Available online at www.sciencedirect.com

ScienceDirect

Health Professions Education 6 (2020) 271–281

www.elsevier.com/locate/hpe

Impact of Structured Feedback on Examiner Judgements in Objective Structured Clinical Examinations (OSCEs) Using Generalisability Theory

Wai Yee Amy Wong [a,*], Chris Roberts [b], Jill Thistlethwaite [c]

[a] *School of Education & Faculty of Medicine, The University of Queensland, QLD 4072, Australia*
[b] *Sydney Medical School, Faculty of Medicine and Health, The University of Sydney, NSW 2006, Australia*
[c] *Faculty of Health, University of Technology Sydney, NSW 2007, Australia*

Received 16 October 2019; revised 18 February 2020; accepted 20 February 2020
Available online 12 March 2020

Abstract

Purpose: In the context of health professions education, the objective structured clinical examination (OSCE) has been implemented globally for assessment of clinical competence. Concerns have been raised about the significant influence of construct irrelevant variance arising from examiner variability on the robustness of decisions made in high-stakes OSCEs. An opportunity to explore an initiative to reduce examiner effects was provided by a secondary analysis of data from a large-scale summative OSCE of the final-year students ($n > 350$) enrolled in a graduate-entry four-year Bachelor of Medicine/Bachelor of Surgery (MBBS) program at one Australian research-intensive university. The aim of this study was to investigate the impact of providing examiners with structured feedback on their stringency and leniency on assessing the final-year students' clinical competence in the pre-feedback (P1) OSCE and post-feedback (P2) OSCE.

Method: This study adopted a quasi-experimental design to analyse the scores from 141 examiners before feedback was provided for the P1 OSCE, and 111 examiners after feedback was provided for the P2 OSCE. This novel approach used generalisability theory to quantify and compare the examiner stringency and leniency variance (V_j) contributing to the examiners' scores before and after feedback was provided. Statistical analyses conducted were controlled for differences in the examiners and OSCE stations.

Results: Comparing the scores of the 51 examiners who assessed students in both P1 and P2 OSCEs, the V_j reduced by 35.65% and its contribution to the overall variation in their scores decreased by 7.43%. The results were more noticeable in the 26 examiners who assessed students in both OSCEs and in at least one station common across both OSCEs. The V_j reduced by 40.56% and its contribution to the overall variation in their scores was also decreased by 7.72%.

Conclusion: The findings might be suggested that providing examiners with structured feedback could reduce the examiner stringency and leniency variation contributing to their scores in the subsequent OSCE, whilst noting limitations with the quasi-experimental design. More definitive research is required prior to making recommendations for practice.

Keywords: Assessment; Faculty development; Feedback; Generalisability theory; Judgement; OSCE

* Corresponding author. School of Nursing and Midwifery, Queen's University Belfast, BT9 7BL, UK.
E-mail addresses: A.Wong@qub.ac.uk (W.Y.A. Wong), christopher.roberts@sydney.edu.au (C. Roberts), Jill.Thistlethwaite@uts.edu.au (J. Thistlethwaite).

Peer review under responsibility of AMEEMR: the Association for Medical Education in the Eastern Mediterranean Region

https://doi.org/10.1016/j.hpe.2020.02.005

1. Introduction

The objective structured clinical examination (OSCE) is a widely used assessment strategy in both undergraduate and postgraduate medical and health professions education.[1,2] A dominant reason for the widespread use of OSCE is that it is perceived as an objective and standardised measure of student clinical competence.[3,4,5] In maintaining the quality assurance of assessments, it is essential to ascertain the variance in examiners' scores awarded to students, and find ways of reducing sources of unwanted construct-irrelevant variance [6] in future iterations of the OSCE. The aim of this study was to investigate the impact of structured feedback by comparing the examiner stringency and leniency variance in their judgements of the final-year students' clinical competence before feedback was provided for the pre-feedback (P1) OSCE, and shortly after feedback was provided for the post-feedback (P2) OSCE.

The OSCE in this study was a large-scale summative assessment of the final-year students (n > 350) enrolled in a four-year graduate-entry Bachelor of Medicine/Bachelor of Surgery (MBBS) program at one Australian research-intensive university. The focus of the initiative in this study to reduce unwanted construct-irrelevant variance was the examiner stringency and leniency. It is defined as the tendency of examiners to use either the top or bottom end of the rating scale consistently. This definition is adapted from the study of Roberts et al.[6] on interviewer stringency and leniency.

The significance of the influence of examiner stringency and leniency on the consistency of examiner judgements in high-stakes clinical examinations such as OSCEs has received considerable attention in the literature.[7–11] Harasym et al.[9] analysed the extent of the influence of examiner stringency and leniency on the communication skill scores of 190 medical students at their family medicine clerkship end-of-rotation OSCE. Results showed that the examiner stringency and leniency contributed 44.2% to the variance in the students' scores, whereas student ability only amounted to 10.3%.

More recently, Hope and Cameron[12] explored the changes in examiner stringency in the scores of 278 third-year undergraduate medical students in a summative OSCE. Two days were required to allow all students to complete the eight face-to-face stations. Results showed that the examiners were most lenient at the start of the two-day OSCE. When comparing the scores of the students who undertook the OSCE in the first and last group, there was approximately 3.3% difference in the effect of the examiner stringency and leniency on the student scores. Although the difference was relatively small, it would have affected the scores for the borderline students. Examiner training was emphasised as a crucial means to assure that examiner stringency and leniency did not vary over time in future iterations of the OSCE, due to the fact that examiners assessed an increasing number of successful students.[12]

Results from these two studies[9,12] highlighted the importance of acquiring empirical evidence on effective strategies to minimise the influence of unwanted sources of examiner variance, particularly in high-stakes summative assessments judged by a sole examiner.[13] This is necessary to guide initiatives aimed at reducing unwanted sources of variance, which may have a significant and direct impact on the robustness of decisions about student progression, certification, and ultimately affect the quality of patient care delivered by future doctors.[14]

