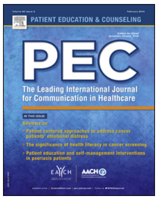
Patient Education and Counseling 100 (2017) 1230–1234

Contents lists available at ScienceDirect

Patient Education and Counseling

journal homepage: www.elsevier.com/locate/pateducou



Short communication

Examiner characteristics and interrater reliability in a communication OSCE

Achim Mortsiefer[a,*], André Karger[b], Thomas Rotthoff[c], Bianca Raski[d], Michael Pentzek[a]

[a] Heinrich-Heine-University Düsseldorf, Medical Faculty, Institute of General Practice, Werdener Str. 7, 40227 Düsseldorf, Germany
[b] Heinrich-Heine-University Düsseldorf, Medical Faculty, Clinical Institute of Psychosomatic Medicine and Psychotherapy, Moorenstr. 5, 40225, Düsseldorf, Germany,
[c] Heinrich-Heine-University Düsseldorf, Medical Faculty, Deanery of Study and Department for Endocrinology, Diabetes, and Rheumatology, Moorenstr. 5, 40225 Düsseldorf, Germany
[d] Heinrich-Heine-University Düsseldorf, Medical Faculty, Deanery of Study and Clinical Institute of Psychosomatic Medicine and Psychotherapy, Moorenstr. 5, 40225 Düsseldorf, Germany

ARTICLE INFO

Article history:
Received 20 September 2016
Received in revised form 18 January 2017
Accepted 21 January 2017

Keywords:
Communication competence
Assessment
Examiner factors
OSCE

ABSTRACT

Objective: To identify inter-individual examiner factors associated with interrater reliability in a summative communication OSCE in the 4th study year.

Methods: The OSCE consists of 4 stations assessed with a 4-item 5-point global rating instrument. A bivariate secondary analysis of interrater reliability in relation to 4 examiner factors (gender, profession, OSCE experience, examiner training) was conducted. Intraclass correlation coefficients (ICC) were calculated and compared between examiner dyads of different similarity.

Results: 169 pairwise ratings from 19 different examiners in 16 dyads were analysed. Interrater reliability is significantly higher in examiner dyads of same vs. different gender (ICC = 0.76 (95%CI = 0.65-0.83) vs. ICC = 0.41 (95%CI = 0.21-0.57)), in dyads of two clinicians vs. non-clinical/mixed professions (ICC = 0.72 (95%CI = 0.56-0.83) vs. ICC = 0.57 (95%CI = 0.41-0.69)), and in dyads with high vs. low/mixed OSCE experience (ICC = 0.73 (95%CI 0.50-0.87) vs. ICC = 0.56 (95%CI = 0.41-0.69)). Participation in recent examiner training had no influence on ICCs.

Conclusion: Better concordance of ratings between clinically active examiners might be a hint for context specificity of good communication. Higher interrater reliability between examiners with same gender may indicate gender-specific communication concepts.

Practice implications: Medical faculties introducing summative assessment of communication competence should focus the influence of examiner characteristics on interrater reliability.

1. Introduction

The “objective structured clinical examination” (OSCE) is a well-established method of assessing a student's clinical skills, including communicative competence [1–4]. The reliability of an OSCE is influenced by various factors [5–7]. However, some authors suggest that even in a well-designed and valid OSCE, examiner factors remain the most important contributors to overall examination error [8,9].

The specific examiner factors that may affect the reliability of an OSCE have not been well studied [9,10], except for a few studies on the variability of individual examiner function over time influenced by fatigue [11,12] or leniency at the start of the OSCE [13].

Wilkinson [8] found that the involvement of examiners in station construction made a positive contribution to interrater reliability (IRR). Many authors have addressed the issue of examiner training, which has been indicated as essential, especially for the use of global ratings [14–16]. However, differences between examiners cannot often be sufficiently eliminated by training programs; therefore, the selection of appropriate examiners should be emphasized [17,18].

