3-1-2020

Impressions on Reliability and Students’ Perceptions of Learning in a Peer-Based OSCE

Rishad Khan
Schulich School of Medicine & Dentistry

Saad Chahine
Schulich School of Medicine & Dentistry

Steven Macaluso
Schulich School of Medicine & Dentistry

Ricardo Viana
Schulich School of Medicine & Dentistry

Caitlin Cassidy
Schulich School of Medicine & Dentistry, caitlin.cassidy@sjhc.london.on.ca

See next page for additional authors

Authors
Rishad Khan, Saad Chahine, Steven Macaluso, Ricardo Viana, Caitlin Cassidy, Thomas Miller, Debra Bartley, and Michael Payne

Medical Science Educator (2020) 30:429–437
https://doi.org/10.1007/s40670-020-00923-2

ORIGINAL RESEARCH

Impressions on Reliability and Students' Perceptions of Learning in a Peer-Based OSCE

Rishad Khan[1] · Saad Chahine[2] · Steven Macaluso[3] · Ricardo Viana[3] · Caitlin Cassidy[3] · Thomas Miller[3] · Debra Bartley[4,5] · Michael Payne[3]

Published online: 18 February 2020

Abstract

Background Peer assessment of performance in the objective structured clinical examination (OSCE) is emerging as a learning instrument. While peers can provide reliable scores, there may be a trade-off with students' learning. The purpose of this study is to evaluate a peer-based OSCE as a viable assessment instrument and its potential to promote learning and explore the interplay between these two roles.

Methods A total of 334 medical students completed an 11-station OSCE from 2015 to 2016. Each station had 1–2 peer examiners (PE) and one faculty examiner (FE). Examinees were rated on a 7-point scale across 5 dimensions: Look, Feel, Move, Special Tests and Global Impression. Students participated in voluntary focus groups in 2016 to provide qualitative feedback on the OSCE. Authors analysed assessment data and transcripts of focus group discussions.

Results Overall, PE awarded higher ratings compared with FE, sources of variance were similar across 2 years with unique variance consistently being the largest source, and reliability (r_{φ}) was generally low. Focus group analysis revealed four themes: Conferring with Faculty Examiners, Difficulty Rating Peers, Insider Knowledge, and Observing and Scoring.

Conclusions While peer assessment was not reliable for evaluating OSCE performance, PE's perceived that it was beneficial for their learning. Insight gained into exam technique and self-appraisal of skills allows students to understand expectations in clinical situations and plan approaches to self-assessment of competence.

Keywords Objective structured clinical examination · Peer assessment · Reliability

Introduction

The objective structured clinical examination (OSCE) is ubiquitous in medical training and assessment [1]. Widely adopted as a method of reliable assessment, its value as a learning tool is emerging with peers as assessors [2]. This paper investigates the scoring variation of a peer examiner (PE) in an OSCE compared with a faculty examiner (FE) and contrasts it to the learning gains identified by students.

The notion of a trade-off between learning and psychometric precision has been previously described [3]. Theoretically, this trade-off is based on the concept of fit for purpose in assessment [4–8]. There may be various purposes for a specific assessment, including gatekeeping, accountability, and learning [9]. These varied purposes may be distilled down to two contrasting goals: (1) assessment of learning (generate valid assessment data on a learner); and (2) assessment for learning (promote a trainee's development) [2, 10–12]. Assessment of learning has been described as summative

✉ Rishad Khan
rkhan2019@meds.uwo.ca

[1] Department of Medicine, Schulich School of Medicine and Dentistry, Western University, 1151 Richmond Street North, London, ON N6A 3K7, Canada

[2] Centre for Education Research and Innovation, Schulich School of Medicine and Dentistry, Western University, 1151 Richmond Street North, London, ON N6A 3K7, Canada

[3] Department of Physical Medicine and Rehabilitation, Schulich School of Medicine and Dentistry, Western University, London, ON N6A 3K7, Canada

[4] Department of Surgery, Schulich School of Medicine and Dentistry, Western University, 1151 Richmond Street North, London, ON N6A 3K7, Canada

[5] Department of Paediatrics, Schulich School of Medicine and Dentistry, Western University, 1151 Richmond Street North, London, ON N6A 3K7, Canada

assessment, or scenarios where a learner's level of knowledge, skills and competency is assigned a grade or rating. When the OSCE has is used for purpose, a great amount of effort is taken to ensure a high level of psychometric precision.

Conversely, when the purpose of assessment is for learning, different approaches are used. For example, practice-tests, peer assessment, or simulated scenarios with targeted feedback and debriefing are intended to be safe learning environments to identify deficits in learning [13–15]. With this purpose in mind, there is less of an emphasis on accurate judgement of skills and more on maximizing learning [16]. More recently, the concept of assessment as learning has been described as a specific kind of assessment for learning, with an emphasis on self-appraisal and metacognition [17]. In this practice, there is often little to separate what is considered assessment and what is considered learning. In both assessments for, and as learning, psychometric precision is less valued in comparison to the learning gains that can be achieved when trainees are able to identify deficits in learning and engage in self-appraisal respectively.