Although recent literature suggested that examiner judgements are inherently subjective and could be based on idiosyncratic reasons,[15,16,17] it is important to provide a fair assessment of student clinical competence taking into account the interactions between students and the specific context including the examiners and the circumstances.[17] Previous empirical studies have attempted to evaluate the impact of examiner training to reduce the unwanted sources of variance in examiner judgements.[18–23] However, results have been inconclusive and difficult to compare as researchers applied different methodologies.[24]

Germane to the aim of providing students with fair assessment, this study addresses the critical challenge of reducing the known impact of the influence of examiner stringency and leniency on the scores awarded to students,[8,9,25] through implementing an examiner feedback system in a high-stakes summative OSCE. The idea of providing examiners with feedback was developed based on the three distinct but related perspectives of examiner cognition in the literature: examiners are *trainable*; examiners are *fallible*; or they are *meaningfully idiosyncratic*.[14] As the provision of feedback could be inferred as an examiner training intervention, this study is closely aligned with the perspective that examiners are *trainable*.[14] The structured feedback created an authentic learning opportunity for the examiners to formally review and reflect on their marking behaviour, and, potentially make subsequent evidenced-based decisions to change their marking practice.

While acknowledging that there are other factors impacting on the examiners' scores such as the station effect, this study focused on exploring the impact of

examiner stringency and leniency underpinning by the below two research questions (RQs). The pre-feedback (P1) OSCE for the final-year medical students was the first year of this study. The P1 OSCE examiners had never had feedback about their marking behaviour. The post-feedback (P2) OSCE for the final-year medical students was the second year of this study. The P2 OSCE examiners received the structured feedback eight weeks prior to assessing students in the P2 OSCE.

RQ 1. What is the contribution of and change in examiner stringency and leniency variance (V_j) for the examiners who assessed students in the pre-feedback (P1) OSCE, received structured feedback, and assessed students again in the post-feedback (P2) OSCE?

RQ 2. What is the contribution of and change in examiner stringency and leniency variance (V_j) for the examiners who assessed students in both the pre-feedback (P1) and post-feedback (P2) OSCEs and in at least one common station across both OSCEs?

2. An analytical framework using generalisability theory

We applied generalisability theory (G theory)[26,27] as the analytical framework which suggests that for a single OSCE station, the student score is a combination of the true score of a student's performance and multiple sources of error variances,[28] such as the examiner stringency and leniency variance (V_j). G theory facilitates the exploration of the impact of structured feedback by computing and comparing the magnitude of V_j contributing to the examiners' scores in the pre-feedback (P1) and post-feedback (P2) OSCEs. We hypothesised that such structured feedback would have a constructive impact on the examiners' marking behaviour when they assessed students in the P2 OSCE, thereby reducing V_j.

3. Context

The final-year OSCE for the four-year graduate-entry Bachelor of Medicine/Bachelor of Surgery (MBBS) students at this Australian research-intensive university is a high-stakes exit assessment as student results have a direct impact on their ability to graduate and thus commence an internship as a qualified medical doctor in the following year. It is a usual practice of this medical school to allocate a single examiner to assess a single student in a station in the final-year OSCE. This medical school was selected as it has had the largest enrolments in Australia since 2010, with nearly 500 final-year students in 2014.[29] Consequently, over 100 volunteer examiners were involved in the annual final-year OSCE to assess students on two consecutive days across different hospital sites. For both P1 and P2 OSCEs, four OSCE sessions (i.e. Saturday morning and afternoon, and Sunday morning and afternoon) were held at one hospital site, whereas only a Saturday morning session was held at the other three sites in the P1 OSCE and two other sites in the P2 OSCE. Examiners were allocated to a specific site based on their availability, whereas students were allocated to the relevant sites based on their geographical locations. The researchers were not involved in the allocation of students and examiners for the OSCEs.

4. Partially-crossed generalisability study design

Based on the G theory analytical framework, we adopted a quasi-experimental pre- and post-design of a generalisability study (G study) as a feasible and

Table 1
The variance components contributed to the examiners' scores in this partially-crossed and unbalanced G study. Adapted from Crossley et al.[31]

Variance Component	Notation Used in Section 8 Statistical Analysis	Explanation
1. Students (*p*)	$Var_{student}$ (V_p)	The consistent differences between student ability across examiners and OSCE stations
2. Stations (*s*)	$Var_{station}$ (V_s)	The consistent differences in OSCE station difficulty across students and examiners
3. Examiners (*j*)	$Var_{examiner}$ (V_j)	The consistent differences in examiner stringency/leniency across students and OSCE stations
4. Interaction between examiners and stations (*j* x *s*)	$Var_{examiner*station}$ (V_{j*s})	The varying case-specific stringency/leniency of examiners between OSCE stations across students
5. Interaction between students and stations (*p* x *s*)	$Var_{student*station}$ (V_{p*s})	The varying case aptitude of students displayed between stations across examiners
6. Measurement error (*e*)	Var_{error} (V_{err})	Any residual variation that cannot explained by other factors

effective way of analysing the secondary assessment data collected in the pre-feedback (P1) OSCE and post-feedback (P2) OSCE. This G study was a quasi-experimental study, as allocating examiners to a control group would not be achievable when the provision of structured feedback might have a real-life impact on students' scores in a high-stakes assessment.

The underlying design adopted was a multifaceted G study design,[30] in which three facets were under investigation: examiners (*j*), students (*p*) and stations (*s*).

