Creative Education
2012. Vol.3, Special Issue, 937-942
Published Online October 2012 in SciRes (http://www.SciRP.org/journal/ce)
DOI:10.4236/ce.2012.326142

Inter-Rater Reliability: Comparison of Checklist and Global Scoring for OSCEs*

Bunmi S. Malau-Aduli[1], Sue Mulcahy[2], Emma Warnecke[1], Petr Otahal[3],
Peta-Ann Teague[4], Richard Turner[1], Cees Van der Vleuten[5]
[1]School of Medicine, Faculty of Health Science, University of Tasmania, Hobart, Australia
[2]Centre for the Advancement of Learning and Teaching, University of Tasmania, Hobart, Australia
[3]Menzies Research Institute, Hobart, Australia
[4]School of Medicine and Dentistry, Faculty of Medicine, Health and Molecular Sciences,
James Cook University, Townsville, Australia
[5]School of Health Professions Education, Faculty of Health, Medicine and Life Sciences, Maastricht University,
Maastricht, Netherlands
Email: bunmi.malauaduli@utas.edu.au, sue.mulcahy@utas.edu.au, emma.warnecke@utas.edu.au,
petr.otahal@utas.edu.au, peta.teague@jcu.edu.au, richard.turner@utas.edu.au,
c.vandervleuten@maastrichtuniversity.nl

Received July 31st, 2012; revised August 28th, 2012; accepted September 14th, 2012

Objective Structured Clinical Examinations (OSCEs) have been used globally in evaluating clinical competence in the education of health professionals. Despite the objective intent of OSCEs, scoring methods used by examiners have been a potential source of measurement error affecting the precision with which test scores are determined. In this study, we investigated the differences in the inter-rater reliabilities of objective checklist and subjective global rating scores of examiners (who were exposed to an online training program to standardise scoring techniques) across two medical schools. Examiners' perceptions of the e-scoring program were also investigated. Two Australian universities shared three OSCE stations in their end-of-year undergraduate medical OSCEs. The scenarios were video-taped and used for on-line examiner training prior to actual exams. Examiner ratings of performance at both sites were analysed using generalisability theory. A single facet, all random persons by raters design [PxR] was used to measure inter-rater reliability for each station, separate for checklist scores and global ratings. The resulting variance components were pooled across stations and examination sites. Decision studies were used to measure reliability estimates. There was no significant mean score difference between examination sites. Variation in examinee ability accounted for 68.3% of the total variance in checklist scores and 90.2% in global ratings. Rater contribution was 1.4% & 0% of the total variance in checklist score and global rating respectively, reflecting high inter-rater reliability of the scores provided by co-examiners across the two schools. Score variance due to interaction and residual error was larger for checklist scores (30.3% vs 9.7%) than for global ratings. Reproducibility coefficients for global ratings were higher than for checklist scores. Survey results showed that the e-scoring package facilitated consensus on scoring techniques. This approach to examiner training also allowed examiners to calibrate the OSCEs in their own time. This study revealed that inter-rater reliability was higher for global ratings than for checklist scores, thus providing further evidence for the reliability of subjective examiner ratings.

Keywords: Objective Structured Clinical Examination; Inter-Rater Reliability; Checklist Scores; Global Ratings

Introduction

The Objective Structured Clinical Examination (OSCE) is recognised by medical educators as an opportunity to evaluate essential clinical skills and competencies necessary for progression in the medical course (Harden & Gleeson, 1979; Hodges, 2003; Newble, 2004). Its widespread use to surmount many of the inherent validity problems of oral clinical examinations is due to its desirable characteristics of objective testing in which examinees are exposed to the same test conditions (Harden et al., 1975; Kirby & Curry, 1982; Downing & Yudkowsky, 2009).

The OSCE format comprises a student rotating through a series of time limited clinical "stations". At each station the student is faced with a simulated scenario, usually involving a simulated patient (SP). The student has to perform the required clinical task under the direct observation of a clinical assessor (examiner), who scores student performance against a checklist and/or global rating scale. There is a body of research on the use of checklists, which describe precisely the occurrence of particular behaviours and global rating scales which describe the quality of a performance, allowing for more interpretation by the examiner (Regehr et al., 1999; Hodges et al., 1999; Hodges et al., 2002). Checklists are designed and incorporated into OSCE to increase the objectivity and reliability of marking by different examiners. However some researchers have criticised the validity of checklists due to their tendency to become objectified and trivial in the evaluation of clinical competence (Van der Vleuten et al., 1991; Cohen et al., 1997; Cunnington