From the perspective of the trainee, the distinction between activities that promote learning (assessment for learning) and those that test what has been learned (assessment of learning) is important. For example, trainees may receive less feedback if they do not disclose a learning deficit for fear of a poor summative rating. This distinction, however, is not always clear, and there may be overlap between assessment of learning and assessment for learning. One scenario where this overlap is important is a peer-based OSCE. In this setting, the trade-off between scoring precision and learning gains is unclear [18–22]. We sought to address this gap through the following two research questions:

1. How do second-year medical students with no prior training in assessment compare with faculty examiners, with respect to score variation, reliability and leniency, when evaluating their peers in a formative OSCE?
2. What are the perceptions of medical students on assessing and being assessed by their classmates in this setting with respect to learning gains?

Methods

This was a sequential mixed methods study conducted from 2015 to 2016 at the Schulich School of Medicine and Dentistry at Western University [23]. In part one, we developed, piloted and implemented a peer-based OSCE over 2 years. In part two, we held focus groups with medical students. Approval was granted by the Western University Research Ethics Board (REB: 106210). All authors reviewed and approved the final manuscript. No changes to the protocol were made after study commencement.

OSCE Design

Students completed an 11-station OSCE as a part of the musculoskeletal (MSK) course at Western University. The same stations and different faculty examiners were used in 2 years. Students were placed into groups of 2–3 and rotated around the 11 stations as a group. At each station, one student was assessed, and the others acted as examiners. The rotation of roles was random. This setup reflected a resource-constrained setting, where it was not feasible to assess each student at each station. Each station was comprised of a clinical scenario related to MSK medicine, with examinees rated on a 7-point scale across 5 dimensions: Look, Feel, Move, Special Tests, and Global Impression. A description of each of these components is listed in Table 1.

Focus Groups

Students participated in focus group interviews on a voluntary basis after the second administration of the OSCE, in 2016. The focus group questions were semi-structured and intended to generate conversation about the experience of participating in the OSCE. The initial question asked whether studies found the OSCE to be beneficial for learning and if so, why. Follow-up questions were guided by focus group participants' responses to the initial question. Some examples include asking what students thought about being assessed by their peers, what they thought they learned from watching their peers complete an OSCE station, and the value of watching faculty give feedback to their peers. The interviews lasted for approximately 1 h.

Data Analysis

The analysis was conducted in two parts. In part 1, all the quantitative data were represented using descriptive statistics and analysed using multivariate analysis of variance (MANOVA). Follow-up generalizability theory (G-theory) analysis was conducted to identify sources of variation and overall consistency of the OSCE. G-theory provides results that are similar to reliability analysis of internal consistency used in multiple choice testing. The final G-coefficient is on a scale of 0–1, where 1 would represent perfect reliability and 0 represents complete randomness. While there is no strict threshold, the typical expectation is 0.60–0.70 for formative assessments and 0.80–1.0 for high stakes assessments [24]. The advantage of G-theory is that the analysis partitions the variance associated with different facets (i.e. examinees, examiners, stations) of the assessment [25]. A greater amount of variation that is associated with the examinees, and less

Table 1 Descriptions of the 5 components of the musculoskeletal examination: Look, Feel, Move, Special Tests, Global Impressions

Look	Look for attitude, swelling, deformity, muscle wasting, and skin changes, at rest and during movements [42].
Feel	Feel for tenderness, swelling, deformity and crepitus with movement and temperature [42].
Move	Move the joint actively, then passively, assessing range of motion and correlating findings from Look and Feel [42]..
Special Tests	There are a range of special tests to further characterize the particular problem, such as tests for shoulder pain [42].
Global Impression	Overall impression of the student's performance and calibrates the students' score for aspects that may not be accounted for by other more discrete components of the assessment [43].

associated with the examiners or the individual OSCE stations, would result in a higher G-coefficient.

There are several different G-theory designs [26]. When examinees are assessed on the same components, this is called a *crossed* design. Conversely, *nested* designs are when examinees are assessed on parts of a component in one set and another part in a different set. In an OSCE setting, different examiners assess examinees, and thus examiners are *nested* within examinees. However, when the number of examiners varies, that is, a different number of examiners are assigned to each examinee, it is considered an *unbalanced* design. In this OSCE, all of the students were assessed using the same dimensions. However, as this is a peer-based OSCE in which some students were in groups of two, and some in groups of three, not all examinees received a the same number of assessments for all stations. From a G-theory perspective, this is considered an *unbalanced* deign. To conduct the analysis, a variance components procedure was used with restricted maximum likelihood estimation (RMLE) to account for the unbalanced design [27]. Three facets were included in the analysis: (1) student, (2) station and (3) examiner type. Each dimension (Look, Feel, Move, Special Test, Global) and year (2015, 2016) were considered separately. Additionally, G-coefficients were calculated for each dimension in each year. Data were analysed using SPSS 24 (IBM, Armonk).