However, to ensure the best estimates of examiner-related variances, this multifaceted G study was modified on account of the partially-crossed and unbalanced dataset.[28] The dataset of students and examiners was partially-crossed because only a proportion of students had the same set of examiners and thus the same set of stations. In addition, not all examiners consented to participate in this study. The dataset of examiners and stations was unbalanced as a number of examiners assessed students in multiple stations within and across different OSCE sessions. This partially-crossed and unbalanced design

Pre-feedback (P1) OSCE		**Post-feedback (P2) OSCE**
P1 OSCE consenting examiners Examiners (*j*) = 141 Students (*p*) = 376 Unique stations (*s*) = 42		*P2 OSCE consenting examiners* Examiners (*j*) = 111 Students (*p*) = 354 Unique stations (*s*) = 28
Analysis 1 *Among the 141 examiners, 51 examined again in the P2 OSCE.* Examiners[1] *(j)* = 51 Students *(p)* = 348 Unique stations *(s)* = 38	⇨ **Feedback provided to examiners eight weeks before P2 OSCE**	**Analysis 1** Examiners[1] *(j)* = 51 Students *(p)* = 322 Unique stations *(s)* = 27
Analysis 2 *Among the 51 examiners, 26 examined in at least one station that was used in both OSCEs.* Examiners[2] *(j)* = 26 Students *(p)* = 251 Unique stations[3] *(s)* = 13	⇨	**Analysis 2** *Among the 51 examiners, 26 examined in at least one station that was used in both OSCEs.* Examiners[2] *(j)* = 26 Students *(p)* = 291 Unique stations[3] *(s)* = 14
		No feedback group Examiners *(j)* = 60 Students *(p)* = 338 Unique stations *(s)* = 27

[1]The composition of the 51 examiners was the same in the P1 and P2 OSCEs in Analysis 1.
[2]The composition of the 26 examiners was different in the P1 and P2 OSCEs in Analysis 2.
[3]A total of 15 P1 OSCE stations were used again in the P2 OSCE. However, only 13 of them were examined by the group of examiners who assessed students in both OSCEs. The additional station in the P2 OSCE was the result of one P1 OSCE station being divided into two stations in the P2 OSCE.

Fig. 1. The number of examiners, students, and stations involved in the P1 and P2 OSCEs for Analysis 1 and 2.

facilitates the calculations of the estimates of the variance components contributed to the examiners' scores shown in Table 1, with the plain English explanations of these variance components adapted from Crossley et al.[31]

5. Participants

The research participants were examiners of the final-year high-stakes summative OSCEs. All the OSCE examiners attended a short briefing (maximum length was 30 minutes) prior to the commencement of the OSCE in each session, which was the only 'on-the-spot' examiner training required. Apart from this, mandatory examiner training was not offered, or required by this medical school. All examiners across all different sites were invited to participate in this study.

In the pre-feedback (P1) OSCE, a total of 159 examiners assessed the final-year medical students across all four sessions; 141 examiners (88.7%) agreed to be research participants and assessed 376 students. Each student was required to complete a full cycle of 12 stations in a single allocated session. There were only 42 unique stations, as six stations were used in more than one session.

In the post-feedback (P2) OSCE, a total of 143 examiners assessed the final-year medical students across all four sessions; 111 examiners (77.6%) agreed to be research participants and assessed 354 students. Each student was required to complete a full cycle of 10 stations in a single allocated session. There were only 28 unique stations, as 12 stations were used in more than one session. As this study focused on the overall OSCE, the total numbers of students, examiners and stations involved in the P1 and P2 OSCEs for Analysis 1 and 2 are presented in Fig. 1.

6. Procedures of examiners scoring student competence

Each OSCE station had a specific marking sheet which followed the same format and had been developed over time by clinicians and medical educators within the medical school. This study focused on the examiners' scores only in Part A of the marking sheet, which listed from three to seven criteria to assess a specific clinical skill or in response to the particular clinical scenario in a station. For each marking criterion, there were checklist points to guide the examiners. Examiners rated each marking criterion of each student's performance based on the following marking standards related to their achievement, the corresponding scores recorded are shown in brackets: *very well* (6)*; well* (4)*; partially* (2)*; poorly* (1)*;* or, *not at all* (0). Part B of the marking sheet was common to all OSCE stations and asked for the examiners' overall impression rating of a student's performance in a station independently of the checklist items for standard-setting purposes. This part was outside the scope of this study, as the majority of

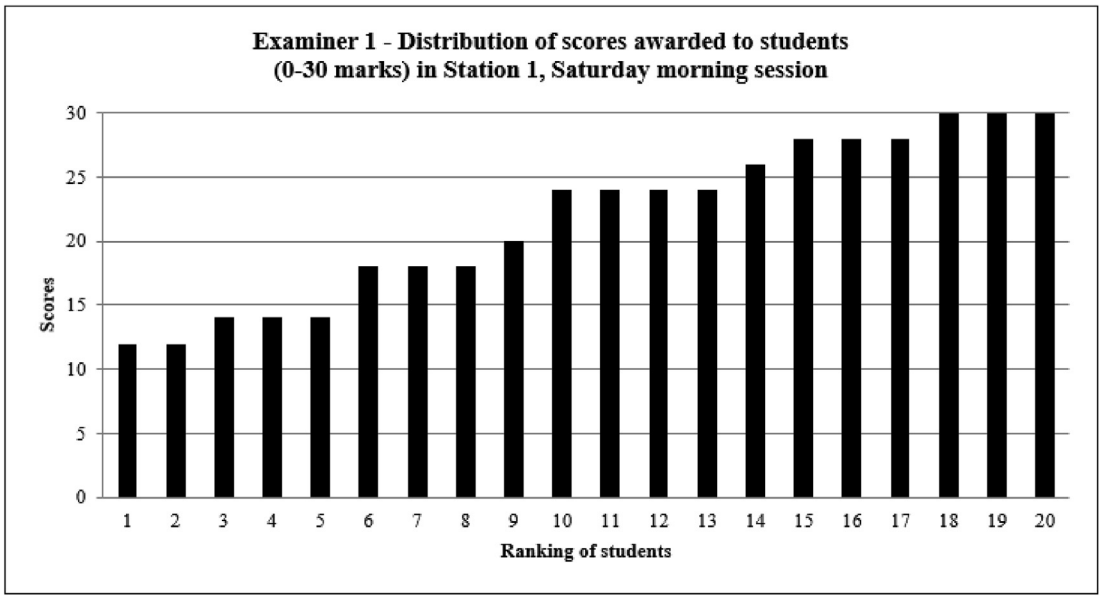

Fig. 2. Distribution of an examiner's scores awarded to students in a station.

