 3	*N*=863	15.4	(13.5 - 16.9)	[5 - 20.0]	15.0	+-2.6	(0.7)
Station 4	*N*=863	12.3	(10.3 - 14.3)	[2.7 - 20.0]	12.2	+-3.0	(0.9)
Station 5	*N*=863	12.7	(10.7 - 14.7)	[4 - 20.0]	12.6	+-2.7	(0.8)
OSCE-C21							
Overall (average)	*N*=747	10.7	(9.4 - 12.0)	[0.0 - 16.2]	10.7	+-1.9	(0.6)
Station 1	*N*=747	11.5	(9.3 - 13.2)	[0.0 - 19.8]	11.3	+-3.1	(0.9)
Station 2	N=747	6.7	(5.0 - 9.5)	[0.0 - 15.8]	7.4	+-3.3	(1.2)
Station 3	*N*=747	10.8	(8.3 - 13.7)	[0.0 - 20.0]	10.7	+-3.9	(1.2)
Station 4	*N*=747	11.5	(9.5 - 13.7)	[0.0 - 19.0]	11.5	+-3.1	(0.9)
Station 5	*N*=747	12.7	(10.2 - 14.8)	[0.0 - 19.8]	12.5	+-3.2	(0.9)

Table 3 Distribution of mean scores given by each rater team for the 2022-2023 academic year OSCEs at Université Paris Cité medical school

Mean score given by each staff team	N	Median (of means)	(Interquartile)	[Min-Max]	Mean (of means)	+-SD	(SEM)
OSCE-C19-1							
Mean of each circuit	*N*=36	11.6	(11.2 - 12.1)	[10.2 - 13.4]	11.7	+-0.8	(0.2)
Station 1	*N*=36	9.5	(8.3 - 11)	[6.3 - 14.1]	9.7	+-1.8	(0.6)
Station 2	*N*=36	13.8	(13.4 - 14.6)	[11.6 - 15.9]	13.9	+-1.0	(0.3)
Station 3	*N*=36	9.1	(7.8 - 10.3)	[5.6 - 15]	9.1	+-1.9	(0.6)
Station 4	*N*=36	12.8	(12.2 - 13.6)	[10.5 - 15.3]	12.9	+-1.1	(0.3)
Station 5	*N*=36	13.1	(12.3 - 14)	[7.9 - 15.8]	12.9	+-1.7	(0.5)
OSCE-C19-2							
Mean of each circuit	*N*=36	11.1	(10.7 - 11.4)	[9.8 - 12.3]	11.0	+-0.6	(0.2)
Station 1	*N*=36	6.0	(5.0 - 6.8)	[3.2 - 9.8]	6.1	+-1.5	(0.6)
Station 2	*N*=36	8.8	(8.0 - 9.7)	[5.5 - 12]	8.8	+-1.4	(0.5)
Station 3	*N*=36	14.7	(14.4 - 15.6)	[14.0 - 17.1]	15.0	+-0.8	(0.2)
Station 4	*N*=36	11.6	(10.7 - 12.5)	[9.5 - 14]	11.6	+-1.1	(0.3)
Station 5	*N*=36	13.8	(12.9 - 14.4)	[11.1 - 15.3]	13.6	+-1.0	(0.3)
OSCE-C18							
Mean of each circuit	*N*=35	13.1	(12.7 - 13.5)	[11.9 - 13.9]	13.1	+-0.5	(0.1)
Station 1	*N*=35	11.4	(10.8 - 11.9)	[8.7 - 14.2]	11.4	+-1.1	(0.3)
Station 2	*N*=35	14.2	(13.5 - 15)	[11.4 - 16.6]	14.2	+-1.2	(0.3)
Station 3	*N*=35	14.8	(14.3 - 15.7)	[12.2 - 17.8]	15.0	+-1.3	(0.3)
Station 4	*N*=35	12.0	(11.4 - 12.9)	[10.0 - 14.4]	12.2	+-1.1	(0.3)
Station 5	*N*=35	12.5	(11.9 - 13.1)	[10.4 - 16]	12.6	+-1.1	(0.3)
OSCE-C21							
Mean of each circuit	*N*=20	10.8	(10.3 - 11.0)	[9.8 - 11.3]	10.7	+-0.5	(0.1)
Station 1	*N*=20	11.2	(10.8 - 11.7)	[9.3 - 13.1]	11.3	+-1.0	(0.3)
Station 2	*N*=20	7.2	(6.9 - 7.6)	[5.1 - 10.1]	7.4	+-1.1	(0.4)
Station 3	*N*=20	10.8	(9.7 - 11.9)	[8.8 - 12.5]	10.8	+-1.2	(0.4)
Station 4	*N*=20	11.5	(11.0 - 11.9)	[10.0 - 13.4]	11.5	+-0.9	(0.3)
Station 5	*N*=20	12.5	(12.0 - 12.8)	[10.9 - 15.9]	12.6	+-1.1	(0.3)

