Article:
Homer, M orcid.org/0000-0002-1161-5938, Fuller, R, Hallam, J orcid.org/0000-0002-1044-0515 et al. (1 more author) (2020) Shining a spotlight on scoring in the OSCE: Checklists and item weighting. Medical Teacher, 42 (9). pp. 1037-1042. ISSN 0142-159X

https://doi.org/10.1080/0142159X.2020.1781072

Shining a spotlight on scoring in the OSCE: checklists and item weighting

Abstract

Introduction

There has been a long running debate about the validity of item-based checklist scoring of performance assessments like OSCEs. In recent years, the conception of a checklist has developed from its dichotomous inception into a more 'key-features' and/or chunked approach, where 'items' have the potential to become weighted differently, but the literature does not always reflect these broader conceptions.

Methods

We consider theoretical, design and (clinically trained) assessor issues related to differential item weighting in checklist scoring of OSCEs stations. Using empirical evidence, this work also compares candidate decisions and psychometric quality of different item-weighting approaches (i.e. a simple 'unweighted' scheme versus a differentially weighted one).

Results

The impact of different weighting schemes affect approximately 30% of the key borderline group of candidates, and 3% of candidates overall. We also find that measures of overall assessment quality are a little better under the differentially weighted scoring system.

Discussion and conclusion

Differentially weighted modern checklists can contribute to valid assessment outcomes, and bring a range of additional benefits to the assessment. Judgment about weighting of particular items should be considered a key design consideration during station development, and must align to clinical assessor expectations of the relative importance of sub-tasks.

Key words

OSCE scoring; checklist design; item weighting; assessment quality

Practice points

- There is a long running debate into scoring in OSCEs, with checklists sometimes presented in the literature as overly reductive and with weaker validity in comparison with global rating scales and other scoring approaches.
- Much literature on checklists presents an over-narrow or confused definition – often only allowing checklist items to be dichotomous in nature.
- Modern conceptions of the checklist allow for a 'key-features' and/or chunked approach to design, which allows items to be differentially weighted, both within and between items.
- Different scoring approaches to modern checklists will produce different student-level pass/fail decisions, and different measures of assessment quality.
- Proper judgment of item weighting produces more valid assessment outcomes, and better measures of assessment quality, and should be seen as integral to the design process in many contexts.

Introduction

For the delivery of high stakes performance assessment (most commonly the OSCE), context specific approaches to scoring are an essential aspect of effective assessment design, yet it is surprising that this practice remains under-developed in the literature in a number of different ways (Khan, Ramachandran, et al. 2013; Khan, Gaunt, et al. 2013; Harden et al. 2015, chap. 11). Current literature sometimes presents an apparently

confusing range of approaches, with differing or vague definitions of scoring instruments (what ultimately is a checklist?), and differing approaches to instrument design, scoring and weighting. In order to help reconcile some of this apparent variation, this paper considers some of the wider practical and theoretical issues that arise when thinking about checklist design and the validity of outcomes and inferences that follow. This work focuses on a key aspect of OSCE scoring design, item weighting, and explores how differently weighted scoring items have an important impact on assessment decision-making, particularly in the pass/fail region. Ultimately, we argue that considerations of checklist item-level weighting/scoring should form an essential part of modern OSCE station development in many contexts.

Overview of scoring in OSCEs

Authoritative sources on assessment practice affirm the importance of theoretical, practical and empirical considerations when designing scoring instruments (Cizek & Bunch 2007, chap. 2; American Educational Research Association 2014, p. 93). Whilst a number of different approaches to scoring in OSCEs are used in medical education assessment, including checklists, domain-based scoring, behaviourally anchored rating scales and key feature formats, little attention is often paid to the appropriate 'marriage' of factors surrounding assessors (expertise and experience, cognitive load), candidates (novice-mastery scale) and the requirements of the clinical encounter being assessed (process/task focused, behavioural/affective components or a mixture).

In the four decades since the original inception of the OSCE, the use of checklists for scoring has continued to attract comment and criticism (Hodges et al. 1999; Wilkinson et al. 2003; Pugh et al. 2016; Wood & Pugh 2019), whilst assumptions about their actual format show evidence of misunderstanding and misconception, rather than awareness of contemporary

design considerations. A relatively recent systematic review comprising 45 studies of validity evidence comparing checklists and global rating scales measuring the same construct (Ilgen et al. 2015) concludes that the evidence in favour of checklists is *not as weak as might have previously been thought* (Regehr et al. 1998; Hodges et al. 1999; Regehr et al. 1999; Ringsted et al. 2003). Another very recent study (Wood & Pugh 2019) found that checklists were as sensitive as rating scales in capturing levels of expertise.