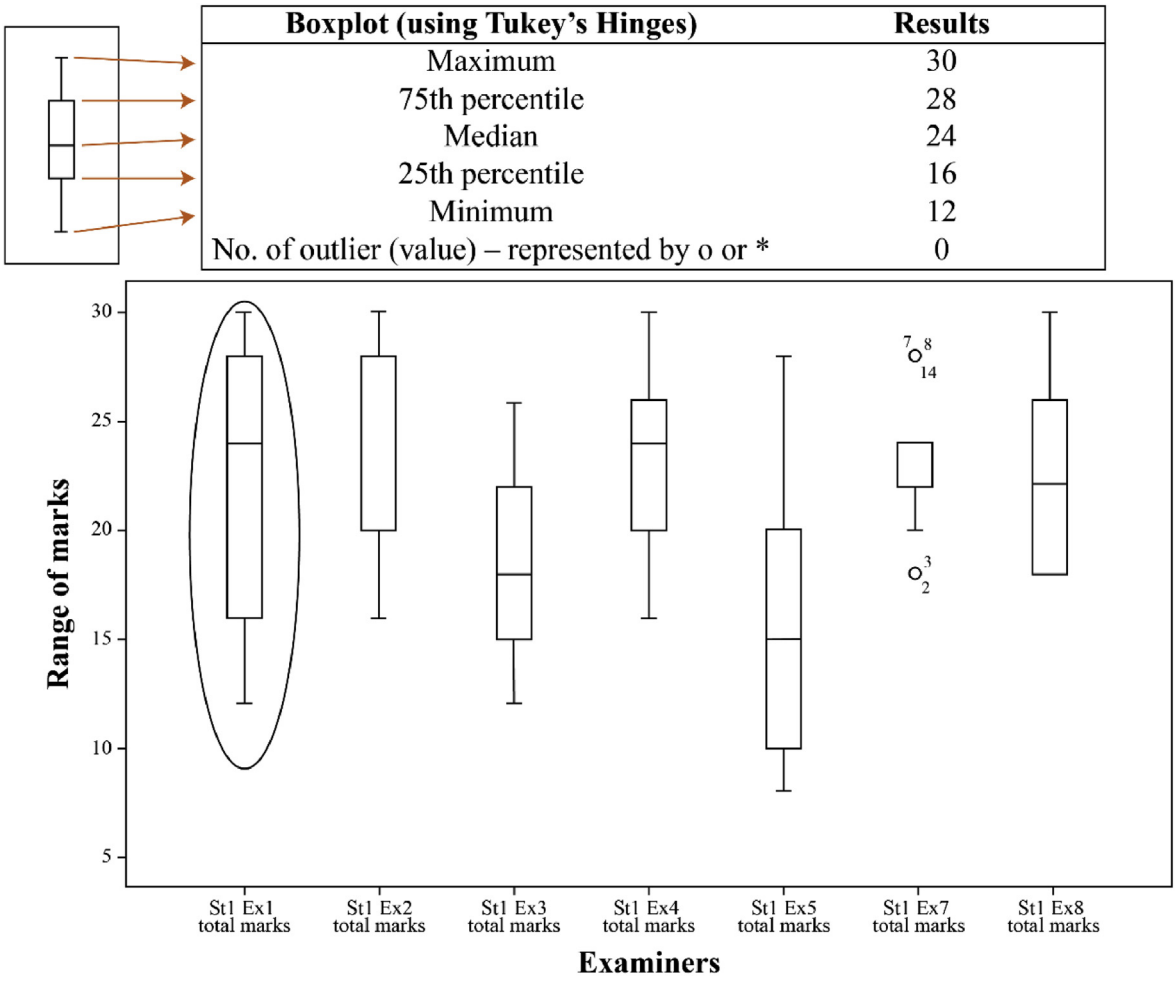

Fig. 3. Comparison of an examiner's scores to these of the other examiners in the same station.

examiners awarded a pass to students across all stations in both OSCEs which provided only limited discrimination of the examiners' marking behaviour in their cohort.

7. Provision of structured feedback as an examiner training strategy

All consenting examiners (n = 141) from the P1 OSCE received a structured feedback report via email approximately eight weeks before the P2 OSCE. This feedback timing was anticipated to provide sufficient time for the examiners to reflect on the feedback prior to assessing students again in the P2 OSCE. The design of the feedback reports aligned with the perspective of examiner cognition that examiners are *trainable*.[14] The purpose of the reports was to provide the examiners with data about the mean and range of scores given for an OSCE station, and comparisons with other examiners' judgements in the same station, as well as in the entire examiner cohort.

The report began by introducing the background of the station in which the examiner was involved, the marking criteria and the total score available for the station. The first part of the report consisted of a graph showing the distribution of an examiner's scores awarded to students in a station (Fig. 2). The y-axis shows the ranking of students in terms of their scores awarded in a descending order. This provided a quick way to show the range of scores given to the number of students within a station.

The second part showed the comparison of an examiner's scores to the other examiners in the same station (Fig. 3).

