Does changing examiner stations during UK postgraduate surgery objective structured clinical examinations influence examination reliability and candidates' scores?

Article (Accepted Version)

Does Changing Examiner Stations During UK Postgraduate Surgery Objective Structured Clinical Examinations Influence Examination Reliability and Candidates' Scores?

Peter A. Brennan, MD, FRCS,[*] David T. Croke, PhD,[†] Malcolm Reed, MD, FRCS,[‡] Lee Smith,[*] Euan Munro, FRCS,[*] John Foulkes, PhD,[*] and Richard Arnett, PhD[†]

[*]Intercollegiate Committee for Basic Surgical Examinations, The Royal College of Surgeons of England, London, UK; [†]Department of Quality Enhancement, The Royal College of Surgeons in Ireland, Dublin, Ireland; and [‡]Dean, Brighton Medical School, Brighton, UK

OBJECTIVE: Objective structured clinical examinations (OSCE) are widely used for summative assessment in surgery. Despite standardizing these as much as possible, variation, including examiner scoring, can occur which may affect reliability. In study of a high-stakes UK postgraduate surgical OSCE, we investigated whether examiners changing stations once during a long examining day affected marking, reliability, and overall candidates' scores compared with examiners who examined the same scenario all day.

DESIGN, SETTING, AND PARTICIPANTS: An observational study of 18,262 examiner-candidate interactions from the UK Membership of the Royal College of Surgeons examination was carried at 3 Surgical Colleges across the United Kingdom. Scores between examiners were compared using analysis of variance. Examination reliability was assessed with Cronbach's alpha, and the comparative distribution of total candidates' scores for each day was evaluated using *t*-tests of unit-weighted *z* scores.

RESULTS: A significant difference was found in absolute scores differences awarded in the morning and afternoon sessions between examiners who changed stations at lunchtime and those who did not ($p < 0.001$). No significant differences were found for the main effects of either broad content area ($p = 0.290$) or station content area ($p = 0.450$). The reliability of each day was not affected by examiner switching ($p = 0.280$). Overall, no difference was found in *z*-score distribution of total candidate scores and categories of examiner switching.

CONCLUSIONS: This large study has found that although the range of marks awarded varied when examiners change OSCE stations, examination reliability and the likely candidate outcome were not affected. These results may have implications for examination design and examiner experience in surgical OSCEs and beyond. (J Surg Ed ▮:▮▮▮-▮▮▮. © 2016 Association of Program Directors in Surgery. Published by Elsevier Inc. All rights reserved.)

KEY WORDS: objective structured clinical examination, surgery, reliability, examiner, scenario

COMPETENCIES: Medical Knowledge, Practice-Based Learning and Improvement, Interpersonal and Communication Skills, Professionalism, System-Based Practice

INTRODUCTION

Undergraduate and postgraduate surgical examinations often utilize objective structured clinical examinations (OSCE) in an attempt to minimize variability and possible examiner bias as well as providing a consistent series of items and tasks for each candidate. The organization and execution of a successful OSCE needs considerable planning and knowledge of examination systems, and collaboration and effective interaction between the organizing Institute and examiners, with many of them undertaking the role in their own time.[1]

A study of undergraduate medical OSCE students found no evidence that the duration of examining in a communication OSCE station influenced examiners and the marks they awarded.[2] However, McLaughlin et al.[3] reported that the point of entry to an OSCE circuit was significantly associated with scoring and could be a factor that may

Correspondence: Inquiries to Peter A. Brennan, Maxillofacial Unit, Queen Alexandra Hospital, Portsmouth, PO6 3LY, UK; fax: (44) 2392-286089; e-mail: Peter.brennan@porthosp.nhs.uk

  1931-7204/$30.00
http://dx.doi.org/10.1016/j.jsurg.2016.01.010

compromise the reliability of the marks awarded and the internal validity of an OSCE examination. Removing the first 2 stations from a candidates' final scoring in an attempt to eliminate examiner "warm up" did not influence this so-called differential rating over time, an effect that might be due to examiner fatigue as the OSCE continues.[3]

It is well known that examiner stringency or leniency, colloquially known as "hawk" and "dove" behavior, might also influence the mark awarded in any particular OSCE station.[4-7] This can be minimized by pairing of examiners in those examinations or stations that are dual manned.[6] However, with a comprehensive examination containing many different stations it is unlikely that this behavior influences the overall outcome for candidates, with these effects essentially canceling each other out.