whereas approximately $2.0/\sqrt{25}=0.4$ was expected ($p<10^{-4}$ by ANOVA). Mixed models were used to disentangle station effects and compare this variability across sessions (using the fraction of staff-explained variance).

Adjustment for rater variability

We applied mixed models with random student, staff, and station effects to correct for staff variability, with or without equalizing station variance (and, therefore, contribution to score and ranking). The overall effect is apparent in Fig. 1, drawn separately for the C19 cohort and the C18 and C21 cohorts. The random-effect model shrinks student evaluations towards a common mean, apparently implicitly attributing part of students' extreme performances to staff effects and/or chance (Fig. 1). Moderate rank changes are apparent when using random effects to evaluate students. Much more scrambling of ranks is apparent when following the C19 class across the OSCE-C19-1 and OSCE-C19-2 sessions.

Each line represents one student, and color grading is based on score-based ranking (for class C19, the color remains based on OSCE-C19-1 to show the extent of stability across sessions). Therefore, if all methods give the same ranking, the gradient should remain unchanged ; the same is true if both C19 sessions give the same ranking. Abbreviations : ASC Average SCore, REU Random Effects with Unequal station contribution, REE Random Effects with Equal station contribution. Sessions are named as in Table 1.

Using these models, we estimated the proportion of variance explained by students, staff, and scripts. Results are shown in Fig. 2 for OSCE-C19-1, Supplementary Figure S1 for OSCE- C19-2, Supplementary Figure S2 for OSCE- C18, and Supplementary Figure S3 for OSCE-C21, in each case for models with or without variance