In order to frame our work as precisely as possible, we next review the literature on how checklists are defined.

What is a checklist?

There is not a commonly accepted consensus in the literature on precisely what a checklist is, and the definition of checklists in the Ilgen *et al.* study (2015) is actually quite narrow, allowing for only a 'dichotomous' response at the item level (i.e. right/wrong, done/not done type response). Other work similarly suggests a checklist item is naturally dichotomous, (Wood & Pugh 2019; Harden et al. 1975; Ma et al. 2012; Daniels et al. 2014), whilst often the actual detail of how items are constructed and scored is not always clear in the literature (Yudkowsky et al. 2014).

It can be argued that defining a checklist scoring system in an overly restrictive (e.g. dichotomous) way is unnecessarily narrow and outdated. In a 'key features' approach to checklist design, originally advocated by Farmer and Page (2005), the scoring instrument focuses entirely on the essential elements of the task, with scoring approaches allowing for differential weighting between items. In a more contemporary realisation of the checklist, groups of items are 'chunked' together to score a group of behaviours that form a meaningful and coherent sub-set of activity within a clinical encounter (Fuller et al. 2013). In such

designs, scoring might involve a hybrid model of key features and behaviourally anchored rating scales for an appropriate task (e.g. a mixture of technical process steps in a procedure, coupled with communication elements). In such models, there is clearly a trade-off between allowing assessors to purposefully discriminate between levels of performance in an individual item, and ensuring that this discrimination is meaningfully reflective of actual levels of performance being assessed, whilst also ensuring that the cognitive load on assessors' is not too high (Lafleur et al. 2015). We develop these scoring/weighting considerations in the next section.

To weight or not to weight?

Development of checklist design presents an opportunity for the exploration of differential weighting, both within the items themselves (e.g. varying weights for anchors within an item), and across items (so the maximum achievable score can be different for different items and stations, e.g. depending on the relative 'importance' of elements of the clinical encounter). Such weights are usually determined via expert judgement of item-writers (Kahraman et al. 2008; American Educational Research Association 2014, p. 93). There is a considerable body of literature, both internal and external to medical education (Wainer 1976; Streiner & Norman 2008; Sandilands et al. 2014) that argues that such differential item weighting schemes make little difference to measures of *overall* assessment quality, and to pass/fail rates at overall cohort levels. Hence, these studies generally contend that, rather than developing 'complex' (Sandilands et al. 2014) weighting schemes, assessment writers would be better off spending their time and effort on other aspects of test development. However, in focusing at the level of *cohort*, these studies sometimes fail to explore the possibility of different decisions at the level of the individual *candidate* through the application of differential item weighting.

Other literature (Kahraman et al. 2008) compares weighting methods in a context where SPs provide judgments of performance (the USMLE Step 2 clinical skills examination). This latter study finds that different weighting schemes produce different levels of reliability, with a regression-based approach to weighting performing well in this regard. The study also finds, that expert judgment of item weights are 'appropriate' too, with outcomes found to be more strongly related to external criteria (i.e. clinical knowledge and clinical skills documentation scores) than are scores under a 'unit' weighting method (Kahraman et al. 2008). The case for re-exploring weighting as part of good item/test design is further supported by the fact that overly detailed scoring rubrics can add to construct-irrelevant variance (i.e. to error in the measurement) (Schuwirth & van der Vleuten 2012).

Much of the evidence in this area rests mainly on psychometric analysis, and sometimes ignores important assessment design and wider validity considerations. In not fully exploring the impact of differential weighting at candidate level, decision-making can be more challenging for those candidates in the critical pass/fail region (Sadler 2009; Homer, Pell, et al. 2017), and for whom maximising the quality of pass/fail decisions is essential.

In an attempt to reconcile these arguments, we undertook a simple empirical study that investigates the impact of different item-weighting schemes on measures of assessment quality and on candidate decisions, focussed on the impact on those in the pass/fail region. In the usual format, we next present an overview of the methodology employed in this study, and then go on to detail results, and finally discuss our findings and what they add to the debates around checklist item scoring in OSCEs.