In this time of increasing scrutiny of examinations, and with recent issues of potential bias around high-volume postgraduate medical examinations, an understanding of the potential role of these factors is needed.[8-10] Candidates presenting themselves for examinations want reassurance of examination validity and that the processes are fair and unbiased.[11]

The OSCE (part B) of the Intercollegiate Membership of the Royal College of Surgeons (MRCS) examination, a practical entry-level examination needed to enter higher surgical training in the United Kingdom and Ireland, consists of 18 stations (10 skills and 8 knowledge), each lasting for 9 minutes. This is a high-volume postgraduate examination delivered by the 4 Royal Colleges of Surgeons in the United Kingdom and Ireland with more than 1500 candidates taking it each year.[12] There are 2 broad content areas of knowledge and skills. Knowledge stations include 3 anatomy, 3 physiology or critical care scenarios, and 2 pathology scenarios. The clinical skills section of the OSCE includes 4 clinical examination stations, 4 communication stations, and 2 procedural stations.[12] For the purposes of quality assurance, standard setting statistics and the establishment of the pass mark, these stations are interpreted as 11 distinct groups (Table 1).

TABLE 1. The 11 Main Content Areas of the MRCS OSCE Examination (in 2 Main Groups). Some Areas Are Assessed in More Than 1 Station as Indicated

Broad Content Area	Content Area	Number of Stations
Knowledge	Anatomy	3
	Surgical pathology	2
	Data interpretation	2
	Critical care	1
Skills	Giving and receiving information	2
	History taking	2
	Physical examination	4
	Procedural skills	2
Total		18

A single examiner is present in most stations but 3 communication skills stations have 2 examiners each, with a final agreed mark being awarded. In all stations, there is an instruction sheet that candidates read to "set the scene" and once in the station the examiners have a pre-prepared list of questions to ask, with marks allocated throughout using a scoring checklist. There is a 1-minute break between candidates to assign the marks and prepare for the next candidate. A cohort of examiners typically works for up to 3 days at a time.

Each circuit lasts 180 minutes with a 20-minute break and a typical day involves 2 sessions of the examination. As the examination takes place for more than a variable number of days at up to 4 sites in the United Kingdom and Ireland, a standard blueprint is used for each examination and drawn from a bank of scenarios all of which are initially piloted. Stringent quality assurance protocols ensure that only questions with acceptable performance statistics are utilized and psychometric analysis is undertaken at individual question, examiner, and venue and session for each examination that is held.

It is recognized that human factors including repetition, tiredness, and boredom may influence examiner behavior during OSCE circuits, with MRCS examiners often having to stay in the same OSCE station all day.[13]

Several of the communication stations (such as history taking) require minimal examiner interaction, thereby increasing the likelihood of fatigue and disengagement. In response to feedback from examiners, it was agreed that examination departments could allow MRCS examiners to change stations at lunchtime, enabling a different scenario to be examined in the afternoon session. This option was not adopted uniformly by all examination departments, with some continuing to require examiners, wherever possible, to examine at the same station all day.

In the present study of more than 18,000 candidate-examiner interactions, we investigated whether marks awarded were comparable between examiners changing OSCE stations at lunchtime after a complete OSCE circuit (18 candidates examined), and those remaining in the same scenario all day. We assessed the effect of this examiner switching on station reliability and the overall scores that candidates obtained per examining day.

MATERIALS AND METHODS

In the absence of a specific ethical committee responsible for postgraduate surgical examinations, the Intercollegiate Committee for Basic Surgical Examinations and Quality Assurance Committee approved the study. Data were collected from 3 examination periods of the MRCS part B OSCE during 2014 and 2015. Each scenario was given a score (of up to 20) by a single examiner in 15 stations, and in the 3 communication skills stations in which both a

clinical and lay examiner are present, the 2 examiners' scores (set independently and assessing separate elements of performance) were combined to achieve a single score of 20. In these 3 stations, the clinical examiner identification (ID) number was used for ID purposes. Standard setting was by the borderline regression method with the pass mark being set at the calculated cut-score plus 0.84 standard error of measurement as required by the General Medical Council of the United Kingdom.