The analysis of our CoMeD–OSCE, which is an assessment of communication competence in challenging doctor-patient encounters [19], showed relatively low IRR according to other

[*] Corresponding author.
E-mail addresses: Andre.Karger@med.uni-duesseldorf.de (A. Karger), Rotthoff@med.uni-duesseldorf.de (T. Rotthoff), Bianca.Raski@med.uni-duesseldorf.de (B. Raski), Pentzek@med.uni-duesseldorf.de (M. Pentzek).

http://dx.doi.org/10.1016/j.pec.2017.01.013
0738-3991/© 2017 Elsevier B.V. All rights reserved.

studies [20,21]. The aim of this exploratory secondary analysis is to identify interindividual examiner factors that may influence IRR in communication skill assessments.

2. Methods

Bivariate secondary analyses of IRR in relation to examiner factors in a communication OSCE were performed.

The Düsseldorf CoMeD undergraduate communication skills training program [22] is followed by a 4-station OSCE in the fourth year when students encounter professional actors trained as typical standardized patients (SP) for the examination of the following types of communication: breaking bad news, sensitive issues (guilt and shame), handling emotions (aggression), and sharing decision-making.

We used the global rating (GR) instrument developed by Hodges, in a German version validated by Scheffer [23,24] containing the following 4 items: empathy, structuring the encounter, and verbal and non-verbal communication. Each item was rated on a 5-point scale with higher scores indicating better performance, resulting in total scores from 4 to 20. Each of the 4 constructs was defined by a short definition of low, medium and excellent behavioural performance. The descriptors have been published elsewhere [23,19].

We ran a 2-h examiner training program 1 week before. The examiners were shown several 8-min videotapes covering scenarios of the OSCE stations that displayed students' good and bad performances. These scenarios were rated by the participants using the GR and then discussed with examiners and trainers to reach a shared reference. A psychometric analysis of the CoMeD–OSCE has been published [19].

We used data from CoMeD-OSCE 2011/12. The examiners were psychotherapists, physicians, or scientists recruited from the Düsseldorf University and a pool of associated lecturers. They were dichotomized as practicing clinical medicine/psychotherapy vs. exclusively working as scientists. Former experience as examiners (≤3 OSCEs "little/no experience" vs. ≥4 OSCEs "more experience") and participation in the examiner training immediately preceding the OSCE (yes/no) were noted.

The 2 simultaneously rating examiners were placed in opposite corners of the examination room to reduce mutual interference. They were advised not to communicate among themselves or with the SP about the examinees' performances.

2.1. Statistical analysis

Wilcoxon tests for dependent samples were performed to compare median GR scores between dissimilar examiners. IRR is presented as intraclass correlation coefficients (ICC; two-way random, single measure, consistency-adjusted). ICCs range from 0.0 to 1.0 (0-0.29, poor; 0.30-0.49, fair; 0.50-0.69, moderate; ≥0.70, strong agreement [25]). Differences between ICCs were established with the cocor procedure [26]. For this explorative analysis in a small sample, we accepted a more lenient one-sided α-level of 0.10 to identify promising aspects for further research [27]. The reference category of the highest postulated dyad similarity was compared with the other categories of lower similarity: (1) examiners of the same gender vs. different gender; (2) both examiners working as clinicians vs. none or only one; (3) both examiners with more OSCE experience vs. none or only one; and (4) both examiners attended recent examiner training vs. none or only one.

Analyses were performed with SPSS 21 (*IBM Corp., Armonk, NY*).

3. Results

A sample of 169 pairwise ratings (=338 OSCE scores) from 19 examiners in 16 dyads were analysed (Table 1**).** Within the OSCE sessions rated by 2 dissimilar examiners, those with greater OSCE experience generally gave more lenient scores. Other examiner characteristics had no effect on global rating scores.

IRR was significantly higher in examiner dyads of the same gender, same professional background, and greater OSCE experience (Table 2). Participation in a recent training session had no influence (rather, it yielded diametrical results). Most ICCs were fair to moderate; strong agreement was only found in dyads with concordance in gender, profession, and OSCE experience (and – counterintuitively – without recent training).

We calculated a simple similarity score for examiner dyads, summing up 1 point per similarity in the 3 significant dyad characteristics from Table 2. Dyads with low similarity had a significantly lower ICC than dyads with medium ($z = 2.449$, $p = 0.007$) or high similarity ($z = 2.964$, $p = 0.002$) (Fig. 1).