Methodological approach

OSCE context and design

The student-level data for this study is the graduating level assessment taken in the final year of a five year undergraduate medicine degree qualification. It uses a sequential examination format (Pell et al. 2013) with the first part of the sequence consisting of 13 stations (the second part is 12 additional stations) with 250-270 candidates randomly assigned to 45-50 separate groups, and assessed across parallel sessions/circuits in four separate test centres, reflecting a two day exam where students will alternately take morning or afternoon slots on different days to ensure 'fairness'. The assessment uses approximately 500 trained OSCE assessors, the vast majority of whom take time out from their clinical practice for a portion of the 2-day period to examine – this is a common approach to OSCE examining in many global contexts. Stations typically integrate higher-level processes (e.g. decision-making, prescribing, case management) at a level of mastery appropriate to that expected of new doctors entering post-graduate training in the UK National Health Service.

Each OSCE 'administration' is designed with a test specification and blueprint across content and skills determined by the programme's assessment group, and mapped against national standards and outcomes for UK graduates (General Medical Council 2018). OSCE scoring has been developed through a programme of synergistic research – (Fuller et al. 2013; Fuller et al. 2017) from dichotomous (and sometimes lengthy) 'lists' of individual, isolated behaviours, towards a modern key features (Farmer & Page 2005) approach of differentially scored and individually weighted and chunked 'items'. In each twelve minute station, the checklist typically consists of 13 items (median 13, minimum 7, maximum 14), some of which are key features, and others are 'chunked' in the sense of combining more simple items into a holistic 'super-item' (e.g. prescribing, decision-making) scored by three

point anchors. A third group of simpler items are scored with two-point anchors. Whilst OSCE scoring sheets are not made publically available (as they are part of a secure test bank), we later give exemplars of select checklist items and how they are weighted (Table 1 below).

Support material is given to assessors at both station and item-level to indicate the level of credit for certain levels of behaviours (e.g. measuring blood pressure to within ±5% for full credit, within ±10% for partial credit, 0 otherwise). Station-level global grades are awarded on five-point scale (0=fail, 1=borderline, 2=pass, 3=good pass; 4=excellent pass).

Within stations, standards are set using borderline regression (BRM) (McKinley & Norcini 2014), and station level cut-scores are aggregated to the test level to give the overall cut-score in the exam on which progression decisions are made (Homer, Fuller, et al. 2017). Robust post hoc analysis assures quality at whole test level, station level and across parallel sessions and circuits (Pell et al. 2010; Pell et al. 2015).

Methods

We compare student outcomes and measures of assessment quality by modelling 'live data' from a recent administration of the OSCE by re-scoring the checklist outcomes. In any OSCE administration, examiners are effectively blinded to the presence or absence of item weighting through the use of anchors which are linked to letters (e.g. XYZ = inadequate/adequate/excellent), allowing us to understand the impact of transforming scoring systems from weighted to 'unweighted' as follows:

From**:**

- the '**actual**' weighting scheme employed in the examination (for example 0, 2 for not done/done or 0, 1, 4 or 0, 2, 4 for not done/done/done well).

To**:**

- a '**simple**' weighting scheme – e.g. not done/done (items score 0,1) or not done/done/done well (e.g. items scored 0, 0, 1)

Examples of how we re-score items from 'actual' to 'simple' are shown in Table 1.

TABLE 1 HERE

For each scoring approach, we employ a BRM methodology for producing station- and test-level pass/fail decisions for each candidate. The key focus of this study is the comparison of overall exam level pass/fail decisions for individual candidates when moving from 'actual' to 'simple' scoring. To understand the impact of weighting on station 'quality', we also compare measures of assessment quality using selected metrics detailed in Pell *et al*. (2010). These are overall exam reliability, Omega total (Revelle & Zinbarg 2008), and then two station level metrics:

- R-squared within stations (a measure of the strength of the relationship between global grades and checklist scores).
- Variation in scores across circuits (a proxy for assessor differences across parallel circuits given random allocation of students and assessors across circuits).

We use the non-parametric related samples Wilcoxson signed rank test to compare the two sets of thirteen station-level metrics across the exam.

Ethical considerations

The co-chairs of the University of Leeds School of Medicine ethics committee confirmed in writing to the authors that formal ethics approval for this study was not required as it involved the use of routinely collected student assessment data which were fully anonymized prior to analysis.