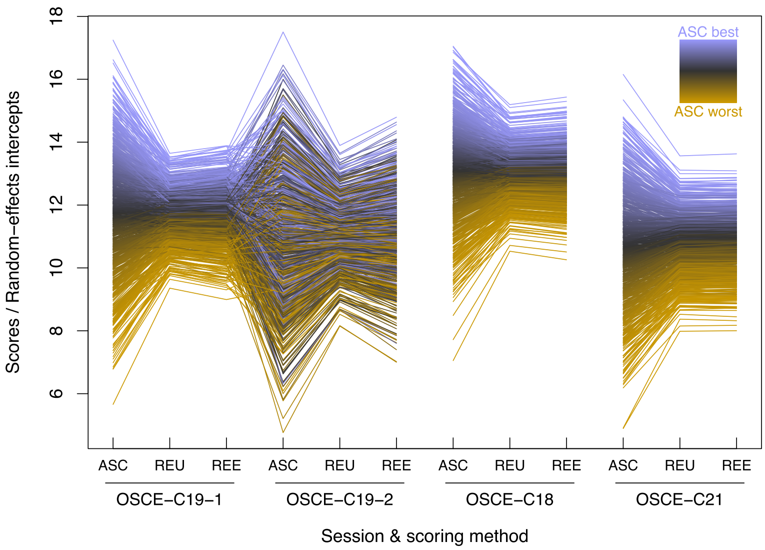

Fig. 1 Score average versus student random effects

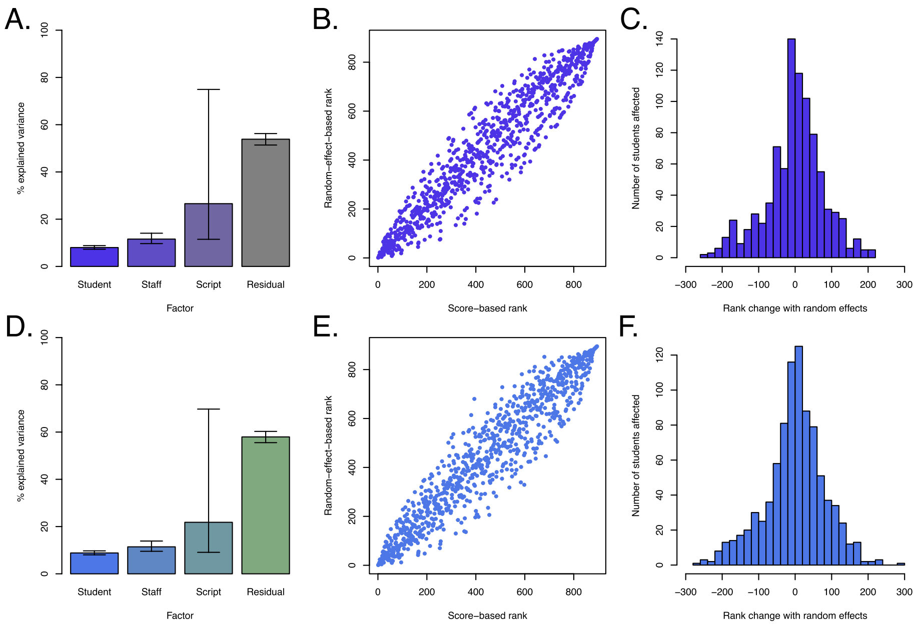

Fig. 2 Using linear mixed-effect models to grade OSCE-C19-1 students, versus score averages

equalization across stations to avoid differential impact of the stations on the final score. With equalized station contributions, staff variability explained 11.4% of score variance (95% confidence interval, 9.5 to 13.8) in OSCE-C19-1 (Fig. 2D, "Staff" bar), then a much lower 4.6% (3.8 to 5.8) in OSCE-C19-2 involving two raters (Supplementary Figure S1D, "Staff bar"), 11.6% (9.7 to 14.1) in OSCE-C18 (Supplementary Figure S2D, "Staff bar"), and again a lower 5.0% (3.9 to 6.6) in OSCE-C21 involving two raters (Supplementary Figure S3D, "Staff bar"), showing the moderating effect of the presence of two raters for OSCE-C19-2 and OSCE-C21. Student variability explained 8.8% (8.0 to 9.7) of score variance in OSCE-C19-1, 9.6% (8.7 to 10.5) in OSCE-C19-2, 11.3% (10.3 to 12.4) in OSCE-C18 and 12.5% (11.3 to 13.8) in OSCE-C21 (Fig. 2 and Supplementary Figures S1 to S3, "Student" bars on panel D for each one). Script effects explained a large amount of variance, which is unsurprising as scripts were more or less challenging, however this did not affect rankings since all students took all five stations. The residual variance was also high, indicating that many other unmeasured factors influence scores.

When using score averages versus student random effects to rank students, significant shifts in rankings are seen, more pronounced in the middle of the distribution (Fig. 2), indicating that the rankings of those students are more vulnerable to staff variability. In OSCE-C19-1, the average absolute student shifts in rankings upon rater adjustment with mixed models were 62 ranks (out of 881) for the variance-equalized models and 59 for the non-equalized. In OSCE- C19-2, the figures were lower, 39 and 30 (out of 871), respectively ($p < 10^{-15}$ for both comparisons with OSCE-C19-1 by the Wilcoxon signed rank paired test). This is as expected since staff-related variability was lower in OSCE-C19-2 (so removing a smaller variability had less effect on rankings). In OSCE-C18, the average absolute rank shifts were 50 (out of 863) with or without station variance equalization. In OSCE-C21, they were 36 and 37 (out of 747) respectively.