4. Discussion and conclusion

4.1. Discussion

Higher IRR is associated with current clinical practice, OSCE experience, and concordant gender of examiners, but is not associated with participation in recent examiner training. Other studies also found that examiner training often yielded no or marginal improvement in reliability of an OSCE [17,18]. Several approaches for examiner training have been reported, but little is known about their effect on examiner performance [28]. However, Wilkinson [8] reported that examiner experience is not associated with IRR when checklist scores are used. Our results suggest that examiner experience is a relevant factor of IRR in communication skills assessed with a global rating instrument.

Wilkinson [8] also found that years of clinical experience had no significant effect on IRR when checklist scores were used. In contrast, Kahn [5] state that the reliability of scores generated by

Table 1
Examiner characteristics.

		Number of raters	Number of ratings	Mean OSCE total score (SD)	p[a]
Total		19	338 (169 OSCE sessions by 2 raters each)	15.5 (2.8)	–
Gender	Male	7	98	15.6 (2.8)	0.665
	Female	12	240	15.4 (2.8)	
Practicing clinical medicine/psychotherapy	Yes	13	190	15.7 (2.8)	0.595
	No (scientists)	6	148	15.3 (2.8)	
OSCE experience	Little/no (≤3 OSCEs)	10	189	14.9 (2.9)	0.027
	More (≥4 OSCEs)	9	149	16.2 (2.6)	
Recent examiner training attendance	Yes	13	220	15.4 (2.8)	0.195
	No	6	118	15.7 (2.8)	

[a] Significance level of the Wilcoxon tests for dependent samples: comparisons of OSCE total scores in the sessions simultaneously rated by 2 dissimilar examiners: female vs. male, practicing vs. non-practicing, little vs. greater OSCE experience, recent examiner training attended vs. non-attendance (Table 2 gives *n*).

Table 2
Interrater reliability of examiner dyads.

		Number of examiner dyads	Number of OSCE sessions rated	Interrater reliability (ICC, 95% CI)[a]	Fisher's z[b]	p[b]
Total		16	169	0.59 (0.48–0.68)	–	
Gender	Same gender	9	87	0.76 (0.65–0.83)	reference	
	Different gender	7	82	0.41 (0.21–0.57)	3.550	<0.001*
Practicing in clinical medicine/ psychotherapy	Both	6	51	0.72 (0.56–0.83)	reference	
	One of the raters	8	88	0.57 (0.41–0.69)	1.457	0.073*
	No rater	2	30	0.53 (0.22–0.75)	1.326	0.093*
OSCE experience	Both with high	4	28	0.73 (0.50–0.87)	reference	
	Only one with high	9	93	0.56 (0.41–0.69)	1.299	0.097*
	Both with little/ no	3	48	0.51 (0.26–0.69)	1.503	0.066*
Recent rater training attendance	Both	6	77	0.49 (0.30–0.64)	reference	
	One of the raters	6	66	0.62 (0.45–0.75)	1.119	0.869
	No rater	4	26	0.72 (0.47–0.87)	1.588	0.944

[a] ICC Intraclass correlation with 95% CI (confidence interval).
[b] Test statistic and significance level of Fisher's test for the comparison with the reference ICC.
* Significant at α-level 0.10.

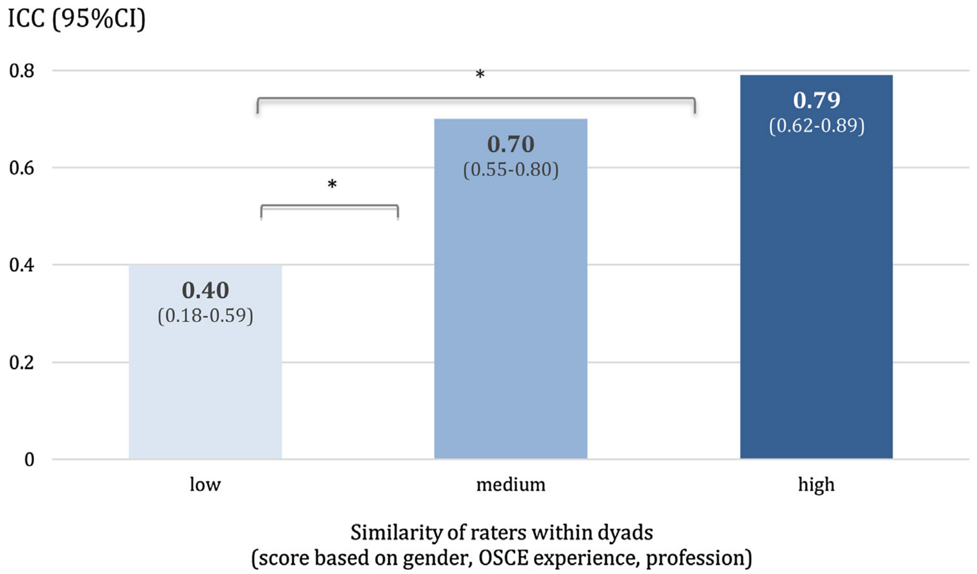


Fig. 1. Interrater reliability (IRR) and examiner similarity.

