OSCEs FOR UNDERGRADUATE CLINICAL EXAMINATION IN ORTHOPAEDICS: INTER-EXAMINER VARIABILITY

B. M. Ndeleva, MMed Ortho (Mak), FCS (Ortho) ECSA, Orthopaedic Surgeon & Lecturer, Kenyatta University, Nairobi, Kenya

Correspondence to: Dr. B. M. Ndeleva, P.O. Box 1413 – 00606, Nairobi, Kenya. Email: benndeleva@gmail.com

ABSTRACT

Background: The traditional clinical examination has fallen into disfavour on account of considerable inter-examiner variability. The OSCE is gaining popularity as it is perceived to be less prone to this.
Objective: To establish whether inter-examiner variability is still a significant factor for the undergraduate orthopaedic clinical examination in our institution.
Method: Thirty three final year students were randomly divided into two groups of 17 and 16 students. Two standardized OSCE questions were administered to each student by four examiners with each group being examined by one lecturer for each of the questions. For the first question, students in Group 1 were examined by Examiner A while those in Group 2 were examined by Examiner B. For the second question, students in Group 1 were examined by Examiner C while those in Group 2 were examined by Examiner D. The scores for each student were tabulated and the range, mean, and pass rate determined for each of the examiners. The Student's t-test was calculated to determine if there was statistically significant inter-examiner variability.
Results: For Question 1, the mean score for students examined by Examiner A (Group 1) was 7.47 marks while that for Examiner B (Group 2) was 5.59 marks. The *p*-value was 0.01367 (95% confidence interval). For Question 2, the mean score for students examined by Examiner C (Group 1) was 7.32 marks while that for Examiner D (Group 2) was 8.625. The *p*-value was 0.001148 (95% confidence interval).
Conclusion: There was statistically significant inter-examiner variability. We recommend that for all OSCE exams, examiners be paired with a deliberate attempt to pair a "Hawk" with a "Dove". Statistical correction of biases is also recommended.

INTRODUCTION

The traditional clinical examination (i.e. long case and short case examination) has fallen into disfavour on account of considerable inter-examiner variability. In 1913, Sir William Osler observed that some examiners (the "Metalics") consistently gave lower scores to candidates while other examiners (the "Molluscoid") were in his own words "so soft and slushy that he has not the heart to reject the man", consistently gave higher scores (1). It is Fleming who is credited with the term "Hawk and Dove" which is used more widely (2). Other studies have demonstrated this same significant variability in examiner stringency afflicting the traditional clinical examination (3,4).

To address this challenge, Harden *et al* (5) developed the Objective Structured Clinical Examination (OSCE) in the late 1970's. The OSCE has been adopted in many of our medical schools. Its raising popularity is because it is perceived to be more reliable and valid (6-8).

However, OSCEs have been shown to also be afflicted by significant inter-examiner variability (9-11). McManus *et al* (12) in probably the largest study on variation in examiner leniency-stringency attributes 12% of the variation in examination scores to differences between examiners in leniency-stringency. Other studies have attributed an even greater contribution to variation in score due to examiner stringency. Hill and colleagues (13) found the contribution of this factor to be 29% while Harasym and colleagues (9) found it to be a much higher at 44.2%.

We sought to establish whether this holds true for our orthopaedic clinical examination administered to undergraduate students. Most of the studies on the variability of OSCEs have been done for postgraduate candidates. Due to the lower expectations for undergraduate students, the conclusions from these studies may not necessarily hold true for undergraduate candidates. The assumption is that the higher standards of the "hawks" may not necessarily apply for undergraduate candidates for whom expectations are lower.

MATERIALS AND METHODS

The study was conducted in the month of April 2016. The OSCE examination was administered on the 5th April 2016. Thirty three final year students were randomly divided into two groups (Group 1 with 17 candidates and Group 2 with 16 candidates). A standardized OSCE question was administered by Examiner A to each of the students in Group 1 and by

Examiner B to each of the students in Group 2. The same groups were retained and a second question administered to the students in Group 1 by Examiner C and to the students in Group 2 by Examiner D. The two questions had been set and moderated by all the four examiners and a marking scheme agreed upon. The examiners were to adhere to the marking scheme. Each station/question was marked out of 10 marks on a prepared standardized score sheet for each candidate.

The scores for each question were tabulated and the range, mean, and pass rate determined for each of the two groups of students based on whom their examiner was. The student *t*-test was calculated to determine if there was statistically significant inter-examiner variability.

RESULTS

The scores for all the students for the two questions are tabulated in Table 1.

Table 1
Score awarded by each of the examiners

Question 1		Question 2	
Examiner A (Group 1)	Examiner B (Group 2)	Examiner C (Group 1)	Examiner D (Group 2)
9	2.5	7.5	10
8	2	5.5	10
8	6	7.5	9.5
8.5	8.5	8.5	9.5
6.5	8	5.5	8
9.5	6	6.5	9
4	6	7.5	9
9.5	8	7	9
6	4.5	8	8
8	5	6	7.5
7	4.5	9.5	9.5
9	5	8.5	9
4	8	7.5	7.5
9	2	8.5	8
7.5	5.5	7	7
4	8	7.5	7.5
9.5		6.5	

The range, mean scores, and pass rates for each of the examiners are tabulated in Table 2.