Results

'Pass/fail' decisions

In this OSCE administration, eight candidates (3.0%) have different decisions about whether they are required to undertake the full sequence of 25 stations, dependent on 'simple' or 'actual' weightings. This is made up of five individuals who 'meet the standard' under the simple scheme but 'fail to meet it' under the 'actual' weighting scheme, and three for whom the opposite is true. This implies that the overall 'pass rate' is slightly higher using simple weighting (89.7% versus 88.9%).

Accepting that the borderline group varies considerably across stations in the OSCE (Homer, Pell, et al. 2017), we wished to explore the impact of item weighting on the performance of this group. In doing so, we make an assumption that the 9th decile of performance includes the majority of the 'borderline' or 'just passing' group of candidates. Under this assumption, we find then that the 3% in the cohort as a whole is equivalent to around 30% of the borderline group having different decisions between the two scoring schemes. This is the major empirical finding of this study which we explore further in the *Discussion*.

Across the 13 stations in the assessment there are 192 station-level 'pass/fail' decisions that are different under the two weighting schemes – a median of 14 different decisions per station, and overall there are 34 more station fails under actual weighting.

Measures of assessment quality

Table 2 presents a summary of three metrics for measuring assessment quality (Pell et al. 2010; Pell et al. 2015).

TABLE 2 HERE

The reliability (i.e. reproducibility of the scores) is marginally higher, and the average R-squared is significantly higher under actual weighting (p=0.005). This latter metric is important as lower values of R-squared can bring into doubt the standard setting process under borderline regression (Pell et al. 2010). The third metric in Table 2 is an estimate of variation in scores by assessors across circuits, and in the final row we see that values of this are lower (i.e. better) under actual weighting but that this is not statistically significant at the standard 5% level.

The overall pattern in Table 2 is clear – improved metrics under actual weighting, which therefore indicates stronger validity evidence in favour of the interpretation and use of outcomes from this scoring scheme compared to those under the simple weighting approach (Kane 2006; Cook et al. 2015).

Discussion

Maximising the quality of borderline decision-making: the impact of item level weighting

Our key empirical finding is that under the two weighting schemes we have compared, a substantial proportion of 'borderline' candidate decisions are different at exam level – 30% in this study (and 28% in an earlier cohort using the same approach). Accurate pass/fail

decision-making at candidate level is obviously one of the most important issues in performance assessments, and in assessment more generally (Cizek & Bunch 2007, chap. 2; Homer, Pell, et al. 2017). Further, station-level differences in decisions, also evidenced in this study, are important in institutions where, in addition to the aggregate cut-score requirement, there is also a minimum number of stations to be passed (usually to minimise excessive compensation across stations).

It is clear that whatever the weighting scheme adopted, some pass/fail decisions will be different and this empirical study using two different models gives an estimate of the scale of the difference (≈30% of borderline cases different). The crucial point is that using professional expertise to judge item weighting will produce more valid assessment outcomes (American Educational Research Association 2014, p. 93; Cook et al. 2015), and this argument is also supported by the evidence of improved psychometrics (Table 2).

Some authors have maintained that the use of weighting requires additional time and expertise at the design stage that is not necessary for both assessment construction and outcome decisions (Sandilands et al., 2014) . However, this is countered by the argument that this process is an integral part of station writing, where maximising authenticity does require a weighting decision irrespective of scoring format (checklist, domain etc) through decisions about what elements to include, and exclude, from scoring rubrics. These decisions may be implicit, and highly aligned to the context of the station, or more explicit with a particular decision to focus on only part of a wider clinical encounter to help 'frame' as an OSCE station. (Fuller et al. 2013; American Educational Research Association 2014, p. 93).