In phase 2, the analysis of the focus group interviews was completed in two steps. First, members of the research team (SC, SM) read the transcripts and thematically coded snippets of text. Second, they met to refine the thematic codes, and one author (RK) recoded the data using the revised themes. The most representative snippets were selected for this paper. Two researchers (RK and SC) agreed on the snippets that provided the richest description of each theme and be included.

Results

A total of 334 students participated in the OSCE. In 2015, 175 participants were assessed and received 949 peer assessments and 360 faculty assessments. In 2016, 159 participants were assessed and received 655 peer assessments and 388 faculty assessments. In 2016, 23 students participated in five separate focus groups.

Peer Vs. Faculty Ratings

In 2015, average PE scores across stations were between 5 and 6 on the 7-point scale (Look = 5.72, Feel = 5.07, Move = 5.78, Spec Test = 4.73, Global = 5.96), while average FE scores were between 4 and 5 (Look = 5.10, Feel = 4.67, Move = 5.00, Spec Test = 4.04, Global = 5.25). In 2016, average PE scores were between 5 and 6 (Look = 5.77, Feel = 5.33, Move = 5.85, Spec Test = 5.03, Global = 6.02) while average FE scores were between 4 and 5 (Look = 5.36, Feel = 4.99, Move = 4.98, Spec Test = 4.34, Global = 5.33). Peer and faculty scores are represented in Fig. 1.

Across all stations and dimensions over the 2 years on average there was a ~ 0.63 point difference between PE and FE scores. MANOVA was conducted to confirm differences between examiner type and stations with a stringent p value of 0.01 and conservatives criterion (Pillai's trace) as the data failed to meet assumptions of homogeneity [28]. There was a significant difference in scores awarded between faculty and peers (2015: $F(5,175) = 33.025$, $p < 0.001$; 2016: $F(5,159) = 36.05$, $p < 0.001$), between stations (2015: $F(5,175) = 8.55$, $p < 0.001$; 2016: $F(50,159) = 9.73$, $p < 0.001$), and in the interaction of examiner type and station (2015: $F(5,175) = 2.41$, $p < 0.001$; 2016: $F(50, 159) = 2.32$, $p < 0.001$).

Generalizability Theory Results

The G-theory analysis identified that the major source of error as unique variance, which represents the variation from sources not captured systematically as a part of the analysis, in 2015 (39.89–50.85%) and 2016 (32.74–49.59%) [29]. Another main facet of variation was whether the examiner was an FE or SE, for Look (14.18%), Move (15.49%) and Global Impression (20.29%) in 2015, and for Move (14.12%) and Global Impression (15.38%) in 2016. This suggests that FE and SEs are less comparable in their scoring. Finally, a key source of variation based on student by station interaction, for Look (15.15%), Feel (17.72%) and Special Tests (13.83%) in 2015, and for Look (14.05%), Feel (13.73%) and Special Tests (22.87%) in 2016. The variances associated with each facet of the assessment were similar when comparing 2015 and 2016 data for all components of

