J Educ Eval Health Prof 2017; 14: 34 • https://doi.org/10.3352/jeehp.2017.14.34

eISSN: 1975-5937
Open Access

The sights and insights of examiners in objective structured clinical examinations

Lauren Chong[1], Silas Taylor[2], Matthew Haywood[3], Barbara-Ann Adelstein[4], Boaz Shulruf[2,5]*

[1]Clinical Skills Teaching Unit, Prince of Wales Hospital, Sydney, Australia
[2]Office of Medical Education, University of New South Wales, Sydney, Australia
[3]University of New South Wales, Sydney, Australia
[4]Prince of Wales Clinical School, University of New South Wales, Sydney, Australia
[5]Centre for Medical and Health Sciences Education, University of Auckland, Auckland, New Zealand

Purpose: The objective structured clinical examination (OSCE) is considered to be one of the most robust methods of clinical assessment. One of its strengths lies in its ability to minimise the effects of examiner bias due to the standardisation of items and tasks for each candidate. However, OSCE examiners' assessment scores are influenced by several factors that may jeopardise the assumed objectivity of OSCEs. To better understand this phenomenon, the current review aims to determine and describe important sources of examiner bias and the factors affecting examiners' assessments. **Methods:** We performed a narrative review of the medical literature using Medline. All articles meeting the selection criteria were reviewed, with salient points extracted and synthesised into a clear and comprehensive summary of the knowledge in this area. **Results:** OSCE examiners' assessment scores are influenced by factors belonging to 4 different domains: examination context, examinee characteristics, examinee-examiner interactions, and examiner characteristics. These domains are composed of several factors including halo, hawk/dove and OSCE contrast effects; the examiner's gender and ethnicity; training; lifetime experience in assessing; leadership and familiarity with students; station type; and site effects. **Conclusion:** Several factors may influence the presumed objectivity of examiners' assessments, and these factors need to be addressed to ensure the objectivity of OSCEs. We offer insights into directions for future research to better understand and address the phenomenon of examiner bias.

Keywords: Bias; Leadership; MEDLINE; Problem solving; Student

Introduction

The objective structured clinical examination (OSCE), introduced by Harden in 1975, is considered to be one of the most robust methods used for clinical assessment across medicine, nursing, exercise physiotherapy, and allied health programs [1-3]. It is most commonly used for summative, high-stakes assessments in medicine, nursing, and clinical psychology education programs [1,2,4,5], and as a selection tool for training and licensure for practice [1,6,7]. An OSCE requires each student to demonstrate specific skills and behaviours, typically in a series of short assessment tasks (stations), each of which is assessed by an examiner using a predetermined objective marking scheme [2]. Whilst OSCEs vary in their specific requirements and process across jurisdictions, the overall design of the OSCE has traditionally been viewed as advantageous, as it standardises the items and tasks for each candidate. Consequently, it has also been considered to minimise the effects of examiner bias through the use of 'identical' patients, structured checklists, and multiple assessor-candidate interactions across a number of stations [1,8]. Despite the intention of this design, OSCEs are in practice prone to high levels of variance [9]. Under ideal circumstances, scores should vary only as a reflection of student capability; however, the evidence shows that a key source of variability originates from the examiner [10-12]. Such examiner effects include assessor stringency or leniency, the halo effect, and a range of pre-existing biases [13,14]. Indeed, up to 29% of score variation may be explained by examiner stringency alone [14,15].

To ensure the validity of the OSCE as an assessment tool, it is crucial to understand and evaluate these sources of examiner bias [6,7].

*Corresponding email: b.shulruf@unsw.edu.au

Editor: Sun Huh, Hallym University, Korea
Received: December 18, 2017; **Accepted:** December 27, 2017;
Published online: December 27, 2017
This article is available from: http://jeehp.org

Traditionally, research in medical, nursing, and allied health education has focused on the reliability of assessments (e.g., items' internal consistency or inter-rater agreement), with less attention given to the effect of examiners' biases on the validity of the assessment [16]. Contemporary studies, however, have focused more on assessors' personal attributes and the nature and validity of the assumptions by which they are guided, and which eventually affect their judgment and scoring [17,18]. This critical review aims to discuss sources of examiner bias, and offers insights into directions for future research to better understand and address this phenomenon.

Methods

Study design

This systematic review used the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines.