*Declaration of Interest: The authors report no conflicts of interest.

et al., 1997; Cushing, 2002). These authors have demonstrated the reliability and validity of global rating scales, thereby providing evidence that subjectivity may not be inherently unreliable. Global ratings have also been reported to better evaluate the performance of advanced students as well as negate some of the nuances associated with checklists (Van der Vleuten et al., 1991; Regehr et al., 1998; Hodges et al., 1999). Some studies have compared the psychometric properties of checklists and global rating scales on OSCEs and concluded that global rating scales scored by experts showed higher inter-station reliability, better construct validity and better concurrent validity than did checklists (Hodges et al., 1997; Regehr et al., 1998).

Intensive examiner training improves inter-rater reliability as it ensures that all raters interprete item descriptions similarly and apply similar standards on students' performance (Williams et al., 2003; Spencer & Silverman, 2004). Although earlier studies have indicated that examiner training varied in effectiveness as a function of medical experience (Newble et al., 1980; Van der Vleuten et al., 1989), more recent studies have demonstrated the high impact of examiner training on the consistency of scoring (Humphrey-Murto et al., 2005; Chesser et al., 2009)

However, establishing excellent examiner training sessions still remains a major problem for medical schools with increasing number of students, difficulty finding sufficient number of experienced examiners for multi-site exams and the challenges of getting time-poor clinicians away from their other activities to attend examiner-training sessions. Innovative and feasible approaches to tackling these tasks are necessary. The primary purpose of this study was to compare the inter-rater reliabilities of checklist and global rating scores of examiners who were exposed to an online training program (to standardise scoring techniques) across two medical schools. The study also examined examiners' perceptions of the feasibility and usability of the e-scoring program.

Methods

Study Context

In November 2010, two Australian medical schools (A and B) participated in a collaborative inter-school study of clinical competence in which three OSCE stations were developed and embedded in the (3rd and 4th years respectively) end-of-year clinical examinations. School A runs a five-year undergraduate medical programme, while School B runs a six-year undergraduate programme. Both schools have similar horizontally and vertically integrated outcomes-based curricula. The selected year groups were chosen because of their comparable levels of intended learning outcomes.

The Shared OSCE Stations

The three OSCEs (chest pain, diabetic foot and gallstones) comprised of eight-minute stations and were administered to a total of 119 third year medical students at School A and 94 fourth year medical students at School B. The three OSCE stations covered a range of core clinical competencies with which examiners at both schools were familiar. Between five to nine task-specific checklist items were developed for each case. The behaviourally anchored 4 - 7-point rating scales assessed degree of coherence, empathy, verbal and non-verbal expressions.

Examination Procedure

The examination at School A was conducted over a two-day period to two different cohorts of students, while at School B it was a one day event with the three shared OSCEs embedded in a 12-OSCE station examination. Two concurrent sessions of each station were conducted at School A and four were conducted at School B, each with one SP and one examiner. Clearance was obtained from the relevant ethics committee for this study.

Examiners

Three examiners were independently selected from each school to serve as external examiners, one on each of the shared stations, and double mark with the internal examiners at the other school. Each external examiner independently double marked a total of 20 student observations. Each examiner rated student performance by first scoring the task-specific checklist and then completing a global rating. The two components were then summed to generate an overall performance score.

Examiner Training

To aid examiner training and standardise marking across the two examination sites, an OSCE e-scoring tool was developed and set up in a secure intranet site, in the on-line Blackboard Learning System Vista environment. The three shared OSCE scenarios were videotaped and used for the on-line examiner training; PGY1 residents (interns) were recruited to role play as medical students and SPs were recruited from the SP pool. Informed consent and confidentiality agreement were obtained from all the video participants.

A total of 24 examiners were involved in the on-line OSCE training program. All the internal (on only the shared OSCEs) and external examiners were invited via email, given login access and instructions on how to use the program; the video clips were made accessible to the examiners one week prior to the examination. The examiners were able to view the recordings in their own time and assess the interns' performances.