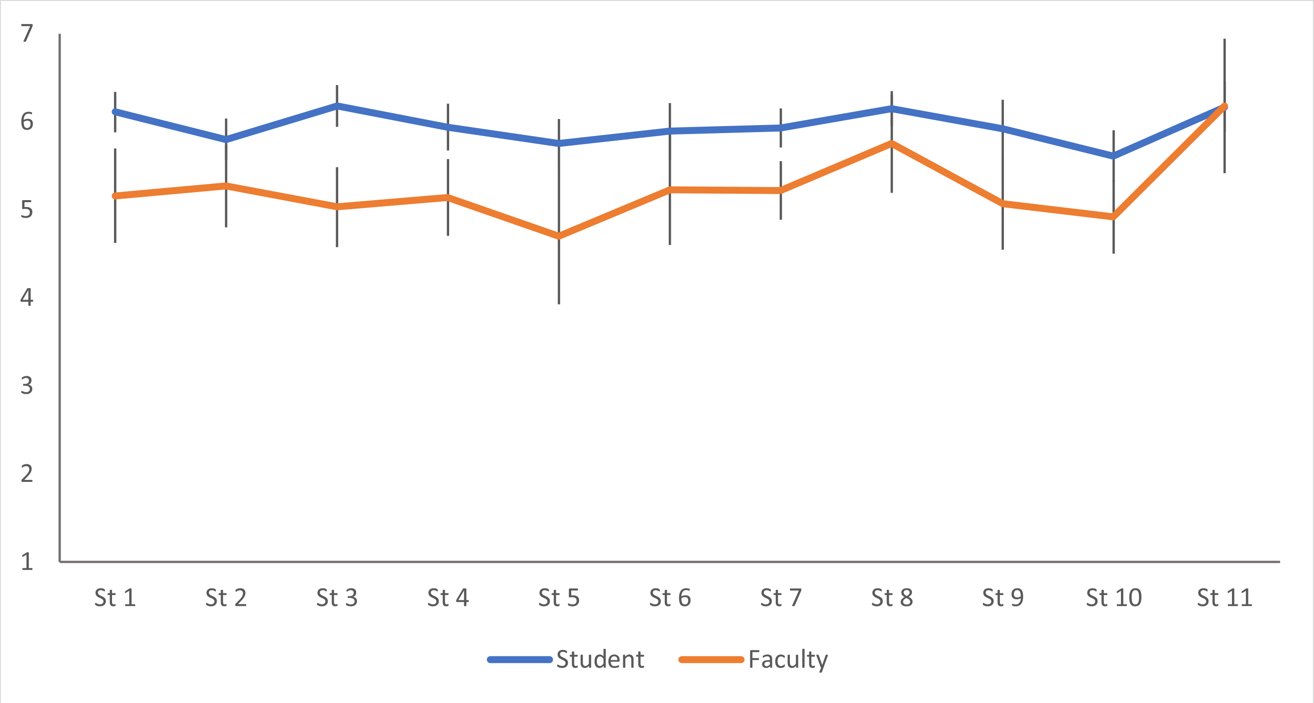

Fig. 1 Mean student and faculty scores with 99% confidence intervals for the Global Impression dimension of the OSCE examination from 2015. The minimum Global Impression score was 1, and the maximum was 7. *St* station

the MSK examination, except for Look. For the Look domain, the examiner type was a much larger source of variance in 2015 (14.05%) than in 2016 (4.96%), while OSCE station was a much smaller source of variance (0.65% in 2015, 6.61% in 2016). All the variances associated with each facet and interaction are summarized in Table 2 and Fig. 2.

Overall the G-coefficients of the dimensions (r_{φ}) were low in 2015, ranging from 0.40 (Look) to 0.66 (Move) and low to moderate in 2016, ranging from 0.39 (Look) to 0.72 (Global Impression). All reliability data are summarized in Table 3.

Focus Groups

Twenty-three students attended the focus groups after the 2016 OSCEs. Several themes arose from the discussions that emphasize students' perceptions about this experience.

Theme 1 Conferring with Faculty Examiners

Watching faculty interact with examinees post-station and discussing the ratings aided in conceptualizing the Look, Feel, Move, and Special Tests of MSK medicine.

> "You got feedback on what you were doing, and there was, again, like we were talking about how it would be cool to go over the entire exam, this kind of was that. If we go over the exam of the hand and try to rule out different types of arthritis, things like that. So I really liked the kind of instant feedback"

Students specifically valued the feedback with the Feel component, which is reliant on tactile knowledge of musculoskeletal anatomy, and Special Tests, which are often complex.

> "What I really liked was feedback on technique, and on palpation technique"
> "The reason why it is nice to have them practice with an examiner in the room is because they (special tests) are easy to mess up in the technique"

Theme 2 Difficulty rating peers

Some students found it difficult marking and rating peers without any prior instruction or knowledge of the rating instruments. They especially struggled with ordinal scales of general performance in a given domain.

> "I don't know what the arbitrary scale of 1 to 7 means."

Some students were uncomfortable giving peers a "satisfactory" rating. They perceived the word satisfactory as meaning a borderline pass, and thought their peers did better than just enough to pass.

Table 2 Variance component analysis for OSCE scores in 2015 and 2016

2015										
	Look		Feel		Move		Special Tests		Global	
Source of variation	Variance component	% of variance	Variance component	% of variance	Variance component	% of variance	Variance component	% of variance	Variance component	% of variance
Student	0.07	5.64	0.24	12.41	0.17	9.08	0.35	12.87	0.15	10.70
Examiner type	0.18	14.18	0.06	3.17	0.30	15.49	0.25	9.28	0.28	20.29
Station	0.01	0.65	0.08	3.93	0.24	12.47	0.38	13.90	0.02	1.32
Student X examiner type	0.11	8.86	0.15	7.81	0.04	2.30	0.19	6.95	0.09	6.52
Student X station	0.19	15.15	0.35	17.72	0.18	9.13	0.37	13.83	0.16	11.50
Examiner type X station	0.06	4.67	0.13	6.59	0.04	2.09	0.09	3.29	0.00	0.22
Student X examiner type X station, error	0.63	50.85	0.95	48.37	0.95	49.45	1.08	39.89	0.68	49.45
2016										
	Look		Feel		Move		Special Tests		Global	
Source of variation	Variance component	% of variance	Variance component	% of variance	Variance component	% of variance	Variance component	% of variance	Variance component	% of variance
Student	0.07	5.79	0.19	7.45	0.10	3.92	0.19	8.52	0.21	14.69
Examiner type	0.06	4.96	0.05	1.96	0.36	14.12	0.18	8.07	0.22	15.38
Station	0.08	6.61	0.18	7.06	0.15	5.88	0.38	17.04	0.11	7.69
Student X examiner type	0.12	9.92	0.15	5.88	0.00	0.00	0.18	8.07	0.08	5.59
Student X station	0.17	14.05	0.35	13.73	0.19	7.45	0.51	22.87	0.08	5.59
Examiner type X station	0.11	9.09	0.08	3.14	0.06	2.35	0.06	2.69	0.07	4.90
Student X examiner type X station, error	0.60	49.59	0.91	35.69	1.15	45.10	0.73	32.74	0.66	46.15