To investigate score differences in individual scenarios that might be attributable to examiners switching stations at lunchtime, only examining days that included both a morning (AM) and afternoon (PM) session were included in the analysis. Furthermore, only scenarios were included in which an examiner stayed in a station for the duration of each session. For example, if an examiner was replaced during an AM or PM session due to illness or some other unforeseen reason then that scenario was excluded. If the examiner ID in the AM session was different to the PM session then switching was deemed to have occurred, whereas if the examiner ID was identical for both the sessions on a particular day then no switching occurred.

For each individual scenario on each day, mean scores were calculated for the AM and PM sessions along with the absolute difference between the 2 sessions. The group means in terms of AM or PM absolute score differences for the 3 examination periods were compared using a one-way analysis of variance (ANOVA) to determine if they could be combined for subsequent statistical analysis.

An independent-samples *t*-test was used on the amalgamated data to compare the absolute score differences in scenarios where examiner switching took place at lunchtime compared with the examiner stayed in the same station for the whole day.

The potential role of station content area was investigated to evaluate whether either the 2 broad content areas (knowledge and skills) or the 11 station groups might have an interactive or additive effect on absolute AM or PM score difference. Two-way ANOVA were used to assess the potential role of broad content and station content area together with examiner switching in absolute score differences.

To investigate whether the effects of switching might affect the total candidate scores, reliability using Cronbach's alpha was calculated for each day (including the scores for both AM and PM sessions). The way in which switching occurred was further classified as either "none" (where no examiner switching took place), "low" (where 5 or less of the 18 stations on the day had switching), and "high" (where more than 5 of the 18 stations in a particular day had examiner switching). Realistically, almost all examiners in the latter group changed scenarios in accordance with the agreed policy of the Intercollegiate Examinations Committee.

The reliabilities for all the examining days in each of these 3 categories were compared using a one-way ANOVA to determine whether lower or higher reliabilities in the categories where switching had taken place might be attributable to examiner switching.

Finally, the comparative distribution of total scores for each day were compared (using *t*-tests of unit-weighted *z* scores) for the candidates who were examined in the AM and PM sessions to determine if differences where switching had taken place might be attributable to examiner switching. It was assumed that candidates were randomly assigned to AM or PM examination sessions. All data were analyzed using R version 3.2.1 (June 18, 2016), R Project, Vienna, Austria. A $p < 0.01$ was used for statistical significance as we wanted any results to be both statistically as well as clinically significant. A $p < 0.01$ provides more substantial evidence more than a $p < 0.05$ that a test is significant.

RESULTS

The 3 MRCS OSCE examinations used in this study comprised 45 examining days—October 2014 ($n = 490$ candidates), February 2015 ($n = 372$), and May 2015 ($n = 453$)—with 1315 candidates and a total number of 23,670 candidate-examiner interactions for potential analysis.

Following data exclusion from days of a single examining session (a half day session at the start of an OSCE examining session in some Colleges, or where fewer candidates meant that a full day of examining was not needed), the final dataset for analysis consisted of 1049 candidates on 29 examining days. The AM and PM score data were used for 505 OSCE scenarios (147 were unique as some scenarios were repeated during the OSCE sessions on different days), giving a total of 18,262 candidate-examiner interactions. There were 9924 candidate-examiner interactions in scenarios where examiner switching had occurred and 8338 candidate-examiner interactions in scenarios where the same examiner had stayed in the station all day.

A summary of the number of candidate-examiner interactions, mean, and standard deviation absolute difference in examiner scores for switching and nonswitching scenarios per OSCE examination period is shown in Table 2. Although the median and the range of absolute AM or PM differences were greater for scenarios in which examiners switched at lunchtime, analysis found no evidence to suggest that the group means for the 3 OSCE examination periods were different, $F(2,502) = 0.23$, $p = 0.790$ (Fig. 1).

