Inter-observer variance of examiner scoring in urology Objective Structured Clinical Examinations (OSCEs)

Naji J. Touma[1], Charles A. Paco[2], Iain MacIntyre[1]
[1]Department of Urology, Queen's University, Kingston, ON, Canada; [2]School of Medicine, Queen's University, Kingston, ON, Canada

Acknowledgement: *QUEST is supported by an unrestricted educational grant from the Canadian Urological Association.*

Cite as: Touma NJ, Paco CA, MacIntyre I. Inter-observer variance of examiner scoring in urology objective structured clinical examinations. *Can Urol Assoc J* 2023 December 21; Epub ahead of print. http://dx.doi.org/10.5489/cuaj.8571

Published online December 21, 2023

Corresponding author: Dr. Naji J. Touma**,** Department of Urology, Queen's University**,** Kingston Health Sciences Centre, Kingston, ON, Canada; naji.touma@kingstonhsc.ca

***

ABSTRACT

Introduction: The Objective Structured Clinical Examination (OSCE) is an attractive tool of competency assessment in a high-stakes summative exam. An advantage of the OSCE is the ability to assess more realistic context, content, and procedures. Each year, the Queen's Urology Exam Skills Training (QUEST) is attended by graduating Canadian urology residents to simulate their upcoming board exams. The exam consists of a written component and an OSCE. The aim of this study was to determine the inter-observer consistency of scoring between two examiners of an OSCE for a given candidate.

Methods: Thirty-nine participants in 2020 and 37 participants in 2021 completed four stations of OSCEs virtually over the Zoom platform. Each candidate was examined and scored independently by two different faculty urologists in a blinded fashion at each station. The OSCE scoring consisted of a checklist rating scale for each question. An intra-class correlation (ICC) analysis was conducted to determine the inter-rater reliability of the two examiners for each of the four OSCE stations in both the 2020 and 2021 OSCEs.

Results: For the 2020 data, the prostate cancer station scores were most strongly correlated (ICC 0.746, 95% CI 0.556–0.862, $p<0.001$). This was followed by the general urology station (ICC 0.688, 95% CI 0.464–0.829, $p<0.001$, the urinary incontinence station (ICC 0.638, 95% CI 0.403–0.794, $p<0.001$), and finally the nephrolithiasis station (ICC 0.472, 95% CI 0.183–0.686,

p<0.001). For the 2021 data, the renal cancer station had the highest ICC at 0.866 (95% CI 0.754–0.930, p<0.001). This was followed by the nephrolithiasis station (ICC 0.817, 95% CI 0.673–0.901, p<0.001), the pediatric station (ICC 0.809, 95% CI 0.660–0.897, p<0.001), and finally the andrology station (ICC 0.804, 95% CI 0.649–0.895, p<0.001). A Pearson correlation coefficient was calculated for all stations, and all show a positive correlation with global exam scores. It is noteworthy that some stations were more predictive of overall performance, but this did not necessarily mean better ICC scores for these stations.

Conclusions: Given a specific clinical scenario in an OSCE exam, inter-rater reliability of scoring can be compromised on occasion. Care should be taken when high-stakes decisions about promotion are made based on OSCEs with limited standardization.

INTRODUCTION

In 1975, Harden introduced the Objective Structured Clinical Examinations (OSCEs) as a tool of assessment in medical education. (1) Other assessments were suited at assessing knowledge, but no test alone could evaluate the combination of knowledge, skills, and behaviours required for functioning in a medical context. (2) The OSCE plays a complementary role in a “test battery” approach for a fulsome evaluation of performance in a simulated context. (3) OSCEs constitute an integral part of the Royal College certifying exam in Urology.

Among other things, OSCEs are thought to address the need for objectivity in medical assessments by minimizing the effects of a patient's performance, examiner bias, a non-standardised marking scheme and the candidate's actual performance. (4) The psychometric validity of OSCEs has been widely evaluated. (5) Issues of reliability, however, remain a concern. With a great deal at stakes in certifying exams, in particular, it is imperative to deliver valid and fair OSCEs that measure competence. Several factors have been found to impact the reliability of OSCEs including: the number of stations and testing time (6), the number of examiners per station (7), content specificity effects (8), and the scoring schema whether it be a global rating or a checklist (9).

One particular potential source of error is related to differences in examiner decision making. This is known colloquially as the “hawks and doves” effect. (10) This effect is mitigated by the overall length of the OSCE, randomization of assessors and students to balance out an individual assessor’s judgement. The extent of this variability in high stakes urology OSCEs is not described. Whereas the Royal College performs psychometric testing on all certifying exams, this data is not published; and therefore, unavailable to educators, and OSCE designers. In addition, a direct comparison of assessor ratings based on the same clinical scenario, the same candidate, and at the same time and location has not previously been described to our knowledge.