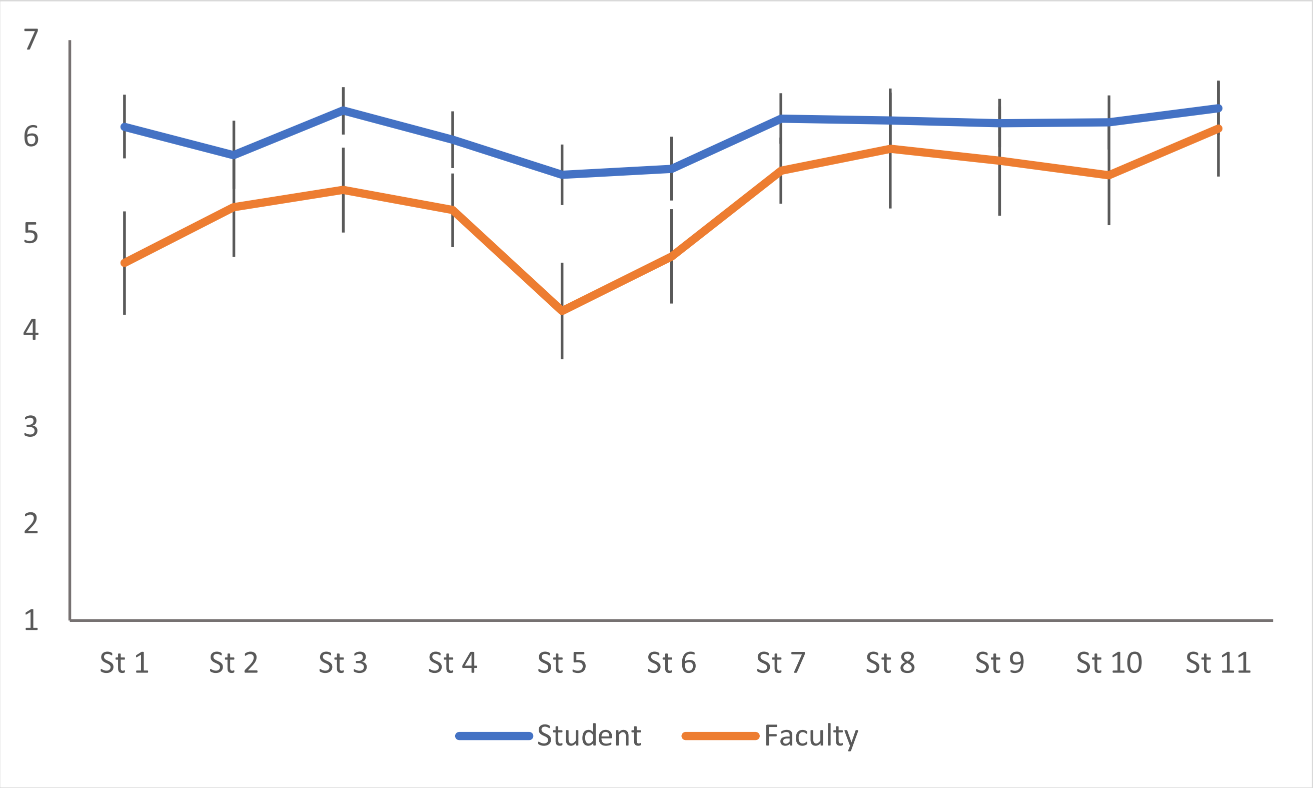

Fig. 2 Mean student and faculty scores with 99% confidence intervals for the Global Impression dimension of the OSCE examination from 2016. The minimum Global Impression score was 1, and the maximum was 7. *St* station

> "I didn't want to circle satisfactory, because I thought the person did good or great, but exceptional seems like you needed to have perfect. So there was kind of a large gap there, so sometimes I would circle both."

Theme 3 Insider Knowledge

Insider knowledge refers to the insight students can gain when they are on the assessor side of an evaluation. In this scenario, students reported that seeing the score sheet used by faculty provided them with understanding of examiners' priorities.

