A systematic review of the reliability of objective structured clinical examination scores

Michael T Brannick,[1] H Tugba Erol-Korkmaz[2] & Matthew Prewett[1]

CONTEXT The objective structured clinical examination (OSCE) is comprised of a series of simulations used to assess the skill of medical practitioners in the diagnosis and treatment of patients. It is often used in high-stakes examinations and therefore it is important to assess its reliability and validity.

METHODS The published literature was searched (PsycINFO, PubMed) for OSCE reliability estimates (coefficient alpha and generalisability coefficients) computed either across stations or across items within stations. Coders independently recorded information about each study. A meta-analysis of the available literature was computed and sources of systematic variance in estimates were examined.

RESULTS A total of 188 alpha values from 39 studies were coded. The overall (summary) alpha across stations was 0.66 (95% confidence interval [CI] 0.62–0.70); the overall alpha within stations across items was 0.78 (95% CI 0.73–0.82). Better than average reliability was associated with a greater number of stations and a higher number of examiners per station. Interpersonal skills were evaluated less reliably across stations and more reliably within stations compared with clinical skills.

CONCLUSIONS Overall scores on the OSCE are often not very reliable. It is more difficult to reliably assess communication skills than clinical skills when considering both as general traits that should apply across situations. It is generally helpful to use two examiners and large numbers of stations, but some OSCEs appear more reliable than others for reasons that are not yet fully understood.

*Medical Education 2011: **45**: 1181–1189*
doi:10.1111/j.1365-2923.2011.04075.x

[1]Department of Psychology, College of Arts and Sciences, University of South Florida, Tampa, Florida, USA
[2]Department of Psychology, Middle East Technical University, Ankara, Turkey

Correspondence: Michael T Brannick, Department of Psychology, PCD 4118G, University of South Florida, Tampa, Florida 33620-7200, USA. Tel: 00 1 813 974 0478; Fax: 00 1 813 974 4617; E-mail: mbrannick@usf.edu

INTRODUCTION

The objective structured clinical examination (OSCE) provides a means of assessing the competence of a broad array of examinees, including medical students, residents and experienced doctors. For the last three decades, OSCEs have been used for the assessment of clinical competence as part of health professional education.[1] The OSCE is 'an approach to the assessment of clinical competence in which the components of competence are assessed in a well-planned or structured way with attention being paid to objectivity'.[2] This type of examination is now widely used to assess clinical competence.[3]

In the OSCE, a series of standardised problems is presented to each examinee. The problems often involve simulated patients (also called standardised patients [SPs]) portrayed by confederates who are trained to play roles. The SP may also evaluate the examinee on aspects of the encounter. Sometimes examinee performance is scored by an external judge who is often a content expert, such as a medical faculty member. In general, a predetermined objective scheme such as a checklist or a Likert-type global evaluation scale is used by the examiner(s) in the rating process.[4]

An advantage of the OSCE over paper-and-pencil tests of knowledge is that the simulations involve more realistic context, content and procedures. For example, in the OSCE, rather than writing an essay about diagnosis, the examinee will encounter an SP and generate a diagnosis based on the clinical interview and examination. An advantage of the OSCE over assessments that use real patients is that the patients are standardised across examinees and thus the patient problems are essentially equivalent and examinee responses and scores are comparable.

Despite the apparent advantages of the OSCE over available alternative assessments, the quality of assessment is not guaranteed simply by assembling some standardised problems. The reliability of the assessment is crucial, particularly when the aim of the OSCE is to provide data for high-stakes decisions, as is often the case in medical school assessments. Ideally, the particular problems chosen for the OSCE should not be terribly influential and examinees who pass or fail a given examination should be expected to also pass or fail an alternative examination, should one be given. In other words, the problems in a given examination are intended to tap a portfolio of skills that should be mastered by the medical practitioner. The reliability of the overall examination represents an estimation of the correlation of scores on the given examination with scores on a hypothetical examination composed of the entire portfolio of problems.

The research questions driving the current study were:

1 What reliability should we expect on average when we develop an OSCE?
2 What is the likely range of such values?
3 What factors appear to influence the expected reliability?

The current paper addresses these questions through a quantitative review, which includes descriptions of the effects of the number of stations, the number of raters, the choice of dimensions to include (aspects of competence), and the purpose of measurement (research versus decision making). We also estimated the variance in reliability that remains after accounting for sampling error and moderator variables. The results of the study can be used to:

1 inform choices about data collection in future research and application (e.g. on the number of stations to include);
2 estimate the likely range of reliability in a given context, and
3 generate hypotheses about the quality of OSCE measures for future research.