Each examiner was asked to watch two unlabelled scenarios (poor and good performance) of the OSCE case which they had been assigned to examine. After watching each scenario, they were required to assess the performance using the marking sheet that was provided in another window. The station information and criteria for marking were also made available. After completing and submitting their marking/scoring sheet, the examiners were then able to view and compare the scores they had given for the checklist task and the global rating with others already submitted. This enabled examiners at both sites to achieve consensus regarding what constituted unsatisfactory, borderline or satisfactory performance. The SPs on the shared OSCE stations were allowed to view the video clips and they discussed face-to-face with the internal examiners about expected performance.

Statistical Analysis

Quantitative Data

Descriptive statistics of the on-line training scores, comparative analysis for checklist scores and global ratings in both schools were calculated using SAS. The difference between internal and external examiners' scores was tested using 2-sample t-test. Generalisability analysis was used to test for inter-rater reliability across sites. Multilevel mixed-effects linear regression in STATA was used to calculate the variance components and to evaluate the magnitude of the different sources of variation affecting the measurement. Different pairs of raters assessed examinees at each of the three stations and the examination at school A was conducted over two days with a different cohort on each day. Due to the disconnected design, variance components for each station within each site were estimated separately and the estimates were pooled across sites to eliminate confounding of the proficiency of examinee groups and the stringency of examiner groups across sites. For both checklist scores and global ratings, a single facet, random, raters/examiners (R) by persons/examinees (P) design [PxR] and the interaction effect of person by rater with residual effect (PxR,e) was used to assess inter-rater reliability. D-study was used to measure reliability estimates.

Qualitative Data

To capture their perceptions of the on-line training/e-scoring program, examiners were prompted to provide anonymous responses to four open-ended on-line survey questions which were administered to them immediately after completing their scoring of the OSCE scenarios. The examiners were asked to 1) comment on aspects they liked most about the e-scoring program; 2) comment on aspects they didn't like; 3) proffer suggestions on improvement of the program and 4) provide their views on the effect of the program on future assessments. The survey data were collated and emerging themes independently coded and confirmed by two researchers. Illustrative quotes are reported verbatim in Appendix 1.

Results

Table 1 portrays the mean checklist scores and global ratings ± the standard deviation (SD) given by co-examiners during the actual examination. There were no statistical differences in the mean scores given by the internal and external examiners in both schools.

The estimated variance components from generalisability analyses for checklist scores and global ratings are presented in **Table 2**. Pooled score variance attributed to student ability was higher on global ratings in comparison to checklist scores

Table 1.
Descriptive statistics for checklist scores and global ratings at both sites (mean scores ± standard deviation).

Station	Examiner	School A checklist score	School B checklist score	School A global rating	School B global rating
Chest pain	Internal	74.3 ± 9.7	70.0 ± 9.9	4.3 ± 1.0	4.3 ± 1.0
N = 20	External	71.15 ± 10.2	72.6 ± 10.6	4.4 ± 0.8	4.3 ± 0.8
Diabetic foot	Internal	65.0 ± 12.3	69.0 ± 15.2	3.5 ± 1.3	3.6 ± 1.4
N = 20	External	63.3 ± 13.1	67.7 ± 14.9	3.4 ± 1.3	3.6 ± 1.4
Gallstones	Internal	72.0 ± 11.9	77.0 ± 12.4	4.2 ± 1.0	4.4 ± 1.0
N = 20	External	72.15 ± 10.4	75.4 ± 11.1	4.1 ± 0.9	4.2 ± 1.2
Total Scores		69.6 ± 11.3	71.9 ± 12.4	3.9 ± 1.1	4.0 ± 1.1

Table 2.
Variance component estimates and G coefficients for checklist scores and global ratings.

		Checklist Scores					Global ratings				
		Variance component estimates*			G coefficients as a function of raters		Variance component estimates*			G coefficients as a function of raters	
School	Station	P	R	PxR, e	1	2	P	R	PxR,e	1	2
A	1	66.88	0	43.35	0.607	0.755	0.733	0	0.075	0.907	0.951
	2	81	0	54.8	0.596	0.747	0.717	0	0.15	0.827	0.905
	3	110.79	5.38	14.87	0.882	0.937	1.553	0.003	0.172	0.9	0.947
	Combined stations	258.66	5.38	113.02	0.696	0.821[a]	3.003	0.003	0.397	0.883	0.938[a]
	% variation	68.60%	1.40%	30.00%			88.20%	0.10%	11.70%		
B	1	69.87	0	53.55	0.566	0.723	0.749	0	0.05	0.937	0.968
	2	78.39	0	54.03	0.592	0.744	0.822	0	0.15	0.846	0.916
	3	124.25	5.38	14.87	0.893	0.944	1.895	0	0.1	0.95	0.974
	Combined stations	272.52	5.38	122.44	0.69	0.817[a]	3.466	0	0.3	0.92	0.959[a]
	% variation	68.10%	1.30%	30.60%			92.00%	0.00%	8.00%		

Note: [a]G-coefficients for this study with 2 raters; *Variance component estimates for persons (P); raters (R); and residual (PR,e), reflecting variance due to person-by-rater interaction (PR) and unidentified sources of error.