(A,B,C) Model built from standard station scores (A) Score variance attributable to each factor (student, staff, script, and residual unexplained variance), with 95% confidence intervals (B) Score average-based versus model-based rankings of students (C) Individual student ranking changes (out of 881 students) when using model-based instead of score-average-based rankings. (D,E,F) Model with variance-normalized station scores, so that each station contributes equally to student effects, panels are otherwise the same as (A,B,C).

Since the same class took the OSCE-C19-1 and OSCE-C19-2 exams six months apart, we evaluated whether ranking changes among students between the two sessions were smaller when ranking was based on random-effect intercepts rather than average scores, possibly better measuring intrinsic student skill. Figure 3 shows that ranking changes, which are comparable in magnitude between score averages and random effects, are more pronounced in the middle of the distribution rather than the tails and are much larger than those obtained when correcting for staff variability (as already suggested in Fig. 1). Absolute rank changes had a median of 179 ranks with score averages (interquartile 84-312, out of 864). With random-effects-based ranks, shifts were 176 (77-318) with unequal station variances (p = 0.22 versus ranks from average scores) and 176 (79-311) ($p = 0.45$) with equalized ones.

The same class C19 took OSCE-C19-1 and OSCE-C19-2, and their within-class rankings changed from one session to the other; these ranking changes are different depending on whether rankings are based on mixed models or on score averages. (A) Changes in rank from OSCE-C19-1 to OSCE-C19-2 using score averages (x-axis) or model-based random effects (y-axis). Colors, as shown in the gradient square, indicate the average rank across the two sessions. The green channel is for the model-based ranking and the blue channel for the score-average-based ranking. Bright cyan thus indicates high-performing students, black low-performing ones, green hue for better-performing students with the model, blue hue for better-performing students with score averages. The equalized station contribution is used. The identity line is also shown ; a different slope for the point cloud would indicate better stability of one or the other ranking metric across sessions. (B) Distribution of rank changes with score averages, much larger than those of Fig. 2 and Supplementary Figures S1, S2 and S3 for within-session shifts upon correction. Colors are arbitrary.