METHODS

Search and inclusion criteria

To identify relevant studies, we searched the PsycINFO and PubMed electronic databases using the keywords 'OSCE' and 'reliability'. Additionally, the reference sections of the studies identified in this search were used to find other potential studies of interest. Our search efforts resulted in the identification of 98 journal articles. Those studies that did not report any empirical reliability values were eliminated from this meta-analysis, which left 64 studies and 457 reliability values to be coded. On average, seven or eight reliability values were reported per study, including Cronbach's alpha, inter-rater agreement, intraclass correlation, Cohen's kappa statistic and generalisability indices.

Alpha may not be the most appropriate estimate of reliability for evaluating the OSCE because it: (i) deals with a single facet of measurement error rather than multiple facets, and (ii) considers the rank order of examinees rather than their absolute deviations from a standard. However, alpha is the most commonly reported index of reliability. Only those studies reporting Cronbach's alpha as the reliability index were meta-analysed for this study, although generalisability estimates are also reported. Overall, 188 alpha values from 39 samples formed the basis of our meta-analytic study.[5–41] The other statistics (e.g. test–retest correlations) were not included in the analyses because the number of estimates of each kind were too few to allow for a meaningful meta-analysis. Different statistical reliability estimates (e.g. alpha and kappa) should not be included in the same meta-analysis because they have different sampling distributions and sometimes have different metrics.[42]

We have also included generalisability coefficients computed across stations as a comparison distribution. Such estimates were not aggregated with the alpha estimates because they have a different meaning.[43] Alpha estimates do not consider differences in means across exercises (or items) as part of the error term because such estimates are concerned solely with the relative standing of examinees. Alternatively, generalisability coefficients do consider differences in means as part of the error term because they are concerned with the absolute standing of examinees. Standard meta-analytic techniques could not be used for the analysis of generalisability coefficients because we could locate neither an expression for their sampling variance nor an appropriate normalising transformation; however, descriptive information about the distribution of generalisability coefficients is presented based on 31 generalisability estimates obtained from 12 different studies.[21,30,44–53]

Of the OSCEs reported in this paper, 88% were used to assess medical students (65%) or residents (23%). An additional 8% of the studies assessed advanced practitioners (e.g. general practitioners, rheumatologists) and the remaining 4% assessed a mix of either medical students and residents, or residents and advanced practitioners. The majority of the OSCEs sampled the broad content that an examination in medical school would be expected to cover (e.g. history taking, communication), but a few contained more specific and focused stations (e.g. one focused on simulated encounters at a blood bank).

Moderators (study characteristics)

A review of the studies revealed that different approaches were taken in the calculation of the alpha indices in different studies. For 100 of the reported alpha values, alpha was calculated based on the number of stations included in the OSCE (referred to here as 'alpha across stations'). For 53 of the reported values, alpha was calculated based on the number of items included in the scale used by the examiners for assessing examinee performance within a station (referred to here as 'alpha across items'). For 35 of the coded alpha values, the authors did not indicate whether they had based their calculations on the number of stations or the number of items. Therefore, those 35 alpha values were eliminated and separate meta-analyses were conducted on the two samples of 'across-stations' and 'across-items' studies.

Note that alpha computed across stations treats differences in cases as error so that alpha is based on the covariances of total scores across stations. For example, alpha might be computed based on several different stations, each with a communication score based on a single case. Alpha estimates computed across stations provide no direct information about differences among the communication items because the items are averaged or totalled before the correlations are computed. An alpha computed across stations would estimate consistency in communication as examinees move from one patient to another. Alpha computed across items estimates the consistency of behaviour within a station (e.g. individual items on a communication scale might concern whether the examinee maintained eye contact, appeared to listen attentively, and so forth). Such estimates are typically computed on the correlations between items within a single station and thus provide no information about the correlation between stations. Such an estimate would consider, for example, whether people who maintain eye contact with a given patient are also likely to listen attentively to that patient. Further, given that the items are subjective and completed by the same judge, their status as truly independent measures is questionable. Thus, although both types of studies report alpha as a reliability estimate, the meaning of this reliability differs. In theory, a generalisability study should simultaneously estimate the effects of both items and stations. However, we did not find any such studies.