The data for all the 3 examination periods were therefore amalgamated. Figure 2 shows the aggregated data for these 3 examination periods. As with the individual OSCE sessions, there was a greater median and range of absolute AM or PM differences in scores between switching and

TABLE 2. Summary of the Number of Candidate-Examiner Interactions, Mean and SD Absolute Difference Between Examiners' Scores for Switching and Nonswitching Scenarios in the 3 OSCE Examination Periods Evaluated (*N* = No Examiner Switching, *Y* = Examiner Changed OSCE Station at Lunchtime)

Examination Session	Switch	Scenarios (*n*)	Interactions	Mean Difference	SD Difference
Oct 14	N	79	2771	1.07	0.90
Oct 14	Y	93	3604	1.79	1.55
Feb 15	N	80	2698	1.09	0.81
Feb 15	Y	77	2826	1.75	1.59
May 15	N	87	2870	1.08	0.75
May 15	Y	89	3494	1.65	1.23
Overall	N	246	8339	1.08	0.82
Overall	Y	259	9924	1.73	1.46

SD, standard deviation.

nonswitching examiners, with a significant difference found between them, t (410) = 6.2, p < 0.001.

Absolute score differences were subjected to a two-way ANOVA having 2 levels of examiner switching (yes and no) and 2 levels of broad content area (knowledge and skills) (Fig. 3). The main effect for examiner switching was significant with an F ratio of F (1,501) = 37.40, p < 0.001. The main effect for broad content area was not significant (F [1,501] = 1.14, p = 0.290) and the interaction effect was also nonsignificant (F [1,501] = 0.13, p = 0.720).

A two-way ANOVA was also done between absolute score differences and 2 levels of examiner switching (yes and no) and 11 levels (Table 1) of station content area (Fig. 4). The main effect for examiner switching was again significant with an F ratio of F (1,483) = 34.42, p < 0.001, whereas the main effect for station content area was not significant, (F [10,483] = 0.99, p =0.450) and the interaction effect was also nonsignificant (F [10,483] = 0.45, p = 0.920).

Table 3 shows the number of examining days in each switching category. A total of 14/29 (48%) examining days had high examiner switching with more than 5/18 stations having examiner switching (with all 18 stations being switched in most cases to allow examiners a change of scenario as agreed by the Intercollegiate Examination Committee). 15/29 examining days (52%) from centers delivering the same OSCE had either low (less than 5/18 stations switched) or no examiner switching.

To investigate whether the effects of switching could affect the total candidate scores, reliability was calculated using Cronbach's alpha for each day and for the AM and PM examining sessions (Fig. 5). No statistically significant differences were found between group means and the different switching categories (none, low, or high), (F [2,26] = 1.33, p = 0.280).

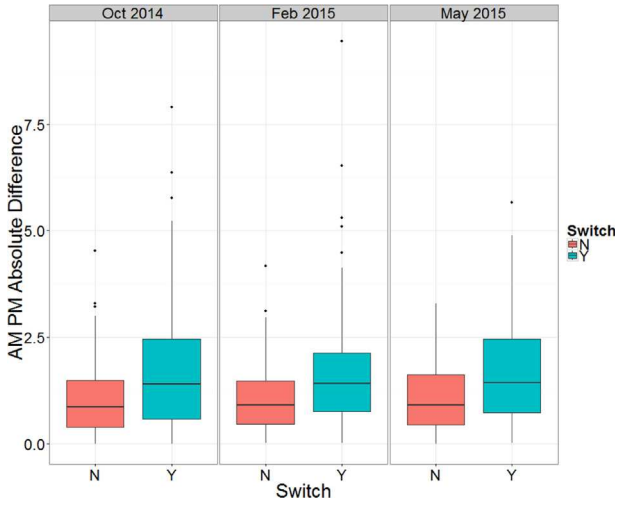

FIGURE 1. Aggregated comparative absolute differences between AM and PM sessions for scenarios in which examiner switching did or did not take place. The median is shown as a horizontal line in the boxes, and the interquartile ranges as vertical lines. Outliers are shown as dots. There were no statistically significant differences found between group means for the 3 examination sessions as determined by one-way ANOVA (F (2,502) = 0.23, p = 0.790).