> "Something else I want to thank PCCM for doing was actually showing us outlines of what an OSCE rubric might look like."
> "We got to see the evaluation sheet, whereas in the OSCEs we never actually kind of see what's on that checklist, so we don't really know, say for the cardio OSCE, we are not really sure what is the perfect score, what things we're supposed to hit, what's emphasized. Whereas today we got a chance to actually glimpse at, for rheumatoid arthritis exam on the hand, for example, what are the things they want you to pick up on the exam? So it was useful to see."

Table 3 Reliability coefficients (G-coefficients) for scores from the five dimensions of the OSCE examination (Look, Feel, Move, Special Tests, Global Impression) in 2015 and 2016

2015		2016	
Look	0.40	Look	0.39
Feel	0.61	Feel	0.56
Move	0.66	Move	0.58
Spec	0.65	Spec	0.52
Global	0.62	Global	0.72

Attached separately

Theme 4 Observing and Scoring

Students reported learning when they were observing and assessing their peers, even though they were not performing in the station themselves.

> "So even if we were not doing the tests at the stations we would still learn what was expected at the station, and it is just kind of a review for us as well, just marking"

This corresponded to students realizing, from observation of their peers, where their own deficiencies lay with respect to MSK examination.

> "But then having now done the OSCEs week 4, I realized I didn't learn enough about the hand, and that I needed to learn a lot more about the deformities, specific deformities and then maybe spend more time practicing, feeling for different things"

Discussion

Conceptions of assessment for learning and assessment of learning may conflict, where one is focused on promoting learning and the other on ensuring reliability. Our study explored the trade-off between these concepts. Overall, PE awarded higher ratings compared with FE and sources of variance were similar across 2 years. Unique variance was the largest source, indicating that a large portion of the variation was due to factors not captured in our analysis. Other major sources of variance were examiner type and content specificity. Reliability (r_{φ}) was generally low, though it was moderate to high in select dimensions. Focus group analysis revealed four themes which captured students' perceptions of this OSCE: Conferring with Faculty Examiners, Difficulty Rating Peers, Insider Knowledge, and Observing and Scoring. These themes highlight participants' increased awareness of the assessment process and gaps in their own learning. Our discussion highlights the tenuous relationship between learning and psychometric accuracy in assessment using the four themes highlighted in the qualitative analysis.

In our data, an important source of variance was the examiner type. This means that the range of scores students received can be partially explained by whether a rating was from a PE or FE. In high stakes settings, minimal variance due to examiner type is desired to provide an objective assessment of a student. We contrast the variance in scores with students' perception of learning from faculty and identifying gaps in their own learning, emphasised in the theme Conferring with Faculty Examiners. Here, peers reported that seeing faculty interact with examinees post-station allowed them to conceptualize the five components of the examination, specifically the more complex Feel and Special Tests. It is possible that variance due to examiner type may have been smaller if faculty provided feedback during, rather than after each station, as this would have allowed PEs to glean more information on what the examinee did or did not do well. Providing concurrent rather than terminal feedback however may have been detrimental for examinees, as terminal feedback has been shown to result in better learning [30]. For the examinee, concurrent feedback and task correction may detract from learning of automaticity when a student is trying to build fluency in the physical examination [31]. Examinees who complete the uninterrupted may also experience more anxiety and apprehension, which can improve performance on high stakes examinations for some students [32–34]. Feedback may therefore be better if provided at the end of the task for the examinee [35]. It is unclear, however, if terminal feedback is beneficial for peer examiners, as they must wait until the end of the exam to appreciate what faculty are thinking and do not reap the potential benefits of the apprehension and anxiety that examinees experience.

The second focus group theme was Difficulty Rating Peers. Students felt that this difficulty was due to (1) ambiguity surrounding Likert scale ratings, and (2) discomfort awarding peers a rating which was perceived as "borderline". Previous studies in peer-based OSCEs have highlighted a similar theme, with students not wishing to be overly critical of their classmates [36, 37]. The difficulty rating peers manifested as inaccurate ratings and mainly low to moderate generalizability ratings, with only one domain having high generalizability (Global Impression) in 2016. Other peer-based OSCE studies, which trained PE, have reported PE awarding similar, and in some cases, lower scores compared with FE. Another previous study has shown that OSCE examiners who undergo training are more consistent in their behaviours while rating students compared with untrained ones [2, 38]. Training peers in our study may have ameliorated confusion around rating scales and yielded more reliable scores. While we found no literature comparing the perceived learning gains of trained versus untrained peers, untrained assessors in our study perceived learning about awarding scores, rubrics and watching faculty assess their peers during the OSCE as beneficial.