(90.2% vs 68.3%). Rater effect accounted for 1.4% and 0% of total variance in checklist score and global rating respectively. Score variance due to interaction and residual error was larger for checklist scores (30.3% vs 9.7%) than for global ratings.

G coefficients for checklist scores and global ratings are also presented in **Table 2**. G coefficients varied from each case, with the lowest values been obtained on the diabetic foot station across the two schools. In addition, reliability estimates for the global ratings were higher than for the checklists.

Survey results showed that examiners valued the process because it gave them an opportunity to see a "dry run" of the station and allowed them to set the "expected standard" for the station prior to the actual exam (Appendix 1). They also indicated that this sort of tool should be used more widely in OSCEs. However, they pointed out that scoring borderline performance, rather than good or poor performance would make the e-scoring process more useful.

Discussion

The observed low variance in rater effect in our study indicates high inter-rater reliability, meaning each rater's scores are consistent across different students. The results also indicate that there are no significant differences in average scores across raters; hence the assessment clearly reveals the competence of each examinee. Our results show higher inter-rater agreement for global ratings in comparison to checklist scores. A growing body of literature has reported that global ratings have higher reliability than checklist scores and are better able to discriminate between examinees (Hodges et al., 1999; Govaerts et al., 2002; Hodges et al., 2003; Wilkinson et al., 2003). The higher examinee and lower residual variance estimates observed in the global ratings in this study in comparison to the checklist scores echoes these findings.

McManus et al. (2006) reported that thorough selection, monitoring and training did not eliminate examiner stringency/ leniency effect. However, our study indicates otherwise, with the observed lower variance due to examiner difference. This might be as a result of the online training, which allowed examiners to agree on the "expected standard" for each station prior to the actual examination. The use of two examiners to reduce examiner bias has been proposed (Norcini, 2002; Wilkinson et al., 2003), but our findings clearly demonstrate that using on-line examiner training, higher reliabilities of 0.7 and above for high stakes examinations can be achieved even with the use of one examiner per station, indicating that there is little or no benefit in using examiners to double mark. Interestingly, our study showed that external examiners gave lower scores than internal examiners; this may indicate the effect of examiner familiarity with candidates as a potential source of bias (Stroud et al., 2011).

Researchers have suggested that variability in performance across cases is not simply related to content variation, but to other factors, such as pattern recognition based on irrelevant contextual features of the case (Govaerts et al., 2002). The observed varying magnitudes of estimated variance components across stations (cases) may indicate that the relative ordering of cases and the specificity of case content have a large effect on the variance. There is therefore the need to explore the magnitude of variance attributable to case, content and/or context specificity.

The survey results showed that the e-scoring program offered training for both quality assurance and appraisal purposes. The examiners valued the process as it allowed them to reach consensus about their scoring techniques and resulted in similar trends of scoring in both schools. Furthermore, given the busy schedule of clinicians and the challenges of getting away from their other activities to attend examiner-training sessions, the e-scoring package allowed examiners to use it in their own time. Most of them found it easy to navigate through the program, but a few expressed difficulties in understanding the technology as well as the statistics generated for comparison of scores.

The examiners also suggested that scoring of borderline performances would be more useful, indicating that it was easier for the examiners to identify and agree on their ratings, particularly for good performance. This is a valid point, given the fact that borderline students are the ones medical educators are most concerned about. It is important for examiners to be able to make accurate pass/fail decisions so that only competent students are allowed to progress academically. On the whole, the examiners concurred on the efficacy and possibility of wider use of the e-scoring program.

The major limitation of this study is the small number of stations used. In addition, the rating of the global scales after the checklists could have affected examiner scoring of student performance. Due to the design of the study, inter-case reliability and the comparison between trained and non-trained examiners could not be determined. Further studies should explore these areas.

