

[Home](#) > [Journal](#) > [Social Sciences & Humanities](#) > [CE](#)[Indexing](#) [View Papers](#) [Aims & Scope](#) [Editorial Board](#) [Guideline](#) [Article Processing Charges](#)[CE](#) > [Vol.3 No.6A, October 2012](#)

OPEN ACCESS

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

PDF (Size: 70KB) PP. 937-942 DOI: 10.4236/ce.2012.326142

Author(s)

Bunmi S. Malau-Aduli, Sue Mulcahy, Emma Warnecke, Petr Otahal, Peta-Ann Teague, Richard Turner, Cees Van der Vleuten

ABSTRACT

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

Cite this paper

Malau-Aduli, B. , Mulcahy, S. , Warnecke, E. , Otahal, P. , Teague, P. , Turner, R. & Vleuten, C. (2012). Inter-Rater Reliability: Comparison of Checklist and Global Scoring for OSCEs. *Creative Education*, 3, 937-942. doi: 10.4236/ce.2012.326142.

References

- [1] 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
- [2] 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
- [3] Cunnington, J. P. W., Neville, A. J., & Norman, G. R. (1997). The risks of thoroughness: Reliability and

- [Open Special Issues](#)
- [Published Special Issues](#)
- [Special Issues Guideline](#)

[CE Subscription](#)[Most popular papers in CE](#)[About CE News](#)[Frequently Asked Questions](#)[Recommend to Peers](#)[Recommend to Library](#)[Contact Us](#)

Downloads: 166,690

Visits: 373,859

Sponsors >>

[The Conference on Information Technology in Education \(CITE 2012\)](#)

- [4] 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
- [5] Downing S., & Yudkowsky R. (2009). *Assessment in health professions education*. London: Routledge.
- [6] 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
- [7] 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
- [8] 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
- [9] 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
- [10] 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
- [11] 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
- [12] 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
- [13] 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
- [14] 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
- [15] 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.
- [16] 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
- [17] 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
- [18] Norcini, J. J. (2002). The death of the long case? *British Medical Journal*, 324, 408-409. doi:10.1136/bmj.324.7334.408
- [19] 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
- [20] 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
- [21] SAS (2009). *Statistical Analysis System*. Cary, CA: SAS Institute.
- [22] 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

- [23] StataCorp. (2011). Stata Statistical Software: Release 12. College Station, TX: StataCorp LP.
- [24] 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
- [25] 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
- [26] 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
- [27] 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
- [28] 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