Helping assessors make good decisions through good design

Tavares and Eva (2013) argue that in order for examiners to be able to make accurate decisions about student performance, and specifically how performance relates to an appropriate level/ standard, they must firstly acquire the relevant knowledge about the assessment and then process it in a meaningful way. Instructional design in assessment is pivotal in ensuring that the cognitive demand on assessors is not excessive - if it is, then accurate rating cannot occur. As the ability to effectively measure cognitive load in OSCE contexts remains a challenge, Tavares and Eva (2013) suggest that a useful move is to reduce the likely burden of this by avoiding the presentation to assessors of a list of individual and isolated behaviours to be scored. Their work therefore supports a re-conceptualisation of scoring formats, moving from fixed conceptions to the utilisation of a design that scores groups of behaviours together. A hybrid approach where items are 'chunked' and the focus is on 'key features' of the encounter, in a more nuanced and meaningful way, guides clinical assessors as to the key elements being assessed; this is not only more representative of a typical clinical encounter, but is also fits with good design principles to better ensure the accuracy of examiner decisions. This is particularly true in some contexts, for example, clinical encounters focusing on technical, practical procedures where practitioners are likely to use their own internalised 'checklist' to deliver care – such as pre-operative safety checks. Our empirical findings imply that without appropriate consideration of item weighting decisions, it is clear that hybrid and/or key features checklists will lack validity as scoring instruments (American Educational Research Association 2014, p. 93; Cook et al. 2015).

What does this means for scoring and checklists in the OSCE?

When considering clinical assessors' affective responses to checklists, one can imagine situations where it could be inappropriate to only allow them to make a dichotomous decision (e.g. competent/not competent or pass/fail) when observing a performance. According to the cognitive load research (Lafleur et al. 2015), the decision-making process of an assessor examining a candidate mimics the cognitive architecture of the relationship between a clinician/teacher and a learner. In other words, the assessor decision-making is a familiar experience in cognitive terms, and this reduces the impact of intrinsic load (Chandler & Sweller 1991) on the examiner. This is clearly another important consideration in good OSCE station design. However, the forcing of dichotomous decisions, in a time limited, high stakes environment, via an old-style checklist can increase the extrinsic load for clinical examiners, and this effect can be exacerbated in the case of more inexperienced examiners (Tavares & Eva 2013). Examiners who are more experienced, both clinically and as assessors, are likely to be more adaptable and to experience lower extrinsic load in such situations. Getting the scoring/weighting correct at the design stage taking into account the nature of the assessor group is therefore a key element in the validity argument (Kane 2006; Cook et al. 2015), and these arguments support the development of checklists beyond the merely dichotomous in a range of contexts.

Study limitations

One important limitation of our study is that the empirical work is based on a simple re-scoring approach (Table 1), rather than on one that directly compares different assessor behaviour when using differently weighted scoring instruments. However, there are considerable logistical and resource challenges in terms of additional assessors required, and possibly ethical issues around different pass/fail decisions for some candidates, that

might be difficult to be overcome in such a study. Whilst we acknowledge this limitation, the design of the OSCE does allow the use of 'live data' to be re-modelled to explore impact. Scoring formats ensure assessors are effectively blinded to whether any form of weighting is being applied in the checklist, as this is only applied after the exam has completed (whether on tablet or optically marked sheet), and is unseen by the examiner.

Conclusion

We have used empirical, theoretical and practical arguments to demonstrate that weighting matters, particularly at the candidate level. We have also developed a more refined argument than that present in the extant literature, which sometimes mis-specifies or neglects details of checklist design, and also the important roles of assessment writers and assessors in the assessment process. Regardless of the weighting scheme employed, some candidate decisions pass/fail will be different, and it is the job of the assessment designers to justify their weighting approach as a key element of the process.

It is important to acknowledge that there are both 'good' and 'bad' scoring formats *irrespective* of whether checklist, or rating/global rating alone. In good assessment design, there needs to be the right marriage of assessor, candidate and task characteristics – and for the latter, scoring and weighting issues are key elements that, in this study, exert powerful effects at candidate level, particularly those in the critical pass/fail area.

Declaration of interest

The authors declare no conflicts of interest.

Biographical note

Matt Homer, BSc, MSc, PhD, PGCE, CStat, is an Associate Professor in the Schools of Education and Medicine at the University of Leeds. Within medical education, he has a research interest in assessment design, standard setting methodologies and psychometrics analysis. He also advises the UK General Medical Council on a range of assessment issues.

Richard Fuller MA, MBChB, FRCP, is a Consultant Geriatrician/Stroke Physician and Deputy Dean of the School of Medicine at the University of Liverpool. His current research focuses on the application of intelligent assessment design in campus and workplace-based assessment formats, assessor behaviours, technology enhanced assessment and the impact of sequential testing methodologies. He advises a number of national and international institutions.

Jennifer Hallam, BSc, PG Dip, MSc, PhD, is an Educational psychometrician in the School of Medicine, University of Leeds. Her current interests include the strategic development of assessment and feedback strategies, specifically for performance based assessments. She also has several national medical education roles which include being on the Board of Directors for the Association for the Study of Medical Education (ASME).