Literature search process: We searched the medical literature using Medline (1946–April 2017) between January and April 2017 for papers that addressed the topic of examiner bias in OSCE settings. Search strategy were as follows:

1. OSCE.mp.
2. Objective structured clinical exam.mp.
3. 1 or 2
4. Bias.mp.
5. 3 and 4
6. Assessor bias.m_titl.
7. Examiner bias.mp.
8. 6 or 7
9. 3 and 8
10. Halo effect.mp.
11. Hawk dove.mp.
12. Hawk dove effect.mp.
13. 11 or 12
14. Examiner fatigue.mp.
15. 1 and 14
16. 5 or 8 or 10 or 13 or 14

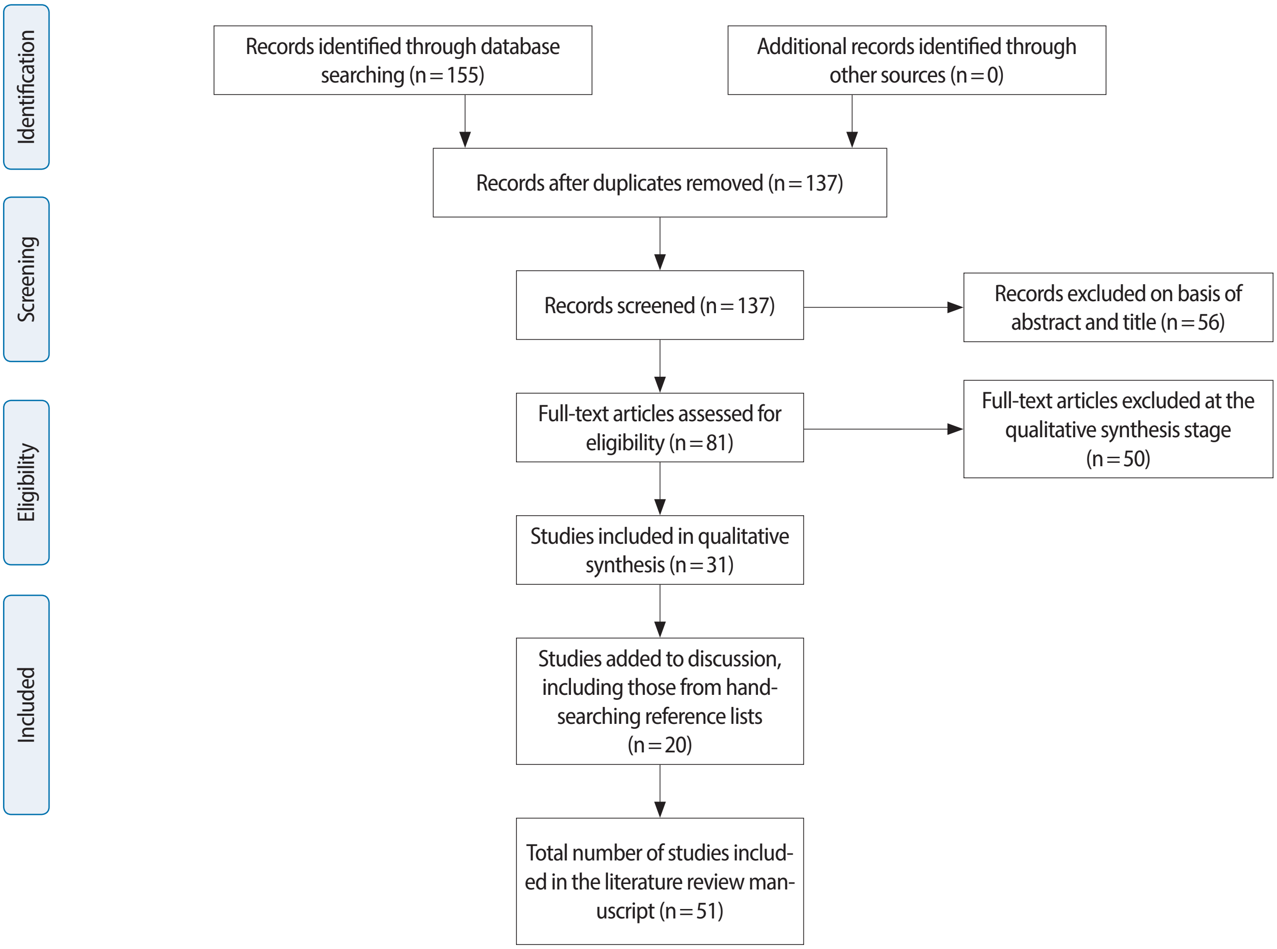
Identification
Screening
Eligibility
Included
Records identified through database searching (n = 155)
Additional records identified through other sources (n = 0)
Records after duplicates removed (n = 137)
Records screened (n = 137)
Records excluded on basis of abstract and title (n = 56)
Full-text articles assessed for eligibility (n = 81)
Full-text articles excluded at the qualitative synthesis stage (n = 50)
Studies included in qualitative synthesis (n = 31)
Studies added to discussion, including those from hand-searching reference lists (n = 20)
Total number of studies included in the literature review manuscript (n = 51)

Fig. 1. PRISMA (preferred reporting items for systematic reviews and meta-analyses) flow diagram. From Moher et al. PLoS Med 2009;6:e1000097 [19].