For each reported alpha, study sample size, number of stations included in the OSCE (or number of items included in the scale), type of examiner (SP, faculty member, trained student, etc.), number of examiners

(one versus two), type of scale for evaluating examinee performance in each station (checklist versus Likert-type global evaluation scale), content of the OSCE (clinical competence versus communication), and study context (research versus high-stakes decision) were coded by one of two authors. A representative subset of five studies was then coded by the third author for reliability. Coders exhibited 100% agreement on most of the study variables (coefficient value, sample size, number of stations, type of examiner, type of scale, content of the OSCE, study context). Raters disagreed once on the number of examiners and once on the number of eligible reliability estimates. These disagreements were resolved by discussion.

Analytic strategy

We generally followed the recommendations of Rodriguez and Maeda[43] for the meta-analysis of coefficient alpha. The alpha value, study sample size and number of stations (or number of items) were used in the calculations. The method of analysis is based on a transformation of alpha and a weighting scheme that involves both the estimated sampling error of each study and the estimated random-effects variance component (i.e. a random-effects method of meta-analysis). The number of stations (or items) was used as a covariate, as recommended by Rodriguez and Maeda.[43]

The estimated population coefficient alphas for across-stations and across-items studies were calculated initially, followed by the moderator analyses. Because of the limited number of studies and missing data, each potential moderator was tested independently using weighted regression analysis.

For the generalisability coefficients, an unweighted average was computed for comparison with the alpha coefficients. Overall distributions of both alpha and generalisability coefficients were illustrated with boxplots (Fig. 1).

RESULTS

Overall results ignoring other moderators are reported first. A total of 100 values of alpha were taken over stations. Their estimated mean (random effects) was 0.66 (95% confidence interval [CI] 0.62–0.70). The 95% credibility (prediction) interval[43] was 0.16–0.99. A total of 49 values of alpha were taken over items. Their mean was 0.78 (95% CI 0.73–0.82). Their credibility interval was 0.36–0.95.

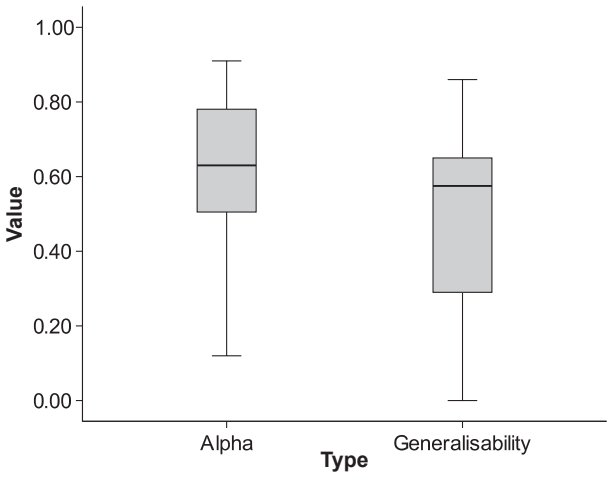

Figure 1 Overall distributions of alpha and generalisability coefficients

The unweighted average of the alpha coefficients was 0.62; the unweighted average of the generalisability coefficients was 0.49. A scatterplot of the joint distribution of effect size and number of stations is shown in Fig. 2 (both alpha and generalisability coefficients are included in the graph). Two unweighted regression lines are also plotted in the graph, each relating the number of stations to either alpha or the generalisability coefficient. As Fig. 2 shows, for both types of coefficient, reliability tends to increase as the number of stations increases, but there is considerable variability in the estimates.

Moderators were modelled separately across stations and across items. The results are shown in Tables 1

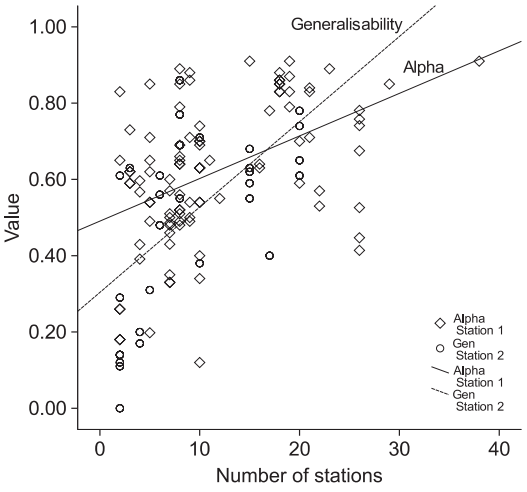

Figure 2 Scatterplot of alpha and generalisability coefficients by number of stations with unweighted regression lines