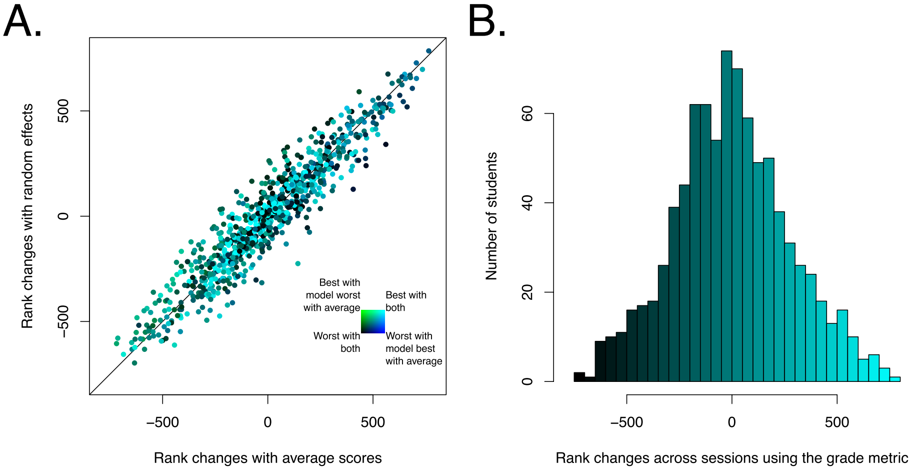

Fig. 3 Stability of model-based versus score-average-based rankings

Discussion

Our study confirms that dual assessment is an interesting strategy to reduce staff variability for OSCEs, and effectively allows to create a more reliable evaluation system. This practice contributes to greater fairness, ensuring that students are assessed based on their actual abilities rather than the idiosyncrasies of a small group of staff involved as standardized patients or raters. These results are in accordance with previously reported observations showing that better reliability was associated with a more significant number of stations and a higher number of raters per station [16]. On the other hand, they contrast with other studies that have assumed that between-rater differences are negligible [2] ; here, we show that staff effects can shift average scores by as much as students' effects themselves (around 11% explained variance for both, with the figure halved by having two raters). Our figure remains lower than what has been reported elsewhere [6, 7]. We suspect this might be due mostly to (i) being a single-institution, single-day setting with a homogenous examiner population and (ii) having a detailed scoring grid, with case-specific items that need to be verbalized by the student and leave less room for interpretation (but which may come at the expense of real-world relevance). The low student-attributable variance is in contrast with what has been observed in the multiple mini-interview evaluation format, where candidate ability has been found to explain about 20% of score variance, versus 13% for rater stringency/leniency [17]. Qualitatively, after our OSCE sessions, students reported fairness concerns, and their direct experiences were in line with our quantitative analysis. Given that OSCEs are expected to be used for competitive examination rankings in the future in Université Paris Cité, methods to correct staff heterogeneity were of interest.

While using two raters is simple (although resource-consuming), computing a rater effect is more challenging. To achieve this, we turned to the use of mixed models. Intuitively, mixed models are statistical models that simultaneously estimate large numbers of individual effects, avoiding inaccuracies of multiple estimations with comparatively scarce data by sharing information between similar quantities of interest (e.g., all "student effects" follow a common distribution, and ever more significant deviation from its mean are considered ever less unlikely, " staff effects" as well, and so on) [18]. This approach worked well, and showed a high degree of staff variability. In addition to individual training, given that staff was assigned to a single script each session, staff-script interactions are included in staff effects. They may merit consideration as a source of staff variability: despite standardized grids, some staff may be more or less qualified to evaluate some aspects of a given station. This constraint is inherent to OSCEs, as it is generally impossible (and perhaps not desirable) to staff them with professionals from the same specialty or background.

While we show that the mixed effect modeling of OSCEs is relatively straightforward, the student effect it computes does not seem more stable than an average score from one session to the next. Ideally, one should compare these two measures of student proficiency (average score versus random effect) to some gold standard for the concept of student skill, for example longitudinal follow-up of early careers, rather than their stability.

Another source of variability we measured is the station script, which explains a more significant fraction of total variance than students and staff. Although not problematic in itself, this emphasizes the need for consistency in assessment practices, and suggests it is not valid to compare students who take different exam sessions with each other, unless some normalization is applied. Considering differences in how stations are written, distinctions between procedural and non-procedural stations, presence or not of standardized patients, or "thinking out loud" station type, these factors can significantly affect student performance and the reliability of assessments. Thus, educators and institutions must strive for a uniform approach when designing station assessments to maintain equity in evaluating students.

Finally, in our study, uncharacterized residuals were the predominant source of variability. The student-script interaction component of these residuals can be interpreted as students entering medical schools with varying levels of competence and interest across different subjects, reflecting the multifaceted nature of the curriculum as well as differences in students' abilities and motives. The staff-student interaction component also merits consideration; qualitative studies would help to define communication styles and how they can be appropriate or inappropriate depending on different interlocutors, as these can affect how an OSCE unfolds [19–21].