This study aims to examine the consistency of assessor scoring on urology OSCE stations. Given a specific clinical scenario and candidate, does the examiner make a difference?

METHODS

The Queen's Urology Exam Skills Training (QUEST) mock examination, has been held annually since 1997 for graduating Urology residents from across Canada. The exam aims to simulate the Royal College of Physicians and Surgeons of Canada (RCPSC) certifying exam with written and OSCE components. Until the pandemic, residents, representing all 12 Canadian urology training programs, travelled from across the country for this in-person event about 3 months prior to the RCPSC exam. The pre-pandemic OSCE consisted of 8 x 15 minutes stations with six out of eight stations being staffed by different examiners guiding the candidates through a clinical scenario, whilst scoring their performance. The other two stations were unstaffed visual recognition stations. Given the pandemic restrictions, the QUEST exam was moved to an online format for the December 2020, and December 2021 iterations. The OSCE component was offered on the ZOOM platform (Zoom Video Conferencing, San Jose, California) with each candidate examined for 1 hour over the same 4 clinical scenarios. This was done for logistical reasons, in order to minimize the risks of technical challenges by a candidate moving from one virtual room to the next. Each clinical scenario was 15 minutes long and scored independently by two examiners. Each examiner conducted two clinical scenarios, but scored all four independently of their peer. Examiners were not allowed to confer about their evaluations during or after each examination. Each examiner administered two stations and sat passively for two. All Four stations were scored in real time by both examiners. *The exam was provided to all examiners on the day before the exam for their review. Clarifications were provided on an ad hoc basis. Although all of the examiners were veteran examiners for QUEST having participated for many years, no other specific training was provided on how to score the stations beyond the checklists within the exam.* All examiners were Royal College Exam certified urologists with many years of experience examining OSCEs in general, and for the QUEST program, specifically. 50% came from community practices, and 50% from academic practices. The practice subspecialty areas of the 16 examiners were as follows: 8 General urology, 2 Uro-Oncology, 2 transplant, 2 Endourology, 1 Andrology, and 1 Reconstructive.

The OSCE scoring consisted of a checklist rating scale for each question. An intra class correlation (ICC) analysis was conducted to determine the inter-rater reliability of the two examiners for each of the four OSCE stations in both the 2020 and 2021 OSCEs.

RESULTS

39 candidates from 12 different Canadian urology post graduate training programs participated in the 2020 exam. The topics of the 4 stations in the 2020 OSCE were: 1. Nephrolithiasis, 2. Urinary Incontinence, 3. General Urology, and 4. Prostate cancer.

In 2021, 37 candidates again from all 12 Canadian programs participated. The topics for 2021 were: 1. Andrology, 2. Pediatrics, 3. Renal cancer, and 4. Nephrolithiasis/Prostate cancer. 16 examiners participated in each OSCE paired in a group of 2 as outlined above. The pairings were different in 2021 compared to 2020. All stations scores are outlined in percentages. The differences between the scores of the two examiners for each station in every exam cohort are

outlined in Table 1, and Table 2 respectively. The median difference in scoring between 2 examiners assessing the same candidate for the same station ranges from 3% to 6.7%. Intraclass correlation (ICC) values less than 0.5 are indicative of poor correlation, values between 0.5 and 0.75 indicate moderate correlation, values between 0.75 and 0.9 indicate good correlation, and values greater than 0.90 indicate excellent correlation. (5)

The nephrolithiasis station in the 2020 OSCE showed poor correlation between the examiners, whereas all other stations in 2020 (Urinary incontinence, general urology, and prostate cancer) showed moderate correlation. All stations in the 2021 OSCE showed good correlation.

The Pearson Correlation Coefficient, Rho, shows the correlation between the performance on a particular station and the overall exam score including the multiple-choice exam and the OSCE score. Rho values vary between -1 and +1. A positive value indicates a positive correlation and all stations showed positive predictability. However, a few stations showed very strong correlations (Rho > 0.75) including the general urology station in 2020, the renal cancer station in 2021, and the nephrolithiasis/Prostate cancer station in 2021. It is noteworthy that the predictability of a station does not necessarily mean better ICC scores.