Table 2
Range of marks awarded by the different examiners as well as their mean scores and pass rates

	Examiner A	Examiner B	Examiner C	Examiner D
Range	4 to 9.5	2 to 8.5	5.5 to 9.5	7 to 10
Mean score	7.47	5.59	7.32	8.625
% Pass rate	82%	68.75 %	100%	100%

The student t-test was performed on these scores to determine if the marks awarded by the different examiners demonstrated statistically significant inter-examiner variation.

For the first set of examiners (Examiners A & B) the *p*-value was 0.01367 (95% confidence interval). This indicates that there was a statistically significant difference in the scores awarded by the two examiners.

For the second set of examiners, (Examiners C & D), the *p*-value was 0.001148 (95% confidence interval). This indicated an even greater level of statistical significance in the difference in scores awarded by Examiners C and D even though both passed all the students that they examined.

DISCUSION

The finding that there was significant inter-examiner variability is similar to what other studies have shown (9-11). This is despite the fact that these studies were done amongst postgraduate students for whom expectations are higher and probably more prone to a broader range of expectations amongst different examiners.

Examiners A and B had more than two decades experience in conducting clinical examination while Examiner C and D had less than a decade experience in conducting clinical examinations. Though we had only 4 examiners, we can infer from our findings that examiner experience did not alter the likelihood of inter-examiner variability as both sets showed significant inter-examiner variability. This is in keeping with the finding that examiner stringency is stable over time though the studies that addressed this covered a short period of time (12,14).

To help mitigate the undesired effect of inter-examiner variability, examiners should be paired as recommended by McManus *et al* (12). An examiner known to be stringent should be paired with one known to be lenient from past experience. For instance, based on the results of this study, Examiner A should be paired with Examiner C and Examiner B paired with Examiner D in future examinations. Brannick and colleagues (15) also recommended the pairing of examiners.

Inter-examiner variability can also be addressed by correcting biases between examiners statistically i.e. standardization of marks (12). McManus *et al* (12) argued that it is more reliable to let examiners mark as they always do and then adjust marks appropriately to correct for any biases statistically rather than attempting to change the way they mark as fixed and unchanging biases can reliably be corrected.

The observation that there was statistically significant variability in the scores by examiner C and D even though all the students they examined passed brings to the fore the need to analyze students scores more critically and not just based on whether they have passed or failed. This is because a more critical

analysis may reveal information that would otherwise pass unnoticed with the more basic analysis.

CONCLUSION AND RECOMMENDATION

OSCEs for undergraduate clinical examination in orthopaedic surgery are afflicted by the problem of inter-examiner variability.

We recommend that for all OSCE examinations, examiners be paired with a deliberate attempt to pair a "Hawk" with a "Dove" based on past observation of individual examiners stringency. Alternatively, marks should be standardized to correct for any biases.

REFERENCES

1. Osler, W. Examinations, examiners and examinees. *Lancet.* 1913; **182**:1047–1050.
2. Fleming, P.R., Manderson, W.G., Matthews, M.B., Sanderson PH and Stokes JF. Evolution of an examination: M.R.C.P. (U.K.) *Br Med J.* 1974; **2**(5910):99–107.
3. Wilkinson, T.J., Campbell, P.J. and Judd, S.J. Reliability of the long case. *Med Educ*. 2008; **42**(9):887–893.
4. Weisse, A.B. The oral examination: awesome or awful? *Perspect Biol Med.* 2002; **45**:569-578.
5. Harden, R.M., Stevenson, M., Downie, W.W. *et al*. Assessment of clinical competence using objective structured examination. *Br Med J.* 1975; **1**(5955):447–451.
6. Sloan, D.A., Donnelly, M.B., Schwartz, R.W., *et al*. The objective structured clinical examination. The new gold standard for evaluating postgraduate clinical performance. *Ann Surg.* 1995; **222**(6):735-742.
7. Cohen, R., Reznick, R.K., Taylor, B.R., *et al*. Reliability and validity of the objective structured clinical examination in assessing surgical residents. *Amer J Surg*. 1990; **160**(3): 302-305.
8. Matsell, D.G.1., Wolfish, N.M. and Hsu, E. Reliability and validity of the objective structured clinical examination in paediatrics. *Med Educ.* 1991; **25**(4):293-299.
9. Harasym, P.H., Woloschuk, W. and 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.
10. Turner, J.L. and Dankoski, M.E. Objective structured clinical exams: a critical review. *Fam Med*. 2008; **40**(8):574– 578.
11. Walters, K., Osborn, D. and Raven, P. The development, validity and reliability of a multimodality objective structured clinical examination in psychiatry. *Med Educ*. 2005; **39(**3):292–298.
12. McManus, I.C., Thompson, M. and 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**:42.
13. Hill, F., Kendall, K., Galbraith, K., *et al*. Implementing the undergraduate mini-CEX: a tailored approach at Southampton University. *Med Educ*. 2009; **43**(4):326–334.
14. McManus, I.C., Elder, A.T. and 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; **13**:103.
15. Brannick, M.T., Erol-Korkmaz, H.T. and Prewett, M. A systematic review of the reliability of objective structured clinical examination scores. *Med Educ*. 2011; **45**(12):1181-1189.