As highlighted in the Insider Knowledge theme, student examiners had the opportunity to view scoring rubrics. Practices such as showing students rating scales encourage them to understand expectations as they develop their clinical skills [16]. This theme, in addition to seeing what faculty physicians perceived as important components of each OSCE station, highlights how students perceived learning what was important both for summative examinations and during clinical practice. In addition, the Observing and Scoring theme highlights that PE's learning went beyond what was gleaned from a rubric. Marking their peers and being an active part of the assessment process, as opposed to passively watching as an observer, provided a review of that station for students. While watching others, students perceived identifying their own deficiencies, a key strategy in self-appraisal of competence [16]. Showing learners rating scales and involving them in the assessment process are effective strategies to implement assessment for learning (AfL) [39], defined as "a part of everyday practice by students, teachers and peers that seeks, reflects upon, and responds to information from dialog, demonstration, and observation in ways that enhance ongoing learning" [40]. Though our PE generally awarded higher scores to FE, our OSCE format allowed for the application of AfL. Components of AfL, such as feedback and questioning, have been credited with having large effect sizes in learning [41].

This study has several limitations. First, focus groups were only carried out in 2016 due to budgetary constraints. Students from the 2015 cohort may have had different opinions than the 2016 cohort. We did, however, feel comfortable applying focus group findings from 2016 to quantitative data sets from both years, as sources of variance were similar between 2015 and 2016. Second, participation in focus groups was voluntary, which may have predisposed findings to being overly optimistic from enthusiastic students. Bias from voluntary participation may be further exacerbated by the fact that a very small percentage of all students, who may not be representative of the entire group, participated in the focus groups. It would have been unadvisable however, to mandate focus group participation as this may have yielded inauthentic responses. Third, we did not ask participants to rate the quality of feedback from peers versus faculty, data which could have helped to clarify the role of peers in assessment for learning. Finally, due to the nature of our OSCE, not all students completed all 11 stations, and thus each students' experience was not standardized.

Future research should be aimed at further clarifying the relationship between reliability and learning in peer assessment. This can be done through randomisation, with one arm of students receiving training and another arm remaining untrained prior to rating peers in an OSCE. PE score reliability and perceived impacts on learning can be compared between the groups. This design can minimize bias through its randomized design and deliver further insight into the reliability-learning trade-off. Additionally, future research can explore whether participating as a peer assessor impacts one's performance on an OSCE.

Conclusion

As the competency-based medical education paradigm encourages further student involvement in assessment practices, there may be a growing interest in peer assessment. In our study, students did not provide reliable or accurate ratings of their peers on an MSK-based OSCE. They did, however, perceive gaining insight into exam technique and a better understanding of gaps in their own knowledge and skills. This may allow students to understand expectations in clinical situations and plan approaches to self-assessment of competence.

Funding Information This project was financially supported by the Schulich Research Opportunities Program.

Compliance with Ethical Standards

Conflict of Interest The authors declare that they have no conflict of interest.

Ethical Approval Ethical approval for this study was granted by the Western University Research Ethics Board (REB: 106210). This research was performed in accordance with the ethical standards as laid down in the 1964 Declaration of Helsinki and its later amendments.

Informed Consent Informed consent was obtained from all individual participants included in the study.

References

1. Carraccio C, Englander R. The objective structured clinical examination: a step in the direction of competency-based evaluation. Arch Pediatr Adolesc Med. 2000;154:736–41.
2. Khan R, Payne MWC, Chahine S. Peer assessment in the objective structured clinical examination: a scoping review. Med Teach. 2017;39:745–56.
3. Wilson M. Towards coherence between classroom assessment and accountability: University of Chicago Press; 2004.
4. Mislevy RJ, Haertel G, Riconscente M, Rutstein DW, Ziker C. Evidence-centered assessment design. Assessing model-based reasoning using evidence-centered design. Berlin: Springer; 2017. p. 19–24.
5. Mislevy RJ, Haertel GD. Implications of evidence-centered design for educational testing. Educational measurement: issues and practice, vol. 25. Hoboken: Wiley Online Library; 2006. p. 6–20.
6. Shepard LA. The role of assessment in a learning culture. Educational researcher, vol. 29. Thousand Oaks: Sage Publications Sage CA; 2000. p. 4–14.
7. Earl L. Assessment as learning. The keys to effective schools: educational reform as continuous improvement; 2007. p. 85–98.
8. Schuwirth LWT, Van der Vleuten CPM. Programmatic assessment: from assessment of learning to assessment for learning. Med Teach. 2011;33:478–85.
9. Nagy P. The three roles of assessment: gatekeeping, accountability, and instructional diagnosis. Can J Educ. 2000:262–79.
10. Schuwirth LWT, Van der Vleuten CPM. Programmatic assessment: from assessment of learning to assessment for learning. Med Teach. 2011;33:478–85.
11. Wiliam D. What is assessment for learning? Studies in educational evaluation, vol. 37. Berlin: Elsevier; 2011. p. 3–14.
12. Black P, Wiliam D. Assessment and classroom learning. Assess Ed. 1998;5:7–74.
13. Stiggins R, Chappuis J. Using student-involved classroom assessment to close achievement gaps. Theory Pract. 2005;44:11–8.
14. Young I, Montgomery K, Kearns P, Hayward S, Mellanby E. The benefits of a peer-assisted mock OSCE, vol. 11. Hoboken: The Clinical Teacher Wiley; 2014. p. 214–8.
15. Grover SC, Scaffidi MA, Khan R, Garg A, Al-Mazroui A, Alomani T, et al. Progressive learning in endoscopy simulation training improves clinical performance: a blinded randomized trial. Gastrointest Endosc. 2017;86(5):881–9.
16. Brown S. Assessment for learning. Learn Teach Higher Ed. 2004;1:81–9.
17. Earl L. Assessment as learning. The keys to effective schools: educational reform as continuous improvement; 2007. p. 85–98.
18. Chenot J-F, Simmenroth-Nayda A, Koch A, Fischer T, Scherer M, Emmert B, et al. Can student tutors act as examiners in an objective structured clinical examination? Med Educ. 2007;41:1032–8.
19. Basehore PM, Pomerantz SC, Gentile M. Reliability and benefits of medical student peers in rating complex clinical skills. Med Teach. 2014;36:409–14.