DISCUSSION

The Intraclass correlations (ICC) between assessors examining the same candidate on the same station show mostly good correlation, but on occasion can only be moderate or even poor. This emphasizes the importance of constant vigilance when creating OSCEs, especially high stakes ones. Metrics that measure the quality of an OSCE have been outlined, and constant review of stations to make improvements is essential. (5) For instance, one aim of a good OSCE is internal consistency, whereby the better students do well across all stations. This is measured by Cronbach's alpha. (5) However, even when the ICC shows good correlation, the median differences between assessors can range from 3% to 5%. Whereas this may be an acceptable variance for a formative OSCE, it can be significant for a summative OSCE. This is especially the case if high stakes decisions, such as promotion to the next stage of a competence by design model, are being made on OSCEs with a low number of stations and poor internal consistency.

There has been some heterogeneity in the literature about the influence of different OSCE circuits, and exam sites on candidates' scores. A pilot study of the Medical Council of Canada (MCC) qualifying exam administered in 2 forms at 4 different sites showed little variance in scores at the different sites. (12) This was confirmed by a study examining the impact of site at the undergraduate level. (13) Conversely, one study asking examiners to rate videos obtained in an OSCE across 2 different sites found inter-examiner agreement at each site but a systemic difference of 6.7% between the two sites. They concluded that this variance may impact pass and fail rates. (14) Another report of an exam of physicians about to enter supervised practice across 21 sites in the USA found that examination site explained between 3.0% and 11.6% of variance. (15) A more recent assessment of the Medical Council of Canada (MCC) data found that site difference explained between 1.5% and 17.1% of score variability. (16) One challenge of all

these studies is that it is difficult to isolate true variance related to students' abilities versus an undesirable variance and a source of error related to different judgements rendered by different examiners. One advantage to our study is that it controls for students.

To address variability brought on by assessor judgements, three different solutions have been suggested: 1) improved faculty training, 2) increased post-test analyses to look for variability among circuits or sites, and, 3) station level enhancement with measurable psychometric improvements. (17) Some helpful tips to improve and standardize assessor judgement include: providing assessor training, refreshing those who were trained previously, providing assessors with support material, improving assessor briefings prior to the exam, and providing dummy runs before the formal assessment. (5) It is important to note that assessors are active information processors who are faced with the cognitive tasks of gathering, interpreting, integrating, and retrieving information for judgement and decision making. This complex process is influenced by their understanding of effective performance, personal biases, and interactions with the student, among other factors. (18)

One issue is whether this variance between assessors is seen in exams using global ratings scales versus the presently used checklist format. This study does not answer this question, but a previous *report suggested that global rating scales may perform better if administered by experts. (9)*

One limitation to this study is the novel method of delivering an OSCE virtually during the COVID-19 pandemic. *Issues of communication, and hearing the candidate properly can be one factor in virtual OSCEs that are not seen in in-person OSCEs.* This mode was new to all assessors, but a previous report suggested a high satisfaction level of this delivery mode by both assessors, and candidates. (19)

CONCLUSIONS

This study demonstrates a potential for variance in assessor evaluation in a urology OSCE. Even, when Intraclass concordance is considered good for a particular OSCE station, a median difference in score of between 3% and 5% can be seen for the same candidate when observed by two different assessors. Care should be taken when high stakes decision about promotion are made based on OSCEs with limited standardization.