Table 1 Moderator analyses for across-stations estimates

Moderator	*k*	Mean	95% CI Lower	95% CI Upper
Content*†				
Communication scale	16	0.55	0.45	0.63
Clinical scale	67	0.69	0.66	0.73
Context†				
High-stakes examination	65	0.65	0.60	0.70
Research study OSCE	35	0.68	0.61	0.74
Number of raters*				
1 rater	90	0.65	0.61	0.68
2 raters	8	0.81	0.73	0.88
Examiner type (1)†				
Faculty member judge	55	0.69	0.64	0.73
SP judge	19	0.54	0.44	0.63
Examiner type (2)†				
Content expert judge	16	0.70	0.62	0.77
SP judge	19	0.54	0.45	0.62
Scale type†				
Checklist scale	44	0.69	0.64	0.71
Likert scale	31	0.59	0.52	0.66

* The moderator was significant ($p < 0.05$) in the weighted regression
† The covariate (number of stations) was significant ($p < 0.05$) in the weighted regression
k = number of coefficients; 95% CI = 95% confidence interval; OSCE = objective structured clinical examination; SP = standardised patient

Table 2 Moderator analyses for across-items estimates

Moderator	*k*	Mean	95% CI Lower	95% CI Upper
Content*†				
Communication scale	14	0.88	0.86	0.90
Clinical scale	14	0.75	0.72	0.79
Format*†				
Likert scale	21	0.88	0.85	0.91
Checklist scale	28	0.67	0.62	0.72
Context				
High-stakes examination	36	0.79	0.74	0.83
Research OSCE	13	0.74	0.64	0.82
Number of raters*				
1 rater	43	0.76	0.71	0.80
2 raters	6	0.89	0.81	0.95
Examiner type (1)				
Faculty member rater	19	0.79	0.71	0.85
SP rater	18	0.77	0.69	0.83
Examiner type (2)*†				
Content expert rater	4	0.61	0.29	0.81
SP rater	18	0.77	0.68	0.83

* The moderator was significant ($p < 0.05$) in the weighted regression
† The covariate (number of items) was significant ($p < 0.05$) in the weighted regression
k = number of coefficients; 95% CI = 95% confidence interval; OSCE = objective structured clinical examination; SP = standardised patient

and 2. For each table, the moderator or potential explanatory variable is listed with its values. For example, in Table 1, the mean alpha was estimated separately for communication scales versus clinical scales. The difference in mean alpha between clinical and content scales was significant, as was the number of stations (weighted regression was used so that the test for the difference between clinical and content means was statistically adjusted for numbers of stations). There were 16 alpha estimates for communication scales and 67 estimates for the rest of the clinical contents. The mean alpha was 0.55 for the communication scales (95% CI 0.45–0.63). For the clinical scales, the mean alpha was 0.69 (95% CI 0.66–0.73). For the results across stations, two moderators were significant: content (clinical versus communication), and number of raters. Both moderators were also significant for results across items. Note that communication scales were more reliable than clinical scales when items were the source of error and less reliable when stations were the source of error. For the results across items, contrasts for scale format (checklist versus Likert scale) and type of examiner (content expert versus SP) were also significant. For both stations and items, mean reliability estimates for research and high-stakes examinations did not differ significantly and neither did estimates for average faculty member and SP ratings.

DISCUSSION

The purpose of this study was to describe the reliability (in terms of both mean reliability and variability of distribution) of the OSCE and to determine whether several features associated with

the administration of the OSCE are related to the reliability of measurement. To that end, we meta-analysed estimates of alpha (internal consistency) across stations and, separately, across items within stations. We also examined the relationships between features of data collection, including numbers of stations, items and examiners, as well as the skill to be evaluated (communication versus clinical), the type of examiner (SP, faculty member, content expert) and the type of rating scale (checklist, Likert scale). Generalisability estimates were included to provide a descriptive comparison.