20. Moineau G, Power B, Pion A-MJ, Wood TJ, Humphrey-Murto S. Comparison of student examiner to faculty examiner scoring and feedback in an OSCE. Med Educ. 2011;45:183–91.
21. Iblher P, Zupanic M, Karsten J, Brauer K. May student examiners be reasonable substitute examiners for faculty in an undergraduate OSCE on medical emergencies? Med Teach. 2015;37:374–8.
22. Burgess A, Clark T, Chapman R, Mellis C. Senior medical students as peer examiners in an OSCE. Med Teach. 2013;35:58–62.
23. Creswell J. A concise introduction to mixed methods research: SAGE Publications; 2014.
24. Bloch R, Norman G. Generalizability theory for the perplexed: a practical introduction and guide: AMEE guide no. 68. Med Teach. 2012;34:960–92.
25. Brennan RL. Generalizability theory. Educ Meas Issues Pract. 1992;11:27–34.
26. Shavelson RJ, Webb NM, Rowley GL. Generalizability theory. 1992;
27. Crossley J, Russell J, Jolly B, Ricketts C, Roberts C, Schuwirth L, et al. "I'm pickin" up good regressions': the governance of generalisability analyses. Med Educ. 2007;41:926–34.
28. Finch H. Comparison of the performance of nonparametric and parametric MANOVA test statistics when assumptions are violated. Methodology. Euro J Res Meth Behav Soc Sci. 2005;1:27.
29. Brennan RL. Generalizability theory. Educ Measure. 1992;11:27–34.
30. Walsh CM, Ling SC, Wang CS, Carnahan H. Concurrent versus terminal feedback: it may be better to wait. Acad Med. 2009;84:S54–7.
31. Hattie J, Timperley H. The power of feedback. Rev Educ Res. 2007;77:81–112.
32. Frierson HT, Hoban D. Effects of test anxiety on performance on the NBME part I examination. J Med Educ. 1987.
33. Cassady JC, Johnson RE. Cognitive test anxiety and academic performance. Contemp Educ Psychol. 2002;27:270–95.
34. Colbert-Getz JM, Fleishman C, Jung J, Shilkofski N. How do gender and anxiety affect students' self-assessment and actual performance on a high-stakes clinical skills examination? Acad Med. 2013;88:44–8.
35. Lefroy J, Watling C, Teunissen PW, Brand P. Guidelines: the do's, don't and don't knows of feedback for clinical education. Perspectn Med Ed. 2015;4:284–99.
36. Cushing A, Abbott S, Lothian D, Hall A, Westwood OMR. Peer feedback as an aid to learning–what do we want? Feedback. When do we want it? Now! Med Teach. 2011;33:e105–12.
37. Cushing AM, Westwood OMR. Using peer feedback in a formative objective structured clinical examination. Med Educ. 2010;44:1144–5.
38. Tan CPL, Azila NMA. Improving OSCE examiner skills in a Malaysian setting. Med Educ. 2007;41:517.
39. Clark I. Formative assessment: 'There is nothing so practical as a good theory.'. Aust J Educ. 2010;54:341–52.
40. Broadfoot PM, Daugherty R, Gardner J, Harlen W, James M, Stobart G. Assessment for learning: 10 principles. Cambridge: University of Cambridge School of Education; 2002.
41. Hattie J. Influences on student learning [internet]. Auckland: University of Auckland; 1999. Available from: http://geoffpetty.com/wp-content/uploads/2012/12/Influencesonstudent2C683.pdf
42. Woolf AD, Akesson K. Primer: history and examination in the assessment of musculoskeletal problems. Nat Clin Pract Rheumatol. 2008;4:26–33.
43. Hodges B, McIlroy JH. Analytic global OSCE ratings are sensitive to level of training. Med Educ. 2003;37:1012–6.