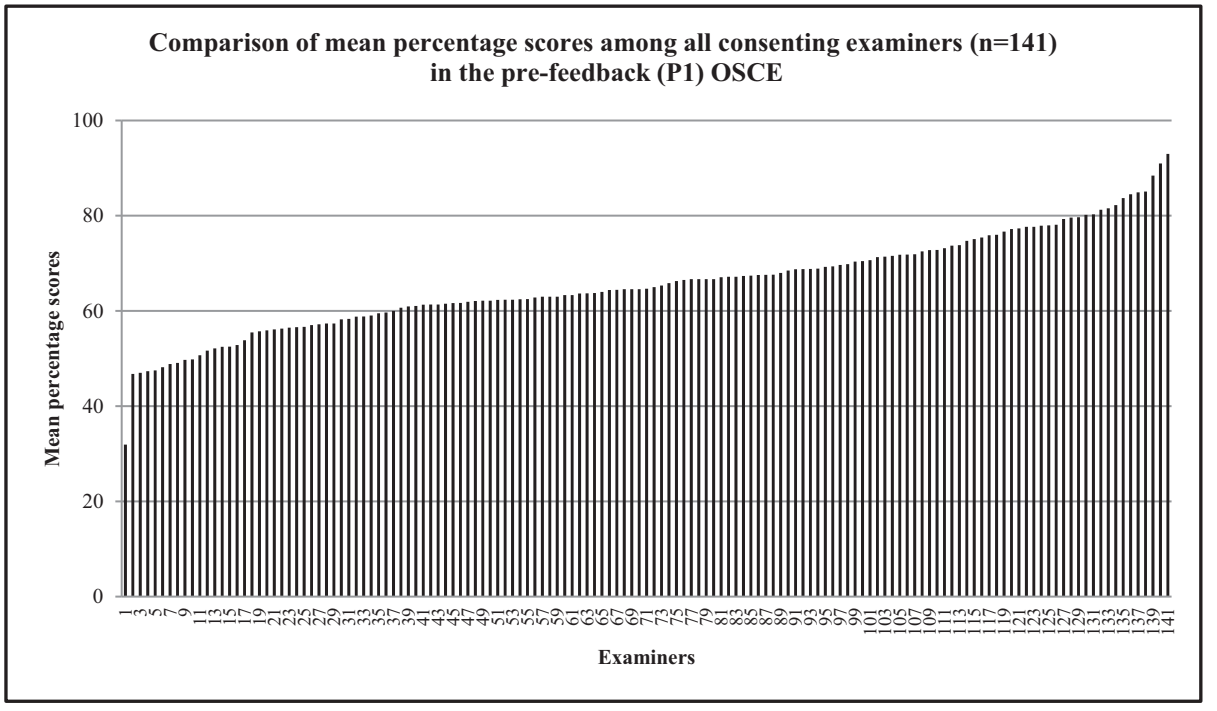

Fig. 4. Comparison of an examiner's mean percentage score among all consenting examiners in the pre-feedback (P1) OSCE.

Table 2
Results for Analysis 1 of the OSCE examiners' scores.

Analysis 1	Pre-feedback (P1) OSCE		Post-feedback (P2) OSCE		Changes	
	Examiners (n = 51)		Examiners (n = 51)			
Variance component	Estimate	% contributed to overall variation	Estimate	% contributed to overall variation	% change in estimate	% change to overall variation
$Var_{student}$ (V_p)	6.72	19.55%	5.18	15.86%	−22.92%	−3.69%
$Var_{station}$ (V_s)	3.17	9.22%	2.27	6.95%	−28.39%	−2.27%
$Var_{examiner}$ (V_j)	**7.91**	**23.01%**	**5.09**	**15.58%**	**−35.65%**	**−7.43%**
$Var_{examiner\ *station}$ (V_{j*s})	0	0%	2.68	8.20%	–	8.20%
$Var_{student*station}$ (V_{p*s})	16.58	48.23%	17.45	53.41%	5.25%	5.19%
Var_{error} (V_{err})	0	0%	0	0%	–	0%

Table 3
Results for Analysis 2 of the OSCE examiners' scores.

Analysis 2	Pre-feedback (P1) OSCE		Post-feedback (P2) OSCE		Changes	
	Examiners (n = 26)		Examiners (n = 26[a])			
Variance component	Estimate	% contributed to overall variation	Estimate	% contributed to overall variation	% change in estimate	% change to overall variation
$Var_{student}$ (V_p)	6.86	17.36%	5.50	15.97%	−19.83%	−1.39%
$Var_{station}$ (V_s)	5.72	14.48%	1.00	2.90%	−82.52%	−11.57%
$Var_{examiner}$ (V_j)	**9.59**	**24.27%**	**5.70**	**16.55%**	**−40.56%**	**−7.72%**
$Var_{examiner\ *station}$ (V_{j*s})	0	0%	2.56	7.43%	–	7.43%
$Var_{student*station}$ (V_{p*s})	17.34	43.89%	19.68	57.14%	13.49%	13.26%
Var_{error} (V_{err})	0	0%	0	0%	–	0%

[a] The composition of the 26 examiners in the P2 OSCE was different from the 26 examiners in the P1 OSCE. This is to ensure that at least one station was common across both OSCEs.

Finally, the third part showed the comparison of an examiner's mean percentage score with those of all the examiners in the P1 OSCE using a bar graph. Each examiner was informed of their rank on the continuum from the most stringent (1st) to the most lenient (141th) examiner (Fig. 4). The feedback was intended to prompt examiners to reflect on their marking behaviour by exploring the patterns of their scores and the comparisons with the cohort.