No restrictions were initially placed on the publication date within our search, although we only included publications within the last decade in our final analysis, resulting in the exclusion of any results pre-dating this period. Appropriate articles (n = 51) were reviewed, and salient points were extracted and synthesised into a clear and comprehensive summary of the knowledge in this area [19] (Fig. 1). LC conducted the initial screening of titles and abstracts and excluded articles that did not fulfil the inclusion criteria. The full texts of the remaining articles were independently reviewed by 2 authors (LC and BS), and studies that met the eligibility criteria were used in the final synthesis. Any discrepancies were resolved via discussion with the author team.

Results

Internal factors affecting OSCE examiners

Halo effect

One of the most studied types of rater effects is the 'halo effect' [7,13,20-22]. The halo effect is a cognitive bias in which an assessor fails to discriminate among independent aspects of behaviour when making a judgement about a student [7,22]. For example, the halo effect may occur when an examiner makes a judgement based on a general impression, such as a first impression, which then influences all subsequent judgements or ratings. Another example of the halo effect occurs when a rater allows an individual's performance in one domain, such as communication, to influence judgements of his or her performance in other domains. This effect is a threat to the validity of inferences made based on performance ratings, as it produces inappropriately similar ratings across items [20].

Hawkishness and dovishness

A potential vulnerability of any clinical examination is that examiners differ in their relative leniency or stringency. This is often termed the 'hawk-dove' effect [14]. Hawks tend to fail more candidates because of having very high standards; doves tend to pass more candidates due to greater leniency. The effect arises from examiners' own perceptions of the standards required for the exam, as well as from personality factors. Variance as high as 45% due to examiner stringency or leniency has been reported, thus making the hawk-dove effect one of the most significant factors influencing student outcomes [9]. In this study, a shift of 11% of OSCE candidates across the pass/fail line was demonstrated when the examiner stringency/leniency effect was removed from communication scores in a 6-station OSCE. At the ends of the examiner leniency distribution curve lie the 'extreme' assessors, defined as individuals giving a mean score greater or less than 3 standard deviations above or below the collective mean score [23]. The extreme nature of their assessments may be due to individual characteristics of an examiner, or less commonly, simple marking errors, for example grading 1/5 as 'excellent' and 5/5 as 'fail' when the opposite is correct [8].

Examiner demographics

Examiner sex and ethnicity were found not to predict score variance among general practitioner trainees in a clinical skills assessment, a finding supported by a similar study showing that examiner demographics (gender, UK or international medical degree, white or other background) explained only 0.2% of performance variance [24,25]. The level of training of the examiner likewise does not affect stringency or leniency [9]; however, both trained and untrained assessors tend to be more lenient and award higher marks to female students, although this interaction may only be slight and not statistically significant [20,26]. The influence of student-patient-examiner gender composition on examiner scores has not been reported, despite evidence from Australian medical schools that the opportunity to practice physical examinations on the opposite gender is limited [27]. Nonetheless, in the specific domain of communication skills assessment, a tendency exists for female students to perform significantly better than males [28]. This may be due to a combination of innately superior communication abilities in females, as well as gender interactions among the student, patient, and examiner. It has been shown that simulated patients tend to rate female students higher in communication skills than males through an effect independent of their own gender [29-31], and, while relatively little data exist on the effect of examiner and student gender interactions, Schleicher et al. [32] reported that male examiners awarded significantly higher communication skills ratings to female examinees.

Despite the above findings, the literature is still not entirely clear. Writing in 2013, Esmail and Roberts [33] commented that "we (cannot) confidently exclude bias from the examiners in the way that they assessed non-white candidates." While it is recognised that students from certain ethnic minorities may perform more poorly on assessments independent of any examiner bias, it is possible that examiner variance may also be up to 4 times greater than that of examinees [9,34]. Concern around issues such as these was sufficiently important to instigate the development of a cultural competence training module at Harvard Medical School [35]. A possible explanation for greater stringency is that people from individualistic cultures such as North America or Western Europe tend to place a higher value on personal independence, whereas people from collectivist cultures such as Asia, the Middle East, or some indigenous groups focus more on interdependence and relatedness to the community [36]. Thus, the latter may be more influenced by 'leadership' bias when multiple examiners are present, adopting the more stringent approach associated with examiners of greater clinical or assessment experience, who are also normally the more senior amongst the OSCE panel members [17]. The effect of a doctor's background on clinical practice has been recognised among international medical graduates who undergo a difficult acculturation process to both the general culture and the healthcare subculture in their host country [37].