Godfrey Pell, BEng, MSc, CStat, is a principal research fellow emeritus at Leeds Institute of Medical Education, who has a strong background in management. His research focuses on quality within the OSCE, including theoretical and practical applications. He acts as an assessment consultant to a number of medical schools.

References

American Educational Research Association. 2014. Standards for Educational and Psychological Testing. Washington, D.C: American Educational Research Association.

Chandler P, Sweller J. 1991. Cognitive Load Theory and the Format of Instruction. Cogn Instr. 8(4):293–332.

Cizek GJ, Bunch MB. 2007. Standard Setting: A Guide to Establishing and Evaluating Performance Standards on Tests. First edition. Thousand Oaks, Calif: SAGE Publications, Inc.

Cook DA, Brydges R, Ginsburg S, Hatala R. 2015. A contemporary approach to validity arguments: a practical guide to Kane's framework. Med Educ. 49(6):560–575.

Daniels VJ, Bordage G, Gierl MJ, Yudkowsky R. 2014. Effect of clinically discriminating, evidence-based checklist items on the reliability of scores from an Internal Medicine residency OSCE. Adv Health Sci Educ. 19(4):497–506.

Farmer EA, Page G. 2005. A practical guide to assessing clinical decision-making skills using the key features approach. Med Educ. 39(12):1188–1194.

Fuller R, Homer M, Pell G. 2013. Longitudinal interrelationships of OSCE station level analyses, quality improvement and overall reliability. Med Teach. 35(6):515–517.

Fuller R, Homer M, Pell G, Hallam J. 2017. Managing extremes of assessor judgment within the OSCE. Med Teach. 39(1):58–66.

General Medical Council. 2018. Outcomes for graduates 2018. London: General Medical Council.

Harden R, Lilley P, Patricio M. 2015. The Definitive Guide to the OSCE: The Objective Structured Clinical Examination as a performance assessment., 1e. 1 edition. Edinburgh ; New York: Churchill Livingstone.

Harden RM, Stevenson M, Downie WW, Wilson GM. 1975. Assessment of clinical competence using objective structured examination. Br Med J. 1(5955):447–451.

Hodges B, Regehr G, McNaughton N, Tiberius R, Hanson M. 1999. OSCE checklists do not capture increasing levels of expertise. Acad Med. 74(10):1129–1134.

Homer M, Fuller R, Pell G. 2017. The benefits of sequential testing: Improved diagnostic accuracy and better outcomes for failing students. Med Teach. 0(0):1–10.

Homer M, Pell G, Fuller R. 2017. Problematizing the concept of the "borderline" group in performance assessments. Med Teach. 0(0):1–7.

Ilgen JS, Ma IWY, Hatala R, Cook DA. 2015. A systematic review of validity evidence for checklists versus global rating scales in simulation-based assessment. Med Educ. 49(2):161–173.

Kahraman N, Clauser BE, Margolis MJ. 2008. A Comparison of Alternative Item Weighting Strategies on the Data Gathering Component of a Clinical Skills Performance Assessment. Acad Med. 83:S72–S75.

Kane MT. 2006. Validation. In: Brennan RL, editor. Educ Meas. Westport, CT: Praeger Publishers; p. 17–64.

Khan KZ, Gaunt K, Ramachandran S, Pushkar P. 2013. The Objective Structured Clinical Examination (OSCE): AMEE Guide No. 81. Part II: Organisation & Administration. Med Teach. 35(9):e1447–e1463.

Khan KZ, Ramachandran S, Gaunt K, Pushkar P. 2013. The Objective Structured Clinical Examination (OSCE): AMEE Guide No. 81. Part I: an historical and theoretical perspective. Med Teach. 35(9):e1437-1446.

Lafleur A, Côté L, Leppink J. 2015. Influences of OSCE design on students' diagnostic reasoning. Med Educ. 49(2):203–214.

Ma IWY, Zalunardo N, Pachev G, Beran T, Brown M, Hatala R, McLaughlin K. 2012. Comparing the use of global rating scale with checklists for the assessment of central venous catheterization skills using simulation. Adv Health Sci Educ. 17(4):457–470.

McKinley DW, Norcini JJ. 2014. How to set standards on performance-based examinations: AMEE Guide No. 85. Med Teach. 36(2):97–110.