8. Statistical analysis

The quasi-experimental pre- and post-design study facilitated the exploration of the examiner stringency and leniency variance (V_j) impacting on the examiners' scores before and after feedback. We applied G theory and generated the estimates of each variance component in the examiners' scores in the P1 and P2 OSCEs using a Minimum Norm Quadratic Unbiased Estimation (MINQUE) procedure in the IBM Statistical Package for the Social Sciences (SPSS) Version 24.0. MINQUE was selected because of the unbalanced dataset[31] used in this study. Analysis 1, which addressed RQ1, explored V*j* of those examiners who assessed students in both P1 and P2 OSCEs, and hence controlled for the differences in the examiners. Analysis 2, which addressed RQ2, explored V*j* of those examiners who assessed students in at least one common station across both P1 and P2 OSCEs, and hence controlled for the differences in the OSCE stations.

9. Results

9.1. Analysis 1: contribution of and change in examiner stringency and leniency (V_j) of those examiners who assessed students in both pre-feedback (P1) and post-feedback (P2) OSCEs

Results for Analysis 1 of the estimates of each variance component in the examiners' scores are presented in Table 2. The first column lists all the variance components contributing to the examiners' scores. The second and third columns list the corresponding estimates and their percentages contributed to the overall variation of the examiners' scores, respectively, in the P1 OSCE. The fourth and fifth columns list the corresponding estimates and their percentages contributed to the overall variation of the same 51 examiners' scores, respectively, in the P2 OSCE. The last two columns show the percentage changes in each of the estimates and in their contribution to the overall variation of the examiners' scores, respectively, after feedback was provided.

Analysis 1 addressed RQ1 by controlling for the differences within the examiner cohort. Results revealed that the magnitude of V_j contributing to the examiners' scores was reduced from 7.91 to 5.09 (% change in estimate=35.65%) after feedback. Its contribution to the overall variation of the examiners' scores also reduced from 23.01% to 15.58% (% change to overall variation=7.43%). Both reductions appeared to be associated with the possible impact of providing structured feedback on decreasing the contribution of the examiner stringency and leniency variance (V_j) to their scores in the subsequent OSCE.

Apart from the impact of V_j, station difficulty and student ability also contributed to the overall variation of the examiners' scores. Results showed that the estimate of station difficulty was 2.27, and its percentage contributing to the overall variation of the examiners' scores was 6.95%, after feedback was provided in the P2 OSCE. This indicated that the consistent differences in OSCE station difficulty contributed less to the examiners' scores compared to V_j (% contributed to overall variation=15.58%) in the P2 OSCE.

Moreover, the estimate of student ability was 5.18, and its percentage contributing to the overall variation of the examiners' scores was 15.86% in the P2 OSCE. This indicated that the consistent differences between student ability contributed to a similar extent to the examiners' scores compared to V_j (% contributed to overall variation=15.58%) in the P2 OSCE.

To further investigate the decrease in the examiner stringency and leniency variance after feedback, we controlled the variance of station difficulty by focusing on the stations that were common across both OSCEs in Analysis 2.

9.2. Analysis 2: contribution of and change in V_j of those examiners who assessed students in at least one common station across both P1 and P2 OSCEs

Results for Analysis 2 of the estimates of each variance component in the examiners' scores are presented in Table 3 which follows the same format as Table 2 in terms of the information presented in each column.

Analysis 2 addressed RQ2 by controlling for the variance of station difficulty to focus on the stations that were common across both OSCEs, the magnitude of V_j contributing to the examiners' scores was reduced from 9.59 to 5.70 (% change in estimate=40.56%) after feedback. Its contribution to the overall variation of the examiners' scores also reduced from 24.27% to 16.55% (% change to overall variation=7.72%). Both reductions

shown appeared to be associated withthe possible impact of structured feedback on decreasing the contribution of the examiner stringency and leniency variance (V_j) to their scores in the subsequent OSCE.

Apart from the impact of V_j, station difficulty and student ability also contributed to the overall variation of the examiners' scores. Results showed that the estimate of station difficulty was 1.00, and its percentage contributing to the overall variation of the examiners' scores was 2.90%, after feedback was provided in the P2 OSCE. This indicated that the consistent differences in OSCE station difficulty contributed less to the examiners' scores compared to V_j (% contributed to overall variation=16.55%) in the P2 OSCE. This was anticipated as the common stations from both years were used in this analysis.

Moreover, the estimate of student ability was 5.50, and its percentage contributing to the overall variation of the examiners' scores was 15.97% in the P2 OSCE. This indicated that the consistent differences between student ability contributed to a similar extent to the examiners' scores as V_j (% contributed to overall variation=16.55%) in the P2 OSCE.

The estimate of error (V_{err}) was equal to zero in both Analysis 1 and 2 because all the errors were redistributed to all other variance components in both analyses. This is the result of using the selected design and analysis model in this study, which specified every variance component. There is no instance where an examiner's score could not be fully described in terms of these five specified variance components, that is, student ability, OSCE station difficulty, examiner stringency/leniency, case-specific stringency and case aptitude (Table 1). Therefore, there should be no residual (error) variance.

10. Discussion

Final-year OSCEs are high-stakes assessments of student results having a direct impact on their progression to internship. The OSCE examiners play a key role as gatekeepers to ensure that only those students who have demonstrated adequate clinical competence are awarded the opportunity to progress their career as medical doctors. This study, aligned with the examiner cognition perspective that examiners are *trainable*,[14] explored the change of the magnitude of examiner stringency and leniency variance (V_j) following the provision of structured feedback to the examiners as a form of training strategy.

When comparing the pre-feedback and post-feedback OSCEs, V_j reduced (from 7.91 to 5.09) for the 51 examiners who assessed students in both OSCEs. The decrease was more obvious (from 9.59 to 5.70) in the 26 examiners who assessed students in both OSCEs and in at least one station common across both OSCEs. It is also worthwhile to note that the contribution of V_j to the overall variation of the examiners' scores was reduced by about 7% in both groups of examiners (last column in Tables 2 and 3) after feedback was provided. These findings were consistent with the research hypothesis that structured feedback reduced examiner variance when they assessed students subsequently. This initial evidence supports the value of providing structured feedback to examiners and suggests ways in which the feedback could be better targeted to initiate and maintain change in examiners' assessment behaviours. Given that there are other possible confounding factors impacting on the examiners' scores, and there is no control group in this study, the results did not intend to make causal inferences. More empirical research is required prior to making recommendations for practice.