Conclusion

The results of this study suggest that global rating scales are a more appropriate summative measure than checklists in assessing examinees on performance based tests, providing further support for the reliability of subjective examiner judgments. This study also indicates possible elimination of examiner variance measurement error with the use of on-line examiner training program. The tool holds great promise for high stakes performance-based assessments conducted across multiple sites and will afford time-poor geographically separated clinicians the opportunity to better engage in the assessment process.

Acknowledgements

The authors would like to thank Jo Hanuszewicz, Di Madden, Kathy Spencer, Felicity Ey, Donna-May Brown, Kaspar Willson, Matt Holmes, Milford McArthur, Theresa Mokry, Stefan Blechinger, Leslee Wells, Gail Richardson and Florence Schaeffer for their contributions to the video recordings and data collation. The authors also acknowledge the contributions of the examiners.

REFERENCES

Chesser, A., Cameron, H., Evans, P., Cleland, J., Boursicot, K., & Mires, G. (2009). Sources of variation in performance on a shared OSCE station across four UK medical schools. *Medical Education, 43,* 526-532. doi:10.1111/j.1365-2923.2009.03370.x

Cohen, D. S., Colliver, J. A., Robbs, R. S., & Swartz, M. H. (1997). A large-scale study of the reliabilities of checklist scores and ratings of interpersonal and communication skills evaluated on a standardised-patient examination. *Advances in Health Science Education, 1,* 209-213. doi:10.1023/A:1018326019953

Cunnington, J. P. W., Neville, A. J., & Norman, G. R. (1997). The risks of thoroughness: Reliability and validity of global ratings and checklists in an OSCE. *Advances in Health Science Education, 1,* 27-33.

Cushing, A. (2002). Assessment of non-cognitive factors. In G. R. Norman, C. P. M. van der Vleuten, & D. I. Newble (Eds.), *International handbook of research in medical education* (pp. 711-755). Dordrecht: Kluwer Academic Publishers. doi:10.1007/978-94-010-0462-6_27

Downing S., & Yudkowsky R. (2009). *Assessment in health professions education.* London: Routledge.

Govaerts, M. J. B., Van der Vleuten, C. P. M., & Schuwirth, L. W. T. (2002). Optimising the reproducibility of a performance-based assessment test in midwifery education. *Advances in Health Science Education, 7,* 133-145. doi:10.1023/A:1015720302925

Harden, R. M., Stevenson, M., Downie, W. W., & Wilson, G. M. (1975). Assessment of clinical competence using objective structured examination. *British Medical Journal, 1,* 447-451. doi:10.1136/bmj.1.5955.447

Harden, R. M., & Gleeson, F. A. (1979). Assessment of clinical competence using an objective structured clinical examination (OSCE). *Medical Education, 13,* 41-54. doi:10.1111/j.1365-2923.1979.tb00918.x

Hodges, B., Regehr, G., Hanson, M., & McNaughton, N. (1997). An objective structured clinical examination for evaluating psychiatric clinical clerks. *Academic Medicine, 72,* 715-721. doi:10.1097/00001888-199708000-00019

Hodges, B., Regehr, G., McNaughton, N., Tiberius, R., & Hanson, M. (1999). Checklists do not capture increasing levels of expertise. *Academic Medicine, 74,* 1129-1134. doi:10.1097/00001888-199910000-00017

Hodges, B., McNaughton, N., Regehr, G., Tiberius, R., & Hanson, M. (2002). The challenge of creating new OSCE measures to capture the characteristics of expertise. *Medical Education, 36,* 742-748. doi:10.1046/j.1365-2923.2002.01203.x

Hodges, B., & McIlroy, J. H. (2003). Analytic global OSCE ratings are sensitive to level of training. *Medical Education, 37,* 1012-1016. doi:10.1046/j.1365-2923.2003.01674.x

Humphrey-Murto, S., Smee, S., Touchie, C., Wood, T. J., & Blackmore, D. E. (2005). A comparison of physician examiners and trained assessors in a high-stakes OSCE setting. *Academic Medicine, 80,* S59-S62. doi:10.1097/00001888-200510001-00017

Kirby, R. L., & Curry, L. (1982). Introduction of an objective structured clinical examination (OSCE) to an undergraduate clinical skills programme. *Medical Education, 16,* 362-364. doi:10.1111/j.1365-2923.1982.tb00951.x

McManus, I. C., Thompson, M., & Mollon, J. (2006). Assessment of examiner leniency and stringency ("hawk-dove effect") in the MRCP (UK) clinical examination (PACES) using multi-facet Rasch modelling. *BMC Medical Education, 6,* 1272-1294.