Duration of examining during an assessment period

Students sitting an OSCE station early in the day receive higher marks on average than those sitting it later [6]. For example, Hope and Cameron [6] found a difference of 3.27% in marks between the first and last students sitting a station during a day, and it was predicted that 2 failing students would have passed had they been assessed in the morning. While this effect is small, it may impact students close to the pass/fail borderline or those in contention for awards. Variation by time of day has been attributed to examiner fatigue as the OSCE continues, as opposed to examiner 'warm-up' in the first few stations [1]. In contrast, some evidence suggests that increasing examiner fatigue over time leads to reduced attention to student errors and thus a tendency to award higher scores later in the day, even when adjusting for the warm-up phenomenon [38]. With regard to prolonged periods of OSCE assessment, assessors tend to be more lenient at the start and become more hawkish with time [6].

The contrast effect

Assessors tend to judge performance comparatively, rather than against fixed standards [39]. They tend to mentally amalgamate previous performances, especially those seen early on, to produce a performance standard to judge against. Thus, examiners who have recently observed and scored good performances give lower scores to borderline candidates than those who recently observed and scored poor performances [6,9,39]. This effect occurs across different parts of the educational curriculum, in non-clinical and clinical exams, different geographical locations, and different formats of examiner response (behavioural and global ratings) [40]. Examiners also show a lack of insight into their susceptibility to this phenomenon [39]. Anchoring bias (originally discussed in the context of diagnostic reasoning) is related to contrast bias and can be regarded as the influence of recent experiences on the examiner's subsequent ratings [41]. The examiner may ascribe disproportionate significance to certain features if exhibited by multiple examinees, thus leading to the award of a higher grade than a candidate deserves if he or she is preceded by a good performance. Hawkishness and dovishness are influenced in a similar way by the performance of recently observed candidates at any level, although the impact of this is less than that of the contrast effect [1,6].

Training of examiners and lifetime experience in assessment

Untrained assessors, as well as those with limited involvement in exam construction, award higher marks than trained assessors [6,13,42]. This may be attributable to a lack of understanding of the rating criteria and a poorer appreciation of the exact purpose, format, and scoring of the assessment [13,21]. Assessor training is therefore arguably an important component of a valid OSCE, as experienced examiners may set higher pass thresholds in OSCEs at least partially as a result of their greater confidence with the marking scheme or understanding of student standards [6,43]. Assessors may also use themselves as a reference point, leading to harsher candidate ratings as they become more experienced. Training is therefore important for both novice and experienced assessors in an attempt to ensure consistency across examiners.

Physician versus non-physician examiners

Good agreement exists between physicians and trained non-physician examiners when scoring against checklists [44]. However, there is poor agreement on pass/fail decisions, and up to 25% of students are misclassified by trained non-physician assessors, suggesting they are not as competent in completing global rating scales as trained physician examiners. This may be because non-physicians lack the medical knowledge to give credit to certain lines of questioning, such as those that ask the candidate to rule out certain differential diagnoses. However, it is interesting to note that among physician examiners, familiarity with a speciality does not influence the marks awarded [45].

Leadership and familiarity with students

If multiple examiners are present, they are influenced by the scores awarded by those with greater expertise or the perceived 'leader' [17]. Furthermore, examiners who are familiar with the students are more generous than those who are not [6,13]. This latter phenomenon may be a product of the 'mere exposure effect' whereby individuals favour things familiar to them [6].

External factors affecting OSCE examiners

Station type

A weak and statistically insignificant relationship was found between examiner scoring and the content area being examined [1]. Communication stations, such as taking a history or breaking bad news, may involve less assessor interaction than clinical examination stations [1]. This may increase the likelihood of assessor fatigue and disengagement, resulting in a higher or lower score than warranted by the performance. Some assessors are also less familiar with communication skill stations than with physical examination skill stations, but training in grading the former has been shown to reduce inter-rater variability [46]. Although the station type may produce bias in OSCE marks, station difficulty and order do not [6,47]. An ongoing tension exists between OSCE performances as determined by global rating scores and more objective, itemised checklist scores, particularly for borderline students [48]. When global and checklist scores are employed within a single station, some evidence indicates that assessors use different traits to inform their impression of these 2 metrics, perhaps due to inadequate assessor training or different levels of experience [48].