10.1. Implications for future research

The impact of feedback on V_j highlights the importance of examiners making their judgements of student clinical competence based on students' ability, instead of being influenced by their own stringency and leniency. To further establish which specific aspects of the feedback were the most impactful in changing examiners' assessment behaviour, we suggest that it is also important to include the examiners' perspective and conduct usability testing in designing an effective feedback report that will enable examiners to better understand their marking behaviour. In addition, to ensure a comprehensive dataset is collected for future naturalistic research of OSCEs, it is crucial that researchers work collaboratively with the academics, clinicians, examiners and professional administrative staff to develop a well-designed examination and data collection plan.

10.2. Strengths and limitations

This study is one of the first studies to have explored the impact of providing structured feedback to examiners, as a form of examiner training intervention, on the magnitude of V_j contributing to the examiners' scores. Previous studies mainly focused on the impact of performance dimension, frame-of-reference and behavioural observation training.[18,20] The findings of this study advance the knowledge in suggesting an association between providing examiners with structured feedback, as a form of training, and its effect on V_j

contributed to their scores. Although the feedback mechanism may well have reduced the examiner stringency and leniency variance, other factors might have contributed to it. For example, as the OSCE examiners gain experience in assessing students, it is possible that they introduce less variance into their scores regardless of the provision of structured feedback about their marking behaviour. Also, different cohorts of students may have different levels and range of abilities and this could potentially have influenced the examiners' judgements. However, it is not possible to have the same cohort of students in the P1 and P2 OSCEs in this study, as the final-year OSCE is only conducted annually.

In addition, there are challenges with the quasi-experimental design in this study. We acknowledge that the stability of the estimates of V_j will need to be demonstrated in other institutions. The primary constraint was that this G study was contingent on the assessment data from large-scale OSCEs in which the examiner judging plan was entirely pragmatic, and not modifiable to gain better estimates of the variance components in the examiners' scores. Additionally, not all the examiners provided consent to participate in this study, which was an agreement to have their scores aggregated for quality improvement purposes, including publications. Therefore, we had to adopt a partially-crossed and unbalanced G study design.[28]

Nevertheless, the large cohorts of examiners and students involved in both OSCEs were a strength of this study, with 141 (88.7%) of the examiners in the pre-feedback (P1) OSCE and 111 (77.6%) of the examiners in the post-feedback (P2) OSCE consenting to participate. These large cohorts facilitated the collection of a reasonable amount of data to compare the examiner stringency and leniency variance (V_j) in subgroups of examiners in Analysis 1 and 2.

11. Conclusions

This study has offered preliminary support to the possible impact of structured feedback on the examiners' marking behaviour in a typical undergraduate OSCE setting using G theory. The findings enhance the understanding of the possible impact of structured feedback, as a form of training, on the magnitude of examiner stringency and leniency variance (V_j) contributing to the examiners' scores before and after feedback. The statistical analyses from the G study suggest that providing feedback to the examiners might be associated with a decrease in the magnitude of V_j contributing to their scores. The outcomes of this study provide a basis to further explore the features of effective feedback to examiners about their marking behaviour. This is particularly important as examiner stringency and leniency in high-stakes assessments impacts not only on student progression, but ultimately, and more importantly, on the delivery of optimal patient care and safety as medical doctors.

Contributors

WYAW and CR led the study conception and contributed to the design, data analysis and interpretation. WYAW wrote the first draft of the paper. JT has contributed to the design of the overall study and made substantial contributions to the interpretation of these data. All authors contributed to the critical revision of the paper and approved the final manuscript for publication.

Ethical approval

This study was approved by The University of Queensland, Behavioural & Social Sciences Ethical Review Committee (approval no: 2013001070).

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declaration of Competing Interest

None.

Acknowledgements

The authors would like to thank Professor Jim Crossley for his invaluable advice on the application of Generalisability Theory in estimating variance components, Associate Professor Karen Moni and Associate Professor Lata Vadlamudi for reviewing previous drafts and providing helpful comments, and the participating OSCE examiners at The University of Queensland.

References

1. Khan KZ, Ramachandran S, Gaunt K, Pushkar P. The objective structured clinical examination (OSCE): AMEE guide no. 81. Part I: an historical and theoretical perspective. *Med Teach*. 2013;35(9):e1437–e1446. https://doi.org/10.3109/0142159X.2013.818634.