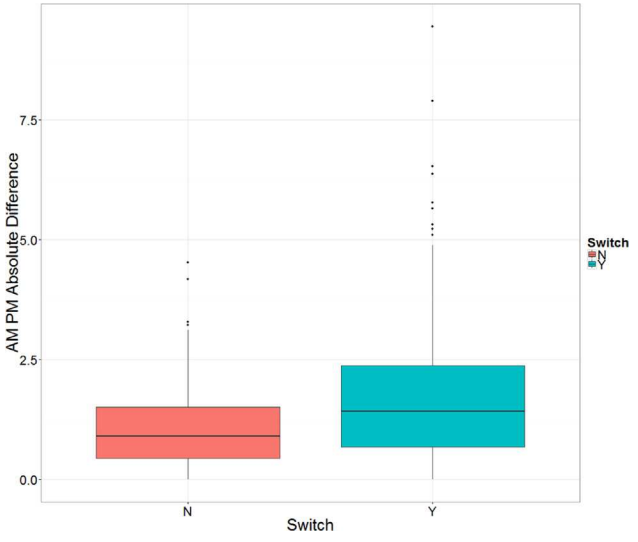

FIGURE 2. Aggregated data for all 3 examination sessions. A statistically significant difference was found in the absolute scores difference in the "No" switching category (mean = 1.08, SD = 0.82) and the "Yes" switching category (M = 1.73, SD = 1.46); t (410) = −6.2, p < 0.001). SD, standard deviation.

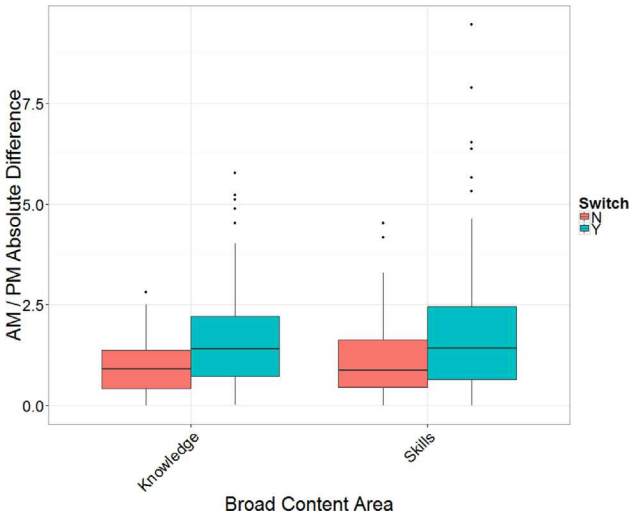

FIGURE 3. Comparison of distribution of AM or PM absolute score differences by broad content area. No significant differences were found between switching and the broad content area of the OSCE station, $F(1,503) = 1.07$, $p = 0.302$.

Comparison of AM and PM *z*-score distribution for examining days with high, low, and no examiner switching are shown in Figure 6. Overall, no significant difference was found in the comparative distribution of total candidate scores and the categories of examiner switching. Interestingly, p values approaching statistical significance at the $p < 0.01$ level ($p = 0.040$, $p = 0.040$, and $p = 0.020$) were found in 3 days—2 of these days were in the low switching category and 1 was on a day when no switching occurred between the AM and PM examining sessions.

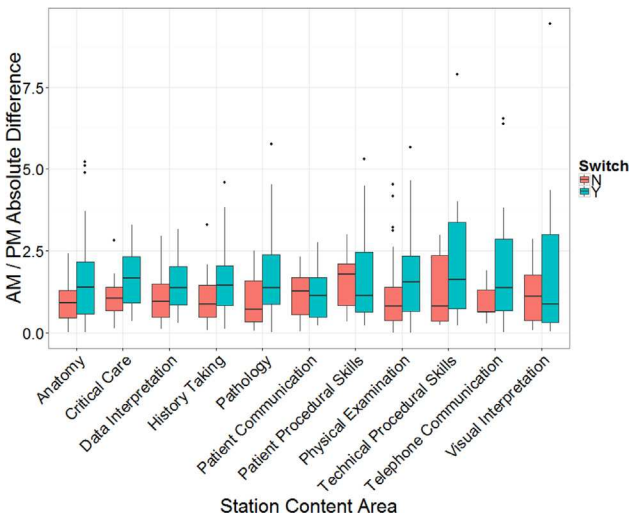

FIGURE 4. Comparison of distribution of AM or PM absolute score differences by OSCE station content area. No significant differences were found between switching and the content area of the OSCE station, $F(10,494) = 0.939$, $p = 0.497$.