Overall reliability

The overall average alpha was reasonable (0.78) for scales within stations, but low (0.66) across stations. It can be shown using the formula for alpha that the reliability of the overall score can be made as large as desired by increasing the number of observations (items or stations, assuming positive correlations among the variables). However, there are limits to the number of variables in practice. Increasing the number of items on a communication scale may simply produce redundancy by increasing the reliability estimate without gaining any real precision in measurement. Increasing the number of stations is expensive. As Fig. 2 shows, although empirical estimates of reliability increase on average with the number of stations, there is surprisingly large variability in reliability at any given number of stations. Note that a similar conclusion follows from the credibility or prediction intervals associated with the overall estimates. Therefore, OSCE designers should not assume that overall scores will be reliable simply because many stations are included in the design.

The mean of the generalisability coefficients was lower than the mean of the alpha coefficients (unweighted means of 0.49 and 0.62, respectively), as would be expected from the meanings and computations of the two values. The generalisability coefficients consider absolute differences to be meaningful, but the alpha coefficients consider only relative differences to be meaningful. Although both values are rather low, which suggests that particular content or the inclusion of specific stations or cases is of vital importance in determining the total score for an examinee, such reliability coefficients do not tell the entire story. Many of the studies reported good estimates of dependability in terms of confidence that examinees who passed or failed the overall OSCE would pass or fail an alternative form despite the low overall reliability value.

Moderators

Because of the variability in reliability estimates, it appears fruitful to examine study characteristics that may explain some or all of the variability in estimates beyond sampling error. Reliability estimates across stations and items are discussed in turn. Regarding reliability across stations, the regression lines relating reliability to stations show positive slopes for both alpha and generalisability estimates, indicating that OSCEs with more stations tend to show higher reliability (e.g. the unweighted mean alpha for OSCEs with ≤ 10 stations was 0.56 and that for OSCEs with > 10 stations was 0.74). Further, in three of six moderator analyses, the covariate (number of stations) was significant, also indicating that OSCEs with more stations tend to show higher reliability. However, there is a lot of variability in the estimates: some studies report a reliability of > 0.80 with < 10 stations and others report a reliability of < 0.80 with > 25 stations.

Based on the current data, having a second rater substantially improves reliability. Although the number of studies that did this was not large (eight across-stations studies and five across-items studies), there were appreciable gains in reliability attributable to adding a second rater. Some authors have argued that adding stations is a better use of resources than adding raters.[47]

Concerning reliability across stations, communication evaluations were less reliable than were measures of clinical skills. We suspect that the evaluation of skill in communication, such as where the rater assesses whether the examinee listened or whether the examinee displayed cultural competence, is more subjective and therefore more idiosyncratic to the judge and that this characteristic results in poorer correlations across stations. Not only does the examiner or judge vary between stations, but so does the SP. There may also be real differences in communication attributable to differences in SPs as well as to the interaction between the student (examinee) and the SP. It is possible, for example, that communication may be relatively good when the SP and doctor are alike in race and sex, and relatively poor otherwise. Alternatively, there may be purely idiosyncratic differences that influence the quality of communication.

With regard to reliability estimated across items (not stations), again the skill to be evaluated and the number of examiners were significant moderators. As with the analyses of the number of stations, superior reliability resulted from using two judges rather than one. By contrast with the across-station analysis,

however, the within-station analysis showed that reliability was *higher* for communication items than for clinical items. We suspect that the difference reflects the fact that the assessment of communication items is more subjective than that of clinical items. Communication items are typically recorded using Likert scales (in 11 of the 14 communication OSCEs), whereas clinical items are usually judged using checklists (12 of the 14 clinical OSCEs). The type of rating scale (checklist versus Likert scale) was also found to be a significant moderator (in the direction consistent with this explanation [Table 2]). Items on a checklist are often rather easily observed (e.g. Did the examinee listen to the chest using the stethoscope under the patient's gown?). Items on a Likert scale are subject to interpretation to a greater degree and call for graded responses to a set of behaviours observed over a longer period of time (e.g. examinees might be evaluated on such items as 'Listened carefully to the SP's complaints' or 'Behaved professionally'). Because a single judge typically rates all of the communication items in a station, any global impression of the examinee's performance in that station is likely to colour all the evaluative ratings for that examinee (i.e. in the manner of the halo effect[54]). Therefore, the high estimates of reliability taken over items in a single station should not be viewed as evidence that scores on the overall OSCE will necessarily be reliable. There are other possible explanations for differences between checklist and Likert scale scores, including the occurrence of ceiling effects for some checklist items, as well as possible differences in underlying causes of behaviour. It may be, for example, that the clinical skills evaluated in the checklist depend upon a greater number of underlying factors than the skills required for communication.