In addition to double grading, which halved staff-attributable variance, our work points to possible strategies for quality control and improvement of OSCEs, mainly to improve their ability to assess station-independent student skills. To that end, techniques derived from psychometrics, using factor loading, could be used to analyze how stations themselves are written and enrich scripts and evaluation grids with items that better capture the consistent skills of students rather than variable effects, e.g., course subjects recently reviewed. A simple version of this would be to find questions that best correlate with the modelled random student effect, overall score, or subsequent performance and qualitatively find how they differ from others. Changing the test structure by making it open-book (i.e., with internet connection allowed and monitored) could also help eliminate sources of variability (i.e., ability for retrieval of rote memorization under pressure) and better capture real-world performance [20].

In conclusion, while healthcare students' assessment using OSCEs is widespread, it is essential to acknowledge the complexity of its reliability. We have therefore explored several factors that influence the fairness of OSCE, shedding light in particular on how dual assessment reduces staff variability issues.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12909-024-05803-6.

Supplementary Material 1.

Supplementary Material 2.

Supplementary Material 3.

Supplementary Material 4.

Acknowledgements
Université Paris Cité OSCE study group: Michael Thy, Anna Pellat, Gilles Soulat, Alexy Tran Dinh, Valentine Ferré, Nathan Peiffer Smadja, Michael Thy, Benjamin Deniau, Bénédicte Oules, Yann Nguyen, Lina Khider, Thibaud Soumagne, Augustin Gaudemer, Idir Ouzaid, Sophie Mazar, Jean Marc Liger, Eric Desrentes, Léonore Muller.

Authors' contributions
The authors would like to thank all Université Paris Cité staff and volunteers, whose large collective efforts made these sessions possible. Data collection: CT NB CL MF PR SB GM HH TM AF DB. Design: CT TM AF DB. Computations and first draft: SH. Data analysis: SH CT DB. Review: SH CT NB. Supervision and funding: CL PR TM AF DB.

Funding
None.

Availability of data and materials
The datasets used and/or analysed during the current study available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate
Our institutional IRB, the "Comité d'Evaluation de l'Ethiquedes projets de Recherche Biomédicale Paris Nord" (IRB00006477), waived the need for ethical review.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Author details
[1]UFR de médecine, Université Paris Cité, Paris, France. [2]Université Paris Cité, INSERM UMR1137, IAME, F-75018 Paris, France. [3]Epidemiology, Biostatistics and Clinical Research Department, Bichat-Claude Bernard Hospital, APHP, Paris, France. [4]Université Paris Cité, INSERM UMR1149, CRI, F-75018 Paris, France. [5]Anesthesia and Intensive Care Department, Louis Mourier Hospital, APHP, Colombes, France. [6]Université Paris Cité, INSERM UMR970, Paris Cardiovascular Research Center (PARCC), Paris, France. [7]Oncology Genetics Department, Fédération de Génétique et de Médecine Génomique, Georges-Pompidou European Hospital, APHP, Paris, France. [8]Université Paris Cité and Université Sorbonne Paris Nord, Inserm, INRAE, Center for Research in Epidemiology and StatisticS (CRESS), Paris, France. [9]Psychiatry Department, Hôtel-Dieu Hospital, APHP, Paris, France. [10]Renal Physiology Department, Bichat-Claude Bernard Hospital, APHP, Paris, France. [11]Gastroenterology and Pancreatology Department, Beaujon Hospital, APHP, Clichy, France. [12]Arterial Hypertension Department, Georges-Pompidou European Hospital, APHP, Paris, France. [13]Pediatrics Department, Robert Debré Hospital, APHP, Paris, France. [14]Université Paris Cité, INSERM UMRS1123, ECEVE, F-75010 Paris, France. [15]Emergency Department, Bichat-Claude Bernard Hospital, APHP, Paris, France.