TABLE 3. Number of Examining Days in Each Switching Category

Examination Session	High	Low	None
May 2015	5	2	3
Feb 2015	4	4	1
Oct 2014	5	4	1
Total	14	10	5

None—no examiner switching.
Low—less than 5/18 stations with examiner switching.
High—greater than 5 stations (in this category almost all examiners switched stations from AM to PM as agreed by the Intercollegiate Examinations Committee.

DISCUSSION

To our knowledge, this by far is the largest study of its kind and from observation by supervising examiners, intercollegiate assessors, and examinations staff, as well as feedback from the examiners themselves, it is clear that a change of scenario during a long examining day results in better morale and engagement of OSCE examiners. We found a difference in absolute score differences between switching and nonswitching examiners during 3 MRCS examination periods for more than 45 days. These scoring differences between examiners evaluating candidates in the same OSCE stations are not unexpected, whereas OSCE have been introduced to provide the same examination for candidates on a given day, there would often be examiner variations because of marking leniency or stringency.[5,6,14,15] There are many possible reasons for this scoring variance including the examiners' own perception of standards, and even personality factors.[7] It is also possible that examiners become more

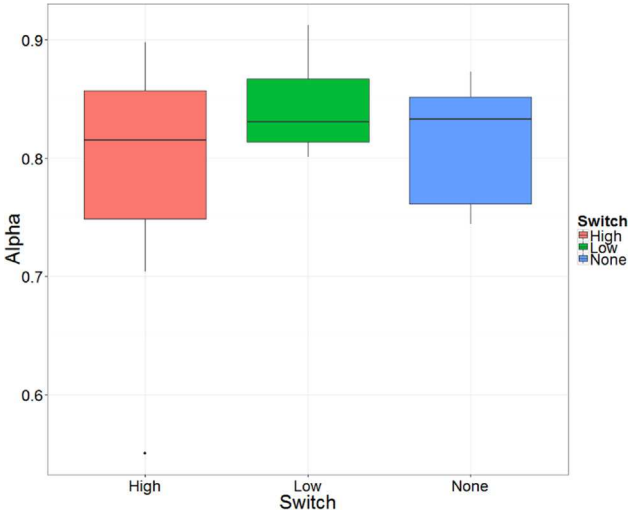

FIGURE 5. Comparative distribution of score reliabilities (Cronbach's alpha) for days classified as having high, low, or no examiner switching. There were no statistically significant differences between group means as determined by one-way ANOVA, ($F(2,26) = 1.33$, $p = 0.280$).

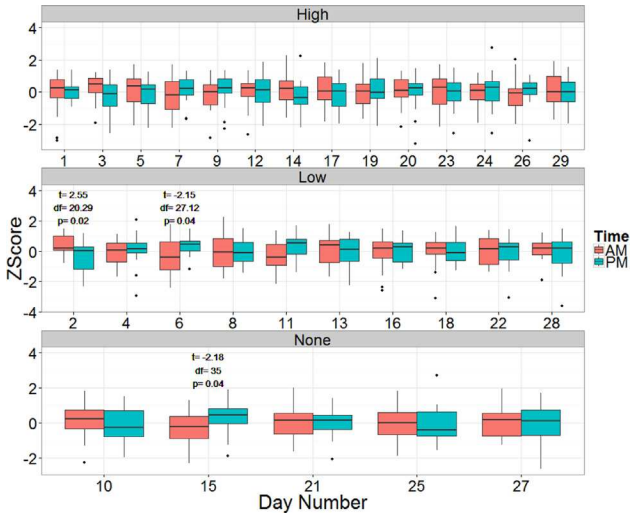

FIGURE 6. Comparison of AM and PM *z*-score distribution for examining days with high, low, and no examiner switching. Days with values approaching significant differences are labeled with a summary of the *t*-test results.

hawkish as their experience grows during a period of examining more than a few days, and as they are influenced by the ability of recently observed candidates.[16] Examiners attend a brief meeting at the start of each period of examining day, and marking issues including consistency are always discussed in an attempt to minimize variation as much as possible.