the examiners depends upon their clinical experience relevant to the clinical case-based OSCE station. Humphrey-Murto [29] found high IRR between physician examiners and trained non-medical assessors (most of them with medical backgrounds) in the use of checklist scores, but not in the use of global ratings. Our finding of a higher IRR of global ratings in clinically active examiners indicate that even an assessment that is not focused on the medical aspects of a clinical case, but solely on the communication process, might be influenced by context specificity [30,31], which may not be sufficiently comprehended by non-medical examiners.

Our results suggest that there might be another gender factor in addition to the examinees' performance [32–34] or the gender of the SPs [34,35]. Although there was no general effect of examiner gender on OSCE scores, the agreement within the same gender proved higher, which is the most robust result of our analysis. The different communication concepts and reference standards held by men and women may be an explanation for this gender effect [36].

In this exploratory study, only bivariate analyses were possible due to the relatively small number of examiner pairs and pairwise ratings. We cannot exclude confounders of the factors under examination. A liberal alpha level of 0.10 was applied due to missing *a priori* sample size calculation. This enabled us to explore effects in this small sample, but increased the risk of over-interpreting random effects. We did not systematically control the behavior of the examiner dyads and their communication during the OSCE sessions. In addition, we did not test whether an extension of our 2-h examiner training would have potentially reduced the effect of examiner characteristics on IRR. Furthermore, the descriptors of the GR introduced by Hodges [23] comprise only context-independent definitions. We did not define precise task-specific behaviors for single OSCE stations (e.g. in a codebook for raters). We used one of several possible ICC interpretations. Individual ICCs will have been interpreted as slightly poorer according to more stringent ICC classifications.

Future research should systematically investigate the examiner characteristics associated with interrater reliability in larger samples and experimental designs.

At the next stage of our study we intend to develop task-specified descriptors in addition to the general descriptors of the GR on basis of cognitive interviews [37] with different examiners. For the following experimental study, we hypothesize that a modified examiner training using additional specified instructions results in higher IRR compared to the existent examiner training using the GR with general descriptors. In an one-way experimental design, two examiner samples will be compared on ratings of the same videotapes (ICC as dependent variable).

5. Conclusion

Our finding of higher rating concordance between examiners of the same gender suggests the hypothesis that unrevealed gender-specific concepts are important in assessing communicative competence. Better concordance of ratings in clinically active examiners hints at context specificity, which is pre-existent even in medical encounters with a focus on communication aspects.

Practice implications

Medical faculties introducing summative assessment of communication competence should focus the influence of examiner characteristics on IRR.

Acknowledgements

We are grateful to the medical students and the examiners for facilitating and supporting our project. Our special thanks go to our CoMeD team and the staff of the Düsseldorf University Hospital Training Centre (TräF) for project organisation.

Appendix A. Global Rating (GR) form used in the CoMeD – OSCE

Original version in English developed and published by Brian Hodges & Jodi Herold McIlroy (Hodges 2003). Translation und validation in German by Scheffer (Scheffer, 2008)

