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

Article (Accepted Version)


This version is available from Sussex Research Online: http://sro.sussex.ac.uk/id/eprint/60522/

This document is made available in accordance with publisher policies and may differ from the published version or from the version of record. If you wish to cite this item you are advised to consult the publisher's version. Please see the URL above for details on accessing the published version.

Copyright and reuse:
Sussex Research Online is a digital repository of the research output of the University.

Copyright and all moral rights to the version of the paper presented here belong to the individual author(s) and/or other copyright owners. To the extent reasonable and practicable, the material made available in SRO has been checked for eligibility before being made available.

Copies of full text items generally can be reproduced, displayed or performed and given to third parties in any format or medium for personal research or study, educational, or not-for-profit purposes without prior permission or charge, provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way.
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,* Evan 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 8·3–4·2. © 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
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.5,10 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).

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
Committee.

with the agreed policy of the Intercollegiate Examinations
in the latter group changed scenarios in accordance
had examiner switching). Realistically, almost all examiners
(where more than 5 of the 18 stations in a particular day
examiner switching took place),

the PM session then switching was deemed to have
morning (AM) and afternoon (PM) session were included in
lunchtime, only examining days that included both a
that might be attributable to examiners switching stations at
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 amalga-
mated 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 = 435)—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 interac-
tions 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 inter-
actions, 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
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), ($F [2,26] = 1.33, p = 0.280$).

### 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]

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

SD, standard deviation.

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.
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.

### TABLE 3. Number of Examining Days in Each Switching Category

<table>
<thead>
<tr>
<th>Examination Session</th>
<th>High</th>
<th>Low</th>
<th>None</th>
</tr>
</thead>
<tbody>
<tr>
<td>May 2015</td>
<td>5</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Feb 2015</td>
<td>4</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>Oct 2014</td>
<td>5</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>Total</td>
<td>14</td>
<td>10</td>
<td>5</td>
</tr>
</tbody>
</table>

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. There are many possible reasons for this scoring variance including the examiners’ own perception of standards, and even personality factors. It is also possible that examiners become more
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. 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, 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.

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 appro-
priate 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. 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.
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 examina-
tion center.

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 tired-
ness, repetition, and even boredom) might influence exam-
iner 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. The current study has
reassuringly found that station reliability remained consist-
tent 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 consider-
ing 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

![Image](http://example.com/image.png)

**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 Most results.
be 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 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