Although these leniency and stringency variations might be a potential issue in communication-related stations,[4,5,17,18] which are generally not dependent on fact (compared with a discipline such as anatomy for which an answer is either correct or not), we did not find a statistical difference in absolute scores between examiners and the content area being examined. In some undergraduate and postgraduate examinations, the use of 2 examiners per scenario is encouraged to minimize inter-examiner variation, though many examinations now rely on just 1 examiner for each OSCE station.[6,14,19]

In the MRCS, quality assurance steps are taken to analyze outlying performance at the examiner, station, session, and venue level, and on occasions adjustments have been made where there is clear evidence of outlying scores because of factors unrelated to candidate performance. In practice, this has not occurred because of perceived examiner behavior (hawk or dove) as although variation is clear, it has not been found to affect individual candidate outcomes.

We found scoring outliers in our study (as indicated in Figures 1-4), explained by replacement of a hawk examiner by a dove (or vice versa) when switching took place. All MRCS examiners' scores are scrutinized by psychometric analysis during quality assurance review after each OSCE examination period (3 times per year) and outliers (both hawks and doves) are identified. Examiners are given a detailed breakdown of their performance relative to their colleagues, and the examinations departments take appropriate action when required.

In stations where examiners did not switch for the second session of the day, variance was still found in absolute score differences, confirming that variability does occur in OSCE.[17] In addition to the reasons discussed earlier, recent work using a NASA task load index rating, identified that excessive workloads occur in OSCE examiners in a similar way that is seen in other professions such as airline pilots.[20] Further research is needed in this relatively unexplored area and other ways to minimize examiner variables need to be considered. For example, in certain situations it might be possible to conduct some OSCE stations using a video facility with an examiner being remote from the examination center.[21]

Each candidate needs reassurance that scoring variance in an examination is minimized and that examiners give them the same attention and concentration as the next candidate. In addition to personality, human factors (including tiredness, repetition, and even boredom) might influence examiner behavior. Recent work has enabled the Royal Colleges' examinations departments an option to allow switching to occur at lunchtime in an attempt to minimize some of these potential "unseen" variables.[13] The current study has reassuringly found that station reliability remained consistent regardless of whether switching occurred or not. Furthermore, reliability was maintained irrespective of the number of examiners changing stations. Additionally, no significant difference was found in the overall candidates' score distribution between AM and PM sessions and the categories of examiner switching, but we did find 3 days on which values approached significance (Fig. 5—days 2, 6, and 15). Interestingly, 1 of these days was where no switching had occurred (day 15) and 2 occurred on low switching days (less than 5 examiners switching). It is possible that on the no switching day, a group of candidates was significantly better than the other cohort, or that the examiners became more hawkish as the day proceeded. The most likely explanation for the 2 days where low examiner switching took place was replacement of hawk with dove examiners (and vice versa).

Weaknesses of the current study include its observational, retrospective, and non-randomized design. When considering randomization, it was clear that examiners would be unlikely to consent if they had a strong preference for switching or vice versa and that it would not be reasonable to ask candidates to participate in a randomized study that could conceivably affect their outcome in a high-stakes examination. Furthermore, we have not assessed examiner experience in the examination. The study was prompted by examiner feedback that is routinely collected, identifying fatigue and repetition as a potential problem and in need of further investigation. We have not included data on the degree of variation between the performance of the same station on different days in different sites (all of which have

been evaluated and shown to result in variation in scores but not overall candidate outcome). Although we have evidence of better examiner morale because of switching OSCE stations at lunchtime from a number of sources, not least the examiners themselves, a follow-up study is required to evaluate the level of morale more fully. Use of our validated questionnaire[13] would assist in further studies. In the meantime, Royal College examination departments now have the option to switch examiners and this practice has been completely adopted by the English College, which has the greatest candidate numbers.

CONCLUSIONS

With no detrimental findings related to station reliability and overall candidate score distribution, our findings have implications for the delivery of OSCE not just in surgery but across medical specialties. It is clearly important to maintain examiner morale and by providing a change of scenario during long examination days, the performance of examiners (both measurable and latent) is likely to be improved. Having identified a potential source of unwanted variance further work is needed in an attempt to reduce this further. Further preparation and calibration might be needed, but it is very difficult to completely standardize an OSCE.