Similarly, the finding that the type of examiner (SP versus content expert) is a significant moderator of across-items reliability estimates may reflect the type of skill being evaluated by the corresponding examiners. A review of this particular subset of studies revealed that all of the four content experts evaluated clinical skills, whereas the majority of the SPs (14 of 18) evaluated communication skills.

Study limitations

Because this study reports a meta-analysis, it is limited in the number of features it can code and analyse, either because of the number of studies of a specific kind or because the published reports do not contain the necessary information. The number of studies was too few to allow for the examination of inter-judge reliability, test–retest reliability or alternative forms of reliability. The sort of information that is rarely provided in studies includes details of the actual content of the stations (i.e. the choice of simulated problems), the characteristics of SPs and the scales used to record raters' judgements. Some of these limiting factors could be addressed by more complete reporting of information, but other questions cannot be addressed without changing the design of the OSCE. For example, it would be useful to have some judges follow the examinee to multiple stations to compare the within-station and between-station reliability of communication items. Using the conventional design, we cannot estimate how much of the between-station drop in reliability reflects characteristics of the judge and how much it reflects SP attributes.

CONCLUSIONS

Although the OSCE does provide a standardised and relatively objective method of evaluating a set of clinical skills in medical personnel, its use does not guarantee reliable scores and accurate decisions about medical students (many estimates of the overall reliability across stations were < 0.60). Based on an analysis of empirical results in the literature, it appears to be more difficult to reliably assess communication skills than clinical skills across stations. Examinations with more stations tend to show higher reliability and using two raters appears preferable to using a single rater. However, there is large variability in estimates of reliability even after sampling error is accounted for, which suggests that design features that could not be analysed in the current paper may be influential determinants of OSCE reliability.

Contributors: MTB and HTE-K contributed to the study design, collected and analysed data, and wrote sections of the manuscript. MP collected and verified data and contributed to the revision of the article. All authors approved the final manuscript for publication.
Acknowledgements: none.
Funding: none.
Conflicts of interest: none.
Ethical approval: this study was decreed exempt from requirements for ethical approval by the University of South Florida Internal Review Board.

REFERENCES

1 Bartfay WJ, Rombough R, Howse E, Leblanc R. Evaluation. The OSCE approach in nursing education. *Can Nurse* 2004;**100** (3):18–23.