Site effect

This multifactorial source of bias is complex and not easily categorised under any of the above domains; however, it is recognised that different medical schools would not award the same score to an identically performing student at an identical OSCE station [18]. Differences in the agreed pass score, scoring criteria, simulated patient behaviour and examiner behaviour, and training have all been implicated and may even be inter-related. For example, a simulated patient's conduct may affect the student's performance directly, as well as influencing the examiner's perception of that performance. Similarly, the local choice of statistical analysis will also influence the proportion of students passing an OSCE. A comparison of 2 statistical analyses on the same data set demonstrated that the borderline regression method resulted in a higher pass mark and a larger difference margin in the failure rate than another common method when analysing smaller groups of students [49].

Discussion

Overall, this comprehensive (but not fully systematic) review identified several factors influencing OSCE examiners' assessment scores. The psychology and impact of the halo [7,13,21,22] and hawk/dove effects [1,9] are well understood, but further research is required into the influence of the contrast effect (and its duration) and the examiner's gender, ethnicity, training, lifetime experience in assessing, leadership, and familiarity with students. In addition, little is known about the effect of the assessment type (e.g., formative or summative), marking criteria, and exam tasks on examiners' judgements [6,13].

The authors propose that the factors discussed in this paper can be categorised into 4 major domains: examination context, examinee characteristics, examinee-examiner interactions, and examiner characteristics. Table 1 summarises the factors that are likely to raise the marks of an OSCE examinee. It should be noted that additional factors may influence the level of error (e.g., whether the examiner is a clinician), but no evidence of bias has been found.

An improved understanding of the potential role of these factors is crucial to reassure candidates and employers of the validity of OSCEs. This is especially true in a time of increased scrutiny surrounding health professional examinations [1]. Addressing these concerns will also have important implications for students close to the pass/fail borderline and those in contention for awards [6,9].

Table 1. Factors likely to raise OSCE marks

Domain	Specific factors increasing the OSCE score
Examination context	Being examined at the beginning of the OSCE day Being examined after a poor examinee
Examinee characteristics	Female gender Having pre-existing good interpersonal skills
Examinee-examiner interaction	Previously acquainted with examiner Culturally matched
Examiner characteristics	Inexperienced or non-expert Similar rank/status to the examinee "Dove" (rather than "hawk") inclination

OSCE, objective structured clinical exam.

While this review comprehensively summarises the biases in OSCE that are known to exist, the next step for researchers is to establish why they exist. Attempts to address examiner subjectivity through measurement standardisation have been largely unsuccessful [50,51], resulting in the recent emergence of rater cognition as a new field of research [52]. It is increasingly understood that assessors are motivated differently and form impressions of candidates dependent upon social interactions and context [51]. Variation in factors such as individuals' concepts of competency, definitions of critical performance aspects, synthesis of information gleaned from observation, production of narrative assessments, and conversion into rating scales are all thought to be key variables that have hitherto received relatively little attention [52]. The challenge is therefore to move away from a focus on rating instruments and raters to a focus on the context of performance assessment, such that assessor cognition can be more fully understood and targeted as part of an ongoing effort to reduce bias.

Limitations

Since this is not a classical systematic review, the authors cannot guarantee the comprehensiveness of the conclusions drawn in this paper. However, medical education is an evolving field and all contemporary evidence was evaluated. Medline contains more than 24 million references to life sciences and biomedical journals, and thus we argue that any relevant publication that was omitted from this paper as a result of not being indexed in Medline is unlikely to represent a substantial body of research not already discussed above.

Factors influencing OSCE examiners' assessment score

Once a stronger understanding of these issues is attained, strategies can then be implemented to address them. However, the challenge will be to achieve a suitable balance once interventions to remedy such biases are put in place. In other words, what does an optimal OSCE look like? We believe that the answer to this question is mostly not to be found within the statistical or psychometrical domains. All statistical analyses and psychometric techniques rely on the data generated by examiners who observe a performance and process that observation with their own skills, knowledge, prejudices, beliefs, and ability to accurately translate their decision into a predefined response or mark [36]. Thus, we urge future researchers to focus on the examiners' cognitive processes during OSCEs [16], an area that hopefully will shed more light on this 'black box' of decision-making and improve our confidence in the well-established OSCE.

ORCID: Lauren Chong: https://orcid.org/0000-0002-1791-1500; Silas Taylor: https://orcid.org/0000-0003-1992-8485; Matthew Haywood: https://orcid.org/0000-0003-3600-7987; Barbara-Ann Adelstein: https://orcid.org/0000-0002-7866-665X; Boaz Shulruf: https://orcid.org/0000-0003-3644-727X

Authors' contributions

Conceptualization: BS, LC, ST. Data curation: LC. Formal analysis: LC, BS. Methodology: BS, LC. Project administration: BS. Writing–original draft: LC, BS, MH. Writing–review & editing: LC, ST, BA, MH, BS.