ACKNOWLEDGMENTS

Authors are grateful to all the examiners who took part in this study, and the heads of examinations in the 3 Surgical Royal Colleges in England and Scotland. We would also like to thank Julia Merchant for her contribution to the introduction section.

REFERENCES

1. Khan KZ, Gaunt K, Ramachandran S, Pushkar P. The objective structured clinical examination (OSCE): AMEE Guide No. 81. Part II: organisation & administration. *Med Teach*. 2013;35(9):e1447-e1463.
2. Humphris GM, Kaney S. Examiner fatigue in communication skills objective structured clinical examinations. *Med Educ*. 2001;35(5):444-449.
3. McLaughlin K, Ainslie M, Coderre S, Wright B, Violato C. The effect of differential rater function over time (DRIFT) on objective structured clinical examination ratings. *Med Educ*. 2009;43(10):989-992.
4. Schwartzman E, Hsu DI, Law AV, Chung EP. Assessment of patient communication skills during OSCE: examining effectiveness of a training program in minimizing inter-grader variability. *Patient Educ Couns*. 2011;83(3):472-477.
5. 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.
6. McManus IC, 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;18(6):42.
7. Finn Y, Cantillon P, Flaherty G. Exploration of a possible relationship between examiner stringency and personality factors in clinical assessments: a pilot study. *BMC Med Educ*. 2014;14:1052.
8. Esmail A, Roberts C. Academic performance of ethnic minority candidates and discrimination in the MRCGP examinations between 2010 and 2012: analysis of data. *Br Med J*. 2013;347:f5662.
9. Denney ML, Freeman A, Wakeford R. MRCGP CSA: are the examiners biased, favouring their own by sex, ethnicity, and degree source? *Br J Gen Pract*. 2013;63(616):718-725.
10. Sokol DK. The exam scam. *Br Med J*. 2014;349:g4808.
11. Nasir AA, Yusuf AS, Abdur-Rahman LO, Babalola OM, Adeyeye AA, Popoola AA, et al. Medical students' perception of objective structured clinical examination: a feedback for process improvement. *J Surg Educ*. 2014;71(5):701-706.
12. Brennan PA, Sherman KP. The MRCS examination—an update on the latest facts and figures. *Br J Oral Maxillofac Surg*. 2014;52(10):881-883.
13. Brennan PA, Konieczny K, Groves J, Parker M, Sherman KP, Foulkes J, et al. Development, validation and initial outcomes of a questionnaire to examine human factors in postgraduate surgical objective structured clinical examinations. *Br J Surg*. 2015;102(4):423-430.
14. Chesser A, Cameron H, Evans P, Cleland J, Boursicot K, Mires G. Sources of variation in performance on a shared OSCE station across four UK medical schools. *Med Educ*. 2009;43(6):526-532.
15. Yeates P, Moreau M, Eva K. Are examiners' judgments in OSCE-style assessments influenced by contrast effects? *Acad Med*. 2015;90(7):975-980.
16. Hope D, Cameron H. Examiners are most lenient at the start of a two-day OSCE. *Med Teach*. 2015;37(1):81-85.

17. Brannick MT, Erol-Korkmaz HT, Prewett M. A systematic review of the reliability of objective structured clinical examination scores. *Med Educ*. 2011;45(12):1181-1189.

18. Wilkinson TJ, Frampton CM, Thompson-Fawcett M, Egan T. Objectivity in objective structured clinical examinations: checklists are no substitute for examiner commitment. *Acad Med*. 2003;78(3):219-223.

19. McManus IC, Elder AT, Dacre J. Investigating possible ethnicity and sex bias in clinical examiners: an analysis of data from the MRCP(UK) PACES and nPACES examinations. *BMC Med Educ*. 2013;30 (13):103.

20. Byrne A, Tweed N, Halligan C. A pilot study of the mental workload of objective structured clinical examination examiners. *Med Educ*. 2014;48(3):262-267.

21. Chan J, Humphrey-Murto S, Pugh DM, Su C, Wood T. The objective structured clinical examination: can physician-examiners participate from a distance? *Med Educ*. 2014;48(4):441-450.