2 Harden RM. What is an OSCE? *Med Teach* 1988;**10** (1):19–22.
3 Gaur L, Skochelak S. STUDENT JAMA. Evaluating competence in medical students. *JAMA* 2004;**291** (17):2143.
4 Harden RM, Stevenson M, Downie WW, Wilson GM. Assessment of clinical competence using objective structured examination. *Br Med J* 1975;**1** (5955):447–51.
5 Amiel GE, Tann M, Krausz MM, Bitterman A, Cohen R. Increasing examiner involvement in an objective structured clinical examination by integrating a structured oral examination. *Am J Surg* 1997;**173**:546–9.
6 Amiel GE, Ungar L, Alperin M, Baharier Z, Cohen R, Reis S. Ability of primary care physicians to break bad news: a performance-based assessment of an educational intervention. *Patient Educ Couns* 2006;**60** (1):10–5.
7 Benbow EW, Harrison I, Dornan TL, O'Neill PA. Pathology and the OSCE: insights from pilot study. *J Pathol* 1998;**184** (1):110–4.
8 Blue AV, Stratton TD, Plymale M, DeGnore LT, Schwartz RW, Sloan DA. The effectiveness of the structured clinical instruction module. *Am J Surg* 1998;**176** (1):67–70.
9 Brailovsky CA, Grand'Maison P. Using evidence to improve evaluation: a comprehensive psychometric assessment of an SP-based OSCE licensing examination. *Adv Health Sci Educ Theory Pract* 2000;**5** (3):207–19.
10 Chesser AM, Laing MR, Miedzybrodzka ZH, Brittenden J, Heys SD. Factor analysis can be a useful standard-setting tool in a high-stakes OSCE assessment. *Med Educ* 2004;**38** (8):825–31.
11 Cohen R, Reznick RK, Taylor BR, Provan J, Rothman A. Reliability and validity of the objective structured clinical examination in assessing surgical residents. *Am J Surg* 1990;**160** (3):302–5.
12 Cohen R, Rothman AI, Bilan S, Ross J. Analysis of the psychometric properties of eight administrations of an objective structured clinical examination used to assess international medical graduates. *Acad Med* 1996;**71** (1 Suppl):22–4.
13 Guiton G, Hodgson CS, Delandshere G, Wilkerson L. Communication skills in standardised patient assessment of final-year medical students: a psychometric study. *Adv Health Sci Educ Theory Pract* 2004;**9** (3): 179–87.
14 Guiton G, Hodgson C, May W, Elliott D, Wilkerson L. Assessing medical students' cross-cultural skills in an objective structured clinical examination. *American Educational Research Association International Conference*, San Diego, CA, 12–16 April 2004.
15 Hodges B, Regehr G, Hanson M, McNaughton N. Validation of an objective structured clinical examination in psychiatry. *Acad Med* 1998;**73** (8):910–2.
16 Hull AL, Hodder S, Berger B, Ginsberg D, Lindheim N, Quan J, Kleinhenz ME. Validity of three clinical performance assessments of internal medicine clerks. *Acad Med* 1995;**70** (6):517–22.
17 Humphris GM. Communication skills knowledge, understanding and OSCE performance in medical trainees: a multivariate prospective study using structural equation modelling. *Med Educ* 2002;**36** (9):842–52.
18 Humphris GM, Kaney S. Examiner fatigue in communication skills objective structured clinical examinations. *Med Educ* 2001;**35** (5):444–9.
19 Junger J, Schafer S, Roth C, Schellberg D, Friedman Ben-David M, Nikendei C. Effects of basic clinical skills training on objective structured clinical examination performance. *Med Educ* 2005;**39** (10):1015–20.
20 Kramer AW, Jansen JJ, Zuithoff P, Dusman H, Tan LH, Grol RP, van der Vleuten CP. Predictive validity of a written knowledge test of skills for an OSCE in postgraduate training for general practice. *Med Educ* 2002;**36** (9):812–9.
21 Kroboth FJ, Hanusa BH, Parker S, Coulehan JL, Kapoor WN, Brown FH, Karpf M, Levey GS. The inter-rater reliability and internal consistency of a clinical evaluation exercise. *J Gen Intern Med* 1992;**7** (2):174–9.
22 Martin JA, Regehr G, Reznick R, MacRae H, Murnaghan J, Hutchison C, Brown M. Objective structured assessment of technical skill (OSATS) for surgical residents. *Br J Surg* 1997;**84** (2):273–8.
23 Matsell DG, Wolfish NM, Hsu E. Reliability and validity of the objective structured clinical examination in paediatrics. *Med Educ* 1991;**25** (4):293–9.
24 McIlroy JH, Hodges B, McNaughton N, Regehr G. The effect of candidates' perceptions of the evaluation method on reliability of checklist and global rating scores in an objective structured clinical examination. *Acad Med* 2002;**77** (7):725–8.
25 Minion DJ, Donnelly MB, Quick RC, Pulito A, Schwartz R. Are multiple objective measures of student performance necessary? *Am J Surg* 2002;**183** (6):663–5.
26 van Nuland M, van den Noortgate W, Degryse J, Goedhuys J. Comparison of two instruments for assessing communication skills in a general practice objective structured clinical examination. *Med Educ* 2007;**41** (7):676–83.
27 Park RS, Chibnall JT, Blaskiewicz RJ, Furman GE, Powell JK, Mohr CJ. Construct validity of an objective structured clinical examination (OSCE) in psychiatry: associations with the clinical skills examination and other indicators. *Acad Psychiatry* 2004;**28** (2):122–8.
28 Regehr G, Freeman R, Robb A, Missiha N, Heisey R. OSCE performance evaluations made by standardised patients: comparing checklist and global rating scores. *Acad Med* 1999;**74** (10 Suppl):135–7.
29 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** (9): 993–7.
30 Reznick RK, Blackmore D, Dauphinee WD, Rothman AI, Smee S. Large-scale high-stakes testing with an OSCE: report from the Medical Council of Canada. *Acad Med* 1996;**71** (1 Suppl):19–21.
31 Roberts J, Norman G. Reliability and learning from the objective structured clinical examination. *Med Educ* 1990;**24** (3):219–23.

32 Robins LS, White CB, Alexander GL, Gruppen LD, Grum CM. Assessing medical students' awareness of and sensitivity to diverse health beliefs using a standardised patient station. *Acad Med* 2001;**76** (1): 76–80.