Conflict of interest

No potential conflict of interest relevant to this article was reported.

Funding

No source of funding relevant to this article was reported.

Supplementary material

Supplement 1. Audio recording of the abstract.

References

1. Brennan PA, Croke DT, Reed M, Smith L, Munro E, Foulkes J, Arnett R. Does changing examiner stations during UK postgraduate surgery objective structured clinical examinations influence examination reliability and candidates' scores? J Surg Educ 2016;73:616-623. https://doi.org/10.1016/j.jsurg.2016.01.010
2. Mitchell ML, Henderson A, Groves M, Dalton M, Nulty D. The objective structured clinical examination (OSCE): optimising its value in the undergraduate nursing curriculum. Nurse Educ Today 2009; 29:398-404. https://doi.org/10.1016/j.nedt.2008.10.007
3. Sakurai H, Kanada Y, Sugiura Y, Motoya I, Wada Y, Yamada M, Tomita M, Tanabe S, Teranishi T, Tsujimura T, Sawa S, Okanishi T. OSCE-based clinical skill education for physical and occupational therapists. J Phys Ther Sci 2014;26:1387-1397. https://doi.org/10.1589/jpts.26.1387
4. Yap K, Bearman M, Thomas N, Hay M. Clinical psychology students' experiences of a pilot objective structured clinical examination. Aust Psychol 2012;47:165-173. https://doi.org/10.1111/j.1742-9544.2012.00078.x
5. Lin CW, Tsai TC, Sun CK, Chen DF, Liu KM. Power of the policy: how the announcement of high-stakes clinical examination altered OSCE implementation at institutional level. BMC Med Educ 2013; 13:8. https://doi.org/10.1186/1472-6920-13-8
6. Hope D, Cameron H. Examiners are most lenient at the start of a two-day OSCE. Med Teach 2015;37:81-85. https://doi.org/10.3109/0142159X.2014.947934
7. Wood TJ. Exploring the role of first impressions in rater-based assessments. Adv Health Sci Educ Theory Pract 2014;19:409-427. https://doi.org/10.1007/s10459-013-9453-9
8. Fuller R, Homer M, Pell G, Hallam J. Managing extremes of assessor judgment within the OSCE. Med Teach 2017;39:58-66. https://doi.org/10.1080/0142159X.2016.1230189
9. Harasym PH, Woloschuk W, Cunning L. Undesired variance due to examiner stringency/leniency effect in communication skill scores assessed in OSCEs. Adv Health Sci Educ Theory Pract 2008;13:617-632. https://doi.org/10.1007/s10459-007-9068-0
10. Clauser BE, Harik P, Margolis MJ, McManus IC, Mollon J, Chis L, Williams S. An empirical examination of the impact of group discussion and examinee performance information on judgments made in the Angoff standard-setting procedure. Appl Meas Educ 2008;22:1-21. https://doi.org/10.1080/08957340802558318
11. Clauser BE, Mee J, Baldwin SG, Margolis MJ, Dillon GF. Judges' use of examinee performance data in an Angoff standard-setting exercise for a medical licensing examination: an experimental study. J Educ Meas 2009;46:390-407. https://doi.org/10.1111/j.1745-3984.2009.00089.x
12. Hurtz GM, Patrick Jones J. Innovations in measuring rater accuracy in standard setting: assessing "Fit" to item characteristic curves. Appl Meas Educ 2009;22:120-143. https://doi.org/10.1080/08957340902754601
13. Stroud L, Herold J, Tomlinson G, Cavalcanti RB. Who you know or what you know?: effect of examiner familiarity with residents on OSCE scores. Acad Med 2011;86(10 Suppl):S8-S11. https://doi.org/10.1097/ACM.0b013e31822a729d
14. Finn Y, Cantillon P, Flaherty G. Exploration of a possible relationship between examiner stringency and personality factors in clinical assessments: a pilot study. BMC Med Educ 2014;14:1052. https://doi.org/10.1186/s12909-014-0280-3
15. Hill F, Kendall K, Galbraith K, Crossley J. Implementing the undergraduate mini-CEX: a tailored approach at Southampton University. Med Educ 2009;43:326-334. https://doi.org/10.1111/j.1365-2923.2008.03275.x
16. Chahine S, Holmes B, Kowalewski Z. In the minds of OSCE examiners: uncovering hidden assumptions. Adv Health Sci Educ Theory Pract 2016;21:609-625. https://doi.org/10.1007/s10459-015-9655-4
17. Shulruf B, Wilkinson T, Weller J, Jones P, Poole P. Insights into the Angoff method: results from a simulation study. BMC Med Educ 2016;16:134. https://doi.org/10.1186/s12909-016-0656-7
18. Chesser A, Cameron H, Evans P, Cleland J, Boursicot K, Mires G. Sources of variation in performance on a shared OSCE station across four UK medical schools. Med Educ 2009;43:526-532. https://doi.org/10.1111/j.1365-2923.2009.03370.x
19. Moher D, Liberati A, Tetzlaff J, Altman DG; PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS Med 2009;6:e1000097. https://doi.org/10.1371/journal.pmed.1000097
20. Iramaneerat C, Yudkowsky R, Myford CM, Downing SM. Quality control of an OSCE using generalizability theory and many-faceted Rasch measurement. Adv Health Sci Educ Theory Pract 2008;13:479-493. https://doi.org/10.1007/s10459-007-9060-8
21. Iramaneerat C, Yudkowsky R. Rater errors in a clinical skills assessment