Received: 22 March 2024 Accepted: 22 July 2024
Published online: 29 July 2024

References

1. Barman A. Critiques on the objective structured clinical examination. Ann Acad Med Singap. 2005;34(8):478–82.
2. Trejo-Mejía JA, Sánchez-Mendiola M, Méndez-Ramírez I, Martínez-González A. Reliability analysis of the objective structured clinical examination using generalizability theory. Med Educ Online. 2016;21:31650.
3. Faherty A, Counihan T, Kropmans T, Finn Y. Inter-rater reliability in clinical assessments: do examiner pairings influence candidate ratings? BMC Med Educ. 2020;20(1):147.
4. Mortsiefer A, Karger A, Rotthoff T, Raski B, Pentzek M. Examiner characteristics and interrater reliability in a communication OSCE. Patient Educ Counsel. 2017;100(6):1230–4.
5. Saal FE, Downey RG, Lahey MA. Rating the ratings: assessing the psychometric quality of rating data. Psychol Bullet. 1980;88(2):413–28.
6. Homer M. Pass/fail decisions and standards: the impact of differential examiner stringency on OSCE outcomes. Adv Health Sci Educ Theory Pract. 2022;27(2):457–73.
7. Homer M. Towards a more nuanced conceptualisation of differential examiner stringency in OSCEs. Adv Health Sci Educ Theory Pract. 2023.
8. Yeates P, O'Neill P, Mann K, Eva K. Seeing the same thing differently: mechanisms that contribute to assessor differences in directly-observed performance assessments. Adv Health Sci Educ Theory Pract. 2013;18(3):325–41.
9. Fuller R, Homer M, Pell G. Longitudinal interrelationships of OSCE station level analyses, quality improvement and overall reliability. Med Teach. 2013;35(6):515–7.
10. Selby C, Osman L, Davis M, Lee M. Set up and run an objective structured clinical exam. BMJ. 1995;310(6988):1187–90.
11. Brown C, Ross S, Cleland J, Walsh K. Money makes the (medical assessment) world go round: The cost of components of a summative final year Objective Structured Clinical Examination (OSCE). Med Teach. 2015;37(7):653–9.
12. Department for Business, Innovation & Skills, UK government. GOV.UK. 2012 [cited 2023 Nov 8]. National Minimum Wage to rise from 1 October 2012. Available from:https://www.gov.uk/government/news/national-minimum-wage-to-rise-from-1-october-2012.
13. Dong T, Saguil A, Artino AR, Gilliland WR, Waechter DM, Lopreaito J, et al. Relationship between OSCE scores and other typical medical school performance indicators: a 5-year cohort study. Mil Med. 2012;177(9 Suppl):44–6.
14. Bouzid D, Mullaert J, Ghazali A, Ferré VM, Mentré F, Lemogne C, et al. eOSCE stations live versus remote evaluation and scores variability. BMC Med Educ. 2022;22(1):861.
15. Song H, Penn State University. Stat 415 syllabus, Lesson 4: Confidence Intervals for Variances. 2021. Cited 2023 Nov 8. Available from:https://online.stat.psu.edu/stat415/book/export/html/810.
16. Brannick MT, Erol-Korkmaz HT, Prewett M. A systematic review of the reliability of objective structured clinical examination scores. Med Educ. 2011;45(12):1181–9.
17. Roberts C, Rothnie I, Zoanetti N, Crossley J. Should candidate scores be adjusted for interviewer stringency or leniency in the multiple mini-interview? Med Educ. 2010;44(7):690–8.
18. Theobald E. Students are rarely independent: when, why, and how to use random effects in discipline-based education research. CBE Life Sci Educ. 2018;17(3):rm2.

19. Blanch DC, Hall JA, Roter DL, Frankel RM. Medical student gender and issues of confidence. Patient Educ Couns. 2008;72(3):374–81.
20. Ibrahim NK, Al-Sharabi BM, Al-Asiri RA, Alotaibi NA, Al-Husaini WI, Al-Khajah HA, et al. Perceptions of clinical years' medical students and interns towards assessment methods used in King Abdulaziz University. Jeddah Pak J Med Sci. 2015;31(4):757–62.
21. Jefferies A, Simmons B, Regehr G. The effect of candidate familiarity on examiner OSCE scores. Med Educ. 2007;41(9):888–91.

Publisher's Note