REFERENCES

1. Harden RM, Gleeson FA. Assessment of clinical competence using an objective structured clinical examination (OSCE). *Med Educ* 1979;13:41. https://doi.org/10.1111/j.1365-2923.1979.tb00918.x
2. Epstein RM. Assessment in medical education. *N Eng J Med* 2007;356:387-96 https://doi.org/10.1056/NEJMra054784
3. Hamdy H, Prasad K, Williams R, Salih FA. Reliability and validity of the direct observation clinical encounter examination (DOCEE). *Med Educ* 2003;37:205-12 https://doi.org/10.1046/j.1365-2923.2003.01438.x
4. Kamran Z. Khan, Ramachandran S, Gaunt K, et al. The Objective Structured Clinical Examination (OSCE): AMEE Guide No. Part I: An historical and theoretical perspective. *Med Teach* 2013;35:e1437-46. https://doi.org/10.3109/0142159X.2013.818634
5. Pell G, Fuller R, Homer M, Roberts T; International Association for Medical Education. How to measure the quality of the OSCE: A review of metrics - AMEE guide no. *Med Teach* 2010;32:802. https://doi.org/10.3109/0142159X.2010.507716
6. Van der Vleuten CPM, Schuwirth LWT. Assessing professional competence: from methods to programmes. *Med Educ* 2005;39:309-17. https://doi.org/10.1111/j.1365-2929.2005.02094.x
7. Van der Vleuten CPM, Swanson DB. Assessment of clinical skills with standardized patients: state of the art. *Teach Learn Med* 1990;2:58-76. https://doi.org/10.1080/10401339009539432
8. Eva KW. On the generality of specificity. *Med Educ* 2003;37:587-8. https://doi.org/10.1046/j.1365-2923.2003.01563.x
9. Regehr G, MacRae H, Reznick RK, Szalay D. Comparing the psychometric properties of checklists and global rating scales for assessing performance on an OSCE-format examination. *Acad Med* 1998;73:993-7. https://doi.org/10.1097/00001888-199809000-00020
10. Harasym P, Woloschuk W, Cunning L. Undesired variance due to examiner stringency/leniency effect in communication skill scores assessed in OSCEs. *Adv Health Sci Educ* 13:617-32. https://doi.org/10.1007/s10459-007-9068-0
11. Koo TK, Li MY. A Guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med* 2016;15:155-63. https://doi.org/10.1016/j.jcm.2016.02.012
12. Reznick R, Smee S, Rothman A, et al. An objective structured clinical examination for the licentiate: report of the pilot project of the Medical Council of Canada. *Acad Med* 1992;67:487-94. https://doi.org/10.1097/00001888-199208000-00001
13. De Champlain AF, MacMillan MK, King AM, et al. Assessing the impacts of intra-site and inter-site checklist recording discrepancies on the reliability of scores obtained in a nationally administered standardized patient examination. *Acad Med* 1999;74:S52-4. https://doi.org/10.1097/00001888-199910000-00038
14. Tamblyn RM, Klass DJ, Schnabl GK, et al. Sources of unreliability and bias in standardized-patient rating. *Teach Learn Med* 1991;3:74-85. https://doi.org/10.1080/10401339109539486

15. Floreck LM, de Champlain AF. Assessing sources of score variability in a multi-site medical performance assessment: an application of hierarchical linear modeling. *Acad Med* 2001;76:S93-5. https://doi.org/10.1097/00001888-200110001-00031
16. Sebok SS, Roy M, Klinger DA, et al. Examiners and content and site: Oh My! A national organization's investigation of score variation in large-scale performance assessments. *Adv Health Sci Educ Theory Pract* 2015;20:581-94. https://doi.org/10.1007/s10459-014-9547-z
17. Fuller R, Homer M, Pell G, et al. Managing extremes of assessor judgment within the OSCE. *Med Teach* 2017;39:58-66. https://doi.org/10.1080/0142159X.2016.1230189
18. Govaerts M, van der Viel M, Schuwirth L, et al. Workplace-based assessment: raters' performance theories and constructs. *Adv Health Sci Educ Theory Pract* 2013;18:375-96 https://doi.org/10.1007/s10459-012-9376-x
19. Jenkins D, Nashed JY, Touma NJ. Virtual OSCE examinations during COVID-19: A 360 satisfaction assessment from examiners and candidates. *Can Urol Assoc J* 2023;17:E315-8. https://doi.org/10.5489/cuaj.8332

FIGURES AND TABLES

Table 1. Variability in assessor evaluation in the 2020 OSCE				
	Nephrolithiasis	**Incontinence**	**General urology**	**Prostate cancer**
Mean (%)	7.7	5.9	7.2	5.5
Median (%)	6.7	4.7	6.3	4.4
SD (%)	5.2	4.6	5	5.3
ICC	0.472	0.638	0.688	0.746
ICC (95% CI)	0.183-0.686	0.403-0.794	0.464-0.829	0.556-0.862
Rho	0.665	0.488	0.775	0.471

Mean, median, and standard deviation (SD) differences between the scores of 2 assessors for the same candidate. ICC: intra-class correlation, confidence interval is provided to $p<0.001$. OSCE: Objective Structured Clinical Examination; Rho: Pearson correlation coefficient.

Table 2. Variability in assessor scoring in the 2021 OSCE				
	Andrology	**Pediatrics**	**Renal cancer**	**Nephrolithiasis and prostate cancer**
Mean (%)	5.2	6	5.4	5.2
Median (%)	4.5	5	4	3
SD (%)	4.5	5.4	5.1	5.3
ICC	0.804	0.809	0.866	0.817
ICC (95% CI)	0.649-0.895	0.660-0.897	0.754-0.930	0.673-0.901
Rho	0.422	0.527	0.827	0.915

Mean, median, and standard deviation (SD) differences between the scores of 2 assessors for the same candidate. ICC: intra-class correlation, confidence interval is provided to $p<0.001$. OSCE: Objective Structured Clinical Examination; Rho: Pearson correlation coefficient.