Response to patient's feelings and needs (empathy)				
1 Does not respond to obvious patient cues (verbal and non-verbal) and / or responds inappropriately	2	3 Responds to patient's needs and cues, but not always effectively	4	5 Responds consistently in a perceptive and genuine manner to the patient's needs and cues
		Degree of coherence in the interview		
1 No recognisable plan to the interaction; the plan does not demonstrate cohesion or the patient must determine the direction of the interview	2	3 Organisational approach is formulaic and minimally flexible and / or control of the interview is inconsistent	4	5 Superior organisation, demonstrating command of cohesive devises, flexibility, and consistent control of the interview
		Verbal expression		
1 Communicates in manner that interferes with and / or prevents understanding by patient, or communicates inappropriately with the patient	2	3 Exhibits sufficient control of expression to be understood by an active, engaged listener (patient)	4	5 Exhibits command of expression (fluency, diction, grammar, vocabulary, tone, volume and modulation of voice, rate of speech, pace and pronunciation
		Non-verbal expression		
1 Fails to engage, frustrates and / or antagonises the patient	2	3 Exhibits enough control of non-verbal expression to engage a patient willing to overlook deficiencies such as passivity,self-consciousness or inappropriate aggressiveness	4	5 Exhibits finesse and command of non-verbal expression (eye-contact, gesture, posture, use of silence, etc.)
Overall assessment of the knowledge and skills demonstrated in the interview				
A = Incompetent Responds inappropriately and ineffectively to the task, indicating a lack of knowledge and / or undeveloped interpersonal and interviewing skills		B = Borderline Responds effectively to some components of the task, some development of interpersonal and interviewing skills		C=Competent Responds precisely and perceptively to the task, consistently integrating all components

References

[1] F.D. Duffy, G.H. Gordon, G. Whelan, K. Cole-Kelly, R. Frankel, N. Buffone, S. Lofton, M. Wallace, L. Goode, L. Langdon, Assessing competence in communication and interpersonal skills: the Kalamazoo II report, Acad. Med. 79 (2004) 495–507.

[2] M. Deveugele, A. Derese, S. De Maesschalck, S. Willems, M. Van Driel, J. De Maeseneer, Teaching communication skills to medical students, a challenge in the curriculum? Patient Educ. Couns. 58 (2005) 265–270.

[3] E.A. Rider, M.M. Hinrichs, B.A. Lown, A model for communication skills assessment across the undergraduate curriculum, Med. Teach. 28 (2006) e127–34.

[4] K.Z. Khan, S. Ramachandran, K. Gaunt, P. Pushkar, The objective structured clinical examination (OSCE): AMEE guide No. 81. part I: an historical and theoretical perspective, Med. Teach. 35 (2013) e1437–46.

[5] K.Z. Khan, K. Gaunt, S. Ramachandran, P. Pushkar, The objective structured clinical examination (OSCE): AMEE guide No. 81. part II: organisation & administration, Med. Teach. 35 (2013) e1447–63.

[6] S.M. Downing, Reliability: on the reproducibility of assessment data, Med. Educ. 38 (2004) 1006–1012.

[7] S. Smee, Skill based assessment, BMJ 326 (2003) 703–706.

[8] T.J. Wilkinson, C.M. Frampton, M. Thompson-Fawcett, T. Egan, Objectivity in objective structured clinical examinations: checklists are no substitute for examiner commitment, Acad. Med. 78 (2003) 219–223.

[9] P.R. Jeffries, A framework for designing, implementing, and evaluating simulations used as teaching strategies in nursing, Nurs. Educ. Perspect. 26 (2005) 96–103.

[10] M.K. Burns, How to establish interrater reliability, Nursing 44 (2014) 56–58.

[11] G.M. Humphris, S. Kaney, Examiner fatigue in communication skills objective structured clinical examinations, Med. Educ. 35 (2001) 444–449.

[12] K. McLaughlin, M. Ainslie, S. Coderre, B. Wright, C. Violato, The effect of differential rater function over time (DRIFT) on objective structured clinical examination ratings, Med. Educ. 43 (2009) 989–992.

[13] D. Hope, H. Cameron, Examiners are most lenient at the start of a two-day OSCE, Med. Teach. 37 (2015) 81–85.

[14] J. Crossley, G. Humphris, B. Jolly, Assessing health professionals, Med. Educ. 36 (2002) 800–804.

[15] J. Van Dalen, C.J. Prince, A.J. Scherpbier, C.P. Van Der Vleuten, Evaluating communication skills, Adv. Health Sci. Educ. Theory Pract. 3 (1998) 187–195.

[16] J.A. Spencer, J. Silverman, Communication education and assessment: taking account of diversity, Med. Educ. 38 (2004) 116–118.

[17] D.I. Newble, J. Hoare, P.F. Sheldrake, The selection and training of examiners for clinical examinations, Med. Educ. 14 (1980) 345–349.