2. Fuller R, Homer M, Pell G, Hallam J. Managing extremes of assessor judgment within the OSCE. *Med Teach*. 2017;39(1):58–66. https://doi.org/10.1080/0142159X.2016.1230189.
3. Downing SM, Yudkowsky R. *Assessment in health professions education*. New York, NY: Routledge; 2009. https://epdf.pub/queue/assessment-in-health-professions-education.html. Accessed September 12, 2019.
4. Harden RM, Lilley P, Patricio M. *The definitive guide to the OSCE: the objective structured clinical examination as a performance assessment*. Edinburgh: Elsevier; 2016.
5. Daniels VJ, Pugh D. Twelve tips for developing an OSCE that measures what you want. *Med Teach*. 2018;40(12):1208–1213. https://doi.org/10.1080/0142159X.2017.1390214.
6. 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):690–698. https://doi.org/10.1111/j.1365-2923.2010.03689.x.
7. Williams RG, Klamen DA, McGaghie WC. Special article: cognitive, social and environmental sources of bias in clinical performance ratings. *Teach Learn Med*. 2003;15(4):270–292. https://doi.org/10.1207/S15328015TLM1504_11.
8. McManus I, Thompson M, Mollon J. Assessment of examiner leniency and stringency ('hawk-dove effect') in the MRCP(UK) clinical examination (PACES) using multi-facet Rasch modelling. *BMC Med Educ*. 2006;6(1):42. https://doi.org/10.1186/1472-6920-6-42.
9. Harasym PH, Woloschuk W, Cunning L. Undesired variance due to examiner stringency/leniency effect in communication skill scores assessed in OSCEs. *Adv Health Sci Educ Theory Pract*. 2008;13(5):617–632. https://doi.org/10.1007/s10459-007-9068-0.
10. Bartman I, Smee S, Roy M. A method for identifying extreme OSCE examiners. *Clin Teach*. 2013;10(1):27–31. https://doi.org/10.1111/j.1743-498X.2012.00607.x.
11. 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):325–341. https://doi.org/10.1007/s10459-012-9372-1.
12. Hope D, Cameron H. Examiners are most lenient at the start of a two-day OSCE. *Med Teach*. 2014;37(1):81–85. https://doi.org/10.3109/0142159X.2014.947934.
13. Berendonk C, Stalmeijer RE, Schuwirth LWT. Expertise in performance assessment: assessors' perspectives. *Adv Health Sci Educ Theory Pract*. 2013;18(4):559–571. https://doi.org/10.1007/s10459-012-9392-x.
14. Gingerich A, Kogan J, Yeates P, Govaerts M, Holmboe E. Seeing the 'black box' differently: assessor cognition from three research perspectives. *Med Educ*. 2014;48(11):1055–1068. https://doi.org/10.1111/medu.12546.
15. Van der Vleuten CPM, Schuwirth LWT, Driessen EW, et al. A model for programmatic assessment fit for purpose. *Med Teach*. 2012;34(3):205–214. https://doi.org/10.3109/0142159X.2012.652239.
16. Hodges B. Assessment in the post-psychometric era: learning to love the subjective and collective. *Med Teach*. 2013;35(7):564–568. https://doi.org/10.3109/0142159X.2013.789134.
17. Ten Cate O, Regehr G. The power of subjectivity in the assessment of medical trainees. *Acad Med*. 2019;94(3):333–337. https://doi.org/10.1097/ACM.0000000000002495.
18. Holmboe ES, Hawkins RE, Huot SJ. Effects of training in direct observation of medical residents' clinical competence: a randomized trial. *Ann Intern Med*. 2004;140(11):874–881. https://doi.org/10.7326/0003-4819-140-11-200406010-00008.
19. Pell G, Homer M, Roberts TE. Assessor training: its effects on criterion-based assessment in a medical context. *Int J Res Method Educ*. 2008;31(2):143–154. https://doi.org/10.1080/17437270802124525.
20. Cook DA, Dupras DM, Beckman TJ, Thomas KG, Pankratz VS. Effect of rater training on reliability and accuracy of mini-CEX scores: a randomized, controlled trial. *J Gen Intern Med*. 2009;24(1):74–79. https://doi.org/10.1007/s11606-008-0842-3.
21. Malau-Aduli BS, Mulcahy S, Warnecke E, et al. Inter-rater reliability: comparison of checklist and global scoring for OSCEs. *Creativ Educ*. 2012;3:937–942. https://doi.org/10.4236/ce.2012.326142. special issue.
22. Weitz G, Vinzentius C, Twesten C, Lehnert H, Bonnemeier H, König IR. Effects of a rater training on rating accuracy in a physical examination skills assessment. *GMS Z Med Ausbild*. 2014;31(4):doc41. https://doi.org/10.3205/zma000933.
23. 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):1230–1234. https://doi.org/10.1016/j.pec.2017.01.013.
24. Reid K, Smallwood D, Collins M, Sutherland R, Dodds A. Taking OSCE examiner training on the road: reaching the masses. *Med Educ Online*. 2016;21(1):32389. https://doi.org/10.3402/meo.v21.32389.
25. Crossley J, Davies H, Humphris G, Generalisability Jolly B. A key to unlock professional assessment. *Med Educ*. 2002;36(10):972–978. https://doi.org/10.1046/j.1365-2923.2002.01320.x.
26. Cronbach LJ, Gleser GC, Nanda H, Rajaratnam N. *The dependability of behavioral measurements: theory of generalizability for scores and profiles*. New York: Wiley; 1972.
27. Brennan RL. Generalizability theory and classical test theory. *Appl Meas Educ*. 2010;24(1):1–21. https://doi.org/10.1080/08957347.2011.532417.
28. Bloch R, Norman G. Generalizability theory for the perplexed: a practical introduction and guide: AMEE guide no. 68. *Med Teach*. 2012;34(11):960–992. https://doi.org/10.3109/0142159X.2012.703791.
29. *Medical deans Australia and New Zealand. Student statistics tables web site*; 12 September 2019. https://medicaldeans.org.au/data/.
30. Marcoulides GA. Generalizability theory. In: Tinsley HEA, Brown SD, eds. *Handbook of applied multivariate statistics and mathematical modeling*. San Diego, CA, US: Academic Press; 2000:527–551. https://doi.org/10.1016/B978-012691360-6/50019-7. . Accessed September 12, 2019.
31. Crossley J, Russell J, Jolly B, et al. 'I'm pickin' up good regressions': the governance of generalisability analyses. *Med Educ*. 2007;41(10):926–934. https://doi.org/10.1111/j.1365-2923.2007.02843.x.

Dr Wai Yee Amy Wong obtained her PhD in medical education through the School of Education and Faculty of Medicine at The University of Queensland in May 2019, and is currently working as a Research Fellow in the School of Nursing and Midwifery at Queen's University Belfast.

Associate Professor Chris Roberts is a health professions educator and researcher at Sydney Medical School, The University of Sydney.

Professor Jill Thistlethwaite is a health professions education consultant, and is affiliated with the Faculty of Health at University of Technology, Sydney.