of medical students. Eval Health Prof 2007;30:266-283. https://doi.org/10.1177/0163278707304040

22. Wood TJ, Chan J, Humphrey-Murto S, Pugh D, Touchie C. The influence of first impressions on subsequent ratings within an OSCE station. Adv Health Sci Educ Theory Pract 2017;22:969-983. https://doi.org/10.1007/s10459-016-9736-z
23. Bartman I, Smee S, Roy M. A method for identifying extreme OSCE examiners. Clin Teach 2013;10:27-31. https://doi.org/10.1111/j.1743-498X.2012.00607.x
24. Denney M, Wakeford R. Do role-players affect the outcome of a high-stakes postgraduate OSCE, in terms of candidate sex or ethnicity?: results from an analysis of the 52,702 anonymised case scores from one year of the MRCGP clinical skills assessment. Educ Prim Care 2016;27:39-43. https://doi.org/10.1080/14739879.2015.1113724
25. Denney ML, Freeman A, Wakeford R. MRCGP CSA: are the examiners biased, favouring their own by sex, ethnicity, and degree source? Br J Gen Pract 2013;63:e718-725. https://doi.org/10.3399/bjgp13X674396
26. McManus IC, Elder AT, Dacre J. Investigating possible ethnicity and sex bias in clinical examiners: an analysis of data from the MRCP(UK) PACES and nPACES examinations. BMC Med Educ 2013;13:103. https://doi.org/10.1186/1472-6920-13-103
27. Taylor S, Shulruf B. Australian medical students have fewer opportunities to do physical examination of peers of the opposite gender. J Educ Eval Health Prof 2016;13:42. https://doi.org/10.3352/jeehp.2016.13.42
28. Casey M, Wilkinson D, Fitzgerald J, Eley D, Connor J. Clinical communication skills learning outcomes among first year medical students are consistent irrespective of participation in an interview for admission to medical school. Med Teach 2014;36:640-642. https://doi.org/10.3109/0142159X.2014.907880
29. Swygert KA, Cuddy MM, van Zanten M, Haist SA, Jobe AC. Gender differences in examinee performance on the step 2 clinical skills data gathering (DG) and patient note (PN) components. Adv Health Sci Educ Theory Pract 2012;17:557-571. https://doi.org/10.1007/s10459-011-9333-0
30. Cuddy MM, Swygert KA, Swanson DB, Jobe AC. A multilevel analysis of examinee gender, standardized patient gender, and United States medical licensing examination step 2 clinical skills communication and interpersonal skills scores. Acad Med 2011;86(10 Suppl):S17-S20. https://doi.org/10.1097/ACM.0b013e31822a6c05
31. Graf J, Smolka R, Simoes E, Zipfel S, Junne F, Holderried F, Wosnik A, Doherty AM, Menzel K, Herrmann-Werner A. Communication skills of medical students during the OSCE: gender-specific differences in a longitudinal trend study. BMC Med Educ 2017;17:75. https://doi.org/10.1186/s12909-017-0913-4
32. Schleicher I, Leitner K, Juenger J, Moeltner A, Ruesseler M, Bender B, Sterz J, Schuettler KF, Koenig S, Kreuder JG. Examiner effect on the objective structured clinical exam - a study at five medical schools. BMC Med Educ 2017;17:71. https://doi.org/10.1186/s12909-017-0908-1
33. Esmail A, Roberts C. Academic performance of ethnic minority candidates and discrimination in the MRCGP examinations between 2010 and 2012: analysis of data. BMJ 2013;347:f5662. https://doi.org/10.1136/bmj.f5662
34. Stupart D, Goldberg P, Krige J, Khan D. Does examiner bias in undergraduate oral and clinical surgery examinations occur? S Afr Med J 2008;98:805-807.
35. White AA 3rd, Hoffman HL. Culturally competent care education: overview and perspectives. J Am Acad Orthop Surg 2007;15 Suppl 1:S80-S85. https://doi.org/10.5435/00124635-200700001-00018
36. Shulruf B, Hattie J, Dixon R. Factors affecting responses to Likert type questionnaires: introduction of the ImpExp, a new comprehensive model. Soc Psychol Educ 2008;11:59-78. https://doi.org/10.1007/s11218-007-9035-x
37. Sciolla AF, Lu FG. Cultural competence for international medical graduate physicians: a perspective. In: Rao NR, Roberts LW, editors. International medical graduate physicians: a guide to training. Cham: Springer International Publishing; 2016. p. 283-303.
38. McLaughlin K, Ainslie M, Coderre S, Wright B, Violato C. The effect of differential rater function over time (DRIFT) on objective structured clinical examination ratings. Med Educ 2009;43:989-992. https://doi.org/10.1111/j.1365-2923.2009.03438.x
39. Yeates P, Moreau M, Eva K. Are Examiners' judgments in OSCE-style assessments influenced by contrast effects? Acad Med 2015;90:975-980. https://doi.org/10.1097/ACM.0000000000000650
40. Schuwirth LW, van der Vleuten CP. General overview of the theories used in assessment: AMEE guide no. 57. Med Teach 2011;33:783-797. https://doi.org/10.3109/0142159X.2011.611022
41. Sibbald M, Panisko D, Cavalcanti RB. Role of clinical context in residents' physical examination diagnostic accuracy. Med Educ 2011;45:415-421. https://doi.org/10.1111/j.1365-2923.2010.03896.x
42. Pell G, Homer MS, Roberts TE. Assessor training: its effects on criterion-based assessment in a medical context. Int J Res Method Educ 2008;31:143-154. https://doi.org/10.1080/17437270802124525
43. Boursicot KA, Roberts TE, Pell G. Using borderline methods to compare passing standards for OSCEs at graduation across three medical schools. Med Educ 2007;41:1024-1031. https://doi.org/10.1111/j.1365-2923.2007.02857.x
44. Humphrey-Murto S, Smee S, Touchie C, Wood TJ, Blackmore DE. A comparison of physician examiners and trained assessors in a high-stakes OSCE setting. Acad Med 2005;80(10 Suppl):S59-S62. https://doi.org/10.1097/00001888-200510001-00017
45. Wong ML, Fones CS, Aw M, Tan CH, Low PS, Amin Z, Wong PS, Goh PS, Wai CT, Ong B, Tambyah P, Koh DR. Should non-expert clinician examiners be used in objective structured assessment of communication skills among final year medical undergraduates? Med Teach 2007;29:927-932. https://doi.org/10.1080/01421590701601535
46. Schwartzman E, Hsu DI, Law AV, Chung EP. Assessment of patient communication skills during OSCE: examining effectiveness of a training program in minimizing inter-grader variability. Patient Educ Couns 2011;83:472-477. https://doi.org/10.1016/j.pec.2011.04.001