Newble, D. (2004). Techniques for measuring clinical competence: Objective structured clinical examinations. *Medical Education, 38,* 199-203. doi:10.1111/j.1365-2923.2004.01755.x

Newble, D. I., Hoare, J., & Sheldrake, P. F. (1980). The selection and training of examiners for clinical examinations. *Medical Education, 14,* 345-349. doi:10.1111/j.1365-2923.1980.tb02379.x

Norcini, J. J. (2002). The death of the long case? *British Medical Journal, 324,* 408-409. doi:10.1136/bmj.324.7334.408

Regehr, G., Freeman, R., Hodges, B., & Russell, L. (1999). Assessing the generalisability of OSCE measures across content domains. *Academic Medicine, 74,* 1320-1322. doi:10.1097/00001888-199912000-00015

Regehr, G., MacRae, H., Reznick, R. K., & Szalay, D. (1998). Comparing the psychometric properties of checklists and global rating scales for assessing performance on an OSCE-format examination. *Academic Medicine, 73,* 993-997. doi:10.1097/00001888-199809000-00020

SAS (2009). *Statistical Analysis System*. Cary, CA: SAS Institute.

Spencer, J. A., & Silverman, J. (2004). Communication education and assessment: Taking account of diversity. *Medical Education, 38,* 116-118. doi:10.1111/j.1365-2923.2004.01801.x

StataCorp. (2011). *Stata Statistical Software: Release 12.* College Station, TX: StataCorp LP.

Stroud, L., Herold, J., Tomlinson, G., & Cavalcanti, R. B. (2011). Who you know or what you know? Effect of examiner familiarity with residents on OSCE scores. *Academic Medicine, 86,* S8-S11. doi:10.1097/ACM.0b013e31822a729d

Van der Vleuten, C. P. M., Van Luyk, S. J., Van Ballegooijen, A. M. J., & Swanson, D. B. (1989). Training and experience of examiners. *Medical Education, 23,* 290-296. doi:10.1111/j.1365-2923.1989.tb01547.x

Van der Vleuten, C. P. M., Norman, G. R., & De Graaff, E. (1991). Pitfalls in the pursuit of objectivity: Issues of reliability. *Medical Education, 25,* 110-118. doi:10.1111/j.1365-2923.1991.tb00036.x

Wilkinson, T. J., Frampton, C. M., Thompson-Fawcett, M., & Egan, T. (2003). Objectivity in objective structured clinical examinations: Checklists are no substitute for examiner commitment. *Academic Medicine, 78,* 219-223. doi:10.1097/00001888-200302000-00021

Williams, R. G., Klamen, D. A., & McGaghie, W. C. (2003). Cognitive, social and environmental sources of bias in clinical competence ratings. *Teaching and Learning Medicine, 15,* 270-292. doi:10.1207/S15328015TLM1504_11

Appendix 1

Survey Findings

Aspects liked most about the e-scoring program:

- Having the opportunity to see a "dry run" of the station;
- Easy to view DVD;
- Reasonably easy to work, and
- Scenario Information was well presented prior to the actual case.

Aspects not liked about the e-scoring program:

- A bit tricky to understand the technology but once I had worked it out it was fine;
- Really poor student and really good one—might be better to have one in between,
- I think a discussion with other examiners immediately after marking both candidates would be beneficial for me;
- Difficulty juggling the various windows.

Suggestions on improvement:

- Great idea—nice to see scenarios and grade them before the day of the exam, takes away the issue of taking the first few scenarios to get comfortable with it;
- I would have found it more useful to have candidates that were borderline in performance, rather than see candidates that were clearly very good or clearly very poor;
- Make the feedback in pictorial form i.e., this is where you are on the graph;
- Start with the good candidate for better standardisation.

Effect of program on future assessments:

- Hope to use this sort of tool more widely in OSCEs;
- Helps set the expected standard;
- I found it very useful to reflect on my assessment of students, particularly how I would approach a candidate who was really better than expected with his verbal communication and approach to patient-focused examination, but might not necessarily have got all the marks he deserved because of time constraints or the criteria of the marking sheet-I guess this is where the global score comes into it;
- It may obviate the need for time-poor examiners being available real-time for OSCEs—if all stations could be filmed. It would be much more preferable than spending all Saturday in a stuffy clinic cubicle!
- Very little as I haven't understood the feedback.