33 Schwartz RW, Witzke DB, Donnelly MB, Stratton T, Blue AV, Sloan DA. Assessing residents' clinical performance: cumulative results of a four-year study with the objective structured clinical examination. *Surgery* 1998;**124** (2):307–12.

34 Searle J. Defining competency – the role of standard setting. *Med Educ* 2000;**34** (5):363–6.

35 Sloan DA, Donnelly MB, Schwartz RW, McGrath PC, Kenady DE, Wood DP, Strodel WE. Measuring the ability of residents to manage oncologic problems. *J Surg Oncol* 1997;**64** (2):135–42.

36 Sloan DA, Donnelly MB, Schwartz RW, Strodel WE. The objective structured clinical examination. The new gold standard for evaluating postgraduate clinical performance. *Ann Surg* 1995;**222** (6):735–42.

37 Teresi JA, Ramirez M, Ocepek-Welikson K, Cook MA. The development and psychometric analyses of ADEPT: an instrument for assessing the interactions between doctors and their elderly patients. *Ann Behav Med* 2005;**30** (3):225–42.

38 Volkan K, Simon SR, Baker H, Todres ID. Psychometric structure of a comprehensive objective structured clinical examination: a factor analytic approach. *Adv Health Sci Educ Theory Pract* 2004;**9** (2):83–92.

39 Wessel J, Williams R, Finch E, Gemus M. Reliability and validity of an objective structured clinical examination for physical therapy students. *J Allied Health* 2003;**32** (4):266–9.

40 Wilkinson TJ, Fontaine S. Patients' global ratings of student competence. Unreliable contamination or gold standard? *Med Educ* 2002;**36** (12):1117–21.

41 Wilkinson TJ, Newble DI, Wilson PD, Carter JM, Helms RM. Development of a three-centre simultaneous objective structured clinical examination. *Med Educ* 2000;**34** (10):798–807.

42 Lipsey MW, Wilson DB. *Practical Meta Analysis*. Thousand Oaks, CA: Sage Publications 2001;34–7.

43 Rodriguez MC, Maeda Y. Meta-analysis of coefficient alpha. *Psychol Methods* 2006;**11** (3):306–22.

44 Lee SJ, Wilkinson SL, Battles JB, Hynan LS. An objective structured clinical examination to evaluate health historian competencies. *Transfusion* 2003;**43** (1):34–41.

45 Bergus GR, Kreiter CD. The reliability of summative judgements based on objective structured clinical examination cases distributed across the clinical year. *Med Educ* 2007;**41** (7):661–6.

46 Gorter S, Rethans JJ, van der Heijde D, Scherpbier A, Houben H, van der Vleuten C, van der Linden S. Reproducibility of clinical performance assessment in practice using incognito standardised patients. *Med Educ* 2002;**36** (9):827–32.

47 Govaerts MJ, van der Vleuten CP, Schuwirth LW. Optimising the reproducibility of a performance-based assessment test in midwifery education. *Adv Health Sci Educ Theory Pract* 2002;**7** (2):133–45.

48 Boulet JR, Rebbecchi TA, Denton EC, McKinley DW, Whelan GP. Assessing the written communication skills of medical school graduates. *Adv Health Sci Educ Theory Pract* 2004;**9** (1):47–60.

49 Walters K, Osborn D, Raven P. The development, validity and reliability of a multimodality objective structured clinical examination in psychiatry. *Med Educ* 2005;**39** (3):292–8.

50 Verhoeven BH, Hamers JG, Scherpbier AJ, Hoogenboom RJ, van der Vleuten CP. The effect on reliability of adding a separate written assessment component to an objective structured clinical examination. *Med Educ* 2000;**34** (7):525–9.

51 Hodges B, Turnbull J, Cohen R, Bienenstock A, Norman G. Evaluating communication skills in the OSCE format: reliability and generalisability. *Med Educ* 1996;**30** (1):38–43.

52 Petrusa ER, Blackwell TA, Ainsworth MA. Reliability and validity of an objective structured clinical examination for assessing the clinical performance of residents. *Arch Intern Med* 1990;**150** (3):573–7.

53 Newble DI, Swanson DB. Psychometric characteristics of the objective structured clinical examination. *Med Educ* 1988;**22** (4):325–34.

54 Cooper W. Ubiquitous halo. *Psychol Bull* 1981;**90**:218–44.

Received 19 January 2011; editorial comments to authors 21 March 2011; accepted 15 June 2011