Pell G, Fuller R, Homer M, Roberts T. 2010. How to measure the quality of the OSCE: A review of metrics - AMEE guide no. 49. Med Teach. 32(10):802–811.

Pell G, Fuller R, Homer M, Roberts T. 2013. Advancing the objective structured clinical examination: sequential testing in theory and practice. Med Educ. 47(6):569–577.

Pell G, Homer M, Fuller R. 2015. Investigating disparity between global grades and checklist scores in OSCEs. Med Teach. 37(12):1106–1113.

Pugh D, Halman S, Desjardins I, Humphrey-Murto S, Wood TJ. 2016. Done or Almost Done? Improving OSCE Checklists to Better Capture Performance in Progress Tests. Teach Learn Med. 28(4):406–414.

Regehr G, Freeman R, Robb A, Missiha N, Heisey R. 1999. OSCE performance evaluations made by standardized patients: Comparing checklist and global rating scores. Acad Med. 74(10):S135–S137.

Regehr G, MacRae H, Reznick RK, Szalay D. 1998. Comparing the psychometric properties of checklists and global rating scales for assessing performance on an OSCE-format examination. Acad Med J Assoc Am Med Coll. 73(9):993–997.

Revelle W, Zinbarg RE. 2008. Coefficients Alpha, Beta, Omega, and the glb: Comments on Sijtsma. Psychometrika. 74(1):145.

Ringsted C, Østergaard D, Ravn L, Pedersen JA, Berlac PA, Vleuten CPMVD. 2003. A feasibility study comparing checklists and global rating forms to assess resident performance in clinical skills. Med Teach. 25(6):654–658.

Sadler DR. 2009. Indeterminacy in the use of preset criteria for assessment and grading. Assess Eval High Educ. 34(2):159–179.

Sandilands D (Dallie), Gotzmann A, Roy M, Zumbo BD, De Champlain A. 2014. Weighting checklist items and station components on a large-scale OSCE: Is it worth the effort? Med Teach. 36(7):585–590.

Schuwirth LWT, van der Vleuten CPM. 2012. Programmatic assessment and Kane's validity perspective. Med Educ. 46(1):38–48.

Streiner DL, Norman GR. 2008. Health measurement scales: a practical guide to their development and use. 4th ed. Oxford; New York: Oxford University Press.

Tavares W, Eva KW. 2013. Exploring the Impact of Mental Workload on Rater-Based Assessments. Adv Health Sci Educ. 18(2):291–303.

Wainer H. 1976. Estimating coefficients in linear models: It don't make no nevermind. Psychol Bull. 83(2):213–217.

Wilkinson TJ, Frampton CM, Thompson-Fawcett M, Egan T. 2003. Objectivity in objective structured clinical examinations: Checklists are no substitute for examiner commitment. Acad Med. 78(2):219–223.

Wood TJ, Pugh D. 2019. Are rating scales really better than checklists for measuring increasing levels of expertise? Med Teach. 0(0):1–6.

Yudkowsky R, Park YS, Riddle J, Palladino C, Bordage G. 2014. Clinically discriminating checklists versus thoroughness checklists: improving the validity of performance test scores. Acad Med J Assoc Am Med Coll. 89(7):1057–1062.

Table 1

Item description	Item type	Actual weighting	Simple weighting
Candidate explores carer burden and acknowledges carer strain	Key feature	0, 2, 4	0, 0, 1
Explores falls history and pressure ulcer risk factors (including frailty, environment, meds)	Chunked	0, 1, 4	0, 0, 1
Adequate hand hygiene	Standard	0, 2	0, 1

Table 1: Examples of checklist item scoring and re-scoring

Table 2

Metric	Description	Simple weighting	Actual weighting	Related samples Wilcoxson signed rank test p-value	Metric better under which weighting method?
Omega total (Revelle & Zinbarg 2008)	Score reliability	0.68	0.69	NA	Actual – reliability slightly higher
Median R-squared across stations	Proportion of shared variance between checklist score and global grade in station	0.51	0.53	0.005	Actual – r-squared values significantly higher
Median variation in scores across circuits	Variation across student groups in station – a proxy for assessor differences in stringency across parallel circuits	29.9%	27.0%	0.221	Actual – lower median variation across circuits (but not significantly)

Table 2: Comparison of metrics between the two weighting approaches