47. Monteiro SD, Walsh A, Grierson LE. OSCE circuit performance effects: does circuit order influence scores? Med Teach 2016;38:98-100. https://doi.org/10.3109/0142159X.2015.1075647

48. Pell G, Homer M, Fuller R. Investigating disparity between global grades and checklist scores in OSCEs. Med Teach 2015;37:1106-1113. https://doi.org/10.3109/0142159X.2015.1009425

49. Malau-Aduli BS, Teague PA, D'Souza K, Heal C, Turner R, Garne DL, van der Vleuten C. A collaborative comparison of objective structured clinical examination (OSCE) standard setting methods at Australian medical schools. Med Teach 2017;39:1261-1267. https://doi.org/10.1080/0142159X.2017.1372565

50. Gingerich A, Kogan J, Yeates P, Govaerts M, Holmboe E. Seeing the 'black box' differently: assessor cognition from three research perspectives. Med Educ 2014;48:1055-1068. https://doi.org/10.1111/medu.12546

51. Govaerts MJ, van der Vleuten CP, Schuwirth LW, Muijtjens AM. Broadening perspectives on clinical performance assessment: rethinking the nature of in-training assessment. Adv Health Sci Educ Theory Pract 2007;12:239-260. https://doi.org/10.1007/s10459-006-9043-1

52. Gauthier G, St-Onge C, Tavares W. Rater cognition: review and integration of research findings. Med Educ 2016;50:511-522. https://doi.org/10.1111/medu.12973