[18] C.P. van der Vleuten, S.J. van Luyk, A.M. van Ballegooijen, D.B. Swanson, Training and experience of examiners, Med. Educ. 23 (1989) 290–296.

[19] A. Mortsiefer, J. Immecke, T. Rotthoff, A. Karger, R. Schmelzer, B. Raski, J.I. Schmitten, A. Altiner, M. Pentzek, Summative assessment of undergraduates' communication competence in challenging doctor-patient encounters. Evaluation of the Dusseldorf CoMeD-OSCE, Patient Educ. Couns. 95 (2014) 348–355.

[20] P.K. Han, K. Joekes, G. Elwyn, K.M. Mazor, R. Thomson, P. Sedgwick, J. Ibison, J.B. Wong, Development and evaluation of a risk communication curriculum for medical students, Patient Educ. Couns. 94 (2014) 43–49.

[21] M.T. Brannick, H.T. Erol-Korkmaz, M. Prewett, A systematic review of the reliability of objective structured clinical examination scores, Med. Educ. 45 (2011) 1181–1189.

[22] A. Mortsiefer, T. Rotthoff, R. Schmelzer, J. Immecke, B. Ortmanns, J. der Schmitten, A. Altiner, A. Karger, Implementation of the interdisciplinary curriculum Teaching and Assessing Communicative Competence in the fourth academic year of medical studies (CoMeD), GMS Z Med Ausbild (2012) 29 (Doc06).

[23] B. Hodges, J.H. McIlroy, Analytic global OSCE ratings are sensitive to level of training, Med. Educ. 37 (2003) 1012–1016.

[24] S. Scheffer, I. Muehlinghaus, A. Froehmel, H. Ortwein, Assessing students' communication skills: validation of a global rating, Adv. Health Sci. Educ. Theory Pract. 13 (2008) 583–592.

[25] L.G. Portney, M.P. Watkins, Foundations of Clinical Research Applications to Practice, Prentice Hall Inc., New Jersey, 2000.

[26] B. Diedenhofen, J. Musch, cocor: a comprehensive solution for the statistical comparison of correlations, PLoS One 10 (2015) e0121945.

[27] R.E. Henkel, Tests of Significance, Sage Publications, Beverly Hills, 1976.

[28] K. Boursicot, L. Etheridge, Z. Setna, A. Sturrock, J. Ker, S. Smee, E. Sambandam, Performance in assessment: consensus statement and recommendations from the Ottawa conference, Med. Teach. 33 (2011) 370–383.

[29] S. Humphrey-Murto, S. Smee, C. Touchie, T.J. Wood, D.E. Blackmore, A comparison of physician examiners and trained assessors in a high-stakes OSCE setting, Acad. Med. 80 (2005) S59–62.

[30] L.A. Baig, C. Violato, R.A. Crutcher, Assessing clinical communication skills in physicians: are the skills context specific or generalizable, BMC Med. Educ. 9 (2009) 22.

[31] E. Keely, K. Myers, S. Dojeiji, Can written communication skills be tested in an objective structured clinical examination format? Acad. Med. 77 (2002) 82–86.

[32] R.C. Smith, J.S. Lyles, J.A. Mettler, A.A. Marshall, L.F. Van Egeren, B.E. Stoffelmayr, G.G. Osborn, V. Shebroe, A strategy for improving patient satisfaction by the intensive training of residents in psychosocial medicine: a controlled, randomized study, Acad. Med. 70 (1995) 729–732.

[33] U. Holm, K. Aspegren, Pedagogical methods and affect tolerance in medical students, Med. Educ. 33 (1999) 14–18.

[34] C.M. Wiskin, T.F. Allan, J.R. Skelton, Gender as a variable in the assessment of final year degree-level communication skills, Med. Educ. 38 (2004) 129–137.

[35] R. Gispert, M. Rue, J. Roma, J.M. Martinez-Carretero, Gender, sequence of cases and day effects on clinical skills assessment with standardized patients, Med. Educ. 33 (1999) 499–503.

[36] D.L. Roter, J.A. Hall, Y. Aoki, Physician gender effects in medical communication: a meta-analytic review, JAMA 288 (2002) 756–764.

[37] G.B. Willis (Ed.), Cognitive Interviewing. A Tool for Improving Questionnaire Design, Sage Publications, Thousand Oaks, 2005.